The MEMS Handbook Second Edition

MEMS

Introduction and Fundamentals

Mechanical Engineering Series

Frank Kreith and Roop Mahajan - Series Editors

Published Titles

Distributed Generation: The Power Paradigm for the New Millennium
Anne-Marie Borbely & Jan F. Kreider
Elastoplasticity Theory
Vlado A. Lubarda
Energy Audit of Building Systems: An Engineering Approach
Moncef Krarti
Engineering Experimentation
Euan Somerscales
Entropy Generation Minimization
Adrian Bejan
Finite Element Method Using MATLAB, 2nd Edition
Young W. Kwon & Hyochoong Bang
Fluid Power Circuits and Controls: Fundamentals and Applications
John S. Cundiff
Fundamentals of Environmental Discharge Modeling
Lorin R. Davis
Heat Transfer in Single and Multiphase Systems
Greg F. Naterer
Introductory Finite Element Method
Chandrakant S. Desai & Tribikram Kundu
Intelligent Transportation Systems: New Principles and Architectures
Sumit Ghosh & Tony Lee
Mathematical & Physical Modeling of Materials Processing Operations
Olusegun Johnson Ilegbusi, Manabu Iguchi & Walter E. Wahnsiedler
Mechanics of Composite Materials
Autar K. Kaw
Mechanics of Fatigue
Vladimir V. Bolotin
Mechanics of Solids and Shells: Theories and Approximations
Gerald Wempner & Demosthenes Talaslidis
Mechanism Design: Enumeration of Kinematic Structures According to Function
Lung-Wen Tsai
The MEMS Handbook, Second Edition
MEMS: Introduction and Fundamentals
MEMS: Design and Fabrication
MEMS: Applications
Mohamed Gad-el-Hak
Nonlinear Analysis of Structures
M. Sathyamoorthy
Practical Inverse Analysis in Engineering
David M. Trujillo & Henry R. Busby
Pressure Vessels: Design and Practice
Somnath Chattopadhyay
Principles of Solid Mechanics
Rowland Richards, Jr.
Thermodynamics for Engineers
Kau-Fui Wong
Vibration and Shock Handbook
Clarence W. de Silva
Viscoelastic Solids
Roderic S. Lakes

The MEMS Handbook Second Edition

MEMS

Introduction and Fundamentals

Edited by
Mohamed Gad-el-Hak

Taylor & Francis
Taylor & Francis Group
Boca Raton London New York

A CRC title, part of the Taylor & Francis imprint, a member of the
Taylor & Francis Group, the academic division of T&F Informa plc.

Foreground: A 24-layer rotary varactor fabricated in nickel using the Electrochemical Fabrication (EFAB®) technology. See Chapter 6, *MEMS: Design and Fabrication*, for details of the EFAB® technology. Scanning electron micrograph courtesy of Adam L. Cohen, Microfabrica Incorporated (www.microfabrica.com), U.S.A.

Background: A two-layer surface macromachined, vibrating gyroscope. The overall size of the integrated circuitry is 4.5 × 4.5 mm. Sandia National Laboratories' emblem in the lower right-hand corner is 700 microns wide. The four silver rectangles in the center are the gyroscope's proof masses, each 240 × 310 × 2.25 microns. See Chapter 4, *MEMS: Applications* (0-8493-9139-3), for design and fabrication details. Photograph courtesy of Andrew D. Oliver, Sandia National Laboratories.

Published in 2006 by
CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

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

International Standard Book Number-10: 0-8493-9137-7 (Hardcover)
International Standard Book Number-13: 978-0-8493-9137-8 (Hardcover)
Library of Congress Card Number 2005050111

Library of Congress Cataloging-in-Publication Data

MEMS : introduction and fundamentals / edited by Mohamed Gad-El-Hak.
p. cm. -- (Mechanical engineering series)
Includes bibliographical references and index.
ISBN 0-8493-9137-7 (alk. paper)
1. Microelectronics. 2. Nanotechnology. I. Gad-el-Hak, M. II. Mechanical engineering series (Boca Raton, Fla.)

TK7874.M3762 2005
621.381--dc22 2005050111

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

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com
and the CRC Press Web site at
http://www.crcpress.com

Preface

In a little time I felt something alive moving on my left leg, which advancing gently forward over my breast, came almost up to my chin; when bending my eyes downward as much as I could, I perceived it to be a human creature not six inches high, with a bow and arrow in his hands, and a quiver at his back. … I had the fortune to break the strings, and wrench out the pegs that fastened my left arm to the ground; for, by lifting it up to my face, I discovered the methods they had taken to bind me, and at the same time with a violent pull, which gave me excessive pain, I a little loosened the strings that tied down my hair on the left side, so that I was just able to turn my head about two inches. … These people are most excellent mathematicians, and arrived to a great perfection in mechanics by the countenance and encouragement of the emperor, who is a renowned patron of learning. This prince has several machines fixed on wheels, for the carriage of trees and other great weights.

(From *Gulliver's Travels—A Voyage to Lilliput*, by Jonathan Swift, 1726.)

In the Nevada desert, an experiment has gone horribly wrong. A cloud of nanoparticles — micro-robots — has escaped from the laboratory. This cloud is self-sustaining and self-reproducing. It is intelligent and learns from experience. For all practical purposes, it is alive.

It has been programmed as a predator. It is evolving swiftly, becoming more deadly with each passing hour.

Every attempt to destroy it has failed.

And we are the prey.

(From Michael Crichton's techno-thriller *Prey*, HarperCollins Publishers, 2002.)

Almost three centuries apart, the imaginative novelists quoted above contemplated the astonishing, at times frightening possibilities of living beings much bigger or much smaller than us. In 1959, the physicist Richard Feynman envisioned the fabrication of machines much smaller than their makers. The length scale of man, at slightly more than 10^0 m, amazingly fits right in the middle of the smallest subatomic particle, which is approximately 10^{-26} m, and the extent of the observable universe, which is of the order of 10^{26} m. Toolmaking has always differentiated our species from all others on Earth. Close to 400,000 years ago, archaic *Homo sapiens* carved aerodynamically correct wooden spears. Man builds things consistent with his size, typically in the range of two orders of magnitude larger or smaller than himself. But humans have always striven to explore, build, and control the extremes of length and time scales. In the voyages to Lilliput and Brobdingnag in *Gulliver's Travels*, Jonathan Swift speculates on the remarkable possibilities which diminution or magnification of physical dimensions provides. The Great Pyramid of Khufu was originally 147 m high when completed around 2600 B.C., while the Empire State Building constructed in 1931 is presently 449 m high. At the other end of the spectrum of manmade artifacts, a dime is slightly less than 2 cm in diameter. Watchmakers have practiced the art of miniaturization since the 13th century. The invention of the microscope in the 17th century opened the way for direct observation of microbes and plant and animal cells. Smaller things were manmade in the latter half of the 20th century. The transistor in today's integrated circuits has a size of 0.18 micron in production and approaches 10 nanometers in research laboratories.

Microelectromechanical systems (MEMS) refer to devices that have characteristic length of less than 1 mm but more than 1 micron, that combine electrical and mechanical components, and that are fabricated using integrated circuit batch-processing technologies. Current manufacturing techniques for MEMS include surface silicon micromachining; bulk silicon micromachining; lithography, electrodeposition, and plastic molding; and electrodischarge machining. The multidisciplinary field has witnessed explosive growth during the last decade and the technology is progressing at a rate that far exceeds that of our understanding of the physics involved. Electrostatic, magnetic, electromagnetic, pneumatic and thermal actuators, motors, valves, gears, cantilevers, diaphragms, and tweezers of less than 100 micron size have been fabricated. These have been used as sensors for pressure, temperature, mass flow, velocity, sound and chemical composition, as actuators for linear and angular motions, and as simple components for complex systems such as robots, lab-on-a-chip, micro heat engines and micro heat pumps. The lab-on-a-chip in particular is promising to automate biology and chemistry to the same extent the integrated circuit has allowed large-scale automation of computation. Global funding for micro- and nanotechnology research and development quintupled from $432 million in 1997 to $2.2 billion in 2002. In 2004, the U.S. National Nanotechnology Initiative had a budget of close to $1 billion, and the worldwide investment in nanotechnology exceeded $3.5 billion. In 10 to 15 years, it is estimated that micro- and nanotechnology markets will represent $340 billion per year in materials, $300 billion per year in electronics, and $180 billion per year in pharmaceuticals.

The three-book *MEMS set* covers several aspects of microelectromechanical systems, or more broadly, the art and science of electromechanical miniaturization. MEMS design, fabrication, and application as well as the physical modeling of their materials, transport phenomena, and operations are all discussed. Chapters on the electrical, structural, fluidic, transport and control aspects of MEMS are included in the books. Other chapters cover existing and potential applications of microdevices in a variety of fields, including instrumentation and distributed control. Up-to-date new chapters in the areas of microscale hydrodynamics, lattice Boltzmann simulations, polymeric-based sensors and actuators, diagnostic tools, microactuators, nonlinear electrokinetic devices, and molecular self-assembly are included in the three books constituting the second edition of *The MEMS Handbook*. The 16 chapters in *MEMS: Introduction and Fundamentals* provide background and physical considerations, the 14 chapters in *MEMS: Design and Fabrication* discuss the design and fabrication of microdevices, and the 15 chapters in *MEMS: Applications* review some of the applications of micro-sensors and microactuators.

There are a total of 45 chapters written by the world's foremost authorities in this multidisciplinary subject. The 71 contributing authors come from Canada, China (Hong Kong), India, Israel, Italy, Korea, Sweden, Taiwan, and the United States, and are affiliated with academia, government, and industry. Without compromising rigorousness, the present text is designed for maximum readability by a broad audience having engineering or science background. As expected when several authors are involved, and despite the editor's best effort, the chapters of each book vary in length, depth, breadth, and writing style. These books should be useful as references to scientists and engineers already experienced in the field or as primers to researchers and graduate students just getting started in the art and science of electromechanical miniaturization. The Editor-in-Chief is very grateful to all the contributing authors for their dedication to this endeavor and selfless, generous giving of their time with no material reward other than the knowledge that their hard work may one day make the difference in someone else's life. The talent, enthusiasm, and indefatigability of Taylor & Francis Group's Cindy Renee Carelli (acquisition editor), Jessica Vakili (production coordinator), N. S. Pandian and the rest of the editorial team at Macmillan India Limited, Mimi Williams and Tao Woolfe (project editors) were highly contagious and percolated throughout the entire endeavor.

Mohamed Gad-el-Hak

Editor-in-Chief

Mohamed Gad-el-Hak received his B.Sc. (summa cum laude) in mechanical engineering from Ain Shams University in 1966 and his Ph.D. in fluid mechanics from the Johns Hopkins University in 1973, where he worked with Professor Stanley Corrsin. Gad-el-Hak has since taught and conducted research at the University of Southern California, University of Virginia, University of Notre Dame, Institut National Polytechnique de Grenoble, Université de Poitiers, Friedrich-Alexander-Universität Erlangen-Nürnberg, Technische Universität München, and Technische Universität Berlin, and has lectured extensively at seminars in the United States and overseas. Dr. Gad-el-Hak is currently the Inez Caudill Eminent Professor of Biomedical Engineering and chair of mechanical engineering at Virginia Commonwealth University in Richmond. Prior to his Notre Dame appointment as professor of aerospace and mechanical engineering, Gad-el-Hak was senior research scientist and program manager at Flow Research Company in Seattle, Washington, where he managed a variety of aerodynamic and hydrodynamic research projects.

Professor Gad-el-Hak is world renowned for advancing several novel diagnostic tools for turbulent flows, including the laser-induced fluorescence (LIF) technique for flow visualization; for discovering the efficient mechanism via which a turbulent region rapidly grows by destabilizing a surrounding laminar flow; for conducting the seminal experiments which detailed the fluid–compliant surface interactions in turbulent boundary layers; for introducing the concept of targeted control to achieve drag reduction, lift enhancement and mixing augmentation in wall-bounded flows; and for developing a novel viscous pump suited for microelectromechanical systems (MEMS) applications. Gad-el-Hak's work on Reynolds number effects in turbulent boundary layers, published in 1994, marked a significant paradigm shift in the subject. His 1999 paper on the fluid mechanics of microdevices established the fledgling field on firm physical grounds and is one of the most cited articles of the 1990s.

Gad-el-Hak holds two patents: one for a drag-reducing method for airplanes and underwater vehicles and the other for a lift-control device for delta wings. Dr. Gad-el-Hak has published over 450 articles, authored/edited 14 books and conference proceedings, and presented 250 invited lectures in the basic and applied research areas of isotropic turbulence, boundary layer flows, stratified flows, fluid–structure interactions, compliant coatings, unsteady aerodynamics, biological flows, non-Newtonian fluids, hard and soft computing including genetic algorithms, flow control, and microelectromechanical systems. Gad-el-Hak's papers have been cited well over 1000 times in the technical literature. He is the author of the book "*Flow Control: Passive, Active, and Reactive Flow Management*," and editor of the books "*Frontiers in Experimental Fluid Mechanics,*" "*Advances in Fluid Mechanics Measurements*," "*Flow Control: Fundamentals and Practices*," "*The MEMS Handbook,*" and "*Transition and Turbulence Control.*"

Professor Gad-el-Hak is a fellow of the American Academy of Mechanics, a fellow and life member of the American Physical Society, a fellow of the American Society of Mechanical Engineers, an associate fellow of the American Institute of Aeronautics and Astronautics, and a member of the European Mechanics

Society. He has recently been inducted as an eminent engineer in Tau Beta Pi, an honorary member in Sigma Gamma Tau and Pi Tau Sigma, and a member-at-large in Sigma Xi. From 1988 to 1991, Dr. Gad-el-Hak served as Associate Editor for *AIAA Journal*. He is currently serving as Editor-in-Chief for *e-MicroNano.com*, Associate Editor for *Applied Mechanics Reviews* and *e-Fluids*, as well as Contributing Editor for Springer-Verlag's *Lecture Notes in Engineering* and *Lecture Notes in Physics*, for McGraw-Hill's Year Book of Science and Technology, and for CRC Press' *Mechanical Engineering Series*.

Dr. Gad-el-Hak serves as consultant to the governments of Egypt, France, Germany, Italy, Poland, Singapore, Sweden, United Kingdom and the United States, the United Nations, and numerous industrial organizations. Professor Gad-el-Hak has been a member of several advisory panels for DOD, DOE, NASA and NSF. During the 1991/1992 academic year, he was a visiting professor at Institut de Mécanique de Grenoble, France. During the summers of 1993, 1994 and 1997, Dr. Gad-el-Hak was, respectively, a distinguished faculty fellow at Naval Undersea Warfare Center, Newport, Rhode Island, a visiting exceptional professor at Université de Poitiers, France, and a Gastwissenschaftler (guest scientist) at Forschungszentrum Rossendorf, Dresden, Germany. In 1998, Professor Gad-el-Hak was named the Fourteenth ASME Freeman Scholar. In 1999, Gad-el-Hak was awarded the prestigious Alexander von Humboldt Prize — Germany's highest research award for senior U.S. scientists and scholars in all disciplines — as well as the Japanese Government Research Award for Foreign Scholars. In 2002, Gad-el-Hak was named ASME Distinguished Lecturer, as well as inducted into the Johns Hopkins University Society of Scholars.

Contributors

Ronald J. Adrian
Department of Mechanical and
Aerospace Engineering
Arizona State University
Tempe, Arizona, U.S.A.

Ramesh K. Agarwal
Department of Mechanical and
Aerospace Engineering
Washington University in St. Louis
St. Louis, Missouri, U.S.A.

Ali Beskok
Department of Mechanical
Engineering
Texas A&M University
College Station, Texas, U.S.A.

Thomas R. Bewley
Department of Mechanical and
Aerospace Engineering
University of California, San Diego
La Jolla, California, U.S.A.

Kenneth S. Breuer
Division of Engineering
Brown University
Providence, Rhode Island, U.S.A.

Hsueh-Chia Chang
Center for Microfluidics and
Medical Diagnostics
University of Notre Dame
Notre Dame, Indiana, U.S.A.

Mohamed Gad-el-Hak
Department of Mechanical
Engineering
Virginia Commonwealth University
Richmond, Virginia, U.S.A.

J. William Goodwine
Department of Aerospace and
Mechanical Engineering
University of Notre Dame
Notre Dame, Indiana, U.S.A.

Nicolas G. Hadjiconstantinou
Department of Mechanical
Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts, U.S.A.

George Em Karniadakis
Center for Fluid Mechanics
Brown University
Providence, Rhode Island, U.S.A.

Robert M. Kirby
School of Computing
University of Utah
Salt Lake City, Utah, U.S.A.

Kartikeya Mayaram
Department of Electrical and
Computer Engineering
Oregon State University
Corvallis, Oregon, U.S.A.

Oleg Mikulchenko
Advanced Mixed Signal Development
Intel Corporation
Sacramento, California, U.S.A.

Joshua I. Molho
Caliper Life Sciences Incorporated
Mountain View, California, U.S.A.

Alexander Oron
Department of Mechanical
Engineering
Technion—Israel Institute of
Technology
Haifa, Israel

Juan G. Santiago
Department of Mechanical
Engineering
Stanford University
Stanford, California, U.S.A.

Mihir Sen
Department of Aerospace and
Mechanical Engineering
University of Notre Dame
Notre Dame, Indiana, U.S.A.

Kendra V. Sharp
Department of Mechanical and
Nuclear Engineering
Pennsylvania State University
University Park, Pennsylvania, U.S.A.

William N. Sharpe, Jr.
Department of Mechanical
Engineering
The Johns Hopkins University
Baltimore, Maryland, U.S.A.

Robert H. Stroud
The Aerospace Corporation
Sterling, Virginia, U.S.A.

William Trimmer
Belle Mead Research, Inc.
Hillsborough, New Jersey, U.S.A.

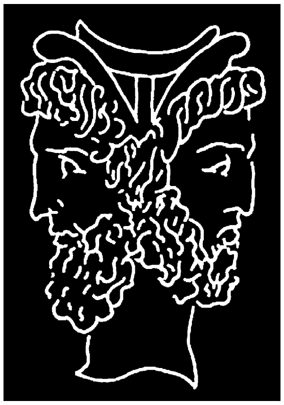
Keon-Young Yun
Research & Development Center
Samhongsa Co., Ltd.
Seoul, Korea

Table of Contents

The farther backward you can look,
the farther forward you are likely to see.

(Sir Winston Leonard Spencer Churchill, 1874–1965)

Janus, Roman god of
gates, doorways and all
beginnings, gazing both
forward and backward.



As for the future, your task is not to foresee, but to enable it.

(Antoine-Marie-Roger de Saint-Exupéry, 1900–1944,
in Citadelle [*The Wisdom of the Sands*])

1

Introduction

Mohamed Gad-el-Hak
Virginia Commonwealth University

How many times when you are working on something frustratingly tiny, like your wife's wrist watch, have you said to yourself, "If I could only train an ant to do this!" What I would like to suggest is the possibility of training an ant to train a mite to do this. What are the possibilities of small but movable machines? They may or may not be useful, but they surely would be fun to make.

(From the talk "There's Plenty of Room at the Bottom," delivered by Richard P. Feynman at the annual meeting of the American Physical Society, Pasadena, California, December 1959.)

Toolmaking has always differentiated our species from all others on Earth. Aerodynamically correct wooden spears were carved by archaic *Homo sapiens* close to 400,000 years ago. Man builds things consistent with his size, typically in the range of two orders of magnitude larger or smaller than himself, as indicated in Figure 1.1. Though the extremes of length-scale are outside the range of this figure, man, at slightly more than 10^0 m, amazingly fits right in the middle of the smallest subatomic particle, which is

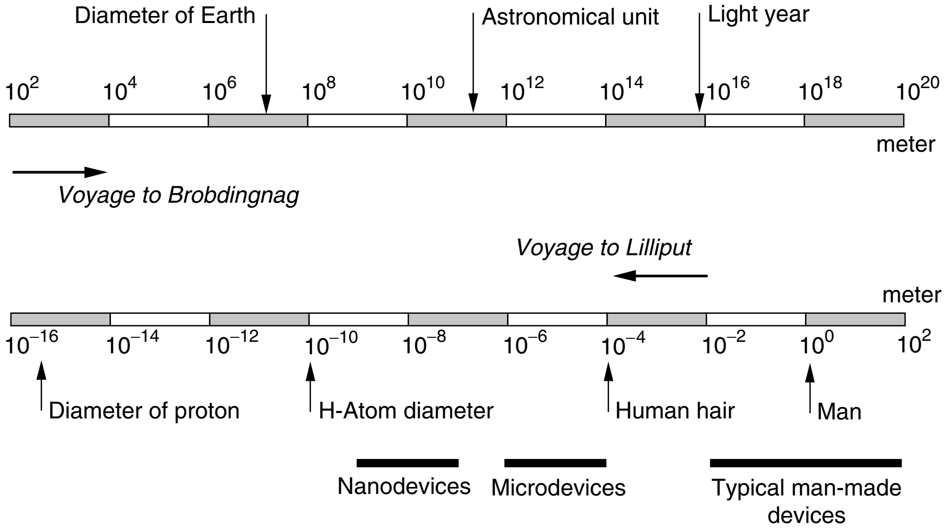


FIGURE 1.1 Scale of things, in meters. Lower scale continues in the upper bar from left to right. One meter is 10^6 microns, 10^9 nanometers, or 10^{10} Angstroms.

approximately 10^{-26} m, and the extent of the observable universe, which is of the order of 10^{26} m (15 billion light years); neither geocentric nor heliocentric, but rather egocentric universe. But humans have always striven to explore, build, and control the extremes of length and time scales. In the voyages to Lilliput and Brobdingnag of *Gulliver's Travels*, Jonathan Swift (1726) speculates on the remarkable possibilities which diminution or magnification of physical dimensions provides.[1] The Great Pyramid of Khufu was originally 147 m high when completed around 2600 B.C., while the Empire State Building constructed in 1931 is presently — after the addition of a television antenna mast in 1950 — 449 m high. At the other end of the spectrum of manmade artifacts, a dime is slightly less than 2 cm in diameter. Watchmakers have practiced the art of miniaturization since the 13th century. The invention of the microscope in the 17th century opened the way for direct observation of microbes and plant and animal cells. Smaller things were manmade in the latter half of the 20th century. The transistor — invented in 1947 — in today's integrated circuits has a size[2] of 0.18 micron (180 nanometers) in production and approaches 10 nm in research laboratories using electron beams. But what about the miniaturization of mechanical parts — machines — envisioned by Feynman (1961) in his legendary speech quoted above?

Manufacturing processes that can create extremely small machines have been developed in recent years (Angell et al., 1983; Gabriel et al., 1988, 1992; O'Connor, 1992; Gravesen et al., 1993; Bryzek et al., 1994; Gabriel, 1995; Ashley, 1996; Ho and Tai, 1996, 1998; Hogan, 1996; Ouellette, 1996, 2003; Paula, 1996; Robinson et al., 1996a, 1996b; Tien, 1997; Amato, 1998; Busch-Vishniac, 1998; Kovacs, 1998; Knight, 1999; Epstein, 2000; O'Connor and Hutchinson, 2000; Goldin et al., 2000; Chalmers, 2001; Tang and Lee, 2001; Nguyen and Wereley, 2002; Karniadakis and Beskok, 2002; Madou, 2002; DeGaspari, 2003; Ehrenman, 2004; Sharke, 2004; Stone et al., 2004; Squires and Quake, 2005). Electrostatic, magnetic, electromagnetic, pneumatic and thermal actuators, motors, valves, gears, cantilevers, diaphragms, and tweezers of less than 100 μm size have been fabricated. These have been used as sensors for pressure, temperature, mass flow, velocity, sound, and chemical composition, as actuators for linear and angular motions, and as simple components for complex systems, such as lab-on-a-chip, robots, micro-heat-engines and micro heat pumps (Lipkin, 1993; Garcia and Sniegowski, 1993, 1995; Sniegowski and Garcia, 1996; Epstein and Senturia, 1997; Epstein et al., 1997; Pekola et al., 2004; Squires and Quake, 2005).

Microelectromechanical systems (MEMS) refer to devices that have characteristic length of less than 1 mm but more than 1 micron, that combine electrical and mechanical components, and that are fabricated using integrated circuit batch-processing technologies. The books by Kovacs (1998) and Madou (2002) provide excellent sources for microfabrication technology. Current manufacturing techniques for MEMS include surface silicon micromachining; bulk silicon micromachining; lithography, electrodeposition, and plastic molding (or, in its original German, *Lithographie Galvanoformung Abformung, LIGA*); and electrodischarge machining (EDM). As indicated in Figure 1.1, MEMS are more than four orders of magnitude larger than the diameter of the hydrogen atom, but about four orders of magnitude smaller than the traditional manmade artifacts. Microdevices can have characteristic lengths smaller than the diameter of a human hair. Nanodevices (some say NEMS) further push the envelope of electromechanical miniaturization (Roco, 2001; Lemay et al., 2001; Feder, 2004).

The famed physicist Richard P. Feynman delivered a mere two, albeit profound, lectures[3] on electromechanical miniaturization: "There's Plenty of Room at the Bottom," quoted above, and "Infinitesimal Machinery," presented at the Jet Propulsion Laboratory on February 23, 1983. He could not see a lot of use for micromachines, lamenting in 1959 that "(small but movable machines) may or may not be useful, but they surely would be fun to make," and 24 years later said, "There is no use for these machines, so I still don't

[1] *Gulliver's Travels* were originally designed to form part of a satire on the abuse of human learning. At the heart of the story is a radical critique of human nature in which subtle ironic techniques work to part the reader from any comfortable preconceptions and challenge him to rethink from first principles his notions of man.

[2] The smallest feature on a microchip is defined by its smallest linewidth, which in turn is related to the wavelength of light employed in the basic lithographic process used to create the chip.

[3] Both talks have been reprinted in the *Journal of Microelectromechanical Systems*, vol. 1, no. 1, pp. 60–66, 1992, and vol. 2, no. 1, pp. 4–14, 1993.

understand why I'm fascinated by the question of making small machines with movable and controllable parts." Despite Feynman's demurring regarding the usefulness of small machines, MEMS are finding increased applications in a variety of industrial and medical fields with a potential worldwide market in the billions of dollars.

Accelerometers for automobile airbags, keyless entry systems, dense arrays of micromirrors for high-definition optical displays, scanning electron microscope tips to image single atoms, micro heat exchangers for cooling of electronic circuits, reactors for separating biological cells, blood analyzers, and pressure sensors for catheter tips are but a few of the current usages. Microducts are used in infrared detectors, diode lasers, miniature gas chromatographs, and high-frequency fluidic control systems. Micropumps are used for ink jet printing, environmental testing, and electronic cooling. Potential medical applications for small pumps include controlled delivery and monitoring of minute amount of medication, manufacturing of nanoliters of chemicals, and development of artificial pancreas. The much sought-after lab-on-a-chip is promising to automate biology and chemistry to the same extent the integrated circuit has allowed large-scale automation of computation. Global funding for micro- and nanotechnology research and development quintupled from $432 million in 1997 to $2.2 billion in 2002. In 2004, the U.S. National Nanotechnology Initiative had a budget of close to $1 billion, and the worldwide investment in nanotechnology exceeded $3.5 billion. In 10 to 15 years, it is estimated that micro- and nanotechnology markets will represent $340 billion per year in materials, $300 billion per year in electronics, and $180 billion per year in pharmaceuticals.

The multidisciplinary field has witnessed explosive growth during the past decade. Several new journals are dedicated to the science and technology of MEMS; for example *Journal of Microelectromechanical Systems*, *Journal of Micromechanics and Microengineering*, *Microscale Thermophysical Engineering*, *Microfluidics and Nanofluidics Journal*, *Nanotechnology Journal*, and *Journal of Nanoscience and Nanotechnology*. Numerous professional meetings are devoted to micromachines; for example Solid-State Sensor and Actuator Workshop, International Conference on Solid-State Sensors and Actuators (Transducers), Micro Electro Mechanical Systems Workshop, Micro Total Analysis Systems, and Eurosensors. Several web portals are dedicated to micro- and nanotechnology; for example, <http://www.smalltimes.com>, <http://www.emicronano.com>, <http://www.nanotechweb.org/>, and <http://www.peterindia.net/NanoTechnologyResources.html>.

The three-book *MEMS set* covers several aspects of microelectromechanical systems, or more broadly, the art and science of electromechanical miniaturization. MEMS design, fabrication, and application as well as the physical modeling of their materials, transport phenomena, and operations are all discussed. Chapters on the electrical, structural, fluidic, transport and control aspects of MEMS are included in the books. Other chapters cover existing and potential applications of microdevices in a variety of fields, including instrumentation and distributed control. Up-to-date new chapters in the areas of microscale hydrodynamics, lattice Boltzmann simulations, polymeric-based sensors and actuators, diagnostic tools, microactuators, nonlinear electrokinetic devices, and molecular self-assembly are included in the three books constituting the second edition of *The MEMS Handbook*. The 16 chapters in *MEMS: Introduction and Fundamentals* provide background and physical considerations, the 14 chapters in *MEMS: Design and Fabrication* discuss the design and fabrication of microdevices, and the 15 chapters in *MEMS: Applications* review some of the applications of microsensors and microactuators.

There are a total of 45 chapters written by the world's foremost authorities in this multidisciplinary subject. The 71 contributing authors come from Canada, China (Hong Kong), India, Israel, Italy, Korea, Sweden, Taiwan, and the United States, and are affiliated with academia, government, and industry. Without compromising rigorousness, the present text is designed for maximum readability by a broad audience having engineering or science background. As expected when several authors are involved, and despite the editor's best effort, the chapters of each book vary in length, depth, breadth, and writing style. The nature of the books — being handbooks and not encyclopedias — and the size limitation dictate the noninclusion of several important topics in the MEMS area of research and development.

Our objective is to provide a current overview of the fledgling discipline and its future developments for the benefit of working professionals and researchers. The three books will be useful guides and references to the explosive literature on MEMS and should provide the definitive word for the fundamentals and applications of microfabrication and microdevices. Glancing at each table of contents, the reader may rightly sense an overemphasis on the physics of microdevices. This is consistent with the strong conviction of the Editor-in-Chief that the MEMS technology is moving too fast relative to our understanding of the unconventional physics involved. This technology can certainly benefit from a solid foundation of the underlying fundamentals. If the physics is better understood, less expensive, and more efficient, microdevices can be designed, built, and operated for a variety of existing and yet-to-be-dreamed applications. Consistent with this philosophy, chapters on control theory, distributed control, and soft computing are included as the backbone of the futuristic idea of using colossal numbers of microsensors and microactuators in reactive control strategies aimed at taming turbulent flows to achieve substantial energy savings and performance improvements of vehicles and other manmade devices.

I shall leave you now for the many wonders of the small world you are about to encounter when navigating through the various chapters of these volumes. May your voyage to Lilliput be as exhilarating, enchanting, and enlightening as Lemuel Gulliver's travels into "Several Remote Nations of the World." *Hekinah degul!* Jonathan Swift may not have been a good biologist and his scaling laws were not as good as those of William Trimmer (see Chapter 2 of *MEMS: Introduction and Fundamentals*), but Swift most certainly was a magnificent storyteller. *Hnuy illa nyha majah Yahoo!*

References

Amato, I. (1998) "Formenting a Revolution, in Miniature," *Science* **282**, no. 5388, 16 October, pp. 402–405.

Angell, J.B., Terry, S.C., and Barth, P.W. (1983) "Silicon Micromechanical Devices," *Faraday Transactions I* **68**, pp. 744–748.

Ashley, S. (1996) "Getting a Microgrip in the Operating Room," *Mech. Eng.* **118**, September, pp. 91–93.

Bryzek, J., Peterson, K., and McCulley, W. (1994) "Micromachines on the March," *IEEE Spectrum* **31**, May, pp. 20–31.

Busch-Vishniac, I.J. (1998) "Trends in Electromechanical Transduction," *Phys. Today* **51**, July, pp. 28–34.

Chalmers, P. (2001) "Relay Races," *Mech. Eng.* **123**, January, pp. 66–68.

DeGaspari, J. (2003) "Mixing It Up," *Mech. Eng.* **125**, August, pp. 34–38.

Ehrenman, G. (2004) "Shrinking the Lab Down to Size," *Mech. Eng.* **126**, May, pp. 26–29.

Epstein, A.H. (2000) "The Inevitability of Small," *Aerospace Am.* **38**, March, pp. 30–37.

Epstein, A.H., and Senturia, S.D. (1997) "Macro Power from Micro Machinery," *Science* **276**, 23 May, p. 1211.

Epstein, A.H., Senturia, S.D., Al-Midani, O., Anathasuresh, G., Ayon, A., Breuer, K., Chen, K.-S., Ehrich, F.F., Esteve, E., Frechette, L., Gauba, G., Ghodssi, R., Groshenry, C., Jacobson, S.A., Kerrebrock, J.L., Lang, J.H., Lin, C.-C., London, A., Lopata, J., Mehra, A., Mur Miranda, J.O., Nagle, S., Orr, D.J., Piekos, E., Schmidt, M.A., Shirley, G., Spearing, S.M., Tan, C.S., Tzeng, Y.-S., and Waitz, I.A. (1997) "Micro-Heat Engines, Gas Turbines, and Rocket Engines — The MIT Microengine Project," AIAA Paper No. 97-1773, AIAA, Reston, Virginia.

Feder, T. (2004) "Scholars Probe Nanotechnology's Promise and Its Potential Problems," *Phys. Today* **57**, June, pp. 30–33.

Feynman, R.P. (1961) "There's Plenty of Room at the Bottom," in *Miniaturization*, H.D. Gilbert, ed., pp. 282–296, Reinhold Publishing, New York.

Gabriel, K.J. (1995) "Engineering Microscopic Machines," *Sci. Am.* **260**, September, pp. 150–153.

Gabriel, K.J., Jarvis, J., and Trimmer, W., eds. (1988) *Small Machines, Large Opportunities: A Report on the Emerging Field of Microdynamics, National Science Foundation*, published by AT&T Bell Laboratories, Murray Hill, New Jersey.

Gabriel, K.J., Tabata, O., Shimaoka, K., Sugiyama, S., and Fujita, H. (1992) "Surface-Normal Electrostatic/Pneumatic Actuator," in *Proc. IEEE Micro Electro Mechanical Systems '92*, pp. 128–131, 4–7 February, Travemünde, Germany.

Garcia, E.J., and Sniegowski, J.J. (1993) "The Design and Modelling of a Comb-Drive-Based Microengine for Mechanism Drive Applications," in *Proc. Seventh International Conference on Solid-State Sensors and Actuators (Transducers '93)*, pp. 763–766, Yokohama, Japan, 7–10 June.

Garcia, E.J., and Sniegowski, J.J. (1995) "Surface Micromachined Microengine," *Sensor. Actuator. A* **48**, pp. 203–214.

Goldin, D.S., Venneri, S.L., and Noor, A.K. (2000) "The Great out of the Small," *Mech. Eng.* **122**, November, pp. 70–79.

Gravesen, P., Branebjerg, J., and Jensen, O.S. (1993) "Microfluidics — A Review," *J. Micromech. Microeng.* **3**, pp. 168–182.

Ho, C.-M., and Tai, Y.-C. (1996) "Review: MEMS and Its Applications for Flow Control," *J. Fluids Eng.* **118**, pp. 437–447.

Ho, C.-M., and Tai, Y.-C. (1998) "Micro–Electro–Mechanical Systems (MEMS) and Fluid Flows," *Annu. Rev. Fluid Mech.* **30**, pp. 579–612.

Hogan, H. (1996) "Invasion of the Micromachines," *New Sci.* **29**, June, pp. 28–33.

Karniadakis, G.E., and Beskok A. (2002) *Microflows: Fundamentals and Simulation*, Springer-Verlag, New York.

Knight, J. (1999) "Dust Mite's Dilemma," *New Sci.* **162**, no. 2180, 29 May, pp. 40–43.

Kovacs, G.T.A. (1998) *Micromachined Transducers Sourcebook*, McGraw-Hill, New York.

Lemay, S.G., Janssen, J.W., van den Hout, M., Mooij, M., Bronikowski, M.J., Willis, P.A., Smalley, R.E., Kouwenhoven, L.P., and Dekker, C. (2001) "Two-Dimensional Imaging of Electronic Wavefunctions in Carbon Nanotubes," *Nature* **412**, 9 August, pp. 617–620.

Lipkin, R. (1993) "Micro Steam Engine Makes Forceful Debut," *Sci. News* **144**, September, p. 197.

Madou, M. (2002) *Fundamentals of Microfabrication*, second edition, CRC Press, Boca Raton, Florida.

Nguyen, N.-T., and Wereley, S.T. (2002) *Fundamentals and Applications of Microfluidics*, Artech House, Norwood, Massachusetts.

O'Connor, L. (1992) "MEMS: Micromechanical Systems," *Mech. Eng.* **114**, February, pp. 40–47.

O'Connor, L., and Hutchinson, H. (2000) "Skyscrapers in a Microworld," *Mech. Eng.* **122**, March, pp. 64–67.

Ouellette, J. (1996) "MEMS: Mega Promise for Micro Devices," *Mech. Eng.* **118**, October, pp. 64–68.

Ouellette, J. (2003) "A New Wave of Microfluidic Devices," *Ind. Phys.* **9**, no. 4, pp. 14–17.

Paula, G. (1996) "MEMS Sensors Branch Out," *Aerospace Am.* **34**, September, pp. 26–32.

Pekola, J., Schoelkopf, R., and Ullom, J. (2004) "Cryogenics on a Chip," *Phys. Today* **57**, May, pp. 41–47.

Robinson, E.Y., Helvajian, H., and Jansen, S.W. (1996a) "Small and Smaller: The World of MNT," *Aerospace Am.* **34**, September, pp. 26–32.

Robinson, E.Y., Helvajian, H., and Jansen, S.W. (1996b) "Big Benefits from Tiny Technologies," *Aerospace Am.* **34**, October, pp. 38–43.

Roco, M.C. (2001) "A Frontier for Engineering," *Mech. Eng.* **123**, January, pp. 52–55.

Sharke, P. (2004) "Water, Paper, Glass," *Mech. Eng.* **126**, May, pp. 30–32.

Sniegowski, J.J., and Garcia, E.J. (1996) "Surface Micromachined Gear Trains Driven by an On-Chip Electrostatic Microengine," *IEEE Electron Device Lett.* **17**, July, p. 366.

Squires, T.M., and Quake, S.R. (2005) "Microfluidics: Fluid Physics at the Nanoliter Scale," *Rev. Mod. Phys.* **77**, pp. 977–1026.

Stone, H.A., Stroock, A.D., and Ajdari, A. (2004) "Engineering Flows in Small Devices: Microfluidics Toward a Lab-on-a-Chip," *Annu. Rev. Fluid Mech.* **36**, pp. 381–411.

Swift, J. (1726) *Gulliver's Travels*, 1840 reprinting of *Lemuel Gulliver's Travels into Several Remote Nations of the World*, Hayward & Moore, London, Great Britain.

Tang, W.C., and Lee, A.P. (2001) "Military Applications of Microsystems," *Ind. Phys.* **7**, February, pp. 26–29.

Tien, N.C. (1997) "Silicon Micromachined Thermal Sensors and Actuators," *Microscale Thermophys. Eng.* **1**, pp. 275–292.

2

Scaling of Micromechanical Devices

William Trimmer
Belle Mead Research, Inc.

Robert H. Stroud
The Aerospace Corporation

2.1 Introduction

A revolution in understanding and utilizing micromechanical devices is starting. The utility of these devices will be enormous, and with time they will fill the niches of our lives as pervasively as electronics. What form will these microdevices take? What will actuate them, and how will they interact with their environment? We cannot foresee where the developing technology will take us.

How, then, do we start to design this world of the micro? As you will discover in this book, there are a large number of ways to fabricate microdevices and a vast number of designs. The number of options is greater than we could possibly pursue. Should we just start trying different approaches until something works? Perhaps there is a better way.

Scaling theory is a valuable guide to what may work and what will not. By understanding how phenomena behave and change as their scale size changes, we can gain some insight and better understand the profitable approaches. This chapter examines how things change with size, and it will develop a mathematics that helps find the profitable approaches.

Three general scale sizes will be discussed: astronomical objects; the normal objects we deal with, called macro-objects; and very small objects, called micro-objects. Things that are effective at one of these scale sizes often are insignificant at another scale size. As an example, gravitational forces dominate on an astronomical scale. The motions of our planet around the sun and of our sun around the galaxy are driven mostly by gravitational forces. Yet on the macroscale of my desk top, the gravitational force between two objects such as my tape dispenser and stapler is insignificant. A few simple scaling calculations later in this chapter will tell us this: on astronomical scales, be concerned with gravity; on smaller scales, look to other forces to move objects.

What is obvious on an astronomical-scale size or on a macroscale size is often not obvious on the microscale. For example, take the case of an electric motor. It is really a magnetic motor, and almost all macrosized electric actuators use magnetic fields to generate forces. Hence, one's first intuition would be to use magnetic motors in designing microdevices. In practice, however, most of the common micromotor designs use electrostatic fields instead of magnetic fields. The reasons for this will become obvious in the following discussion of how forces scale.

The field of micromechanical devices is extremely broad. It encompasses all of the traditional science and engineering disciplines, only on a smaller scale. Try to think of a traditional science or engineering discipline that does not have a microequivalent. What we are about in our new discipline is replicating the macroscience and macroengineering on a microscale. As a result, technical people from all science and engineering disciplines can make important contributions to this field.

The time scale from conception to utilization has been collapsing. Alessandro Volta and Andre Marie Ampere developed the basic concepts of electricity, and about 100 years later, Nikola Tesla and Thomas Alva Edison developed practical electric generators and motors. In contrast, the micro-comb-drive motor was described in 1989 and currently is being used in automobiles as an airbag sensor [Tang et al., 1989]. Volta and Ampere's ideas took 100 years to culminate in practical implementation, but the micro-comb-drive motor took less than a dozen years from conception to full-scale implementation.

One of the marvelous things about nature is its widely varying scale sizes. Section 2.2 will discuss this broad range of scales. Section 2.3 will show how scaling theory can be used as a guide to understand how phenomena change with scale. We hope the material that follows encourages you to explore the broad scope of this new field.

2.2 The Log Plot

As the scale, or size, of a system changes by several orders of magnitude, the system tends to behave differently. Consider, for example, a glass of water that is about 5 cm on a side. Pour the glass of water onto a table and notice how the water flows and runs off the edge of the table. If the size of the glass is decreased by two orders of magnitude, or a factor of 100, the glass is now 0.05 cm (or 0.5 mm) on a side. Pour this glass onto the table and see how the surface tension pulls the water into a drop that sticks to the table. Turn the table on its side and observe that it is difficult to make the drop flow to the edge of the table. In each case, the substance is the same, water, and the table is the same, but changing the water's scale size makes it behave very differently.

Decreasing the size of the glass by another two orders of magnitude, the glass is now 0.0005 cm, or 5 μm, on a side. If you try to pour a drop this size onto the table, it most likely will not even reach the table. Some air current will entrain the drop and carry it away like mist flowing through the city at night. Again, the behavior of the water is dramatically different because of its size. Even the act of pouring the glass over the table is different. The 5 cm glass pours, whereas water in the 0.05 cm and 0.0005 cm glasses is constrained by surface tension. Different physical effects manifest themselves differently because of the system size.

Figure 2.1 shows the full range of sizes available to us, from atoms to the universe. Atoms are the smallest mechanical system we will manipulate in the near future; their size is several angstroms (10^{-10} m). The universe is the largest mechanical system we can observe. Depending upon the particular astronomical theory, the universe is about 10^{37} m in diameter. Hence, the full range available for us to investigate and use is about 10^{47} m, or 47 orders of magnitude.

The horizontal axis in Figure 2.1 represents the size of the system. The short vertical lines in the center of the plot represent a factor-of-10 change in the system size. The long vertical lines represent a change of 100,000, or five orders of magnitude. Along the top, the size of the system is given in meters, and in the central band the size of the system is given in angstroms. Figure 2.1 is plotted as a log plot for two reasons: (1) to enable everything to be depicted on the same piece of paper, and (2) to easily portray the different size domains.

One can get a sense of the size of things by looking at the ant, the human, and the whale. These familiar objects span about five orders of magnitude. Several orders of magnitude smaller than the ant are bacteria and viruses. Going to larger systems, the U.S. road system is about five orders of magnitude larger than the whale, and the earth's orbit is about five orders of magnitude larger than the U.S. road system. Increasing another five or six orders of magnitude brings us to interstellar distances.

The bottom portion of Figure 2.1 shows the units we use to measure things. The angstrom, micron, millimeter, meter, kilometer, and mile are familiar units, but then we see a gap of about a dozen orders of magnitude before we reach the astronomical units of the light year and parsec.

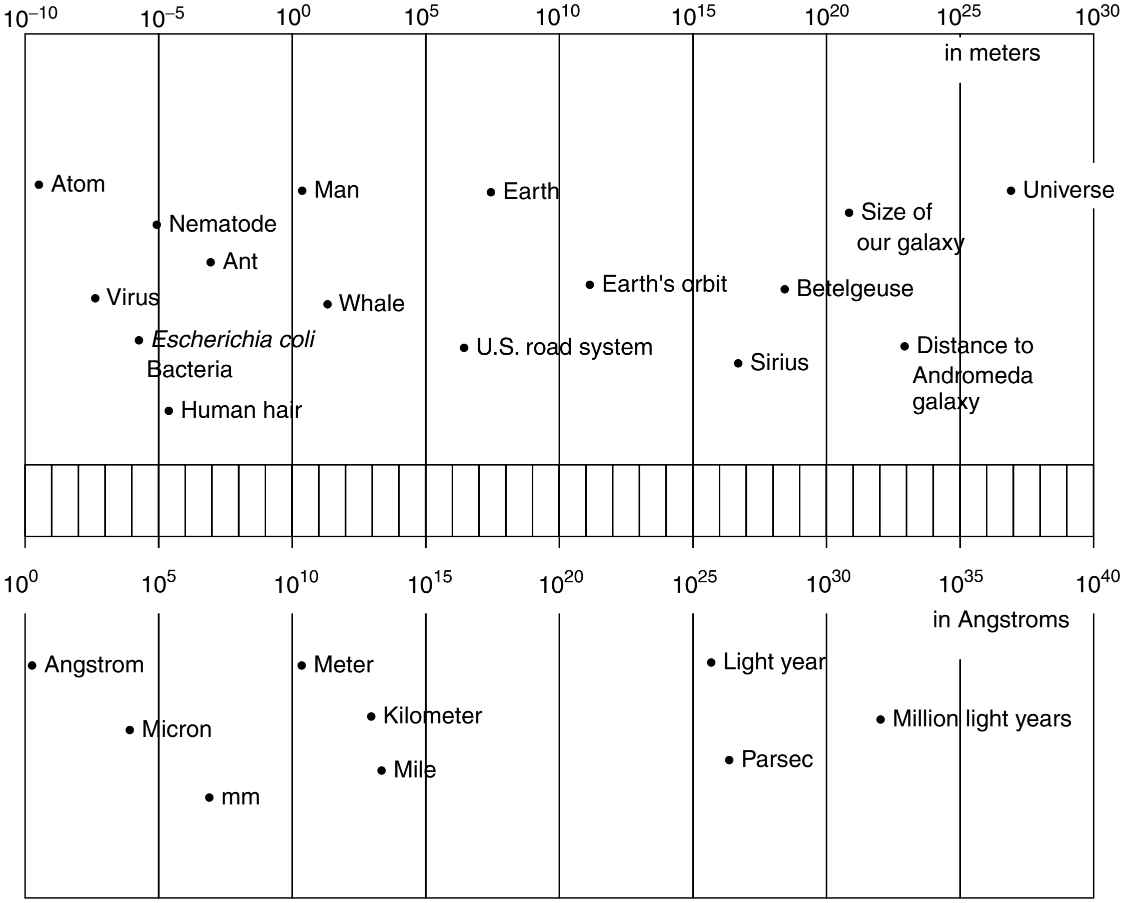


FIGURE 2.1 Log plot of all mechanical systems available for exploration.

The microregion of interest to this chapter ranges from about a millimeter to an angstrom (from about 10^{-3} to 10^{-10} meters). This region comprises roughly a fifth of the full range of domains available for us to explore and may seem like a small portion, but consider that the U.S. roadway system is one of the largest mechanical systems we will build for quite a while. Buildings and ships are probably the largest self-contained mechanical systems we will construct in the near future. Most of the larger domains are so large that they simply are not accessible to us. Thus, the microregion represents the majority of the new domains available for exploration.

This microdomain is enticing. Part of its charm is that conventional designs do not work well, and ingenuity is needed to make new designs. For example, macrodevices and microdevices that transfer water tend to use different physical principles. An open ditch works at one scale, and a capillary works at a smaller scale. Because microdesigners are left without the conventional solutions, they have the opportunity to find their own solutions.

2.3 Scaling of Mechanical Systems

As the size of a system changes, its physical parameters also change, often in a dramatic way [Trimmer et al., 1989; Madou, 1997]. To understand how these parameters change, consider the scale factor S. This scale factor is similar to the small notation on the corner of a mechanical drawing that might say the scale of the drawing is 1:10. The actual object to be made is 10 times the size of the drawing. A scale of 1:100 means the actual object is 100 times larger. In the microdomain, the scale might be 100:1, meaning the object is 100 times smaller than the drawing. When the scale size changes, all the dimensions of the object change by exactly the same amount S such that 1:S.

This scale factor S can be used to describe how physical phenomena change. All the lengths of the drawing scale by the factor S, but other parameters such as the volume scale differently. Volume V is length L times width W times height H, or

$$V = L \cdot W \cdot H \tag{2.1}$$

When the scale changes by 1/100 (that is, decreases by a factor of 100), the length and width and height all change by 1/100, and the volume decreases by $(1/100)^3$ or 1/1,000,000. The volume decreases by a factor of a million when the scale size decreases by a factor of a hundred. Volume is an example of a parameter that scales as S^3. The force due to surface tension scales as S^1; the force due to electrostatics scales as S^2; the force due to certain magnetic forces scales as S^3; and gravitational forces scale as S^4. Now, if the size of the system decreases from a meter to a millimeter, this is a decrease of a factor of a thousand, $S = 1/1000$. The surface tension force decreases by a factor of a thousand, $S^1 = (1/1000)^1$; the electrostatic force decreases by a factor of a million, $S^2 = (1/1000)^2 = 1/1{,}000{,}000$; the magnetic force decreases by a factor of a billion, $S^3 = (1/1000)^3 = 1/1{,}000{,}000{,}000$; and the gravitational force decreases by a factor of a trillion, $S^4 = (1/1000)^4 = 1/1{,}000{,}000{,}000{,}000$. Indeed, changing the size of a mechanical system changes which forces are important.

Knowing how a physical phenomenon scales, whether as S^1 or S^2 or S^3 or S^4 or some other power of S, guides our understanding of how to design small mechanical systems. As an example, consider a water bug. The weight of the water bug scales as the volume, or S^3, while the force used to support the bug scales as the surface tension (S^1) times the distance around the bug's foot (S^1), and the force on the bug's foot scales as $S^1 \times S^1 = S^2$. When the scale size, S, decreases, the weight decreases more rapidly than the surface tension forces. Changing from a 2 m man to a 2 mm bug decreases the weight by a factor of a billion while the surface tension force decreases by a factor of only a million. Hence, the bug can walk on water.

Scaling provides a good guide to how things behave and offers insight into small systems, but scaling is just that — a good guide. It usually does not provide exact solutions. For example, the scaling above does not take into account the difference between the water bug's foot and a person's foot. Water bug's feet are designed for water, and we expect superior performance. Creativity and intuition are what make an excellent design; scaling is a guide to understanding which design elements are important.

A mathematical notation captures the different scaling laws in a convenient form. This notation shows many different scaling laws at once and can be used to easily understand what happens to the different terms and parameters of an equation as the scale size changes.

Consider four different force laws, $F = S^1$, $F = S^2$, $F = S^3$, $F = S^4$, and collect these different cases into a vertical Trimmer bracket:

$$F = \begin{bmatrix} S^1 \\ S^2 \\ S^3 \\ S^4 \end{bmatrix} \tag{2.2}$$

The topmost element of this bracket refers to the case where the force scales as S^1, the next element down refers to the case where the force scales as S^2, and so on.

To continue, let us do something with this bracket. Work W is force F times distance D, or

$$W = F \cdot D \tag{2.3}$$

and, extending our notation,

$$W = F \cdot D = \begin{bmatrix} S^1 \\ S^2 \\ S^3 \\ S^4 \end{bmatrix} \begin{bmatrix} S^1 \\ S^1 \\ S^1 \\ S^1 \end{bmatrix} = \begin{bmatrix} S^2 \\ S^3 \\ S^4 \\ S^5 \end{bmatrix} \tag{2.4}$$

or

$$W = \begin{bmatrix} S^2 \\ S^3 \\ S^4 \\ S^5 \end{bmatrix} \tag{2.5}$$

Note that distance D always scales as S^1, and its bracket consists of all S^1's. In the top case where the force scales as S^1, the distance scales as S^1, and their product scales as S^2. In the second element down, the force scales as S^2, the distance scales as S^1, and their product scales as S^3. Here in one notation we have shown how the work scales for the four different force laws. For example, the gravitational force between an object and the earth scales as S^3 (the earth's mass remains constant in this example, and the mass of the object scales as its volume, S^3). Looking at the third element down, we see that a force scaling of S^3 gives us a work, or energy, scaling of S^4. If the size of a system decreases by a factor of a thousand (say, from 10 cm to 0.10 mm), the gravitational energy required to move an object from the bottom to the top of a machine under consideration decreases by $(1/1000)^4 = 1/1{,}000{,}000{,}000{,}000$. The gravitational work decreases by a factor of a trillion. We know this intuitively: drop an ant from ten times its height, and it walks away. Please do not try this with a horse.

How do the acceleration and transit times change for the different force-scaling laws? Acceleration a is equal to force F divided by the mass m:

$$a = \frac{F}{m} = F \cdot m^{-1} \tag{2.6}$$

and we know the mass scales as S^3, and m^{-1} scales as S^{-3}, giving:

$$a = \begin{bmatrix} S^1 \\ S^2 \\ S^3 \\ S^4 \end{bmatrix} \begin{bmatrix} S^3 \\ S^3 \\ S^3 \\ S^3 \end{bmatrix}^{-1} = \begin{bmatrix} S^1 \\ S^2 \\ S^3 \\ S^4 \end{bmatrix} \begin{bmatrix} S^{-3} \\ S^{-3} \\ S^{-3} \\ S^{-3} \end{bmatrix} = \begin{bmatrix} S^{-2} \\ S^{-1} \\ S^0 \\ S^1 \end{bmatrix} \tag{2.7}$$

This is an interesting result. When the force scales as S^1, the acceleration scales as S^{-2}. If the size of the system decreases by a factor of 100, the acceleration increases by $(1/100)^{-2} = 10{,}000$. As the system becomes smaller, the acceleration increases. A predominance of the forces we use in the microdomain scales as S^2. For these forces, the acceleration scales as S^{-1}, and decreasing the size by a factor of 100 increases the acceleration by a factor of 100, still a nice increase in acceleration. In general, small systems tend to accelerate very rapidly. Where the force scales as S^3, the acceleration remains constant, $(1/100)^0 = 1$, and the acceleration decreases for forces that scale as S^4.

The transit time t to move from point A to B in our scalable drawing can be calculated as:

$$x = \frac{1}{2} a t^2 \tag{2.8}$$

$$t = \sqrt{\frac{2x}{a}} = \sqrt{2} \cdot (x)^{0.5} \cdot (a)^{-0.5} \tag{2.9}$$

and

$$t = \begin{bmatrix} S^0 \\ S^0 \\ S^0 \\ S^0 \end{bmatrix} \begin{bmatrix} S^1 \\ S^1 \\ S^1 \\ S^1 \end{bmatrix}^{0.5} \begin{bmatrix} S^{-2} \\ S^{-1} \\ S^0 \\ S^1 \end{bmatrix}^{-0.5} = \begin{bmatrix} S^0 \\ S^0 \\ S^0 \\ S^0 \end{bmatrix} \begin{bmatrix} S^{0.5} \\ S^{0.5} \\ S^{0.5} \\ S^{0.5} \end{bmatrix} \begin{bmatrix} S^1 \\ S^{0.5} \\ S^0 \\ S^{-0.5} \end{bmatrix} = \begin{bmatrix} S^{1.5} \\ S^1 \\ S^{0.5} \\ S^0 \end{bmatrix} \tag{2.10}$$

$$t = \begin{bmatrix} S^{1.5} \\ S^1 \\ S^{0.5} \\ S^0 \end{bmatrix} \tag{2.11}$$

For the case where the force scales as S^2, transit time t scales as S^1. If the system decreases by a factor of 100, the transit time decreases by a factor of 100. Again, we know this intuitively; small things tend to be fast.

Depending upon the equation and variables of interest, the Trimmer brackets can be configured differently. To continue the above example, we might be interested in how things will behave if the acceleration instead of the force scales in different ways. We could write:

$$a = \begin{bmatrix} S^1 \\ S^2 \\ S^3 \\ S^4 \end{bmatrix} \tag{2.12}$$

From above:

$$t = \sqrt{\frac{2x}{a}} = \sqrt{2} \cdot (x)^{0.5} \cdot (a)^{-0.5} \tag{2.13}$$

and

$$t = \begin{bmatrix} S^0 \\ S^0 \\ S^0 \\ S^0 \end{bmatrix} \begin{bmatrix} S^1 \\ S^1 \\ S^1 \\ S^1 \end{bmatrix}^{0.5} \begin{bmatrix} S^1 \\ S^2 \\ S^3 \\ S^4 \end{bmatrix}^{-0.5} = \begin{bmatrix} S^0 \\ S^0 \\ S^0 \\ S^0 \end{bmatrix} \begin{bmatrix} S^{0.5} \\ S^{0.5} \\ S^{0.5} \\ S^{0.5} \end{bmatrix} \begin{bmatrix} S^{-0.5} \\ S^{-1} \\ S^{-1.5} \\ S^{-2} \end{bmatrix} = \begin{bmatrix} S^0 \\ S^{-0.5} \\ S^{-1} \\ S^{-1.5} \end{bmatrix} \tag{2.14}$$

$$t = \begin{bmatrix} S^0 \\ S^{-0.5} \\ S^{-1} \\ S^{-1.5} \end{bmatrix} \tag{2.15}$$

The top element in this bracket describes how time scales when the acceleration scales as S^1. (In the earlier discussion, the top element describes how time scales when the force scales as S^1.) We can arrange these brackets to fit the problem at hand. We need not even use integer exponents. For example, we could have defined the acceleration as:

$$a = \begin{bmatrix} S^{0.1} \\ S^{0.2} \\ S^{0.3} \\ S^2 \\ S^4 \end{bmatrix} \tag{2.16}$$

and then calculated the transit times for these five new scaling functions.

Let us examine the gravitational example in the introduction to this chapter. As we will see in a moment, gravitational forces scale as S^4 and are a dominant force in large systems but not in small systems. The force between two objects is

$$F = G\frac{M_1 \cdot M_2}{r^2} \tag{2.17}$$

where F is the force; G is the gravitational constant ($= 6.670 \times 10^{-11}$ N m^{2}kg^{-2}), which does not change with scale size; M_1 and M_2 are the masses of the two objects; and r is the separation. The masses scale as:

$$M = \rho \cdot V = S^0 \cdot S^3 = S^3 \tag{2.18}$$

where the density ρ is assumed constant (S^0), and V is the volume (S^3). Now force F scales as:

$$F = S^0\frac{S^3 \cdot S^3}{S^2} = S^4 \tag{2.19}$$

Now, let us make a different assumption and suppose the density is not constant with scale size. The density could be represented as:

$$\rho = \begin{bmatrix} S^0 \\ S^{-1} \\ S^{-2} \\ S^{-3} \end{bmatrix} \tag{2.20}$$

and force *F* becomes:

$$F = G\frac{M_1 \cdot M_2}{r^2} = G\frac{\rho \cdot V_1 \cdot \rho V_2}{r^2} = G \cdot \rho^2 \cdot V_1 \cdot V_2 \cdot R^{-2} \tag{2.21}$$

$$F = S^0 \begin{bmatrix} S^0 \\ S^{-1} \\ S^{-2} \\ S^{-3} \end{bmatrix} S^3 S^3 S^{-2} = S^0 \begin{bmatrix} S^0 \\ S^{-2} \\ S^{-4} \\ S^{-6} \end{bmatrix} S^3 S^3 S^{-2} = \begin{bmatrix} S^4 \\ S^2 \\ S^0 \\ S^{-2} \end{bmatrix} \tag{2.22}$$

From the top element, where the density does not change with size, the force scales as S^4. From the third element down, when the density scales as S^{-2}, the gravitational force remains constant as the scale size changes. That is, if astronomical objects become less dense as they become larger (as $\rho = S^{-2}$), then the gravitational force between objects remains constant ($F = S^0$) among differently sized astronomical systems.

It is useful to understand how different forces scale. A more complete listing of forces and their scaling is given below,

$$F = \begin{bmatrix} S^1 \\ S^2 \\ S^3 \\ S^4 \end{bmatrix} = \begin{bmatrix} \text{Surface tension} \\ \text{Electrostatic, Pressure, Biological, Magnetic } (J = S^{-1}) \\ \text{Magnetic } (J = S^{-0.5}) \\ \text{Gravitational, Magnetic } (J = S^0) \end{bmatrix} \tag{2.23}$$

Surface tension has the propitious scaling of S^1 and increases rapidly relative to other forces as a system becomes smaller; however, changing the surface tension usually requires changing the temperature, adding a surfactant, or altering some other parameter that is usually difficult to control. Most forces currently used by microdesigners scale as S^2. These include electrostatic forces, forces generated by pressures, and biological forces (the force an animal can exert generally depends upon the cross-section of the muscle). How magnetic forces scale depends upon how the current density (current per unit area of the coils) scales. If the current density J in the coil remains constant (S^0), the magnetic force between two coils scales as S^4, and in this case the magnetic forces become weak in the microdomain; however, we can remove heat much more efficiently from a small volume, and the current density of a microcoil can be much higher than in a large coil. If the current density scales as S^{-1} when the system decreases by a factor of ten, the current density increases by a factor of ten. In this case, the coil has much higher resistive losses, but the force scales much more advantageously as S^2.

References

Madou, M. (1997) *Fundamentals of Microfabrication*, CRC Press, Boca Raton, pp. 405–12.

Tang, W.C., Nguyen, T.-C.H., and Howe, R.T. (1989) "Laterally Driven Polysilicon Resonant Microstructures," *Proceedings of the IEEE Micro Electro Mechanical Systems Workshop*, February 1989; reprinted in *Micromechanics and MEMS: Classic and Seminal Papers to 1990*, W. Trimmer, ed., Institute of Electrical and Electronics Engineers, New York, 1997, pp. 187–93.

Trimmer, W.S.N.T. (1989) "Microrobots and Micromechanical Systems," *Sensor. Actuator.*, September; reprinted in *Micromechanics and MEMS: Classic and Seminal Papers to 1990*, W. Trimmer, ed., Institute of Electrical and Electronics Engineers, New York, 1997, pp. 96–116.

3

Mechanical Properties of MEMS Materials

William N. Sharpe, Jr.
The Johns Hopkins University

3.1 Introduction

New technologies tend to originate with new materials and manufacturing processes that are used to create new products. In the early stages, the emphasis is on novel devices and systems as well as on ways of making them. Studies of fundamental issues such as mechanical properties and design procedures come later. For example, in 1830 there were 23 miles of railroad track in the U.S., and by 1870 there were 53,000 miles of track. The Bessemer steelmaking process, however, did not originate until 1856, and the American Society for Testing and Materials (ASTM) was not organized until 1898.

The same is true for microelectromechanical systems (MEMS). The emphasis over the past dozen or so years has been on new materials, new manufacturing processes, and new microdevices — and rightfully so. These technological advances have been paralleled by an increasing interest in mechanical testing of materials used in MEMS. More researchers are becoming involved, with the topic appearing in symposia sponsored by the Society for Experimental Mechanics, the American Society of Mechanical Engineers, and the Materials Research Society. Further, the November 2000 ASTM symposium, Mechanical Testing of Structural Films, was an important first step toward standardization of test methods. This increase in MEMS material testing has occurred over the past ten or so years, and this chapter is a review of the current status of the field.

Mechanical properties of interest fall into three general categories: elastic, inelastic, and strength. The designer of a microdevice needs to know its elastic properties in order to predict the amount of deflection from an applied force, or vice versa. If the material is ductile and the deformed structure does not need to return to its initial state, then the designer must know the device's inelastic behavior. The strength of the material must be known so that the allowable operating limits can be set. The manufacturer of a MEMS device needs to understand the relation between the processing and the properties of the material.

The importance of mechanical properties was recognized early on by a leader in the MEMS field, Richard Muller, who wrote in 1990, "Research on the mechanical properties of the electrical materials forming microdynamic structures (which previously had exclusively electrical uses), on the scaling of mechanical design, and on the effective uses of computer aids is needed to provide the engineering base that will make it possible to exploit fully this technology" [Muller, 1990]. Later, expanded conclusions and recommendations were made in a 1997 report of a National Research Council committee that Muller chaired [Muller, 1997]. One conclusion was, "Test-and-characterization methods and metrologies are required to (1) help fabrication facilities define MEMS materials for potential users, (2) facilitate consistent evaluations of material and process properties at the required scales, and (3) provide a basis for comparisons among materials fabricated at different facilities." One recommendation was, "Studies that address fundamental mechanical properties (e.g., Young's modulus, fatigue strength, residual stress, internal friction) and the engineering physics of long-term reliability, friction, and wear are vitally needed." These rather obvious statements apply to the development of any new technology.

There have been other reviews of the topic. The first on freestanding thin films by Menter and Pashley (1959) is interesting from a historical point of view. This author reviewed existing techniques and introduced new ones in 1996 [Sharpe et al., 1996] and looked at the variation in the mechanical properties of polysilicon as tested by several researchers in 1997 [Sharpe et al., 1997b]. Obermeier (1996) reviewed test methods for mechanical and thermophysical properties. Ballarini (1998) prepared a report for the Air Force Research Laboratory that reviewed pertinent experimental and theoretical work up until then. Yi and Kim (1999a) published a review article, "Measurement of Mechanical Properties for MEMS Materials," on just this topic. Schweitz and Ericson (1999) reviewed the state of the art and offered some interesting conclusions and advice. Chang and Sharpe (1999) wrote an introductory chapter on the subject, and Spearing (2000) wrote a comprehensive exposition from a materials aspect.

This chapter is intended to be a comprehensive survey focusing on both the test methods and the properties that have been measured. After briefly defining the mechanical properties of interest, the chapter reviews the current test methods for MEMS materials. Then, a comprehensive set of tables summarizes the properties of the various materials. In almost all cases, these properties are not yet firmly established with the confidence typical in a handbook, so a final table of initial design values completes the chapter as an aid to initial consideration and design of MEMS.

If the reader is interested in the experimental methods, then the review of test methods will lead to the appropriate references. If the reader desires details about mechanical properties of specific materials, then the tables and the references will prove useful. Finally, if the reader wants to know only the typical properties for an initial design concept, the last section provides a succinct answer.

3.2 Mechanical Property Definitions

The properties of interest here are material properties; that is, the measured value is independent of the test method. Implicit is the understanding that the property is also independent of the size of the specimen, but that may not necessarily be the case for MEMS materials. The fabrication process for, say, thin-film silicon carbide is completely different from that of bulk silicon carbide, and it is reasonable to expect different mechanical behavior. The question of specimen-size effect needs to be considered at the appropriate length scale — in this case, whether a 200 × 200 μm cross-section tensile specimen behaves the same way as a 2 × 2 μm specimen. That question is not very easy to answer until test methods exist with sufficient sensitivity and reproducibility to differentiate the material behavior.

The American Society for Testing and Materials defines standard test procedures through a lengthy process of draft and review. Many of the common standards for structural materials were set in the early part of the twentieth century; however, new standards are established to meet the demands of new technologies, and a complete set of standards is issued each year. For example, the field of fracture mechanics as a usable measure of material and structural response emerged in the early 1950s. The first draft of a standard measure of fracture toughness did not appear until 1965, with the first complete standard appearing in

1970 [ASTM, 1970]. It will be some time before standards for measuring the mechanical properties of MEMS materials are established, but it is useful to be guided by the accepted definitions of mechanical properties. The pertinent standards for testing the mechanical properties of metals appear in [ASTM, 2000a] and those for ceramics in [ASTM, 2000b]. ASTM Standard E-8 gives directions for tension testing of metals, while E-9 covers compression testing. ASTM Standards C-1273 and C-1161 cover the tension and creep testing of ceramics. Once the stress–strain curve is obtained, various approximations, or curve-fits, can be used to insert the material behavior into the design process.

Young's modulus is the slope of the linear part of the stress–strain curve of a material; it is a measure of its stiffness. ASTM E-111 specifies that, "The test specimen is loaded uniaxially and load and strain are measured, either incrementally or continuously." It goes on to prescribe how the slope is determined along with a myriad of other details. Poisson's ratio is a measure of the lateral contraction or expansion of a material when subjected to an axial stress within the elastic region. ASTM E-132 requires that, "In this test method, the value of Poisson's ratio is obtained from strains resulting from uniaxial stress only." Note that these elastic properties are defined for isotropic materials only. Neither of these is easy to measure at the MEMS material size scale, as will be seen in the next section. When a material is inelastic (and nonlinear), we need the complete stress–strain curve to specify the material's behavior.

The strength of a material enables us to determine how much force can be applied to a component or structure. ASTM E-6 defines fracture strength as "the normal stress at the beginning of fracture"; it is the useful measure for brittle materials. ASTM C-1161 defines flexural strength as "a measure of the ultimate strength of a specified beam in bending"; note the linking of the strength measure to a particular size and shape of specimen. If the material is inelastic, then yield strength (defined by a prescribed deviation from initial linearity) defines the departure from elastic response, and tensile strength denotes the maximum stress the material will support before complete failure. Compressive strength is more difficult to establish unless the material is brittle.

Fracture toughness is a generic term for various measures of resistance to extension of a crack. The most familiar measure is plane-strain fracture toughness, ASTM E-399, which requires that the test specimen be thick enough to produce a state of plane strain at the tip of the crack. In this case, the value measured is indeed a material property that is independent of specimen size. Perhaps a more appropriate measure for MEMS is plane-stress fracture toughness, ASTM E-561, but it requires either measuring or inferring the actual crack extension. Implicit in all fracture testing is the condition that the radius at the tip of the crack be very small relative to other dimensions; this is a difficult requirement at the MEMS size scale.

The response of a material to cyclic loading is presented as the S–N curve, which is a plot of the applied stress S on the ordinate vs. the number of cycles to failure N on the abscissa. One obtains such a plot by testing many samples at various levels of applied stress and recording the number of cycles until the specimen breaks in two. ASTM E-466 gives the detailed procedures for metals; this is obviously an expensive test.

Creep is the time-dependent increase in strain under applied stress. Although creep is important in systems operating at high temperature, there is no ASTM standard for creep testing of metals. ASTM C-1291 defines procedures for testing ceramics. As in fatigue testing, results are usually presented in the form of plots.

3.3 Test Methods

Measuring mechanical properties of materials manufactured by processes used in MEMS is not easy. We must be able to: (1) obtain and mount a specimen, (2) measure its dimensions, (3) apply force or displacement to deform it, (4) measure the force, and (5) measure the displacement or, preferably, measure the strain. All of these steps are fully developed and standardized for common structural materials where the minimum dimension of the gauge section of a tensile specimen is 2 mm or so. ASTM E-345 does describe procedures for testing metallic foils that are less than 150 μm thick, but the rest of the dimensions are large. ASTM E-8 includes wires and even describes special grips but does not state a minimum diameter. These two standards offer guidance, but neither is completely appropriate for small MEMS specimens.

The preferred way to determine mechanical properties is by direct methods similar to the approaches of ASTM. To obtain Young's modulus, a uniform stress is applied; it is calculated from direct measurements of the applied force and the dimensions of the specimen. Strain is measured directly as the force is applied. The specimen is designed to have a uniform gauge section that is long enough to assure that the stress field is not affected by the grip ends and to permit strain measurement. This is not always possible for MEMS materials; in fact, it is most often neither possible nor practical. It is then necessary to resort to inverse methods using a model (simple or complex) of the test structure. Force is applied to the test structure and displacement is measured with the elastic, inelastic, or strength properties then extracted from the model. A simple example that has been widely used in MEMS material testing is a cantilever beam. If it is sufficiently long and thin, then only the Young's modulus enters as a material property into the formula relating force and displacement. Other examples are resonant structures and bulge tests with pressurized membranes; these are described later. If more than one material property appears in the model, then different geometries must be tested.

The formulas for determining Young's modulus, *E*, by various methods are

Static Beam	Resonant Beam	Bulge Test	Tensile Test
$\frac{4PL^3}{\delta bh^3}$	$\frac{ML^3\omega^2}{2bh^3}$	$\frac{p(1-\nu)a^4}{\delta^3 hc(\nu)}$	$\frac{P}{bh\varepsilon}$

where *h*, *b*, and *L* are the thickness, width, and length of the specimen; *P* and *p* are the applied force and pressure; *M* is the effective mass; ω is the resonant frequency; a is the dimension of a square membrane; and δ and ε are the measured deflection and strain, respectively. The function of Poisson's ratio, $c(\nu)$, depends upon the geometry and is often approximated. The simplicity of the tensile test is an obvious advantage.

Johnson et al. (1999) have compared the uniaxial and bending tests and point out that uncertainty in specimen dimensions is more of a problem in bending tests, while overall elongation is difficult to measure in a tension test. However, if strain can be measured directly, the overall elongation does not need to be measured. Johnson et al. also point out that strength due to misalignment is more of a problem in tension than in bending.

As will be seen later in this chapter, there is an alarming variability among measured values of even so basic a property as Young's modulus for the most widely studied MEMS material — polysilicon. Senturia (1998) attributes this to two primary reasons, "insufficiently precise models used to interpret the data and metrology errors in establishing the geometry of the test devices." He is referring to inverse methods in the first point; whether the boundary conditions of the actual structure actually match the model is a significant question. Senturia's point on metrology applies to all test methods — direct or inverse.

It is useful to distinguish between on-chip test methods and property tests. It is very important in this technology to be able to obtain a measure of mechanical properties from test structures that are on the same chip (or die) as the manufactured MEMS. That usually precludes a direct property measurement on a specimen, which must be larger to allow gripping and pulling even though the size of the gauge section is the same size scale as the microdevice. This is not an issue in mechanical and civil engineering fields, where the required specimen size is smaller than the system or structure. We may regard property tests as basic or baseline and on-chip tests as practical. Obviously, completeness requires direct comparisons of the two approaches with specimens and test structures on the same die.

3.3.1 Specimen and Test Structure Preparation

Microdesign processes cannot take a billet of bulk material and shape it into the final MEMS product as is common for most manufacturing processes. Rather, the microdevice is produced by deposition and etching processes. This means that the specimen or test structure cannot be cut from the bulk material but must be produced by the same processes as the product. A tensile specimen must be designed so that one end remains fixed to the die and the other end accommodates some sort of gripping mechanism. A test structure must be designed so that the boundaries are indeed fixed, and it must incorporate some sort of actuating mechanism to produce force and a sensor to determine displacement or, perhaps, strain.

An early and interesting approach to producing tensile specimens of thin foils was conceived by Neugebauer (1960), who deposited gold films onto oriented rocksalt crystals. The gold film was glued to the grips of a test machine and the test section covered with sealing wax while the salt was dissolved away. These specimens ranged in thickness from 0.05 to 1.5 μm and were 1–2 mm wide and approximately 1 cm long. Neugebauer found the tensile strength to be two to four times higher than for annealed bulk material but observed no dependence on film thickness. The Young's modulus values agreed with those of the bulk material.

This is a simple example of specimen preparation, but it is illustrative of the methods used in the mechanical test methods for MEMS materials. One deposits the material of interest and removes the unwanted portions of the supporting substrate. An additional step patterns the test material through photolithography.

3.3.2 Dimension Measurement

Minimum features in MEMS are usually on the order of 1 μm — a bit larger than in microelectronics. Measuring the 2 × 2 μm cross section of a tensile specimen or the equivalent dimensions of a test structure might seem to be easy, but it is not. The thickness of a layer is well controlled and measured by the manufacturer. Lengths are large enough to measure with sufficient accuracy in an optical microscope. It is the width of small specimens or test structure components that is difficult to determine.

A major problem is that the cross-section is not sharply defined or even rectangular as expected. Figure 3.1 is a scanning electron microscope (SEM) photograph of the end of a polysilicon tensile specimen after testing. This specimen is from the Multi-User MEMS Process (MUMPs) process at Cronos, which deposits a first layer of polysilicon that is 2.0 μm thick and then a second layer that is 1.5 μm thick. The interface between these two layers is visible in the photograph. The designed width is 2.0 μm, which is approximately the case at the bottom of the rectangular. The fact that the cross section is not a perfect rectangle contributes to the uncertainty in the area. The corners are somewhat rounded, which makes it difficult to establish the edges when making a plan-view measurement.

The dimensions of a specimen or test structure are normally established before the experiment, but a more accurate measurement may be made after the specimen is broken. Optical or scanning electron microscopy, mechanical or optical profilometry, and interferometry are possible measurement techniques, but some of these can be quite time consuming and expensive. Johnson et al. (1999) state that it is typical to assign an uncertainty of between ±0.05 and ±0.10 μm to width measurements. A 2 μm wide

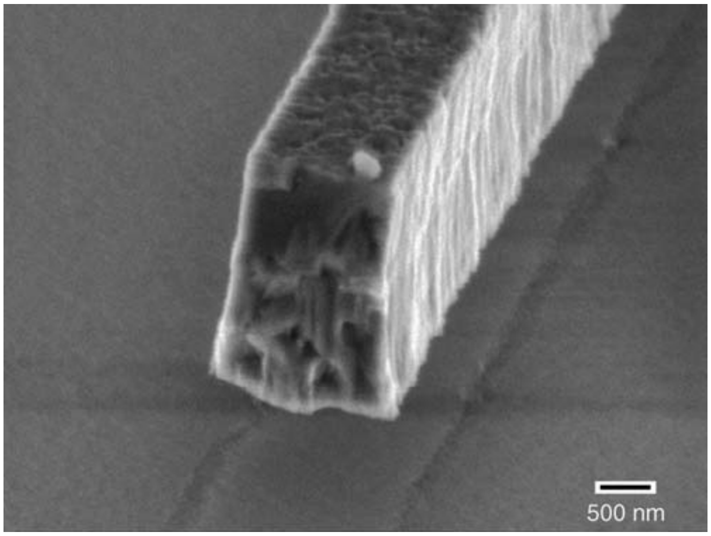


FIGURE 3.1 Scanning electron micrograph of the end of a broken tensile specimen. The specimen is 3.5 mm thick and 2 mm wide at the bottom. (Reprinted with permission from Sharpe et al. [1999a] "Polysilicon Tensile Testing with Electrostatic Gripping," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 518*, pp. 191–96, 15–16 April, San Francisco.).

specimen would therefore have at most ±5% relative uncertainty in its width, which is actually quite reasonable for such small specimens. This reinforces the statement by Senturia (1998) that metrology is a major problem in determining mechanical properties.

3.3.3 Force and Displacement Measurement

Johnson et al. (1999) explain that tensile tests require the measurement of larger forces and smaller overall displacements, while the opposite is true for bending tests. To break a polysilicon tensile specimen $2 \times 2\,\mu$m square with a fracture strength of 2 GPa requires a force of 8 mN. If that same specimen is 50 μm long, fixed at one end, and with a transverse point load at the other end, breaking it requires a force of only 0.05 mN. If that material has a modulus of 160 GPa, the elongation of a tensile specimen is 0.62 μm, while the deflection at the end of a bending specimen is 10.4 μm.

Commercial force transducers are readily available with a range of ±5 g (50 mN) and a resolution of 0.001 g. This author prefers to use the lower range of a ±100-g load cell because it is stiffer relative to a tensile specimen and to calibrate it with weights. This achieves a resolution of 0.01 g with a full-scale uncertainty of ±1% [Sharpe et al., 1999a]. Howard and Fu (1997) review suitable force transducers, and others, such as Greek et al. (1995) and Saif and MacDonald (1996), construct their own.

Commercial capacitance-based displacement transducers can be used to measure the overall displacement of a test system to a resolution of 0.01 μm and full-scale uncertainty of ±1% [Sharpe et al., 1999a]. Schemes to measure mechanical deflections at the optical microscope level are attractive, and Pan and Hsu (1999) present a vernier gauge approach to measure residual stress. This approach can be electrically instrumented with differential capacitance measurement as shown by Que et al. (1999).

3.3.4 Strain Measurement

It is, of course, preferable to measure strain directly, whether the test arrangement is bending or tension; however, this is difficult to do on such small specimens. The author and his colleagues have developed a laser-based strain-measurement system in which two reflective lines are deposited on the gauge section of a tensile specimen during manufacture. These lines are perpendicular to the loading axis, and when they are illuminated with a low-power laser beam, interference fringe patterns are formed. When the specimen is strained, the lines separate and the fringes move; tracking the motion with diode arrays and a computer system enables real-time strain measurement on specimens as narrow as 20 μm. A set of four lines on wider specimens permits measurement of Poisson's ratio; details are given in Sharpe et al. (1997c; 1997d), and the resolution is approximately ±5 microstrain with a relative uncertainty of 5% at 0.5% strain.

Detailed full-field strain measurements at the MEMS size scale are desirable but difficult. Micro-Raman spectroscopy can probe very small areas on the order of 1 μm in diameter on thin films. Analysis of frequency shifts as force is applied to a specimen leads to local strain measurements [Benrakkad et al., 1995; Pinardi et al., 1997; Zhang et al., 1997; Amimoto et al., 1998]. The moiré method using e-beam lithography to write high-frequency line and dot gratings at a small scale has been demonstrated by Dally and Read (1993), but this is a very challenging process. Chasiotis and Knauss (1998) are developing digital image correlation methods to measure strains in tensile specimens; the resolution is 300–500 microstrain. Mazza et al. (1996a) have demonstrated this to be a viable technique on single-crystal silicon specimens. Laser speckle methods can give full-field results and have been demonstrated by Anwander et al. (2000) and Chang et al. (2000). None of these techniques has been applied to extensive studies of mechanical properties of MEMS materials.

3.3.5 Tensile Tests

Three arrangements are used in tensile tests of MEMS materials: specimen in a supporting frame, specimen fixed to a die at one end; and separate specimen. A fourth clever approach was introduced early on by Koskinen et al. (1993) but has not been continued. They deposited a grid of long, thin tensile specimens that were all fastened to larger portions at each end; the appearance was similar to a foil resistance strain gauge. One end of the arrangement was fixed, and the other was attached to a movable grip that could be rotated about an axis perpendicular to the grid. This caused all of the specimens to buckle, each a different amount than its neighbor. When the grip moved, each specimen in turn was straightened and pulled. The recorded force-displacement record enabled measurement of modulus and strength.

3.3.5.1 Specimen in Frame

Read and Dally (1992) introduced a very effective way of handling thin-film specimens in 1992. The tensile specimen is patterned onto the surface of a wafer, and then a window is etched in the back of the wafer to expose the gauge section. The result is a specimen suspended across a rectangular frame, which can be handled easily and placed into a test machine. The two larger ends of the frame are fastened to grips, and the two narrower sides are cut to completely free the specimen. This is an extension of the much earlier approach by Neugebauer (1960) and has been adopted by others [Cunningham et al., 1995; Emery et al., 1997; Ogawa et al., 1997; Sharpe et al., 1997c; Cornella et al., 1998; Yi and Kim, 1999b]. A SEM photograph of such a specimen while still in the frame is shown in Figure 3.2.

3.3.5.2 Specimen Fixed at One End

Tsuchiya introduced the concept of a tensile specimen fixed to the die at one end and gripped with an electrostatic probe at the other end [Tsuchiya et al., 1998]. This approach has been adopted by this author and his students [Sharpe et al., 1998a]; Figure 3.3 is a photograph of this type of specimen. The gauge section is 3.5 μm thick, 50 μm wide, and 2 mm long. The fixed end is topped with a gold layer for electrical contact. The grip end is filled with etch holes, as are the two curved transition regions from the grips

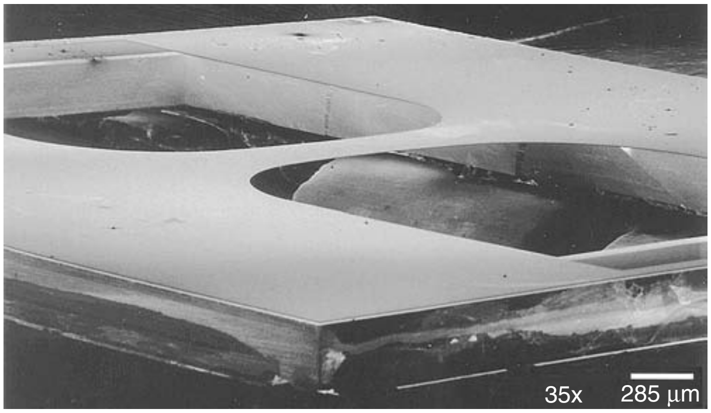


FIGURE 3.2 Scanning electron micrograph of a polysilicon tensile specimen in a supporting single-crystal silicon frame. (Reprinted with permission from Sharpe, W.N., Jr., Yuan, B., Vaidyanathan, R., and Edwards, R.L. [1996] *Proc. SPIE* 2880, pp. 78–91.)

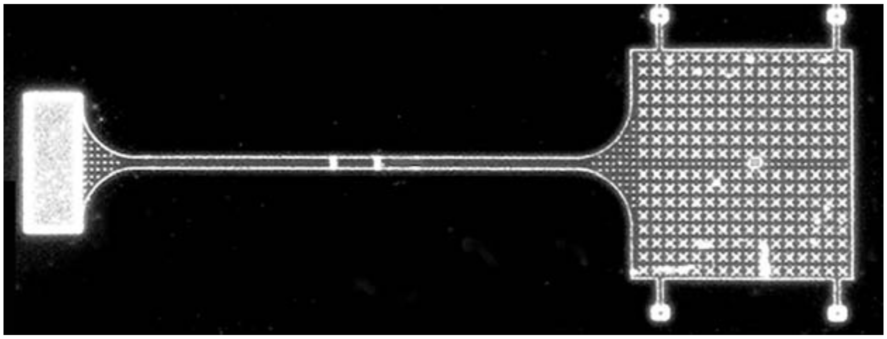


FIGURE 3.3 A tensile specimen fixed at the left end with a free grip end at the right end. (Reprinted with permission from Sharpe, W.N., Jr., and Jackson, K. [2000] *Microscale Systems: Mechanics and Measurements Symposium, Society for Experimental Mechanics,* pp. ix–xiv.)

to the gauge section. The large grip end is held in place during the etch-release process by four anchor straps, which are broken before testing.

Chasiotis and Knauss (2000) have developed procedures for gluing the grip end of a similar specimen to a force/displacement transducer, which enables application of larger forces. A different approach is to fabricate the grip end in the shape of a ring and insert a pin into it to make the connection to the test system. Greek et al. (1995) originated this with a custom-made setup, and LaVan et al. (2000a) use the probe of a nanoindenter for the same purpose.

It is possible to build the deforming mechanism onto or into the wafer, although getting an accurate measure of the forces and deflections can be difficult. Biebl and von Philipsborn (1995) stretched polysilicon specimens in tension with residual stresses in the structure. Yoshioka et al. (1996) etched a hinged paddle in the silicon wafer, which could be deflected to pull a thin single-crystal specimen. Nieva et al. (1998) produced a framed specimen and heated the frame to pull the specimen, as did Kapels et al. (2000).

3.3.5.3 Separate Specimen

The challenge of picking up a tensile specimen only a few microns thick and placing it into a test machine is formidable. However, if the specimens are on the order of tens or hundreds of microns thick, as they are for LIGA-deposited materials, doing so is perfectly possible. This author and his students developed techniques to test steel microspecimens having submillimeter dimensions [Sharpe et al., 1998b]. The steel dog-biscuit-shaped specimens were obtained by cutting thin slices from the bulk material and then cutting out the specimens with a small CNC mill. Electroplated nickel specimens can be patterned into a similar shape in LIGA molds as shown in Figure 3.4. These specimens are released from the substrate by etching, picked up, and put into grips with inserts that match the wedge-shaped ends [Sharpe et al., 1997e].

McAleavey et al. (1998) used the same sort of specimen to test SU-8 polymer specimens. Mazza et al. (1996b) prepared nickel specimens of similar size in the gauge section but with much larger grip ends. Christenson et al. (1998) fabricated LIGA nickel specimens of a more conventional shape; they were approximately 2 cm long with flat grip ends, large enough to test in a commercial table-top electrohydraulic test machine.

3.3.5.4 Smaller Specimens

All of the above methods may appear impressive to the materials test engineer accustomed to common structural materials, but there is a continuing push toward smaller structural components at the nanoscale. Yu et al. (2000) have successfully attached the ends of carbon nanotubes as small as 20 nm in

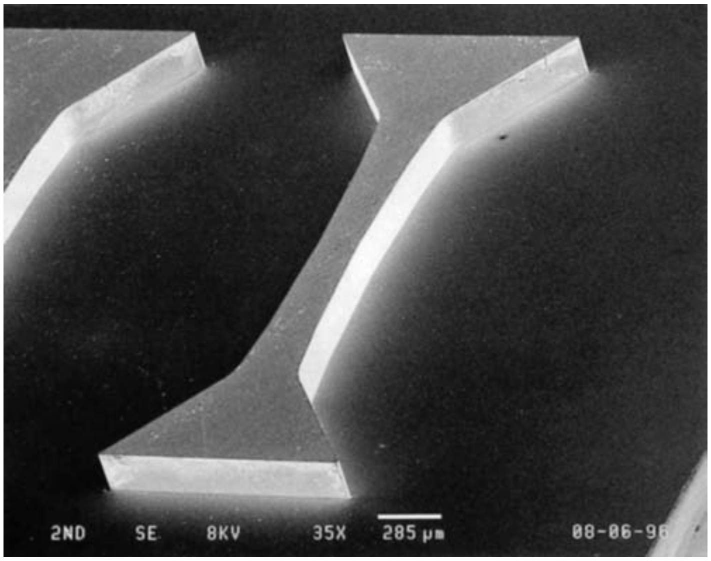


FIGURE 3.4 Nickel microspecimen produced by the LIGA method. The overall length is 3.1 mm, and the width of the specimen at the center is 200 mm. (Reprinted with permission from Sharpe, W.N., Jr., et al. [1997] *Proc. Int. Solid State Sensors and Actuators Conf. — Transducers '97*, pp. 607–10. © 1997 IEEE.)

diameter and a few microns long to atomic force microscopy (AFM) probes. As the probes are moved apart inside a SEM, their deflections are measured and used to extract both the force in the tube and its overall elongation. They report strengths up to 63 GPa and modulus values up to 950 GPa.

3.3.6 Bend Tests

Three arrangements are also used in bend tests of structural films: out-of-plane bending of cantilever beams, beams fastened at both ends, and in-plane bending of beams. Larger specimens, which can be individually handled, can also be tested in bending fixtures similar to those used for ceramics.

3.3.6.1 Out-of-Plane Bending

The approach here is simple. The process patterns long, narrow, and thin beams of the test material onto a substrate and then etches away the material underneath to leave a cantilever beam hanging over the edge. By measuring the force vs. deflection at or near the end of the beam, one can extract Young's modulus via the formula in section 3.3. However, this is difficult because if the beams are long and thin, the deflections can be large, but the forces are small. The converse is true if the beam is short and thick, but then the applicability of simple beam theory comes into question. If the beam is narrow enough, Poisson's ratio does not enter the formula; otherwise, beams of different geometries must be tested to determine it.

Weihs et al. (1988) introduced this method in 1988 by measuring the force and deflection with a nanoindenter having a force resolution of 0.25 μN and a displacement resolution of 0.3 nm. Typical specimens had a thickness, width, and length of 1.0, 20, and 30 μm, respectively. Figure 3.5 shows a cantilever beam deflected by a nanoindenter tip in a later investigation [Hollman et al., 1995].

Biebl et al. (1995a) attracted the end of a cantilever down to the substrate with electrostatic forces and recorded the capacitance change as the voltage was increased to pull more of the beam into contact. Fitting these measurements to an analytical model permitted a determination of Young's modulus.

Krulevitch (1996) proposed a technique for measuring Poisson's ratio of thin films fabricated in the shapes of beams and plates by comparing the measured curvatures. These were two-layer composite

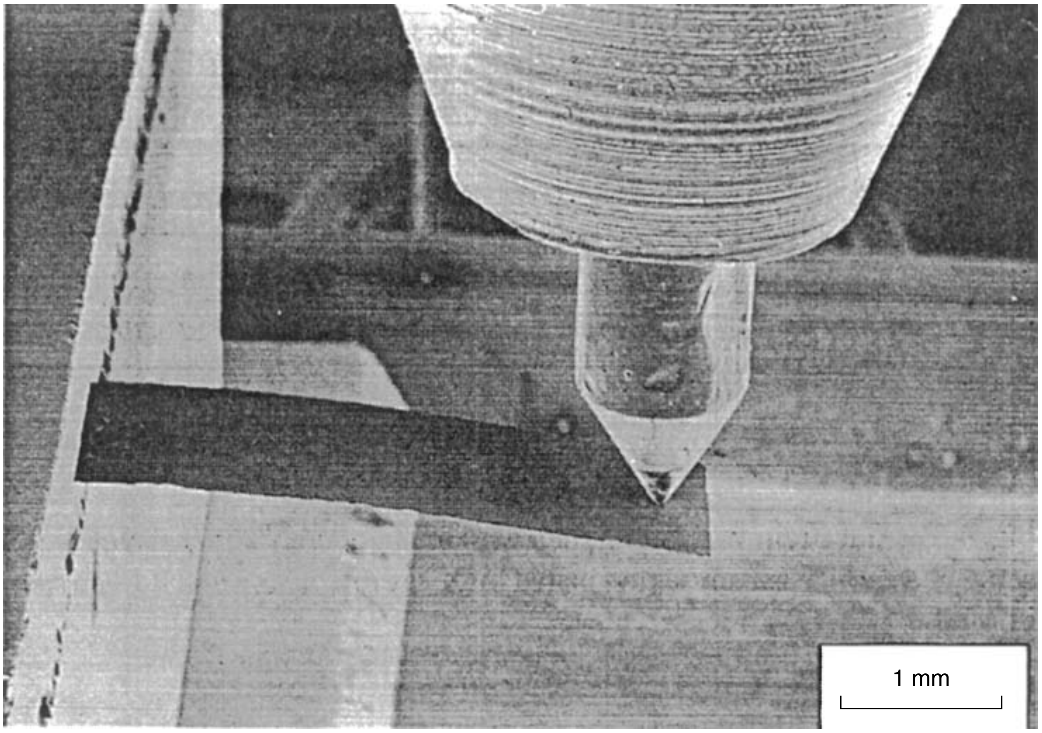


FIGURE 3.5 A cantilever microbeam deflected out of plane by a diamond stylus. The beam was cut from a free-standing diamond film. (Reprinted with permission from Hollman, P., et al. (1995) "Residual Stress, Young's Modulus and Fracture Stress of Hot Flame Deposited Diamond," *Thin Solid Films* **270**, pp. 137–42.)

structures, so the properties of the substrate must be known. Kraft et al. (1998) also tested composite beams by measuring the force-deflection response with a nanoindenter. Bilayer cantilever beams have been tested by Tada et al. (1998), who heated the substrate and measured the curvature.

More sensitive measurements of force and displacement on smaller cantilever beams can be made by using an AFM probe, as shown by Serre et al. (1998), Namazu et al. (2000), Comella and Scanlon (2000), and Kazinczi et al. (2000). A specially designed test machine using an electromagnetic actuator has been developed by Komai et al. (1998).

3.3.6.2 Beams with Fixed Ends

Working with a beam that is fixed at both ends is somewhat easier; the beam is stiffer and more robust. Tai and Muller (1990) used a surface profilometer to trace the shapes of fixed-fixed beams at various load settings. By comparing measured traces and using a finite element analysis of the structure, they were able to determine Young's modulus.

A promising on-chip test structure has been developed over the years by Senturia and his students; it is shown schematically in Figure 3.6. A voltage is applied between the conductive polysilicon beam and the substrate to pull the beam down, and the voltage that causes the beam to make contact is a measure of its stiffness. This concept was introduced early on by Petersen and Guarnieri (1979) and further developed by Gupta et al. (1996). A similar approach and analysis were described by Zou et al. (1995). The considerable advantage here is that the measurements can be made entirely with electrical probing in a manner similar to that used to check microelectronic circuits. This opens the opportunity for process monitoring and quality control.

The fixed ends clearly exert a major influence on the stiffness of the test structure. Kobrinsky et al. (1999) have thoroughly examined this effect and shown its importance. The problem is that a particular manufacturing process, or even variations within the same process, may etch the substrate slightly differently and change the rigidity of the ends. Nevertheless, this is a potentially very useful method for monitoring the consistency of MEMS materials and processes.

Zhang et al. (2000) recently conducted a thorough study of silicon nitride in which microbridges (fixed–fixed beams) were deflected using a nanoindenter with a wedge-shaped indenter. By fitting the measured force-deflection records to their analytical model, they extracted both Young's modulus and residual stress.

3.3.6.3 In-Plane Bending

In-plane bending may be a more appropriate test method in that the structural supports of MEMS accelerometers are subjected to that mode of deformation. Jaecklin et al. (1994) pushed long, thin cantilever beams with a probe until they broke; optical micrographs gave the maximum deflections, from which the fracture

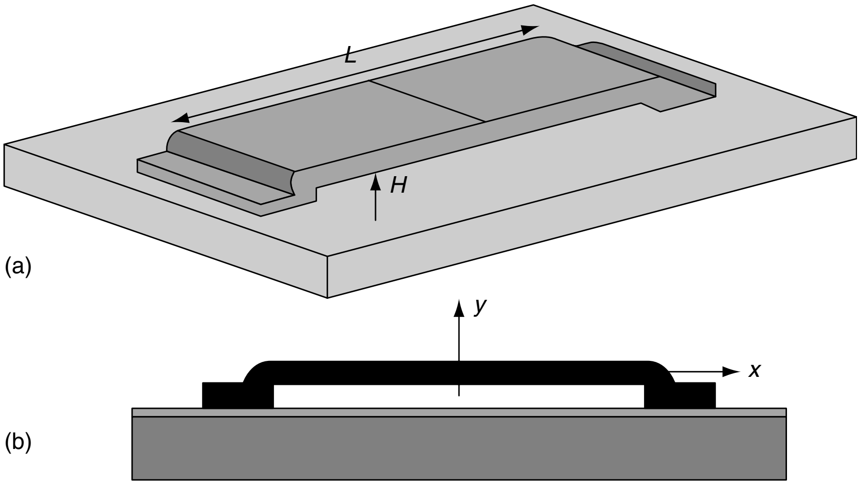


FIGURE 3.6 Schematic of a fixed-fixed beam. (Reprinted with permission from Kobrinsky, M. et al. [1999] "Influence of Support Compliance and Residual Stress on the Shape of Doubly-Supported Surface Micromachined Beams," *MEMS Microelectromechanical Systems* 1, pp. 3–10, ASME, New York.)

strain was determined. Jones et al. (1996) constructed a test structure consisting of cantilever beams of different lengths fastened to a movable shuttle. As the shuttle was pushed, the beams contacted fixed stops on the substrate; the deformed shape was videotaped and the fracture strain determined. Figure 3.7 is a photograph of one of their deformed specimens.

Kahn et al. (1996) developed a double cantilever beam arrangement to measure the fracture toughness of polysilicon and used the measured displacement between the two beams to determine Young's modulus via a finite element model. The beams were separated by forcing a mechanical probe between them and pushing it toward the notched end. Fitzgerald et al. (1998) have taken a similar approach to measure crack growth and fracture toughness in single-crystal silicon, but they use a clever structure that permits opening the beams by compression of cantilever extensions.

3.3.6.4 Bending of Larger Specimens

Microelectromechanical technology is not restricted to thin-film structures, although they are far-and-away predominant. Materials fabricated with thicknesses on the order of tens or hundreds of microns are of current interest and likely to become more important in the future.

Ruther et al. (1995) manufactured a microtesting system using the LIGA process to test electroplated copper. The interesting feature is that the in-plane cantilever beam and the test system are fabricated together on the die; however, this requires a rather complex assembly. Stephens et al. (1998) fabricated rows of LIGA nickel beams sticking up from the substrate and then measured the force applied near the upper tip of the beam while displacing the substrate. The resulting force-displacement curve permitted extraction of Young's modulus, and the recorded maximum force gave a modulus of rupture.

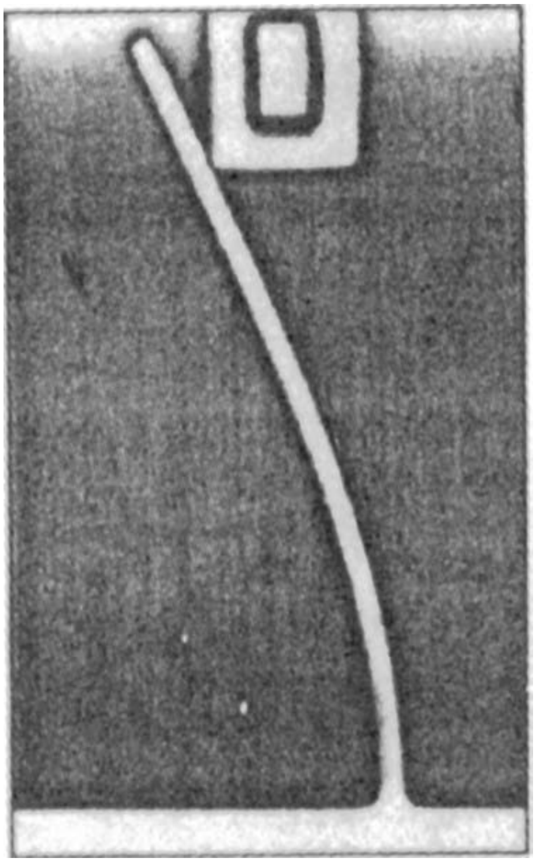


FIGURE 3.7 A polysilicon cantilever beam subjected to in-plane bending. The beam is 2.8 mm wide, and the vertical distance between the fixed end at the bottom and the deflected end at the top is 70 mm. (Reprinted with permission from Sharpe, W.N., Jr., et al. [1998] "Round-Robin Tests of Modulus and Strength of Polysilicon," *Microelectromechanical Structures for Materials Research Symposium*, pp. 56–65.)

Larger structures, such as the microengine under development at the Massachusetts Institute of Technology, have thicknesses on the order of several millimeters. It then becomes necessary to test specimens of similar sizes in what is sometimes called the mesocale region, whose dimensions generally range from 0.1 mm to 1 cm. Single-crystal silicon is the material of interest for initial versions, and Chen et al. (1998) have developed a method for bend testing square plates simply supported over a circular hole and recording the force as a small steel ball is pushed into the center of the plate. Fracture strengths are obtained, and this efficient arrangement permits study of the effects of various manufacturing processes on the load-carrying capability of the material.

3.3.7 Resonant Structure Tests

Frequency and changes in frequency can be measured precisely, and elastic properties of modeled structures can be determined. The microstructures can be very small and excited by capacitive comb-drives, which require only electrical contact. This makes this approach suitable for on-chip testing; in fact, the MUMPs process at Cronos includes a resonant structure on each die. That microstructure moves parallel to the substrate, but others vibrate perpendicularly.

Petersen and Guarnieri (1979) introduced the resonant structure concept in 1979 by fabricating arrays of thin, narrow cantilever beams of various lengths extending over an anisotropically etched pit in the substrate. The die containing the beams was excited by variable frequency electrostatic attraction between the substrate and the beams, and the vibration perpendicular to the substrate was measured by reflection from an incident laser beam, as shown by the schematic in Figure 3.8. Yang and Fujita (1997) used a similar approach to study the effect of resistive heating on U-shaped beams. Commercial AFM cantilevers were tested in a similar manner by Hoummady et al. (1997), who measured the higher resonant modes of a cantilever beam with a mass on the end. Zhang et al. (1991) measured vibrations of a beam fixed at both ends by using laser interferometry. Michalicek et al. (1995) developed an elaborate and carefully modeled micromirror that was excited by electrostatic attraction. Deflection was also measured by laser interferometry, and experiments determined Young's modulus over a range of temperatures as well as validating the model.

Microstructures that vibrate parallel to the plane of the substrate require less processing because the substrate does not have to be removed. Biebl et al. (1995b) introduced this concept, and Kahn et al. (1998) have used a more recent version to study the effects of heating on the Young's modulus of films sputtered

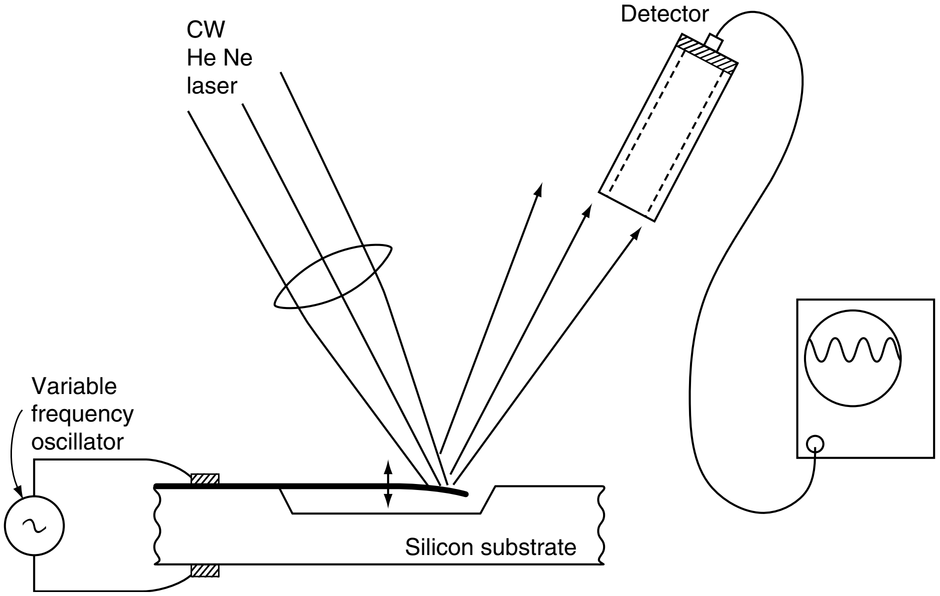


FIGURE 3.8 Schematic of the resonant structure system of Petersen and Guarnieri (1979). (Reprinted with permission from Petersen, K.E., and Guarnieri, C.R. [1979] "Young's Modulus Measurements of Thin Films Using Micromechanics," *J. Appl. Phys.* **50**, pp. 6761–66.)

onto the structure. Figure 3.9 is a SEM image of their structure, which is easy to model. Pads A, B, C, and D are fixed to the substrate; the rest of the structure is free. Electrostatic comb-drives excite the two symmetrical substructures, which consist of four flexural springs and a rigid mass. The resonant frequency of this device is around 47 kHz. Brown et al. (1997) have developed a different approach in which a small notched specimen is fabricated as part of a large resonant fan-shaped component. This resonant structure, shown in Figure 3.10, has been used primarily for fatigue and crack growth studies,

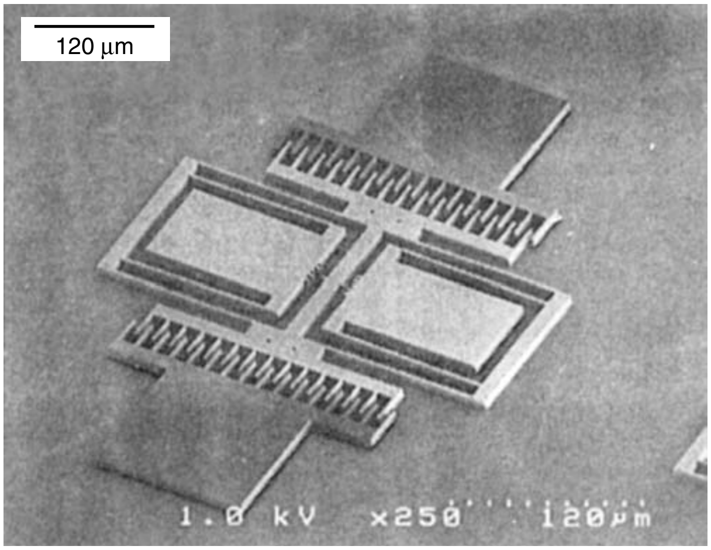


FIGURE 3.9 Scanning electron micrograph of the in-plane resonant structure of Kahn et al. (1998). (Reprinted with permission from Kahn, H. et al. [1998] "Heating Effects on the Young's Modulus of Films Sputtered onto Micromachined Resonators," *Microelectromechanical Structures for Materials Research Symposium,* pp. 33–38.)

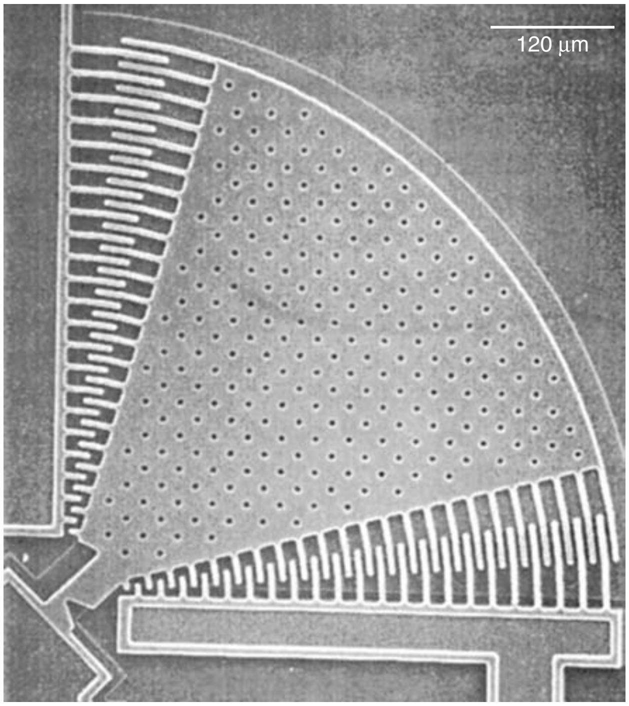


FIGURE 3.10 Scanning electron micrograph of the in-plane resonant structure of Brown et al. (1997). (Reprinted with permission from Brown, S.B. et al. [1997] "Materials Reliability in MEMS Devices," *Proc. Int. Solid-State Sensors and Actuators Conf. — Transducers '97*, pp. 591–93. © 1997 IEEE.)

but Young's modulus of polysilicon has been extracted from its finite element model [Sharpe et al., 1998c].

3.3.8 Membrane Tests

It is relatively easy to fabricate a thin membrane of test material by etching away the substrate; the membrane is then pressurized and the measured deflection can be used to determine the biaxial modulus. An advantage of this approach is that tensile residual stress in the membrane can be measured, but the value of Poisson's ratio must be assumed. This method, often called bulge testing, was first introduced by Beams (1959), who tested thin films of gold and silver and measured the center deflection of the circular membrane as a function of applied pressure. Jacodine and Schlegel (1966) used this approach to measure Young's modulus of silicon oxide. Tabata et al. (1989) tested rectangular membranes whose deflections were measured by observations of Newton's rings, as did Maier-Schneider et al. (1995). The variation of Hong et al. (1990) used circular membranes with force deflection measured at the center with a nanoindenter. Pressurized square membranes with the deflection measured by a stage-mounted microscope were tested by Walker et al. (1990) to study the effect of hydrofluoric acid exposure on polysilicon; a similar approach to determine biaxial modulus, residual stress, and strength was used by Cardinale and Tustison (1992). Vlassak and Nix (1992) eliminated the need to assume a value of Poisson's ratio by testing rectangular silicon nitride films with different aspect ratios. More recently, Jayaraman et al. (1998) used this same approach to measure Young's modulus and Poisson's ratio of polysilicon.

3.3.9 Indentation Tests

A nanoindenter is, in the fewest words, simply a miniature and highly sensitive hardness tester. It measures both force and displacement, and modulus and strength can be obtained from the resulting plot. Penetration depths can be very small (a few nanometers), and automated machines permit multiple measurements to enhance confidence in the results and also to scan small areas for variations in properties.

Weihs et al. (1989) measured the Young's modulus of an amorphous silicon oxide film and a nontextured gold film with a nanoindenter and obtained only limited agreement with their microbeam deflection experiments. The modulus measured by indentation was consistently higher, and the large pressure of the indenter tip was the probable cause. Taylor (1991) used nanoindenter measurements restricted to penetrations of 200 nm into silicon nitride films 1 μm thick to study the effects of processing on mechanical properties. Young's modulus decreased with decreasing density of the films.

Bhushan and Li (1997) have studied the tribological properties of MEMS materials, and Li and Bhusan (1999) used a nanoindenter to measure the modulus and a microhardness tester to measure the fracture toughness of thin films. Measurements of Young's modulus of polysilicon showed a wide scatter. Bucheit et al. (1999) examined the mechanical properties of LIGA-fabricated nickel and copper by using a nanoindenter as one of the tools. In most cases, Young's modulus from nanoindenter measurements were higher than from tension tests, but the nanoindenter does allow looking at both sides of the thin film as well as at sectioned areas.

3.3.10 Other Test Methods

The readily observed buckling of a column-like structure under compression can be used to measure forces in specimens; if the specimen breaks, the fracture strength can be estimated. Tai and Muller (1988) fabricated long, thin polysilicon specimens with one end fixed and the other enclosed in slides. The movable end was pushed with a micromanipulator, and its displacement when the structure buckled was used to determine the strain (not stress) at fracture. Ziebart and colleagues have analyzed thin films with various boundary conditions ranging from fixed along two sides [Ziebart et al., 1997] to fixed on all four sides [Ziebart et al., 1999]. The first arrangement permitted the measurement of Poisson's ratio when the side supports were compressed, and the second determined prestrains induced by processing. Beautiful patterns are obtained, but the analysis and the specimen preparation can be time consuming.

Another clever approach based on buckling is described by Cho et al. (1997). They etched away the silicon substrate under an overhanging strip of diamond-like carbon film and used the buckled pattern to determine the residual stress in the film. A more traditional creep test was used by Teh et al. (1999) to study creep in $2 \times 2 \times 100\,\mu$m polysilicon strips fixed at each end. As current passed through the specimens, they heated up, and their buckled deflection over time at a constant current was used to extract a strain-vs.-time creep curve. This approach is complicated by the nonuniformity of the strain in the specimen.

Although torsion is an important mode of deformation in certain MEMS, such as digital mirrors, few test methods have been developed. Saif and MacDonald (1996) introduced a system to twist very small (10 µm long and 1 µm on a side) pillars of single-crystal silicon and measure both the force and deflection. Larger (300 µm long with side dimensions varying from 30 to 180 µm) of both silicon and LIGA nickel were tested by Schiltges et al. (1998). Emphasis was on the elastic properties only with the shear modulus values agreeing with expected bulk values.

Nondestructive measurements of elastic properties of thin films can be accomplished with laser-induced ultrasonic surface waves. A laser pulse generates an impulse in the film, and a piezoelectric transducer senses the surface wave. In principle, Young's modulus, density, and thickness can be determined, but this cannot be achieved for all combinations of film and substrate materials. Schneider and Tucker (1996) describe this test method and present results for a wide range of films; the Young's modulus values generally agree with other thin-film measurements. A drawback here is the planar size of the film; the input and output must be several millimeters apart. A related technique uses Brillouin scattering as described in Monteiro et al. (1996).

3.3.11 Fracture Tests

Single-crystal silicon and polysilicon are both brittle materials, and it is therefore natural to want to measure their fracture toughness. This is even more difficult than measuring their fracture strength because of the need for a crack with a tip radius that is small relative to the specimen dimensions.

Photolithography processes for typical thin films have a minimum feature radius of approximately 1 mm. Fan et al. (1990), Sharpe et al. (1997f) and Tsuchiya et al. (1998) have tested polysilicon films in tension using edge cracks, center cracks, and edge cracks, respectively. Kahn et al. (1999) modeled a double-cantilever specimen with a long crack and wedged it open with an electrostatic actuator.

Fitzgerald et al. (1999) prepared sharp cracks in double-cantilever silicon crystal specimens by etching, and Suwito et al. (1997) modeled the sharp corner of a tensile specimen to measure the fracture toughness. Van Arsdell and Brown (1999) introduced cracks at notches in polysilicon with a diamond indenter. A promising new approach using a focused ion beam (FIB) can prepare cracks with tip radii of 30 nm according to K. Jackson (pers. comm.).

3.3.12 Fatigue Tests

Many MEMS operate for billions of cycles, but that kind of testing is conducted on microdevices, such as digital mirrors instead of the more basic reversed bending or push–pull tests so familiar to the metal fatigue community. Brown and his colleagues have developed a fan-shaped, electrostatically driven notched specimen that has been used for fatigue and crack growth studies [Brown et al., 1993, 1997; Van Arsdell and Brown, 1999]. Minoshima et al. (1999) have tested single-crystal silicon in bending fatigue, and Sharpe et al. (1999) reported some preliminary tension–tension tests on polysilicon. As noted earlier, fatigue data are reported as stress-vs.-life plots, and Kapels et al. (2000) present a plot that looks much like one would expect for a metal; the allowable applied stress decreases from 2.9 GPa for a monotonic test to 2.2 GPa at one million cycles.

3.3.13 Creep Tests

Some MEMS are thermally actuated, so the possibility of creep failure exists. No techniques similar to the familiar dead-weight loading to produce strain-vs.-time curves exist. Teh et al. (1999) have observed the buckling of heated fixed-end polysilicon strips.

3.3.14 Round-Robin Tests

Mechanical testing of MEMS materials presents unique challenges as the above review shows. Convergence of test methods into a standard is still far in the future, but progress in that direction usually begins with a round-robin program in which a common material is tested by the method-of-choice in participating laboratories. That first step was taken in 1997/1998 with the results reported at the Spring 1998 meeting of the Materials Research Society [Sharpe et al., 1998c]. Polysilicon from the MUMPs 19 and 21 runs of Cronos were tested in bending (Figure 3.7), resonance (Figure 3.10), and tension (Figure 3.3). Young's modulus was measured as 174 ± 20 GPa in bending, 137 ± 5 GPa in resonance, and 139 ± 20 GPa in tension. Strengths in bending were 2.8 ± 0.5 GPa, in resonance 2.7 ± 0.2 GPa, and in tension 1.3 ± 0.2 GPa. These variations were alarming but in retrospect perhaps not too surprising given the newness of the test methods at that time.

A more recent interlaboratory study of the fracture strength of polysilicon manufactured at Sandia has been arranged by LaVan et al. (2000b). Strengths measured on similar tensile specimens by Tsuchiya in Japan and at Johns Hopkins were 3.23 ± 0.25 and 2.85 ± 0.40 GPa respectively. LaVan tested in tension with a different approach and obtained 4.27 ± 0.61 GPa. It seems clear that more effort needs to be devoted to the development of test methods that can be used in a standardized manner by anyone who is interested.

3.4 Mechanical Properties

This section lists in tabular form the results of measurements of mechanical properties of materials used in MEMS structural components. Its intent is not only to provide values of mechanical properties but also to supply references on materials and test methods of interest. Because as yet no standard test method exists and such a wide variety in the values is obtained for supposedly identical materials, readers with a strong interest in the mechanical behavior of a particular material can use the tables to identify pertinent references.

Almost all the data listed comes from experiments directly related to free-standing structural films. The only exceptions are the results from ultrasonic measurements by Schneider and Tucker (1996) because they tested a number of materials of interest. Including information on the processing conditions for each reference proved too cumbersome, but the short comments in the tables should be useful. Many of the results are average values of multiple replications, and the standard deviations are included when they are available. Most of the materials used in MEMS are ceramics and show linear and brittle behavior, in which case only the fracture strength is listed. The tables for ductile materials show both yield and ultimate strengths. Also note that the values in the tables are edited from a larger list. Some of the same values have been presented in two different venues (e.g., a conference publication and a journal paper), in which case the more archival version was referenced. A limited number of studies have been conducted on the effects of environment (temperature, hydrofluoric acid, saltwater, etc.) on MEMS materials, but that area of research is in its infancy and is not included.

First, typical stress–strain curves are plotted in Figure 3.11 to compare the mechanical behavior of MEMS materials with a common structural steel, A533-B, which is moderately strong (yield strength of 440 MPa) but ductile and tough. Polysilicon is linear and brittle and much stronger. LIGA nickel is ductile and considerably stronger than bulk pure nickel. One must test materials as they are produced for MEMS instead of relying on bulk material values.

The microstructure of these MEMS materials is also different from that of bulk materials. The physics of the thin-film deposition process cause the grains to be columnar in a direction perpendicular to the film as shown in Figure 3.12. The result is similar to the cross-section of a piece of bamboo or wood, and the material is transversely isotropic. Test methods are not sensitive enough to measure the anisotropic constants.

Table 3.1 lists metal films tested in a free-standing manner such as would be appropriate for use in MEMS. Only aluminum is currently used in that fashion, but the other materials are commonly used in the electronics industry and may be of interest. Note that all of the materials are ductile; the complete stress–strain curves are included in many of the references. The values of Young's modulus as measured for pure bulk materials are listed for reference.

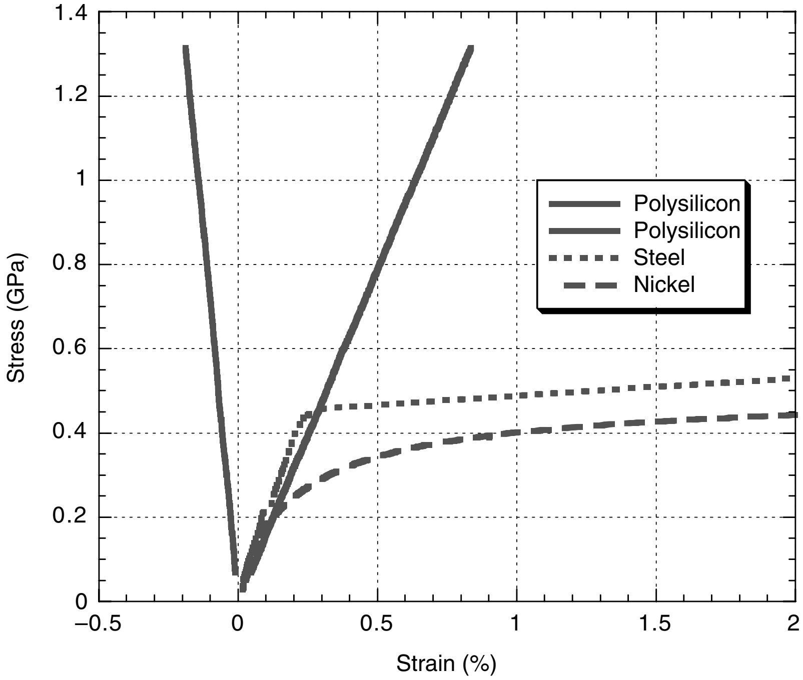


FIGURE 3.11 Representative stress–strain curves of polysilicon, electroplated nickel, and A-533B steel. These are from microspecimens tested in the author's laboratory.

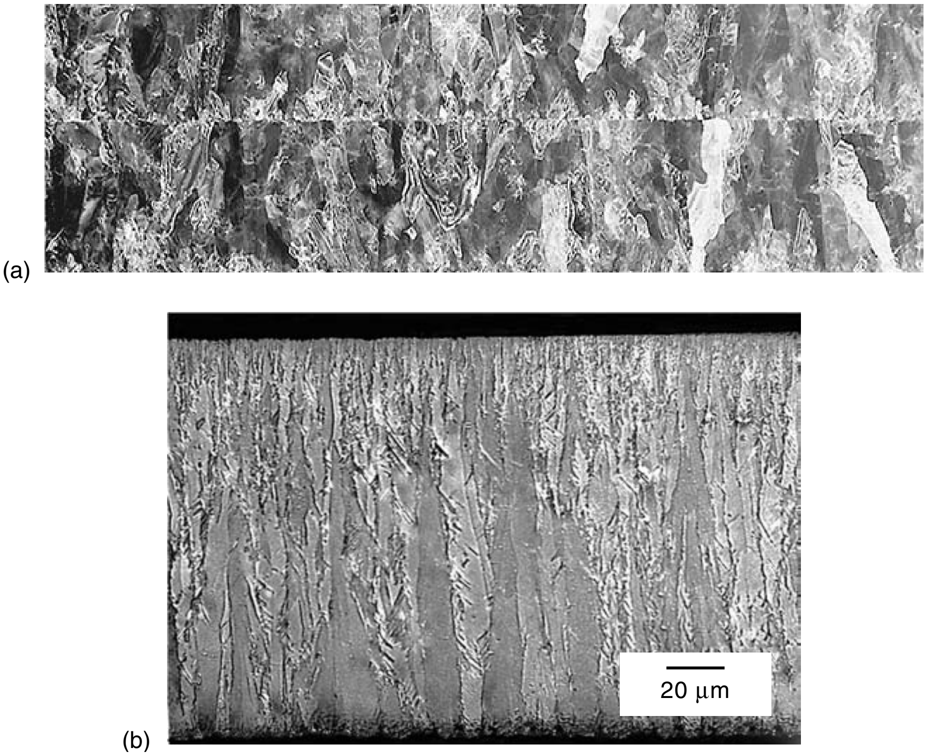


FIGURE 3.12 Microstructure of two common MEMS materials. Note the columnar grain structure perpendicular to the plane of the film. (a) Polysilicon deposited in two layers; the bottom layer is 2.0 μm thick and the top one is 1.5 μm thick. (Reprinted with permission from Sharpe et al. [1998c] "Round-Robin Tests of Modulus and Strength of Polysilicon," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 518*, pp. 56–65, 15–16 April, Francisco. © 1998 IEEE.) (b) Nickel electroplated into LIGA molds. (Reprinted with permission from Sharpe et al. [1997d] "Measurements of Young's Modulus, Poisson's Ratio, and Tensile Strength of Polysilicon," *Proc. IEEE Tenth Annual Int. Workshop on Micro Electro Mechanical Systems*, pp. 424–29, 26–30 January, Nagoya, Japan. © 1998 IEEE.)

TABLE 3.1 Metals

Metals	Young's Modulus (GPa)	Yield Strength (GPa)	Ultimate Strength (GPa)	Method	Comments	Ref.
Aluminum	8–38	—	0.04–0.31	Tension	110–160 µm thick	Hoffman (1989)
modulus of bulk material = 69 GPa	40	—	0.15	Tension	1.0 µm thick	Ogawa et al. (1996)
	69–85	—	—	Bending	Various lengths	Comella and Scanlon (2000)
Copper	86–137	0.12–0.24	0.33–0.38	Tension	Plated; annealed	Buchheit et al. (1999)
modulus of bulk material = 117 GPa	108–145	—	—	Indentation	Various locations	Buchheit et al. (1999)
	98 ± 4	—	—	Tension	Laser speckle	Anwander et al. (2000)
Gold	40–80	—	0.2–0.4	Tension	0.06–16 µm thick	Neugebauer (1960)
modulus of bulk material = 74 GPa	57	0.26	—	Bending	~1 µm thick	Weihs et al. (1988)
	74	—	—	Indentation	~1 µm thick	Weihs et al. (1988)
	82	—	0.33–0.36	Tension	0.8 µm thick	Emery et al. (1997)
	—	—	0.22–0.27	Bending	Composite beam	Kraft et al. (1998)
Titanium						
modulus of bulk material = 110 GPa	96 ± 12	—	0.95 ± 0.15	Tension	0.5 µm thick	Ogawa et al. (1997)
Ti–Al–Ti	—	0.07–0.12	0.14–0.19	Tension	Composite film	Read and Dally (1992)

TABLE 3.2 Diamond-Like Carbon

Young's Modulus (GPa)	Fracture Strength (GPa)	Method	Comments	Ref.
600–1100	0.8–1.8	Bending	Hot flame deposited	Hollman et al. (1995)
800–1140	—	Ultrasonic	CVD diamond	Schneider and Tucker (1996)
150–800	—	Ultrasonic	Laser arc deposited	Schneider and Tucker (1996)
580	—	Brillouin	CVD diamond	Monteiro et al. (1996)
94–128	—	Buckling	Poisson's ratio = 0.22	Cho et al. (1998)
—	8.5 ± 1.4	Tension	Amorphous diamond	LaVan et al. (2000a)

Carbon can be deposited to form an amorphous or crystalline structure that is often referred to as diamond-like carbon, (DLC). Diamond itself has a very high stiffness and strength as well as a low coefficient of friction; for these reasons DLC offers exciting possibilities in MEMS. The very limited results to date, shown in Table 3.2, support this line of reasoning although they are far too sparse to be conclusive.

Electroplated nickel and nickel–iron MEMS, usually manufactured via the LIGA process, offer the possibility of larger and stronger actuators and connectors. The microstructure and mechanical properties of an electroplated material are highly dependent upon the composition of the plating bath and on the current and temperature. Similarly, the composition of a nickel–iron alloy significantly affects its characteristics. Young's modulus and strength values are listed in Tables 3.3 and 3.4 for nickel and nickel–iron respectively. The modulus of bulk nickel is around 200 GPa, and the yield strength of pure fine-grained nickel is approximately 60 MPa [ASM, 1990]. Table 3.3 shows that the modulus of nickel is generally somewhat lower and the strength considerably higher. Nickel–iron has a smaller modulus, as expected, but can be a very strong material as seen from the limited results in Table 3.4.

TABLE 3.3 Nickel

Young's Modulus (GPa)	Yield Strength (GPa)	Ultimate Strength (GPa)	Method	Comments	Ref.
202	0.40	0.78	Tension	Vibration for modulus	Mazza et al. (1996b)
~200	—	—	Ultrasonic	3–75 µm thick	Schneider and Tucker (1996)
168–182	0.1 ± 0.01	—	FE Model	Microgrippers	Basrour et al. (1997)
205	—	—	Resonance	Also fatigue	Dual et al. (1997)
68*	—	—	Torsion	*Shear modulus	Dual et al. (1997)
176 ± 30	0.32 ± 0.03	0.55	Tension	~200 µm thick	Sharpe et al. (1997e)
131–160	0.28–0.44	0.46–0.76	Tension	Varied current	Christenson et al. (1998)
231 ± 12	1.55 ± 05	2.47 ± 0.07	Tension	6 µm thick	Greek and Ericson (1998)
180 ± 12	—	—	Resonance	Film on resonator	Kahn et al. (1998)
181 ± 36	0.33 ± 0.03	0.44 ± 0.04	Tension	LIGA 3 films	Sharpe and McAleavey (1998)
158 ± 22	0.32 ± 0.02	0.52 ± 0.02	Tension	LIGA 4 films	Sharpe and McAleavey (1998)
182 ± 22	0.42 ± 0.02	0.60 ± 0.01	Tension	HI-MEMS films	Sharpe and McAleavey (1998)
153 ± 14	—	1.28 ± 0.24*	Bending	*Modulus of rupture	Stephens et al. (1998)
156 ± 9	0.44 ± 0.03	—	Tension	Current = 20 ma/cm^2	Buchheit et al. (1999)
92	0.06/0.16*	—	*Tension/ compression	Annealed	Buchheit et al. (1999)
160 ± 1	0.28/0.27*	—	*Tension/ compression	Current = 50 ma/cm^2	Buchheit et al. (1999)
146–184	—	—	Indentation	Various locations	Buchheit et al. (1999)
194	—	—	Tension	Laser speckle	Anwander et al. (2000)

TABLE 3.4 Nickel–Iron

Young's Modulus (GPa)	Yield Strength (GPa)	Ultimate Strength (GPa)	Method	Comments	Ref.
65	—	—	Fixed ends	80% Ni–20% Fe	Chung and Allen (1996)
119	0.73	1.62	Tension	50% Ni–50% Fe	Dual et al. (1997)
115	—	—	Resonance	50% Ni–50% Fe	Dual et al. (1997)
15–54*	—	—	Torsion	*Shear modulus	Dual et al. (1997)
155	—	2.26	Tension	Electroplated	Greek and Ericson (1998)
—	1.83–2.20	2.26–2.49	Tension	HI-MEMS films	Sharpe and McAleavey (1998)

The most common MEMS material, polysilicon, is also the most tested, as Table 3.5 demonstrates. The stiffness coefficients of single-crystal silicon are well established, and the modulus in different directions can vary from 125 to 180 GPa [Sato et al., 1997]. Aggregate theories predict that randomly oriented polycrystalline silicon should have a Young's modulus between 163 and 166 GPa [Guo et al., 1992; Jayaraman et al., 1999]. Most of the modulus values in Table 3.5 are near or within this range, but some vary widely, especially when a test method is first used. An estimate of what the fracture strength should be is more difficult as it depends on the flaws in the material. Even though strength is easier to measure than modulus (one needs to measure only force), there are fewer entries. This is because many of the bending, resonance, and bulge tests do not lead to failure in the specimen.

Single-crystal silicon has also been studied extensively, as Table 3.6 shows. The modulus values are measured along particular crystallographic directions, so they should not be expected to compare with the polysilicon values.

Silicon carbide holds promise for MEMS because of its expected high stiffness, strength, and chemical and temperature stability; and Sarro (2000) provides a thorough overview of its potential. Bulk silicon carbide is commonly available, but manufacturing processes for thin, free-standing films are still in development. Table 3.7 lists results from the few tests to date; note that no strength values appear.

TABLE 3.5 Polysilicon

Young's Modulus (GPa)	Fracture Strength (GPa)	Method	Comments	Ref.
160	—	Bulge	Obtains residual stress	Tabata et al. (1989)
123	—	Fixed ends	Heavily doped	Tai and Muller (1990)
190–240	—	Bulge	Various etches	Walker et al. (1990)
164–176	2.86–3.37	Tension	Varied grain size	Koskinen et al. (1993)
—	2.11–2.77	Bending	CMOS process	Biebl et al. (1995a)
147 ± 6	—	Resonance	Temperature effects	Biebl et al. (1995b)
170	—	Bending	Varied doping	Biebl and Philipsborn (1995)
—	0.57-0.77	Tension	Weibull analysis	Greek et al. (1995)
151–162	—	Bulge	Various anneals	Maier-Schneider et al. (1995)
163	—	Resonance	Temperature effects	Michalicek et al. (1995)
171–176	—	Fixed ends	Pull-in voltage	Zou et al. (1995)
149 ± 10	—	Fixed ends	Pull-in voltage	Gupta et al. (1996)
150 ± 30	—	Resonance	10 µm thick	Kahn et al. (1996)
140*	0.70	Tension	*Approximate	Read and Marshall (1996)
152–171	—	Ultrasonic	0.4 µm thick	Schneider and Tucker (1996)
176–201	—	Indentation	Different depths	Bhushan and Li (1997)
160–167	1.08–1.25	Tension	Weibull analysis	Greek and Johansson (1997)
178 ± 3	—	Fixed ends	Ph.D. thesis	Gupta (1997)
169 ± 6	1.20 ± 0.15	Tension	Poisson's ratio = 0.22 ± .01	Sharpe et al. (1997d)
174 ± 20	2.8 ± 0.5	Bending	Tested by Jones et al.	Sharpe et al. (1998c)
132	—	Tension	Tested by Chasiotis et al.	Sharpe et al. (1998c)
137 ± 5	2.7 ± 0.2	Resonance	Tested by Brown et al.	Sharpe et al. (1998c)
140 ± 14	1.3 ± 0.1	Tension	Tested by Sharpe et al.	Sharpe et al. (1998c)
172 ± 7	1.76	Tension	10 µm thick	Greek and Ericson (1998)
162 ± 4	—	Bulge	Poisson's ratio = 0.19 ± .03	Jayaraman et al. (1998)
168 ± 4	—	Resonance	0.45–0.9 µm thick	Kahn et al. (1998)
135 ± 10	—	Bending	AFM	Serre et al. (1998)
95–167	—	Indentation	Also wear tests	Sundararajan and Bhushan (1998)
167	2.0–2.7	Tension	Modulus from bulge; P-doped	Tsuchiya et al. (1998a)
163	2.0–2.8	Tension	Modulus from bulge; undoped	Tsuchiya et al. (1998a)
—	1.8–3.7	Tension	Different sizes and anneals	Tsuchiya et al. (1998b)
95/175	—	Indentation	Doped and undoped	Li and Bhushan (1998)
198	—	Bending	Capacitive device	Que et al. (1999)
166 ± 5	1.0 ± 0.1	Tension	Force-displacement	Chasiotis and Knauss (2000)
—	4.27 ± 0.61	Tension	By LaVan et al.	LaVan et al. (2000b)
—	2.85 ± 0.40	Tension	By Sharpe et al.	LaVan et al. (2000b)
—	3.23 ± 0.25	Tension	By Tsuchiya et al.	LaVan et al. (2000b)
158 ± 8	1.56 ± 0.25	Tension	Size effects	Sharpe and Jackson (2000)
159 and 169	—	Tension	Two specimens from Sharpe	Yi (pers. comm.)
—	3.2 ± 0.3	Bending	Assumed E = 160 GPa	Jones et al. (2000)
—	2.9 ± 0.5	Tension	4 µm thick	Kapels et al. (2000)
—	3.4 ± 0.5	Bending	4 µm thick	Kapels et al. (2000)

Silicon nitride commonly appears in both MEMS and in microelectronics as an insulating layer, and interest in its use as a structural material is growing. Table 3.8 lists its properties. Silicon oxide is also typically included in a MEMS or microelectronics process, but it is less likely to be used as a structural component because of its low stiffness and strength, as shown in Table 3.9.

To date, the main application of the polymer SU-8 is as a mask material for thicker electroplated metal MEMS. Its use as a structural component is possible, but the values of stiffness and strength in Table 3.10 are very low.

Fracture toughness values have been measured for polysilicon; Table 3.11 lists the results. Note that this is not the plane-strain fracture toughness that is a materials property; care is needed, as some authors list this value as KIc.

TABLE 3.6 Silicon Crystals

Young's Modulus (GPa)	Fracture Strength (GPa)	Method	Comments	Ref.
177 ± 18	2.0–4.3	Bending	⟨110⟩	Johansson et al. (1988)
188	—	Indentation		Weihs et al. (1989)
163	>3.4	Bending	⟨110⟩	Weihs et al. (1989)
122 ± 2	—	Bending	⟨110⟩	Ding et al. (1989)
125 ± 1	—	Resonance	⟨110⟩	Ding et al. (1989)
131	—	Resonance		Zhang et al. (1991)
173 ± 13	—	Bending	⟨110⟩	Osterberg et al. (1994)
147	0.26–0.82	Tension	⟨110⟩	Cunningham et al. (1995)
—	8.5–20	Torsion	Shear and normal	Saif and MacDonald (1996)
60–200	—	Indentation	Various doping	Bhushan and Li (1997)
130	—	Resonance	⟨100⟩	Dual et al. (1997)
75	—	Torsion	Shear modulus	Dual et al. (1997)
125–180	1.3–2.1	Tension	Three orientations	Sato et al. (1997)
—	9.5–26.4	Bending	Various etches	Chen et al. (1998)
—	0.7–3.0	Bending	Measured roughness	Chen et al. (1999)
142 ± 9	1.73	Tension	⟨100⟩	Greek and Ericson (1998)
165 ± 20	2–8	Bending	Fatigue tests also	Komai et al. (1998)
168	—	Indentation	⟨100⟩	Li and Bhushan (1999)
—	0.59 ± 0.02	Tension	⟨100⟩	Mazza and Dual (1999)
—	2–6	Bending	Fatigue also	Minoshima et al. (1999)
169.2 ± 3.5	0.6–1.2	Tension	Various etches	Yi and Kim (1999b)
115–191	—	Tension	Three orientations	Yi and Kim (1999c)
164.9 ± 4	—	Tension	Laser speckle	Anwander et al. (2000)
169.9	0.5–17	Bending	Various sizes	Namazu et al. (2000)

TABLE 3.7 Silicon Carbide

Young's Modulus (GPa)	Fracture Strength (GPa)	Method	Comments	Ref.
394	—	Bulge	3C–SiC	Tong and Mehregany (1992)
88 ± 10 to 242 ± 30	—	Bulge + indentation	Amorphous SiC	El Khakani et al. (1993)
694	—	Resonance	3C–SiC	Su and Wettig (1995)
100–150	—	Ultrasonic	0.2–0.3 μm thick	Schneider and Tucker (1996)
331	—	Bulge	3C–SiC; assumed	Mehregany et al. (1997)
n = 0.25 196 and 273	—	Acoustic microscopy	Amorphous SiC	Cros et al. (1997)
395	—	Indentation	3C–SiC	Sundararajan and Bhushan (1998)
470 ± 10	—	Bending	3C–SiC	Serre et al. (1999)

Poisson's ratio is an important materials property when the stress state is biaxial, but only a very limited number of measurements have been made. Those are listed in the comments columns of the tables.

The question of the effect of size on the strength of MEMS materials often arises. This is because MEMS structural components can be on the same size scale as fine single-crystal "whiskers" of materials, which can have very high strengths, the premise being that they have fewer imperfections. However, there are no dramatic increases in strength because the materials still have fine grains relative to the specimen size. Tsuchiya et al. (1998) found an increase in the tensile strength of polysilicon specimens 2.0 μm thick as their length increased from 30 to 300 μm, but the gain was only 30%. Recent results show that the modulus of polysilicon does not vary with specimen size, but the strength increases from 1.21 to 1.65 GPa with decreasing specimen size [Sharpe et al., 2001]. From a practical point of view, the effect of size on strength for common MEMS structural components is not a concern.

On the other hand, Namazu et al. (2000) tested silicon crystal beams ranging in width from 0.2 to 1.04 mm, in thickness from 0.25 to 0.52 mm and in length from 6 to 9.85 mm. The beams were prepared by anisotropic etching; the smallest were tested using an atomic force microscope, and the largest with a

TABLE 3.8 Silicon Nitride

Young's Modulus (GPa)	Fracture Strength (GPa)	Method	Comments	Ref.
130–146 ± 20%	—	Resonance	~0.3 μm thick	Petersen and Guarnieri (1979)
230 and 330	—	Bulge	Different processing	Hong et al. (1990)
373	—	Fixed ends	Low stress	Tai and Muller (1990)
101–251*	—	Indentation	*Assume Poisson's ratio = 0.27	Taylor (1991)
110 and 160*	0.39–0.42	Bulge	*Biaxial modulus	Cardinale and Tustison (1991)
222 ± 3	—	Bulge	Poisson's ratio = 0.28 ± 0.05	Vlassak and Nix (1992)
216 ± 10	—	Indentation		Vlassak and Nix (1992)
230–265	—	Ultrasonic	0.2–0.3 μm thick	Schneider and Tucker (1996)
192	—	Resonance		Buchaillot et al. (1997)
194.25 ± 1%	—	Resonance		Hoummady et al. (1997)
130	—	Buckling		Ziebart et al. (1999)
290	7.0 ± 0.9	Bending		Kuhn et al. (2000)
202.57 ± 15.80	12.26 ± 1.69*	Fixed ends	*Bending strength	Zhang et al. (2000)
255 ± 3	6.4 ± 1.1	Tension	Poisson's ratio = 0.23 ± 0.01	G. Coles (pers. comm.)

TABLE 3.9 Silicon Oxide

Young's Modulus (GPa)	Fracture Strength (GPa)	Method	Comments	Ref.
66*	—	Bulge	*Assumed n = 0.18	Jaccodine and Schlegel (1966)
57–92 ± 20%	—	Resonance	Various depositions	Petersen and Guarnieri (1979)
64	>0.6	Indentation		Weihs et al. (1988)
83	—	Bending		Weihs et al. (1988)
—	0.6–1.9	Tension	In vacuum and in air	Tsushiya et al. (1999)

TABLE 3.10 SU-8

Young's modulus (GPa)	Yield Strength (GPa)	Ultimate Strength (GPa)	Method	Comments	Ref.
~3	—	0.12–0.13	Tension		McAleavey et al. (1998)
1.5–3.1	0.03–0.05	0.05–0.08	Tension	Strain by SIEM	Chang et al. (2000)

TABLE 3.11 Fracture Toughness Values

Fracture Toughness (MPa-m1/2)	Test Method	Material	Ref.
1.8 ± 0.3	Tension; edge crack	Silicon nitride; two kinds	Fan et al. (1990)
1.2	Indentation	Silicon crystal	DeBoer et al. (1993)
0.96–1.65	Double cantilever	Silicon crystal	Fitzgerald et al. (1999)
1.4 ± 0.6	Tension; center crack	Polysilicon	Sharpe et al. (1997f)
1.9–4.5	Tension; edge crack	Polysilicon	Tsuchiya et al. (1997)
3.5–5.0	Notched specimen	Polysilicon; various dopings	Ballarini et al. (1998)
1.1–2.7	Notched specimen	Polysilicon; various dopings	Kahn et al. (1999)
1.2 ± 0.3	Sharp precrack	Polysilicon	Kahn et al. (2000)
1.6 ± 0.3	Tension; corner	Polysilicon	K. Jackson (pers. comm.)
1.0 ± 0.1	Surface crack	Polysilicon	J. Bagdahn (pers. comm.)

microhardness tester. The mean bending strengths covered an astonishing range from 0.47 to 17.5 GPa — a factor of 37.

3.5 Initial Design Values

If the manufacturing and testing technology for materials used in MEMS were as fully developed as those associated with common structural materials, such as aluminum, for example, then this entire chapter

TABLE 3.12 Initial Design Values

Material	Young's Modulus (GPa)	Poisson's Ratio	Yield Strength (GPa)	Ultimate or Fracture Strength (GPa)
Aluminum	70	—	—	0.15
Copper	120	—	0.15	0.35
Gold	70	—	—	0.30
Nickel	180	—	0.30	0.50
Nickel–iron	120	—	0.70	1.60
Diamond-like carbon	800	0.22	—	8.0
Polysilicon	160	0.22	—	1.2–3.0
Silicon crystal	125–180	—	—	>1.0
Silicon carbide	400	0.25	—	—
Silicon nitride	250	0.23	—	6.0
Silicon oxide	70	—	—	1.0

could have been reduced to a one-page table. However, that is not the case; the materials themselves are new, and the test methods are still in their infancy. It may be useful to list "best guesses" at the material properties of MEMS materials to be used in an initial design, and Table 3.12 does that. These are only estimates, and the actual properties resulting from a particular manufacturing process may be quite different from these nominal values.

Aluminum, copper, and gold have essentially the same modulus values as the bulk materials, but the ultimate strengths are slightly higher than those found for commercially pure materials. Young's modulus for thin-film nickel can vary depending upon the deposition parameters, but it is conservative to assume that it will be lower (at 180 GPa) than the 200 GPa expected for bulk pure nickel. There are fewer results for nickel–iron, so the modulus of 120 GPa is only a rough estimate. However, it is clear that thin-film nickel and nickel–iron alloys are quite a bit stronger than one would expect from knowledge of bulk behavior.

The values listed for diamond-like carbon are only an optimistic guide. There are many variations of this material, and very few test results. These properties are included because such a material would be very attractive if it could be realized.

Polysilicon has certainly been thoroughly tested and is widely used, but there still is no "standard" value — at least for Young's modulus. The explanation for this is, of course, the difficulty in testing at this size scale, but there is a clear trend toward a modulus in the neighborhood of 160 GPa. An assumption of that number ±10 GPa can be used with confidence in the initial design of a microdevice. It is also clear that the strength can vary depending upon the manufacturing process but will fall in the range of 1.2 to 3.0 GPa.

Single-crystal silicon has been thoroughly characterized to the point that it has been used as a "standard material" to validate test systems. The modulus depends on orientation, and the strength range is enormous with some extremely high values being reported.

Silicon carbide is widely promoted as a MEMS material, but conclusive measurements of its modulus have yet to be made, and there are no measurements of strength. One should use the modulus value with caution. The situation is better for silicon nitride, as it has been more widely used and tested.

Although Table 3.12 lists numbers to three significant figures, the reader will surely appreciate their unreliability and wonder as to their value. But many other uncertainties occur between the initial design and the product. Dimensions may not result as specified, and that can have a profound effect on the stiffnesses of small components. Boundary conditions may not be as specified either, due to variations in etch release processes. Nevertheless, the values in Table 3.12 offer a starting point. Users should certainly refer to the more detailed information in the other tables and probably should consult the appropriate references.

Acknowledgments

The author is grateful for the interactions with students and colleagues over the past five years that have provided the background for this chapter. Vanessa Coleman's assistance with preparation is appreciated.

The effort was sponsored by the Defense Advanced Research Projects Agency (DARPA) and Air Force Research Laboratory, Air Force Materiel Command, USAF, under agreement number F30602-99-2-0553. The U.S. government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon.

References

Amimoto, S.T., Chang, D.J., and Birkitt, A. (1998) "Stress Measurements in MEMS Using Raman Spectroscopy," *Proc. SPIE* **3512**, pp. 123–29.

Anwander, M., Kaindl, G., Klein, M., and Weiss, B. (2000) "Noncontacting Laser Based Techniques for the Determination of Elastic Constants of Thin Foils," *Micromat 2000*, pp. 1100–3, 17–19 April, Berlin, Germany.

ASM (1990) *Metals Handbook*, 10th ed., vol. 2, ASM International, Materials Park, OH, p. 1143.

ASTM (1970) *Review of Developments in Plane Strain Fracture Toughness Testing,* STP 463, American Society for Testing and Materials, New York.

ASTM (2000a) *Metals: Mechanical Testing; Elevated and Low-Temperature Tests; Metallography*, vol. 03.01, Annual Book of ASTM Standards, American Society for Testing and Materials, New York.

ASTM (2000b) *Refractories: Carbon and Graphite Products; Activated Carbon*, vol. 15.01, Annual Book of ASTM Standards, American Society for Testing and Materials, New York.

Ballarini, R. (1998) "The Role of Mechanics in Microelectromechanical Systems (MEMS) Technology," AFRL-ML-WP-TR-1998-4209, Air Force Research Laboratory, Wright-Patterson Air Force Base, OH.

Ballarini, R., Mullen, R.L., Kahn, H., and Heuer, A.H. (1998) "The Fracture Toughness of Polysilicon Microdevices," in *Microelectromechanical Structures for Materials Research*, *Materials Research Society Symposium 518*, pp. 137–42, 15–16 April, San Francisco.

Basrour, S., Robert, L., Ballandras, S., and Hauden, D. (1997) "Mechanical Characterization of Microgrippers Realized by LIGA Technique," in *Proc. Int. Solid-State Sensors and Actuators Conf. — Transducers '97*, pp. 599–602, 16–19 June, Chicago.

Beams, J.W. (1959) "Mechanical Properties of Thin Films of Gold and Silver," in *Proc. Int. Conf. Sponsored by Air Force Office of Scientific Research, Air Research and Development Command and The General Electric Research Laboratory '59*, pp. 183–92, 9–11 September, Bolton Landing, NY.

Benrakkad, M.S., Benitex, M.A., Esteve, J., Lopez-Villegas, J.M., Samitier, J., and Morante, J.R. (1995) "Stress Measurement by Microraman Spectroscopy of Polycrystalline Silicon Structures," *J. Micromech. Microeng.* **5**, pp. 132–35.

Bhushan, B., and Li, X. (1997) "Micromechanical and Tribological Characterization of Doped Single-Crystal Silicon and Polysilicon Films for Microelectromechanical Systems Devices," *J. Mater. Res.* **12**, pp. 54–63.

Biebl, M., and von Philipsborn, H. (1995) "Fracture Strength of Doped and Undoped Polysilicon," in *Proc. Int. Solid-State Sensors and Actuators Conf. — Transducers '95*, pp. 72–75, 25–29 June, Stockholm.

Biebl, M., Brandl, G., and Howe, R.T. (1995a) "Young's Modulus of In-Situ Phosphorus-Doped Silicon," in *Proc. Int. Solid-State Sensors and Actuators Conf. — Transducers '95*, pp. 80–83, 25–29 June, Stockholm.

Biebl, M., Scheiter, T., Hierold, C., Philipsborn, H., and Klose, H. (1995b) "Micromechanics Compatible with an 0.8 μm CMOS Process," *Sensor. Actuator. A (Phys.)* **46–47**, pp. 593–97.

Brown, S.B., Povirk, G., and Connally, J. (1993) "Measurement of Slow Crack Growth in Silicon and Nickel Mechanical Devices," in *Proc. IEEE Micro Electro Mechanical Systems*, pp. 99–104, 7–10 February, Fort Lauderdale.

Brown, S.B., Van Arsdell, W., and Muhlstein, C.L. (1997) "Materials Reliability in MEMS Devices," in *Proc. Int. Solid-State Sensors and Actuators Conf. — Transducers '97*, pp. 591–93, 16–19 June, Chicago.

Buchaillot, L., Farnault, E., Hoummady, M., and Fujita, H. (1997) "Silcon Nitride Thin Films Young's Modulus Determination by an Optical Non Destructive Method," *Jpn. J. Appl. Phys.*, part 2 (Lett.) **36**, pp. L794–L797.

Buchheit, T.E., Christenson, T.R., Schmale, D.T., and Lavan, D.A. (1999) "Understanding and Tailoring the Mechanical Properties of LIGA Fabricated Materials," in *Materials Science of Microelectromechanical Systems (MEMS) Devices, Materials Research Society Symposium 546*, pp. 121–126, 1–2 December, Boston.

Cardinale, G.F., and Tustison, R.W. (1992) "Fracture Strength and Biaxial Modulus Measurement of Plasma Silicon Nitride Films," *Thin Solid Films* **207**, pp. 126–30.

Chang, D.J., and Sharpe, W.N., Jr. (1999) "Mechanical Analysis and Properties of MEMS Materials," *Microengineering for Aerospace Systems*, The Aerospace Press of the Aerospace Corporation, El Segundo, CA, pp. 73–118.

Chang, S., Warren, J., and Chiang, F.P. (2000) "Mechanical Testing of EPON SU-8 with SIEM," in *Microscale Systems: Mechanics and Measurements Symposium,*" Society for Experimental Mechanics, pp. 46–49, 8 June, Orlando.

Chasiotis, I., and Knauss, W.G. (1998) "Mechanical Properties of Thin Polysilicon Films by Means of Probe Microscopy," *Proc. SPIE* **3512**, pp. 66–75.

Chasiotis, I., and Knauss, W.G. (2000) "Instrumentation Requirements in Mechanical Testing of MEMS Materials," in *Microscale Systems: Mechanics and Measurements Symposium, Society for Experimental Mechanics*, pp. 56–61, 8 June, Orlando.

Chen, K.S., Ayon, A., and Spearing, M. (1998) "Silicon Strength Testing for Mesoscale Structural Applications," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 518*, pp. 123–30, 15–16 April, San Francisco.

Chen, K.S., Ayon, A., Lohner, K.A., Kepets, M.A., Melconian, T.K., and Spearing, S.M. (1999) "Dependence of Silicon Fracture Strength and Surface Morphology on Deep Reactive Ion Etching Parameters," in *Materials Science of Microelectromechanical Systems (MEMS) Devices, Materials Research Society Symposium 546*, pp. 21–26, 1–2 December, Boston.

Cho, S.J., Lee, K.R., Eun, K.Y., Han, J.H., and Ko, D.H. (1997) "A Method for Independent Measurement of Elastic Modulus and Poisson's Ratio of Diamond-Like Carbon Films," in *Thin-Films-Stresses and Mechanical Properties VII, Materials Research Society Symposium 505*, pp. 33–38, 1–5 December, Boston.

Cho, S.J., Kwang-Ryeol, L., Eun, K.Y., and Ko, D.H. (1998) "Measurement of Elastic Modulus and Poisson's Ratio of Diamond-Like Carbon Films," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 518*, pp. 203–8, 15–16 April, San Francisco.

Christenson, T.R., Buchheit, T.E., Schmale, D.T., and Bourcier, R.J. (1998) "Mechanical and Metallographic Characterization of LIGA Fabricated Nickel and 80%Ni-20%Fe Permalloy," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 518*, pp. 185–190, 15–16 April, San Francisco.

Chung, C.C., and Allen, M.G. (1996) "Measurement of Mechanical Properties of Electroplated Nickel–Iron Alloys," ASME Dynamic Syst. Contr. Div. DSC 59, pp. 453–457.

Comella, B.T., and Scanlon, M.R. (2000) "The Determination of the Elastic Modulus of Microcantilever Beams Using Atomic Force Microscopy," *J. Mater. Sci.* **35**, pp. 567–72.

Connally, J.A., and Brown, S.B. (1993) "Micromechanical Fatigue Testing," *Exp. Mech.* **33**, pp. 81–90.

Cornella, G., Vinci, R.P., Iyer, R.S., and Dauskardt, R.H. (1998) "Observations of Low-Cycle Fatigue of Al Thin Films for MEMS Applications," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 518*, pp. 81–86, 15–16 April, San Francisco.

Cros, B., Gat, E., and Saurel, J. (1997) "Characterization of the Elastic Properties of Amorphous Silicon Carbide Thin Films by Acoustic Microscopy," *J. Non-Crystalline Solids* **209**, pp. 273–82.

Cunningham, S.J., Wan, S., and Read, D.T. (1995) "Tensile Testing of Epitaxial Silicon Films," in *Proc. Int. Solid-State Sensors and Actuators Conf. — Transducers '95*, pp. 96–99, 25–29 June, Stockholm.

Dally, J.W., and Read, D.T. (1993) "Electron Beam Moire," *Exp. Mech.* **33**, pp. 270–77.

DeBoer, M.P., Huang, H., Nelson, J.C., Jiang, Z.P., and Gerberich, W. (1993) "Fracture Toughness of Silicon and Thin Film Micro Structures by Wedge Indentation," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 308*, pp. 647–52, 30 November–2 December, Boston.

Ding, X., Ko, W.H., and Mansour, J.M. (1989) "Residual Stress and Mechanical Properties of Boron-Doped-Silicon Films," in *5th Int. Conf. on Solid-State Sensors and Actuators and Eurosensors III*, pp. 866–71, 25–30 June, Montreux, Switzerland.

Dual, J., Mazza, E., Schiltges, G., and Schlums, D. (1997) "Mechanical Properties of Microstructures: Experiments and Theory," *Proc. SPIE* **3225**, pp. 12–22.

El Khakani, M.A., Chaker, M., Jean, A., Boily, S., and Kieffer, J.C. (1993) "Hardness and Young's Modulus of Amorphous a-SiC Thin Films Determined by Nanoindentation and Bulge Tests," *J. Mater. Res.* **9**, pp. 96–103.

Emery, R.D., Lenshek, D.X., Behin, B., Gherasimova, M., and Povrik, G.L. (1997) "Tensile Behavior of Free-Standing Gold Films," in *Thin-Films-Stresses and Mechanical Properties VII, Materials Research Society Symposium 505*, pp. 361–66, New Haven.

Fan, L.S., Howe, R.T., and Muller, R.S. (1990) "Fracture Toughness Characterization of Brittle Thin Films," *Sensor. Actuator.* **A21–A23**, pp. 872–74.

Fitzgerald, A.M., Iyer, R.S., Dauskart, R.H., and Kenny, T.W. (1998) "Fracture and Sub-Critical Crack Growth Behavior of Micromachined Single Crystal Silicon Structures," ASME Dynamic System Contr. Div. DSC 66, pp. 395–99.

Fitzgerald, A.M., Iyer, R.S., Dauskardt, R.H., and Kenny, T.W. (1999) "Fracture Toughness and Crack Growth Phenomena of Plasma-Etched Single Crystal Silicon," in *Transducers '99: 10th Int. Conf. on Solid-State Sensors and Actuators*, pp. 194–99, 7–10 June, Sendai, Japan.

Greek, S., and Ericson, F. (1998) "Young's Modulus, Yield Strength, and Fracture Strength of Microelements Determined by Tensile Testing," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 518*, pp. 51–56, 15–16 April, San Francisco.

Greek, S., and Johansson, S. (1997) "Tensile Testing of Thin Film Microstructures," *Proc. SPIE* **3224**, pp. 344–51.

Greek, S., Ericson, F., Johansson, S., and Schweitz, J. (1995) "In Situ Tensile Strength Measurement of Thick-Film and Thin-Film Micromachined Structures," in *Proc. Int. Solid-State Sensors and Actuators Conf. — Transducers '95*, pp. 56–59, 25–29 June, Stockholm.

Guo, S., Daowen, Z., and Wang, W. (1992) "Theoretical Calculation for the Young's Modulus of Poly-Si and a-Si Films," in *Smart Materials Fabrication, Materials Research Society Symposium 276*, pp. 233–38, 28–30 April, San Francisco.

Gupta, R. (1997) Electrosatic Pull-In Test Structure Design for In-Situ Mechanical Property Measurements of Microelectromechanical Systems (MEMS), Ph.D. thesis, Massachusetts Institute of Technology.

Gupta, R.K., Osterberg, P.M., and Senturia, S.D. (1996) "Material Property Measurements of Micromechanical Polysilicon Beams," *Proc. SPIE* **2880**, pp. 39–45.

Hoffman, R.W. (1989) "Nanomechanics of Thin Films: Emphasis, Tensile Properties," in *Thin Films: Stresses and Mechanical Properties Symposium, Materials Research Society Symposium 130*, pp. 295–307, 28–30 November, Boston.

Hollman, P., Alahelisten, A., Olsson, M., and Hogmark, S. (1995) "Residual Stress, Young's Modulus and Fracture Stress of Hot Flame Deposited Diamond," *Thin Solid Films* **270**, pp. 137–42.

Hong, S., Weihs, T.P., Bravman, J.C., and Nix, W.D. (1990) "Measuring Stiffnesses and Residual Stresses of Silicon Nitride Thin Films," *J. Electron. Mater.* **19**, pp. 903–9.

Hoummady, M., Farnault, E., Kawakatsu, H., and Masuzawa, T. (1997) "Applications of Dynamic Techniques for Accurate Determination of Silicon Nitride Young's Moduli," in *Proc. Int. Solid-State Sensors and Actuators Conf. — Transducers '97*, pp. 615–18, 16–19 June, Chicago.

Howard, L.P., and Fu, J. (1997) "Accurate Force Measurements for Miniature Mechanical Systems: A Review of Progress," *Proc. SPIE* **3225**, pp. 2–11.

Jaccodine, R.J., and Schlegel, W.A. (1966) "Measurement of Strains at Si-O2 Interface," *J. Appl. Phys.* **37**, pp. 2429–34.

Jaecklin, V.P., Linder, C., Brugger, J., and deRooij, N.F. (1994) "Mechanical and Optical Properties of Surface Micromachined Torsional Mirrors in Silicon, Polysilicon, and Aluminum," *Sensor. Actuator. A (Phys.)* **43**, pp. 269–75.

Jayaraman, S., Edwards, R.L., and Hemker, K. (1998) "Determination of the Mechanical Properties of Polysilicon Thin Films Using Bulge Testing: Thin-Films-Stresses and Mechanical Properties VII," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 505*, pp. 623–28, 1–5 December, Boston.

Jayaraman, S., Edwards, R.L., and Hemker, K. (1999) "Relating Mechanical Testing and Microstructural Features of Polysilicon Thin Films," *J. Mater. Res.* **14**, pp. 688–97.

Johansson, S., Schweitz, J.A., Tenerz, L., and Tiren, J. (1988) "Fracture Testing of Silicon Microelements In Situ in a Scanning Electron Microscope," *J. Appl. Phys.* **63**, pp. 4799–803.

Johnson, G.C., Jones, P.T., and Howe, R.T. (1999) "Materials Characterization for MEMS: A Comparison of Uniaxial and Bending Tests," *Proc. SPIE* **3874**, pp. 94–101.

Jones, P.T., Johnson, G.C., and Howe, R.T. (1996) "Micromechanical Structures for Fracture Testing of Brittle Thin Films," ASME Dynamic System Contr. Div. DSC 59, pp. 325–30.

Jones, P.T., Johnson, G.C., and Howe, R.T. (2000) "Statistical Characterization of Fracture of Brittle MEMS Materials," in *Proc. SPIE* **3880**, pp. 20–29.

Kahn, H., Stemmer, S., Nandakumar, K., Heuer, A.H., Mullen, R.L., Ballarini, R., and Huff, M.A. (1996) "Mechanical Properties of Thick, Surface Micromachined Polysilicon Films," in *Proc. Ninth Int. Workshop on Micro Electromechanical Systems*, pp. 343–48, 11–15 February, San Diego.

Kahn, H., Huff, M.A., and Heuer, A.H. (1998) "Heating Effects on the Young's Modulus of Films Sputtered onto Micromachined Resonators," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 518*, pp. 33–38, 15–16 April, San Francisco.

Kahn, H., Ballarini, R., Mullen, R.L., and Heuer, A.H. (1999) "Electrostatically Actuated Failure of Microfabricated Polysilicon Fracture Mechanics Specimens," in *Proc. R. Soc. London,* Ser. A 455, pp. 3807–23.

Kahn, H., Tayebi, N., Ballarina, R., Mullen, R.L., and Heur, A.H. (2000) "Fracture Toughness of Polysilicon MEMS Devices," *Sensor. Actuator. A (Phys.)* **82**, pp. 274–80.

Kapels, H., Aigner, R., and Binder, J. (2000) "Fracture Strength and Fatigue of Polysilicon Determined by a Novel Thermal Actuator," *IEEE Trans. Electron Devices* **47**, pp. 1522–28.

Kazinczi, R., Mollinger, J.R., and Bossche, A. (2000) "Versatile Tool for Characterising Long-Term Stability and Reliability of Micromechanical Structures," *Sensor. Actuator. A (Phys.)* **85**, pp. 84–89.

Kobrinsky, M., Deutsch, E., and Senturia S. (1999) "Influence of Support Compliance and Residual Stress on the Shape of Doubly-Supported Surface Micromachined Beams," *MEMS Microelectromech. Syst.* **1**, pp. 3–10.

Komai, K., Minoshima, K., and Inoue, S. (1998) "Fracture and Fatigue Behavior of Single Crystal Silicon Microelements and Nanoscopic AFM Damage Evaluation," *Microsyst. Technol.* **5**, pp. 30–7.

Koskinen, J., Steinwall, J.E., Soave, R., and Johnson, H.H. (1993) "Microtensile Testing of Free-Standing Polysilicon Fibers of Various Grain Sizes." *J. Micromech. Microeng.* **3**, pp. 13–17.

Kraft, O., Schwaiger, R., and Nix, W.D. (1998) "Measurement of Mechanical Properties in Small Dimensions by Microbeam Deflection," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 518*, pp. 39–44, 15–16 April, San Francisco.

Krulevitch, P. (1996) "Technique for Determining the Poisson's Ratio of Thin Films," ASME Dynamic Syst. Contr. Div. DSC 59, pp. 319–23.

Kuhn, J., Fettig, R.K., Moseley, S.H., Kutyrev, A.S., and Orloff, J. (2000) "Fracture Tests of Etched Components Using a Focused Ion Beam Machine," NASA/Goddard Space Flight Center, Greenbelt, MD (prepublication report).

LaVan, D.A., Bucheit, T. E., and Kotula, P.G. (2000a) "Mechanical and Microstructural Characterization of Critical Features of MEMS Materials," in *Microscale Systems: Mechanics and Measurements Symposium*, Society for Experimental Mechanics, pp. 41–45.

LaVan, D.A., Tsuchiya, T., and Coles, G. (2000b) "Cross Comparison of Direct Tensile Testing Techniques on Summit Polysilicon Films," in *Mechanical Properties of Structural Films,* ASTM STP 1413, American Society for Testing and Materials, submitted for publication.

Li, X., and Bhushan, B. (1998) "Measurement of Fracture Toughness of Ultra-Thin Amorphous Carbon Films," *Thin Solid Films* **315**, pp. 214–21.

Li, X., and Bhushan, B. (1999) "Micro/Nanomechanical Characterization of Ceramic Films for Microdevices," *Thin Solid Films* **340**, pp. 210–17.

Maier-Schneider, D., Maibach, J., Obemeier, E., and Schneider, D. (1995) "Variations in Young's Modulus and Intrinsic Stress of LPCVD-Polysilicon Due to High-Temperature Annealing," *J. Micromech. Microeng.* **5**, pp. 121–24.

Mazza, E., and Dual, J. (1999) "Mechanical Behaviour of a mm-Sized Single Crystal Silicon Structure with Sharp Notches," *J. Mech. Phys. Solids* **47**, pp. 1795–821.

Mazza, E., Danuser, G., and Dual, J. (1996a) "Light Optical Deformation Measurements in Microbars with Nanometer Resolution," *Microsyst. Technol.* **2**, pp. 83–91.

Mazza, E., Abel, S., and Dual, J. (1996b) "Experimental Determination of Mechanical Properties of Ni and Ni-Fe Microbars," *Microsyst. Technol.* 2, pp. 197–202.

McAleavey, A., Coles, G., Edwards, R.L., and Sharpe, W.N. (1998) "Mechanical Properties of SU-8," in *Microelectromechanical Structures for Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 546*, pp. 213–18, 1–2 December, Boston.

Mehregany, M., Tong, L., Matus, L.G., and Larkin, D.J. (1997) "Internal Stress and Elastic Modulus Measurements on Micromachined 3C-SiC Thin Films," *IEEE Trans. Electron. Devices* **44**, pp. 74–79.

Menter, J.W., and Pashley, D.W. (1959) "The Microstructure and Mechanical Properties of Thin Films," in *Proc. Int. Conf. Sponsored by Air Force Office of Scientific Research, Air Research and Development Command and The General Electric Research Laboratory '59*, pp. 111–50, 9–11 September, Bolton Landing, NY.

Michalicek, A.M., Sene, D.E., and Bright, V.M. (1995) "Advanced Modeling of Micromirror Devices," in *Proc. Int. Conf. Integrated Micro/Nanotechnology for Space Applications,* pp. 214–29, 30 October–3 November, Houston.

Minoshima, K., Inoue, S., Terada, T., and Komai, K. (1999) "Influence of Specimen Size and Sub-Micron Notch on the Fracture Behavior of Single Crystal Silicon Microelements and Nanoscopic AFM Damage Evaluation," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 546*, pp. 15–20, 1–2 December, Boston.

Monteiro, O.R., Brown, I.G., Sooryakumar, R., and Chirita, M. (1996) "Elastic Properties of Diamond Like Carbon Thin Films: A Brillouin Scattering Study," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 444*, pp. 93–98, 4–5 December, Boston.

Muller, R.S. (1990) "Microdynamics," *Sensor. Actuator. A (Phys.)* **A21–A23**, pp. 1–8.

Muller, R.S. (1997) *Microelectromechanical Systems*, National Academy of Sciences, Washington, D.C.

Namazu, T., Isono, Y., and Tanaka, T. (2000) "Nano-Scale Bending Test of Si Beam for MEMS," in *Proc. IEEE Thirteenth Annual Int. Conf. on Micro Electro Mechanical Systems*, pp. 205–10, 23–27 January, Miyazaki, Japan.

Neugebauer, G. (1960). "Tensile Properties of Thin, Evaporated Gold Films," *J. Appl. Phys.* **31**, pp. 1096–101.

Nieva, P., Tada, H., Zavracky, P., Adams, G., Miaoulis, I., and Wong, P. (1998) "Mechanical and Thermophysical Properties of Silicon Nitride Thin Films at High Temperatures Using In Situ MEMS Temperature Sensors," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 546*, pp. 97–102, 1–2 December, Boston.

Obermeier, E. (1996) "Mechanical and Thermophysical Properties of Thin Film Materials for MEMS: Techniques and Devices," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 444*, pp. 39–57, 4–5 December, Boston.

Ogawa, H., Ishikawa, Y., and Kitahara, T. (1996) "Measurements of Stress–Strain Diagrams of Thin Films by a Developed Tensile Machine," *Proc. SPIE* **2880**, pp. 272–79.

Ogawa, H., and Suzuki, K. et al. (1997) "Measurements of Mechanical Properties of Microfabricated Thin Films," in *Proc. IEEE Tenth Annual Int. Workshop on Micro Electro Mechanical Syst.*, pp. 430–35, 26–30 January, Nagoya, Japan.

Osterberg, P.M., Gupta, R.K., Gilbert, J.R., and Senturia, S.D. (1994) "Quantitative Models for the Measurement of Residual Stress, Poisson Ratio and Young's Modulus Using Electrostatic Pull-In of

Beams and Diaphragms," in *Technical Digest Solid-State Sensor and Actuator Workshop*, pp. 184–88, 13–16 June, Hilton Head Island, SC.

Pan, C.S., and Hsu, W. (1999) "A Microstructure for In Situ Determination of Residual Strain," *J. Microelectromech. Syst.* **8**, pp. 200–7.

Petersen, K.E., and Guarnieri, C.R. (1979) "Young's Modulus Measurements of Thin Films Using Micromechanics," *J. Appl. Phys.* **50**, pp. 6761–66.

Pinardi, K., Jain, S.C., Maes, H.E., Van Overstraeten, R., Willander, M., and Atkinson, A. (1997) "Measurement of Nonuniform Stresses in Semiconductors by the Micro-Raman Method," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 505*, pp. 507–12, 1–5 December, Boston.

Que, L., Li, L., Chu, L., and Gianchandani, Y.B. (1999) "A Micromachined Strain Sensor with Differential Capacitive Readout," in *Proc. 12th Int. Workshop on Micro Electro Mechanical Syst.* pp. 552–57, 17–21 January, Orlando.

Read, D.T., and Dally, J.W. (1992). "A New Method for Measuring the Constitutive Properties of Thin Films," *J. Mater. Res.* **8**, pp. 1542–49.

Read, D.T., and Marshall, R.C. (1996) "Measurements of Fracture Strength and Young's Modulus of Surface-Micromachined Polysilicon," *Proc. SPIE* **2880**, pp. 56–63.

Ruther, P., Bacher, W., Feit, K., and Menz, W. (1995) "Microtesting System Made by the LIGA Process to Measure the Young's Modulus in Cantilever Microbeams," ASME Dynamic Syst. Contr. Div. DSC 57–2, pp. 963–67.

Saif, M.T.A., and MacDonald, N.C. (1996) "Micro Mechanical Single Crystal Silicon Fracture Studies: Torsion and Bending," in *Ninth Annual Int. Workshop on Micro Electro Mechanical Syst.*, pp. 105–109, 11–15 February, San Diego.

Saif, T., and MacDonald, N.C. (1998) "Failure of Micron Scale Single Crystal Silicon Bars Due to Torsion Developed by MEMS Micro Instruments," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 518*, pp. 45–49, San Francisco.

Sarro, P. (2000) "Silicon Carbide as a New MEMS Technology," *Sensor. Actuator. A (Phys.)* **82**, pp. 210–18.

Sato, K., Shikida, M., Yoshioka, T., Ando, T., and Kawbata, T. (1997) "Micro Tensile-Test of Silicon Film Having Different Crystallographic Orientations," in *Proc. Int. Solid-State Sensors and Actuators Conf. — Transducers '97*, pp. 595–98, 16–19 June 1997, Chicago.

Schiltges, G., Gsell, D., and Jual, J. (1998) "Torsional Tests on Microstructures: Two Methods to Determine Shear-Moduli," *Microsyst. Technol.* 5, pp. 22–29.

Schneider, D., and Tucker, M.D. (1996) "Non-Destructive Characterization and Evaluation of Thin Films by Laser Induced Ultrasonic Surface Waves," *Thin Solid Films* **290–291**, pp. 305–11.

Schweitz, J.A., and Ericson, F. (1999) "Evaluation of Mechanical Materials Properites by Means of Surface Micromachined Structures," *Sensor. Actuator. A (Phys.)* **74**, pp. 126–33.

Senturia, S.D. (1998) "CAD Challengers for Microsensors, Microactuators, and Microsystems," *Proc. IEEE* **86**, pp. 1611–26.

Serre, C., Gorostiza, P., Perez-Rodriquez, A., Sanz, F., and Morante, J.R. (1998) "Measurement of Micromechanical Properties of Polysilicon Microstructures with an Atomic Force Microscope," *Sensor. Actuator. A (Phys.)* **67**, pp. 215–19.

Serre, C., Perez-Rodriguez, A., Romano-Rodriquez, A., Morante, J.R., Esteve, J., and Acero, M.C. (1999) "Test Microstructures for Measurement of SiC Thin Film Mechanical Properties," *J. Micromech. Microeng.* **9**, pp. 190–93.

Sharpe, W.N., Jr. (1999) "Fatigue Testing of Materials Used in Microelectromechanical Systems," in *Fatigue '99: Proc. Seventh Int. Fatigue Congress*, pp. 1837–44, 8–12 June, Beijing.

Sharpe, W.N., Jr., and Jackson, K. (2000) "Tensile Testing of MEMS Materials," in *Microscale Systems: Mechanics and Measurements Symposium, Society for Experimental Mechanics*, pp. ix–xiv, 8 June, Orlando.

Sharpe, W.N., Jr., and McAleavey, A. (1998) "Tensile Properties of LIGA Nickel," *Proc. SPIE* **3512**, pp. 30–137.

Sharpe, W.N., Jr., Yuan, B., Vaidyanathan, R., and Edwards, R.L. (1996) "New Test Structures and Techniques for Measurement of Mechanical Properties of MEMS Materials," *Proc. SPIE* **2880**, pp. 78–91.

Sharpe, W.N., Jr., LaVan, D.A., and McAleavey, A. (1997a) "Mechanical Testing of Thicker MEMS Materials," ASME Dynamic Syst. Contr. Div. DSC 62, pp. 93–97.

Sharpe, W.N., Jr., Yuan, B., and Edwards, R.L. (1997b) "Variations in Mechanical Properties of Polysilicon," in *43rd Int. Symp. Instrumentation Society of America*, pp. 179–88, 4–8 May, Orlando.

Sharpe, W.N., Jr., Yuan, B., and Edwards, R.L. (1997c) "A New Technique for Measuring the Mechanical Properties of Thin Films," *J. Microelectromech. Syst.* **6**, pp. 193–99.

Sharpe, W.N., Jr., Yuan, B., and Vaidyanathan, R. (1997d) "Measurements of Young's Modulus, Poisson's Ratio, and Tensile Strength of Polysilicon," *Proc. IEEE Tenth Annual Int. Workshop on Micro Electro Mechanical Systems*, pp. 424–29, 26–30 January, Nagoya, Japan.

Sharpe, W.N., Jr., LaVan, D.A., and Edwards, R.L. (1997e) "Mechanical Properties of LIGA-Deposited Nickel for MEMS Transducers," *Proc. Int. Solid-State Sensors and Actuators Conf. — Transducers '97*, pp. 607–10, 16–19 June, Chicago.

Sharpe, W.N., Jr., Yuan, B., and Edwards, R.L. (1997f) "Fracture Tests of Polysilicon Film," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 505*, pp. 51–56, 1–5 December, Boston.

Sharpe, W.N., Jr., Turner, K., and Edwards, R.L. (1998a) "Polysilicon Tensile Testing with Electrostatic Gripping," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 518*, pp. 191–96, 15–16 April, San Francisco.

Sharpe, W.N., Jr., Danley, D., and LaVan, D.A. (1998b) "Microspecimen Tensile Tests of A533-B Steel," in *Small Specimen Test Techniques*, ASTM STP 1329, American Society of Testing and Materials (ASTM), pp. 497–512, 13–14 January, 1997, New Orleans.

Sharpe, W.N., Jr., Brown, S., Johnson, G.C., and Knauss, W. (1998c) "Round-Robin Tests of Modulus and Strength of Polysilicon," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 518*, pp. 56–65, 15–16 April, Francisco.

Sharpe, W.N., Jr., Turner, K.T., and Edwards, R.L. (1999) "Tensile Testing of Polysilicon," *Exp. Mech.* **39**, pp. 162–70.

Sharpe, W.N., Jr., Jackson, K.M., Hemker, K.J., and Xie, Z. (2001) "Effect of Specimen Size on Young's Modulus and Strength of Polysilicon," *J. Microelectromech. Syst.* **10**, pp. 317–26.

Spearing, S.M. (2000) "Materials Issues in Microelectromechanical Systems (MEMS)," *Acta Mater.* **48**, pp. 179–96.

Stephens, L.S., Kelly, K.W., Meletis, E.I., and Simhadri, S. (1998) "Mechanical Property Evaluation of Electroplated High Aspect Ratio Microstructures," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 518*, pp. 173–78, 15–16 April, San Francisco.

Su, C.M., and Wuttig, M. (1995) "Elastic and Anelastic Properties of Chemical Vapor Deposited Epitaxial 3C-SiC," *J. Appl. Phys.* **77**, pp. 5611–15.

Sundararajan, S., and B. Bhushan (1998) "Micro/Nanotribological Studies of Polysilicon and SiC films for MEMS Applications," *Wear* **217**, pp. 251–61.

Suwito, W., Dunn, M.L., and Cunningham, S. (1997) "Strength and Fracture of Micromachined Silicon Structures," ASME Dynamic Syst. Contr. Div. DSC 62, pp. 99–104.

Tabata, O., Kawahata, K., Sugiyama, S., and Igarashi, I. (1989) "Mechanical Property Measurements of Thin Films Using Load-Deflection of Composite Rectangular Membranes," *Sensor. Actuator.* **20**, pp. 135–41.

Tada, H., Nieva, P., Zavracky, P., Miaoulis, N., and Wong, P.Y. (1998) "Determining the High-Temperature Properties of Thin Films Using Bilayered Cantilevers," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 546*, pp. 39–44, 1–2 December, Boston.

Tai, Y.C., and Muller, R.S. (1988) "Fracture Strain of LPCVD Polysilicon," in *1988 Sensor and Actuator Workshop*, pp. 88–91, 6–9 June, Hilton Head Island, SC.

Tai, Y.C., and Muller, R.S. (1990) "Measurement of Young's Modulus on Microfabricated Structures Using a Surface Profiler," in *IEEE Micro Electro Mechanical Systems*, pp. 147–52, 11–14 February, Napa Valley, CA.

Taylor, J.A. (1991). "The Mechanical Properties and Microstructure of Plasma Enhanced Chemical Vapor Deposited Silicon Nitride Thin Films," *J. Vac. Sci. Technol. A* **9**, pp. 2464–68.

Teh, K.S., Lin, L., and Chiao, M. (1999) "The Creep Behavior of Polysilicon Microstructures," in *Transducers '99*, pp. 508–11, 7–10 June, Sendai, Japan.

Tong, L., and Mehregany, M. (1992) "Mechanical Properties of 3C Silicon Carbide," *Appl. Phys. Lett.* **60**, pp. 2992–94.

Tsuchiya, T., Tabata, O., Sakata, J., and Taga, Y. (1998a) "Specimen Size Effect on Tensile Strength of Surface Micromachined Polycrystalline Silicon Thin Films," *J. Microelectromech. Syst.* pp. 106–13.

Tsuchiya, T., Sakata, J., and Taga, Y. (1998b) "Tensile Strength and Fracture Toughness of Surface Micromachined Polycrystalline Silicon Thin Films Prepared Under Various Conditions," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 505*, pp. 285–90, 1–5 December, Boston.

Tsuchiya, T., Inoue, A., and Sakata, J. (1999) "Tensile Testing of Insulating Thin Films: Humidity Effect on Tensile Strength of SiO_2 Films," in *Proc. 10th Int. Conf. on Solid-State Sensors and Actuators — Transducers '99*, pp. 488–91, 7–10 June, Sendai, Japan.

Van Arsdell, W.W., and Brown, S.B. (1999) "Subcritical Crack Growth in Silicon MEMS," *J. Microelectromech. Syst.* **8**, pp. 319–27.

Vlassak, J.J., and Nix, W.D. (1992). "A New Bulge Test Technique for the Determination of Young's Modulus and Poisson's Ratio of Thin Films," *J. Mater. Res.* **7**, pp. 3242–49.

Walker, J.A., Gabriel, K.J., and Mehregany, M. (1990) "Mechanical Integrity of Polysilicon Films Exposed to Hydrofluoric Acid Solutions," *J. Electron. Mater.* **20**, pp. 665–70.

Weihs, T.P., Hong, S., Bravman, J.C., and Nix, W.D. (1988) "Mechanical Deflection of Cantilever Microbeams: A New Technique for Testing the Mechanical Properties of Thin Films," *J. Mater. Res.* **3**, pp. 931–42.

Weihs, T.P., Hong, S., Bravman, J.C., and Nix, W.D. (1989) "Measuring the Strength and Stiffness of Thin Film Materials by Mechanically Deflecting Cantilever Microbeams," in *Thin Films: Stresses and Mechanical Properties Symposium, Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 402*, pp. 87–92, 28–30 November, Boston.

Yang, E.H., and Fujita, H. (1997) "Fabrication and Characterization of U-Shaped Beams for the Determination of Young's Modulus Modification Due to Joule Heating of Polysilicon Microstructures," in *Proc. Int. Solid-State Sensors and Actuators Conf. — Transducers '97*, pp. 603–6, 16–19 June, Chicago.

Yi, T., and Kim, C.J. (1999a) "Measurement of Mechanical Properties for MEMS Materials," *Meas. Sci. Technol.* **10**, pp. 706–16.

Yi, T., and Kim, C.J. (1999b) "Microscale Material Testing: Etchant Effect on the Tensile Strength," in *Transducers '99*, pp. 518–21, 7–10 June, Sendai, Japan.

Yi, T., and Kim, C.J. (1999c) "Tension Test with Single Crystalline Silicon Microspecimen," *MEMS Microelectromech. Syst.* **1**, pp. 81–86.

Yoshioka, T., Yamasaki, M., Shikida, M., and Sato, K. (1996) "Tensile Testing of Thin-Film Materials on a Silicon Chip," in *MHS '96 Proc. Seventh Int. Symp. on Micro Machine and Human Science*, pp. 111–17, 2–4 October, Nagoya, Japan.

Yu, M.F., Lourie, O., Dyer, M.J., Moloni, K., Kelly, T.F., and Ruoff, R.S. (2000) "Strength and Breaking Mechanism of Multiwalled Carbon Nanotubes Under Tensile Load," *Science* **287**, pp. 637–40.

Zhang, L.M., Uttamchandani, D., and Culshaw, B. (1991) "Measurement of the Mechanical Properties of Silicon Microresonators," *Sensor. Actuator. A (Phys.)* **29**, pp. 79–84.

Zhang, X., Zhang, T.Y., Wong, M., and Zohar, Y. (1997) "Effects of High-Temperature Rapid Thermal Annealing on the Residual Stress of LPCVD-Polysilicon Thin Films," in *Proc. IEEE Tenth Annual Int. Workshop on Micro Electro Mechanical Systems*, pp. 535–40, 26–30 January, Nagoya, Japan.

Zhang, T.Y., Su, Y.J., Qian, C.F., Zhao, M.H., and Chen, L.Q. (2000) "Microbridge Testing of Silicon Nitride Thin Films Deposited on Silicon Wafers," *Acta Mater.* **48**, pp. 2843–57.

Ziebart, V., Paul, O., Munch, U., and Baltes, H. (1997) "A Novel Method to Measure Poisson's Ratio of Thin Films," in *Microelectromechanical Structures for Materials Research, Materials Research Society Symposium 505*, pp. 103–8, 1–2 December, Boston.

Ziebart, V., Paul, O., Munch, U., and Baltes, H. (1999) "Strongly Buckled Square Micromachined Membranes," *J. Microelectromech. Syst.* **8**, pp. 423–32.

Zou, Q., Li, Z., et al. (1995). "New Methods for Measuring Mechanical Properties of Thin Films in Micromachining: Beam Pull-In Voltage (VPI) Method and Long Beam Deflection (LBD) Method," *Sensor. Actuator. A (Phys.)* **48**, pp. 137–43.

4

Flow Physics

Mohamed Gad-el-Hak
Virginia Commonwealth University

> *One of the first men who speculated on the remarkable possibilities which magnification or diminution of physical dimensions provides was Jonathan Swift, who, in Gulliver's Travels, drew some conclusions as to what dwarfs and giants would really look like, and what sociological consequences size would have. Some time ago Florence Moog (Scientific American, November 1948) showed that Swift was a "bad biologist," or Gulliver a "poor liar." She showed that a linear reduction in size would carry with it a reduction in the number of brain cells, and hence a reduction in intellectual capacity in Lilliputians, whereas the enormous Brobdingnagians were physically impossible; they could have had physical reality only if their necks and legs had been short and thick. These 90-ton monsters could never have walked on dry land, nor could their tremendous weight have been carried on proportionately-sized feet.*
>
> *Even though Swift, in his phantasy, committed a number of physical errors, because he was not sufficiently aware of the fact that some physical properties of a body are proportional to the linear dimensions (height), whereas others vary with the third power of linear size (such as weight and cell number), yet he surpassed his medieval predecessors in many respects and drew a number of excellent conclusions, bringing both giants and dwarfs close to physical reality.*
>
> **(F. W. Went, "The Size of Man")**

4.1 Introduction

This chapter reviews the status of our understanding of fluid flow physics particular to microdevices. It is an update of the earlier publication by the same author [Gad-el-Hak, 1999]. The coverage here is broad leaving the details to other chapters in the handbook that treat specialized problems in microscale fluid mechanics. Not all MEMS devices involve fluid flows of course, but the present chapter will focus on those that do. Microducts, micronozzles, micropumps, microturbines, and microvalves are examples of small devices involving the flow of liquids and gases. MEMS can also be related to fluid flows indirectly. The availability of inexpensive, batch-processing-produced microsensors and microactuators provides opportunities for targeting small-scale coherent structures in macroscopic turbulent shear flows. Flow control using MEMS promises a quantum leap in control system performance [Gad-el-Hak, 2000]. Additionally, the extremely small sensors made possible by microfabrication technology allow measurements with spatial and temporal resolutions not achievable before. For example, high-Reynolds-number turbulent flow diagnoses are now feasible down to the Kolmogorov scales [Löfdahl and Gad-el-Hak, 1999]. Those indirect topics are also left to other chapters in the book.

4.2 Flow Physics

The rapid progress in fabricating and utilizing microelectromechanical systems during the last decade has not been matched by corresponding advances in our understanding of the unconventional physics involved in the manufacture and operation of small devices [Kovacs, 1998; Knight, 1999; Gad-el-Hak, 1999; Karniadakis and Beskok, 2002; Nguyen and Wereley, 2002; Madou, 2002; Stone et al., 2004; Squires and Quake, 2005]. Providing such understanding is crucial to designing, optimizing, fabricating, and utilizing improved MEMS devices. The present chapter focuses on the physics of fluid flows in microdevices.

Fluid flows in small devices differ from those in macroscopic machines. The operation of MEMS-based ducts, nozzles, valves, bearings, turbomachines, etc., cannot always be predicted from conventional flow models such as the Navier–Stokes equations with no-slip boundary condition at a fluid–solid interface as routinely and successfully applied for larger flow devices. Many questions have been raised when the results of experiments with microdevices could not be explained via traditional flow modeling. The pressure gradient in a long microduct was observed to be nonconstant, and the measured flow rate was higher than that predicted from the conventional continuum flow model. Load capacities of microbearings were diminished and electric currents needed to move micromotors were extraordinarily high. The dynamic response of micromachined accelerometers operating at atmospheric conditions was observed to be overdamped.

In the early stages of development of this exciting new field, the objective was to build MEMS devices as productively as possible. Microsensors were reading something, but not many researchers seemed to know exactly what. Microactuators were moving, but conventional modeling could not precisely predict their motion. After a decade of unprecedented progress in MEMS technology, perhaps the time is now ripe to slow down a bit, take stock, and answer the many questions that arose. The ultimate aim of this long-term exercise is to achieve rational-design capability for useful microdevices and to be able to characterize definitively and with as little empiricism as possible the operations of microsensors and microactuators.

Dealing with fluid flow through microdevices presents the questions of which model to use, which boundary condition to apply, and how to proceed to obtain solutions to the problem at hand. Obviously surface effects dominate in small devices. The surface-to-volume ratio for a machine with a characteristic length of $1\,\text{m}$ is $1\,\text{m}^{-1}$, while that for a MEMS device having a size of $1\,\mu\text{m}$ is $10^6\,\text{m}^{-1}$. The millionfold increase in surface area relative to the mass of the minute device substantially affects the transport of mass, momentum, and energy through the surface. The small length scale of microdevices may invalidate the continuum approximation altogether. Slip flow, thermal creep, rarefaction, viscous dissipation, compressibility, intermolecular forces, and other unconventional effects may have to be taken into account, preferably using only first principles, such as conservation of mass, Newton's second law, and conservation of energy.

This chapter discusses continuum as well as molecular-based flow models and the choices to be made. Computing typical Reynolds, Mach, and Knudsen numbers for the flow through a particular device is a good start to characterize the flow. For gases, microfluid mechanics has been studied by incorporating slip boundary conditions, thermal creep, and viscous dissipation as well as compressibility effects into the continuum equations of motion. Molecular-based models have also been attempted for certain ranges of the operating parameters. Use is made of the well-developed kinetic theory of gases embodied in the Boltzmann equation and of direct simulation methods such as Monte Carlo. Microfluid mechanics of liquids is more

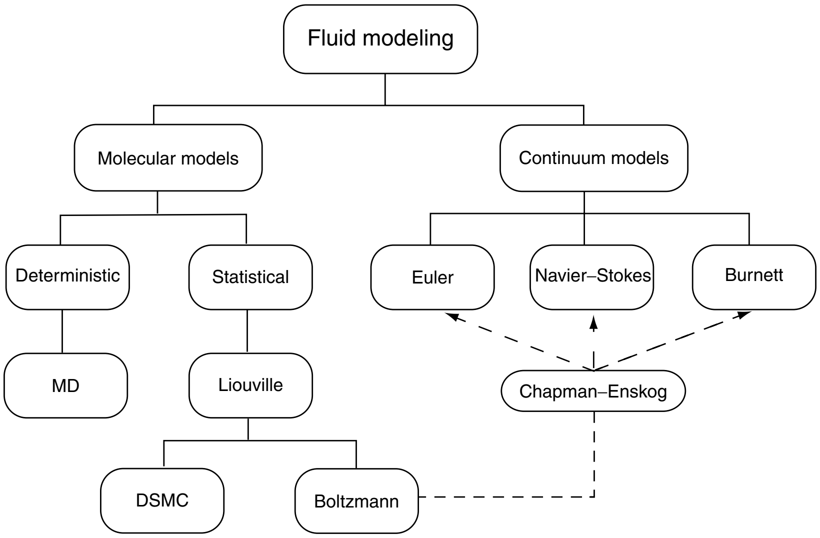


FIGURE 4.1 Molecular and continuum flow models.

complicated. The molecules are much more closely packed at normal pressures and temperatures, and the attractive or cohesive potential between the liquid molecules as well as between the liquid and solid molecules plays a dominant role if the characteristic length of the flow is sufficiently small. In cases when the traditional continuum model fails to provide accurate predictions or postdictions, expensive molecular dynamics simulations seem to be the only first-principle approach available to rationally characterize liquid flows in microdevices. Such simulations are not yet feasible for realistic flow extent or number of molecules. As a consequence, the microfluid mechanics of liquids is much less developed than that for gases.

4.3 Fluid Modeling

There are basically two ways of modeling a flow field, either as the fluid really is — a collection of molecules — or as a continuum where the matter is assumed continuous and indefinitely divisible. The first method is subdivided into deterministic methods and probabilistic ones, while in the second method the velocity, density, pressure, etc., are defined at every point in space and time, and conservation of mass, energy, and momentum leads to a set of nonlinear partial differential equations (Euler, Navier–Stokes, Burnett, etc.). Fluid modeling classification is depicted schematically in Figure 4.1.

The continuum model, embodied in the Navier–Stokes equations, applies to numerous flow situations. It ignores the molecular nature of gases and liquids and regards the fluid as a continuous medium describable in terms of the spatial and temporal variations of density, velocity, pressure, temperature, and other macroscopic flow quantities. For dilute gas flows near equilibrium, the Navier–Stokes equations are derivable from the molecularly based Boltzmann equation but can also be derived independently of that for both liquids and gases. In the case of direct derivation, some empiricism is necessary to close the resulting indeterminate set of equations. The continuum model is easier to handle mathematically (and is also more familiar to most fluid dynamicists) than the alternative molecular models. Continuum models should therefore be used as long as they are applicable. Thus, careful considerations of the validity of the Navier–Stokes equations and the like are in order.

Basically, the continuum model leads to fairly accurate predictions as long as local properties, such as density and velocity, can be defined as averages over elements that are large compared with the microscopic structure of the fluid but small enough in comparison with the scale of the macroscopic phenomena to permit using differential calculus to describe them. Additionally, the flow must not be too far from thermodynamic equilibrium. The former condition is almost always satisfied, but it is the latter that usually restricts the validity of the continuum equations. As the following section shows, the continuum flow equations do not form a determinate set. The shear stress and heat flux must be expressed in terms of lower-order macroscopic quantities such as velocity and temperature, and the simplest (i.e., linear) relations are valid only when the flow is near thermodynamic equilibrium. Worse yet, the traditional no-slip boundary condition at a solid–fluid interface breaks down even before the linear stress–strain relation becomes invalid.

To be more specific, we temporarily restrict the discussion to gases where the concept of mean free path is well defined. Liquids are more problematic and we defer their discussion to a later section. For gases, the mean free path $\mathcal{L}$ is the average distance traveled by molecules between collisions. For an ideal gas modeled as rigid spheres, the mean free path is related to temperature T and pressure p as follows

$$\mathcal{L} = \frac{1}{\sqrt{2}\,\pi n \sigma^2} = \frac{kT}{\sqrt{2}\,\pi p \sigma^2} \tag{4.1}$$

where n is the number density (number of molecules per unit volume), σ is the molecular diameter, and k is the Boltzmann constant (1.38×10^{-23} J/K · molecule).

The continuum Navier–Stokes model is valid when $\mathcal{L}$ is much smaller than a characteristic flow dimension L. As this condition is violated, the flow is no longer near equilibrium, and the linear relation between stress and rate of strain and the no-slip velocity condition are no longer valid. Similarly, the linear relation between heat flux and temperature gradient and the no-jump temperature condition at a solid–fluid interface are no longer accurate when $\mathcal{L}$ is not much smaller than L.

The length-scale L can be some overall dimension of the flow, but a more precise choice is the scale of the gradient of a macroscopic quantity, as for example the density ρ,

$$L = \frac{\rho}{\left|\frac{\partial \rho}{\partial y}\right|} \tag{4.2}$$

The ratio between the mean free path and the characteristic length is known as the Knudsen number

$$Kn = \frac{\mathcal{L}}{L} \tag{4.3}$$

and generally the traditional continuum approach is valid, albeit with modified boundary conditions, as long as $Kn < 0.1$.

The Knudsen number can be expressed in terms of other important dimensionless parameters in fluid mechanics. The Reynolds number is the ratio of inertial forces to viscous forces

$$Re = \frac{v_o L}{\nu} \tag{4.4}$$

where v_o is a characteristic velocity and ν is the kinematic viscosity of the fluid. The Mach number is the ratio of flow velocity to the speed of sound

$$Ma = \frac{v_o}{a_o} \tag{4.5}$$

The Mach number is a dynamic measure of fluid compressibility and may be considered as the ratio of inertial forces to elastic forces. From the kinetic theory of gases, the mean free path is related to the viscosity as follows:

$$\nu = \frac{\mu}{\rho} = \frac{1}{2}\,\mathcal{L}\,\bar{v}_m \tag{4.6}$$

where μ is the dynamic viscosity and $\bar{v}_m$ is the mean molecular speed, which is somewhat higher than the sound speed a_o,

$$\bar{v}_m = \sqrt{\frac{8}{\pi\gamma}}\,a_o \tag{4.7}$$

where γ is the specific heat ratio (i.e., the isentropic exponent). Combining Equations (4.3)–(4.7), we reach the required relation

$$Kn = \sqrt{\frac{\pi\gamma}{2}}\,\frac{Ma}{Re} \tag{4.8}$$

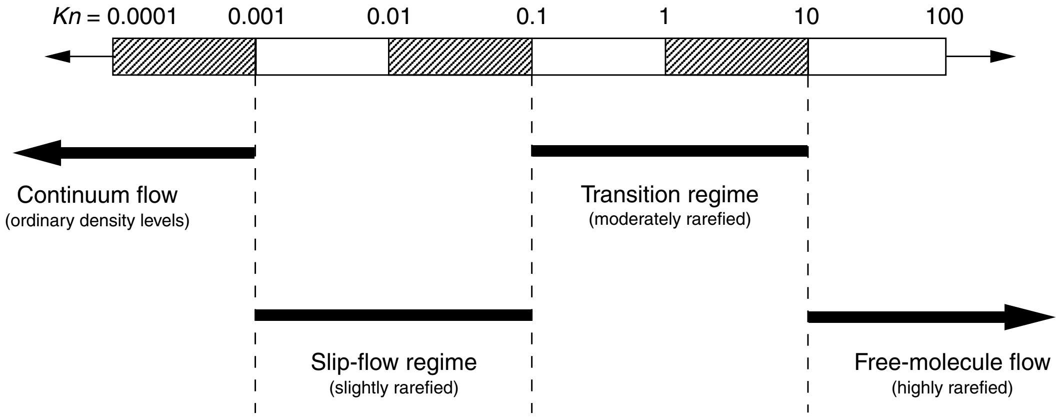


FIGURE 4.2 Knudsen number regimes.

In boundary layers, the relevant length scale is the shear-layer thickness δ, and for laminar flows

$$\frac{\delta}{L} \sim \frac{1}{\sqrt{Re}} \tag{4.9}$$

$$Kn \sim \frac{Ma}{Re_\delta} \sim \frac{Ma}{\sqrt{Re}} \tag{4.10}$$

where Re_δ is the Reynolds number based on the freestream velocity v_o, and the boundary layer thickness δ, and Re is based on v_o and the streamwise length scale L.

Rarefied gas flows are in general encountered in flows in small geometries, such as MEMS devices, and in low-pressure applications, such as high-altitude flying and high-vacuum gadgets. The local value of Knudsen number in a particular flow determines the degree of rarefaction and the degree of validity of the Navier–Stokes model. The different Knudsen number regimes are determined empirically and are therefore only approximate for a particular flow geometry. The pioneering experiments in rarefied gas dynamics were conducted by Knudsen in 1909. In the limit of zero Knudsen number, the transport terms in the continuum momentum and energy equations are negligible, and the Navier–Stokes equations then reduce to the inviscid Euler equations. Both heat conduction and viscous diffusion and dissipation are negligible, and the flow is then approximately isentropic (i.e., adiabatic and reversible) from the continuum viewpoint, while the equivalent molecular viewpoint is that the velocity distribution function is everywhere of the local equilibrium or Maxwellian form. As Kn increases, rarefaction effects become more important, and eventually the continuum approach breaks down altogether. The different Knudsen number regimes are depicted in Figure 4.2, and can be summarized as follows:

Euler equations (neglect molecular diffusion):	$Kn \to 0$ ($Re \to \infty$)
Navier–Stokes equations with no-slip boundary conditions:	$Kn < 10^{-3}$
Navier–Stokes equations with slip boundary conditions:	$10^{-3} \leqslant Kn < 10^{-1}$
Transition regime:	$10^{-1} \leqslant Kn < 10$
Free-molecule flow:	$Kn \geqslant 10$

As an example, consider air at standard temperature (T = 288 K) and pressure (p = 1.01 ×10^5 N/m^2). A cube one micron on a side contains 2.54 × 10^7 molecules separated by an average distance of 0.0034 microns. The gas is considered dilute if the ratio of this distance to the molecular diameter exceeds 7; in the present example this ratio is 9, barely satisfying the dilute gas assumption. The mean free path computed from Equation (4.1) is $\mathcal{L}$ = 0.065 μm. A microdevice with characteristic length of 1 μm would have Kn = 0.065, which is in the slip-flow regime. At lower pressures, the Knudsen number increases. For example, if the pressure is 0.1 atm and the temperature remains the same, Kn = 0.65 for the same 1 μm device, and the flow is then in the transition regime. There would still be more than 2 million molecules in the same 1 μm cube, and the average distance between them would be 0.0074 μm. The same device at 100 km altitude would have $Kn = 3 \times 10^4$, well into the free-molecule flow regime. Knudsen number for the flow of a light gas like helium is about three times larger than that for air flow at otherwise the same conditions.

Consider a long microchannel where the entrance pressure is atmospheric and the exit conditions are near vacuum. As air goes down the duct, the pressure and density decrease while the velocity, Mach number, and Knudsen number increase. The pressure drops to overcome viscous forces in the channel. If isothermal conditions prevail,[1] density also drops and conservation of mass requires the flow to accelerate down the constant-area tube. The fluid acceleration in turn affects the pressure gradient resulting in a nonlinear pressure drop along the channel. The Mach number increases down the tube, limited only by choked-flow condition $Ma = 1$. Additionally, the normal component of velocity is no longer zero. With lower density, the mean free path increases, and Kn correspondingly increases. All flow regimes depicted in Figure 4.2 may occur in the same tube: continuum with no-slip boundary conditions, slip-flow regime, transition regime, and free-molecule flow. The air flow may also change from incompressible to compressible as it moves down the microduct. A similar scenario may take place if the entrance pressure is, say, 5 atm, while the exit is atmospheric. This deceivingly simple duct flow may in fact manifest every single complexity discussed in this section. The following six sections discuss in turn the Navier–Stokes equations, compressibility effects, boundary conditions, molecular-based models, liquid flows, and surface phenomena.

4.4 Navier–Stokes Equations

This section recalls the traditional conservation relations in fluid mechanics. A concise derivation of these equations can be found in Gad-el-Hak (2000). Here, we reemphasize the precise assumptions needed to obtain a particular form of the equations. A continuum fluid implies that the derivatives of all the dependent variables exist in some reasonable sense. In other words, local properties, such as density and velocity, are defined as averages over elements that are large compared with the microscopic structure of the fluid but small enough in comparison with the scale of the macroscopic phenomena to permit the use of differential calculus to describe them. As mentioned earlier, such conditions are almost always met. For such fluids, and assuming the laws of nonrelativistic mechanics hold, the conservation of mass, momentum, and energy can be expressed at every point in space and time as a set of partial differential equations as follows:

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x_k}(\rho u_k) = 0 \tag{4.11}$$

$$\rho\left(\frac{\partial u_i}{\partial t} + u_k \frac{\partial u_i}{\partial x_k}\right) = \frac{\partial \Sigma_{ki}}{\partial x_k} + \rho g_i \tag{4.12}$$

$$\rho\left(\frac{\partial e}{\partial t} + u_k \frac{\partial e}{\partial x_k}\right) = -\frac{\partial q_k}{\partial x_k} + \Sigma_{ki}\frac{\partial u_i}{\partial x_k} \tag{4.13}$$

where ρ is the fluid density, u_k is an instantaneous velocity component (u, v, w), Σ_{ki} is the second-order stress tensor (surface force per unit area), g_i is the body force per unit mass, e is the internal energy, and q_k is the sum of heat flux vectors due to conduction and radiation. The independent variables are time t and the three spatial coordinates x_1, x_2, and x_3 or (x, y, z).

[1] More likely the flow will be somewhere between isothermal and adiabatic, Fanno flow. In that case both density and temperature decrease downstream, the former not as fast as in the isothermal case. None of that changes the qualitative arguments made in the example.

Equations (4.11), (4.12), and (4.13) constitute five differential equations for the 17 unknowns ρ, u_i, Σ_{ki}, e, and q_k. Absent any body couples, the stress tensor is symmetric having only six independent components, which reduces the number of unknowns to 14. Obviously, the continuum flow equations do not form a determinate set. To close the conservation equations, the relation between the stress tensor and deformation rate, the relation between the heat flux vector and the temperature field, and appropriate equations of state relating the different thermodynamic properties are needed. The stress–rate-of-strain relation and the heat-flux–temperature-gradient relation are approximately linear if the flow is not too far from thermodynamic equilibrium. This is a phenomenological result but can be rigorously derived from the Boltzmann equation for a dilute gas assuming the flow is near equilibrium. For a Newtonian, isotropic, Fourier, ideal gas, for example, those relations read

$$\Sigma_{ki} = -p\,\delta_{ki} + \mu\left(\frac{\partial u_i}{\partial x_k} + \frac{\partial u_k}{\partial x_i}\right) + \lambda\left(\frac{\partial u_j}{\partial x_j}\right)\delta_{ki} \tag{4.14}$$

$$q_i = -\kappa\frac{\partial T}{\partial x_i} + \text{Heat flux due to radiation} \tag{4.15}$$

$$de = c_v dT \quad \text{and} \quad p = \rho\mathscr{R}T \tag{4.16}$$

where p is the thermodynamic pressure, μ and λ are the first and second coefficients of viscosity, respectively, δ_{ki} is the unit second-order tensor (Kronecker delta), κ is the thermal conductivity, T is the temperature field, c_v is the specific heat at constant volume, and $\mathscr{R}$ is the gas constant which is given by the Boltzmann constant divided by the mass of an individual molecule $k = m\mathscr{R}$. Stokes' hypothesis relates the first and second coefficients of viscosity thus, $\lambda + \frac{2}{3}\mu = 0$, although the validity of this assumption for other than dilute, monatomic gases has occasionally been questioned [Gad-el-Hak, 1995]. With the above constitutive relations and neglecting radiative heat transfer, Equations (4.11), (4.12), and (4.13) respectively read

$$\frac{\partial\rho}{\partial t} + \frac{\partial}{\partial x_k}(\rho u_k) = 0 \tag{4.17}$$

$$\rho\left(\frac{\partial u_i}{\partial t} + u_k\frac{\partial u_i}{\partial x_k}\right) = -\frac{\partial\rho}{\partial x_i} + \rho g_i + \frac{\partial}{\partial x_k}\left[\mu\left(\frac{\partial u_i}{\partial x_k} + \frac{\partial u_k}{\partial x_i}\right) + \delta k_i\lambda\frac{\partial u_j}{\partial x_j}\right] \tag{4.18}$$

$$\rho\left(\frac{\partial e}{\partial t} + u_k\frac{\partial e}{\partial x_k}\right) = \frac{\partial}{\partial x_k}\left(\kappa\frac{\partial T}{\partial x_k}\right) - p\frac{\partial u_k}{\partial x_k} + \phi \tag{4.19}$$

The three components of the vector Equation (4.18) are the Navier–Stokes equations expressing the conservation of momentum for a Newtonian fluid. In the thermal energy Equation (4.19), ϕ is the always positive dissipation function expressing the irreversible conversion of mechanical energy to internal energy as a result of the deformation of a fluid element. The second term on the right-hand side of (4.19) is the reversible work done (per unit time) by the pressure as the volume of a fluid material element changes. For a Newtonian, isotropic fluid, the viscous dissipation rate is given by

$$\phi = \frac{1}{2}\mu\left(\frac{\partial u_i}{\partial x_k} + \frac{\partial u_k}{\partial x_i}\right)^2 + \lambda\left(\frac{\partial u_j}{\partial x_j}\right)^2 \tag{4.20}$$

There are now six unknowns, ρ, u_i, p, and T, and the five coupled Equations (4.17), (4.18), and (4.19) plus the equation of state relating pressure, density, and temperature. These six equations together with sufficient number of initial and boundary conditions constitute a well-posed, albeit formidable, problem. The system of Equations (4.17)–(4.19) is an excellent model for the laminar or turbulent flow of most fluids, such as air and water, under many circumstances including high-speed gas flows for which the shock waves are thick relative to the mean free path of the molecules.

Considerable simplification is achieved if the flow is assumed incompressible, usually a reasonable assumption provided that the characteristic flow speed is less than 0.3 of the speed of sound. The incompressibility assumption is readily satisfied for almost all liquid flows and many gas flows. In such cases, the density is assumed either a constant or a given function of temperature (or species concentration). The governing equations for such flow are

$$\frac{\partial u_k}{\partial x_k} = 0 \tag{4.21}$$

$$\rho\left(\frac{\partial u_i}{\partial t} + u_k\frac{\partial u_i}{\partial x_k}\right) = -\frac{\partial p}{\partial x_i} + \frac{\partial}{\partial x_k}\left[\mu\left(\frac{\partial u_i}{\partial x_k} + \frac{\partial u_k}{\partial x_i}\right)\right] + \rho gi \tag{4.22}$$

$$\rho c_p\left(\frac{\partial T}{\partial t} + u_k\frac{\partial T}{\partial x_k}\right) = \frac{\partial}{\partial x_k}\left(\kappa\frac{\partial T}{\partial x_k}\right) + \phi_{\text{incomp}} \tag{4.23}$$

where ϕ_{incomp} is the incompressible limit of Equation (4.20). These are now five equations for the five dependent variables u_i, p, and T. Note that the left-hand side of Equation (4.23) has the specific heat at constant pressure c_p and not c_v. It is the convection of enthalpy — and not internal energy — that is balanced by heat conduction and viscous dissipation. This is the correct incompressible-flow limit — of a compressible fluid — as discussed in detail in Section 10.9 of Panton (1996); a subtle point, perhaps, but one that is frequently missed in textbooks.

For both the compressible and the incompressible equations of motion, the transport terms are neglected away from solid walls in the limit of infinite Reynolds number ($Kn \to 0$). The fluid is then approximated as inviscid and nonconducting, and the corresponding equations read (for the compressible case)

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x_k}(\rho u_k) = 0 \tag{4.24}$$

$$\rho\left(\frac{\partial u_i}{\partial t} + u_k\frac{\partial u_i}{\partial x_k}\right) = -\frac{\partial p}{\partial x_i} + \rho g_i \tag{4.25}$$

$$\rho c_v\left(\frac{\partial T}{\partial t} + u_k\frac{\partial T}{\partial x_k}\right) = -p\frac{\partial u_k}{\partial x_k} \tag{4.26}$$

The Euler Equation (4.25) can be integrated along a streamline, and the resulting Bernoulli's equation provides a direct relation between the velocity and pressure.

4.5 Compressibility

The issue of whether to consider the continuum flow compressible or incompressible seems straightforward but is in fact full of potential pitfalls. If the local Mach number is less than 0.3, then the flow of a compressible fluid like air can — according to the conventional wisdom — be treated as incompressible. But the well-known $Ma < 0.3$ criterion is only a necessary criterion, not a sufficient one, to allow a treatment of the flow as approximately incompressible. In other words, in some situations the Mach number can be exceedingly small while the flow is compressible. As is well documented in heat transfer textbooks, strong wall heating or cooling may cause the density to change sufficiently and the incompressible approximation to break down, even at low speeds. Less known is the situation encountered in some microdevices where the pressure may strongly change due to viscous effects even though the speeds may not be high enough for the Mach number to go above the traditional threshold of 0.3. Corresponding to the pressure changes would be strong density changes that must be taken into account when writing the continuum equations of motion. In this section, we systematically explain all situations where compressibility effects must be considered. Let us rewrite the full continuity Equation (4.11) as follows

$$\frac{D\rho}{Dt} + \rho\frac{\partial u_k}{\partial x_k} = 0 \tag{4.27}$$

where

$$\frac{D}{Dt}$$

is the substantial derivative

$$\left(\frac{\partial}{\partial t} + u_k \frac{\partial}{\partial x_k}\right)$$

expressing changes following a fluid element. The proper criterion for the incompressible approximation to hold is that

$$\left(\frac{1}{\rho}\frac{D\rho}{Dt}\right)$$

is vanishingly small. In other words, if density changes following a fluid particle are small, the flow is approximately incompressible. Density may change arbitrarily from one particle to another without violating the incompressible flow assumption. This is the case, for example, in the stratified atmosphere and ocean, where the variable-density/temperature/salinity flow is often treated as incompressible.

From the state principle of thermodynamics, we can express the density changes of a simple system in terms of changes in pressure and temperature,

$$\rho = \rho(p, T) \tag{4.28}$$

Using the chain rule of calculus,

$$\frac{1}{\rho}\frac{D\rho}{Dt} = \alpha\frac{Dp}{Dt} - \beta\frac{DT}{Dt} \tag{4.29}$$

where α and β are respectively the isothermal compressibility coefficient and the bulk expansion coefficient — two thermodynamic variables that characterize the fluid susceptibility to change of volume — which are defined by the following relations

$$\alpha(p, T) \equiv \frac{1}{\rho}\frac{\partial \rho}{\partial p}\bigg|_T \tag{4.30}$$

$$\beta(p, T) \equiv -\frac{1}{\rho}\frac{\partial \rho}{\partial T}\bigg|_p \tag{4.31}$$

For ideal gases, $\alpha = 1/p$ and $\beta = 1/T$. Note, however, that in the following arguments invoking the ideal gas assumption will not be necessary. The flow must be treated as compressible if pressure- and/or temperature-changes — following a fluid element — are sufficiently strong. Equation (4.29) must, of course, be properly nondimensionalized before deciding whether a term is large or small. Here, we follow closely the procedure detailed in Panton (1996).

Consider first the case of adiabatic walls. Density is normalized with a reference value ρ_o, velocities with a reference speed v_o, spatial coordinates and time with respectively L and L/v_o, the isothermal compressibility coefficient and bulk expansion coefficient with reference values α_o and β_o. The pressure is nondimensionalized with the inertial pressure-scale $\rho_o v_o^2$. This scale is twice the dynamic pressure; that is, the pressure change as an inviscid fluid moving at the reference speed is brought to rest.

Temperature changes for adiabatic walls can only result from the irreversible conversion of mechanical energy into internal energy via viscous dissipation. Temperature is therefore nondimensionalized as follows

$$T^* = \frac{T - T_o}{\left(\dfrac{\mu_o v_o^2}{c_{\kappa_o}}\right)} = \frac{T - T_o}{Pr\left(\dfrac{v_o^2}{c_{p_o}}\right)} \tag{4.32}$$

where T_o is a reference temperature, μ_o, κ_o, and c_{p_o} are respectively reference viscosity, thermal, conductivity, and specific heat at constant pressure, and Pr is the reference Prandtl number, $(\mu_o c_{p_o})/\kappa_o$.

In the present formulation, the scaling used for pressure is based on the Bernoulli's equation and therefore neglects viscous effects. This particular scaling guarantees that the pressure term in the momentum equation will be of the same order as the inertia term. The temperature scaling assumes that the conduction, convection, and dissipation terms in the energy equation have the same order of magnitude. The resulting dimensionless form of Equation (4.29) reads

$$\frac{1}{\rho^*}\frac{D\rho^*}{Dt^*} = \gamma_o Ma^2 \left\{ \alpha^* \frac{Dp^*}{Dt^*} - \frac{PrB\beta^*}{A}\frac{DT^*}{Dt^*} \right\} \tag{4.33}$$

where the superscript * indicates a nondimensional quantity, Ma is the reference Mach number (v_o/a_o, where a_o is the reference speed of sound), and A and B are dimensionless constants defined by $A \equiv \alpha_o \rho_o c_{p_o} T_o$ and $B \equiv \beta_o T_o$. If the scaling is properly chosen, the terms having the * superscript in the right-hand side should be of order one, and the relative importance of such terms in the equations of motion is determined by the magnitude of the dimensionless parameters appearing to their left (e.g. Ma, Pr, etc.). Therefore, as $Ma^2 \rightarrow 0$, temperature changes due to viscous dissipation are neglected (unless Pr is very large as, for example, in the case of highly viscous polymers and oils). Within the same order of approximation, all thermodynamic properties of the fluid are assumed constant.

Pressure changes are also neglected in the limit of zero Mach number. Hence, for $Ma < 0.3$ (i.e., $Ma^2 < 0.09$), density changes following a fluid particle can be neglected and the flow can then be approximated as incompressible.[2] However, there is a caveat to this argument. Pressure changes due to inertia can indeed be neglected at small Mach numbers, and this is consistent with the way we nondimensionalized the pressure term above. If, on the other hand, pressure changes are mostly due to viscous effects, as is the case, for example, in a long microduct or a micro-gas-bearing, pressure changes may be significant even at low speeds (low Ma). In that case the term

$$\frac{Dp^*}{Dt}$$

in Equation (4.33) is no longer of order one and may be large regardless of the value of Ma. Density then may change significantly, and the flow must be treated as compressible. Had pressure been nondimensionalized using the viscous scale

$$\left(\frac{\mu_o v_o}{L}\right)$$

instead of the inertial one

$$(\rho_o v_o^2)$$

the revised Equation (4.33) would have Re^{-1} appearing explicitly in the first term in the right-hand side, accentuating this term's importance when viscous forces dominate.

A similar result can be gleaned when the Mach number is interpreted as follows

$$Ma^2 = \frac{\mathrm{v}_o^2}{a_o^2} = v_o^2 \left.\frac{\partial\rho}{\partial p}\right|_s = \frac{\rho_o v_o^2}{\rho_o} \left.\frac{\partial\rho}{\partial p}\right|_s \sim \frac{\Delta p}{\rho_o}\frac{\Delta\rho}{\Delta p} = \frac{\Delta\rho}{\rho_o} \tag{4.34}$$

where s is the entropy. Again, the above equation assumes that pressure changes are inviscid, and therefore small Mach number means negligible pressure and density changes. In a flow dominated by viscous effects — such as that inside a microduct — density changes may be significant even in the limit of zero Mach number.

Identical arguments can be made in the case of isothermal walls. Here strong temperature changes may be the result of wall heating or cooling even if viscous dissipation is negligible. The proper

[2] With an error of about 10% at $Ma = 0.3$, 4% at $Ma = 0.2$, 1% at $Ma = 0.1$, and so on.

temperature scale in this case is given in terms of the wall temperature T_w and the reference temperature T_o as follows

$$\hat{T} = \frac{T - T_o}{T_w - T_o} \tag{4.35}$$

where $\hat{T}$ is the new dimensionless temperature. The nondimensional form of Equation (4.29) now reads

$$\frac{1}{\rho^*}\frac{D\rho^*}{Dt^*} = \gamma_o Ma^2 \alpha^* \frac{Dp^*}{Dt^*} - \beta^* B \left(\frac{T_w - T_o}{T_o}\right)\frac{D\hat{T}}{Dt^*} \tag{4.36}$$

Here we notice that the temperature term is different from that in Equation (4.33). *Ma* no longer appears in this term, and strong temperature changes, that is, large $(T_w - T_o)/T_o$, may cause strong density changes regardless of the value of the Mach number. Additionally, the thermodynamic properties of the fluid are not constant but depend on temperature; as a result the continuity, momentum, and energy equations all couple. The pressure term in Equation (4.36), on the other hand, is exactly as it was in the adiabatic case, and the arguments made before apply: the flow should be considered compressible if $Ma > 0.3$ or if pressure changes due to viscous forces are sufficiently large.

Experiments in gaseous microducts confirm the above arguments. For both low- and high-Mach-number flows, pressure gradients in long microchannels are nonconstant, consistent with the compressible flow equations. Such experiments were conducted by, among others, Prud'homme et al. (1986), Pfahler et al. (1991), van den Berg et al. (1993), Liu et al. (1993, 1995), Pong et al. (1994), Harley et al. (1995), Piekos and Breuer (1996), Arkilic (1997), and Arkilic et al. (1995, 1997a, 1997b). Sample results will be presented in the following section.

In three additional scenarios significant pressure and density changes may take place without inertial, viscous, or thermal effects. First is the case of quasi-static compression/expansion of a gas in, for example, a piston-cylinder arrangement. The resulting compressibility effects are, however, compressibility of the fluid and not of the flow. Two other situations where compressibility effects must also be considered are problems with length-scales comparable to the scale height of the atmosphere and rapidly varying flows as in sound propagation [Lighthill, 1963].

4.6 Boundary Conditions

The continuum equations of motion described earlier require a certain number of initial and boundary conditions for proper mathematical formulation of flow problems. In this section, we describe the boundary conditions at a fluid–solid interface. Boundary conditions in the inviscid flow theory pertain only to the velocity component normal to a solid surface. The highest spatial derivative of velocity in the inviscid equations of motion is first order, and only one velocity boundary condition at the surface is admissible. The normal velocity component at a fluid–solid interface is specified, and no statement can be made regarding the tangential velocity component. The normal-velocity condition simply states that a fluid-particle path cannot go through an impermeable wall. Real fluids are viscous, of course, and the corresponding momentum equation has second-order derivatives of velocity, thus requiring an additional boundary condition on the velocity component tangential to a solid surface.

Traditionally, the no-slip condition at a fluid–solid interface is enforced in the momentum equation, and an analogous no-temperature-jump condition is applied in the energy equation. The notion underlying the no-slip/no-jump condition is that within the fluid there cannot be any finite discontinuities of velocity/temperature. Those would involve infinite velocity/temperature gradients and so produce infinite viscous stress/heat flux that would destroy the discontinuity in infinitesimal time. The interaction between a fluid particle and a wall is similar to that between neighboring fluid particles, and therefore no discontinuities are allowed at the fluid–solid interface either. In other words, the fluid velocity must be zero relative to the surface, and the fluid temperature must be equal to that of the surface. But strictly speaking those two boundary conditions are valid only if the fluid flow adjacent to the surface is in thermodynamic equilibrium. This requires an infinitely high frequency of collisions between the fluid and the solid surface. In practice, the no-slip/no-jump condition leads to fairly accurate predictions as long as

$Kn < 0.001$ (for gases). Beyond that, the collision frequency is simply not high enough to ensure equilibrium, and a certain degree of tangential-velocity slip and temperature jump must be allowed. This is a case frequently encountered in MEMS flows, and we develop the appropriate relations in this section.

For both liquids and gases, the linear Navier boundary condition empirically relates the tangential velocity slip at the wall $\Delta u|_w$ to the local shear

$$\Delta u|_w = u_{\text{fluid}} - u_{\text{wall}} = L_s \frac{\partial u}{\partial y}\bigg|_w \tag{4.37}$$

where L_s is the constant slip length, and

$$\frac{\partial u}{\partial y}\bigg|_w$$

is the strain rate computed at the wall. In most practical situations, the slip length is so small that the no-slip condition holds. In MEMS applications, however, that may not be the case. Once again we defer the discussion of liquids to a later section and focus for now on gases.

Assuming isothermal conditions prevail, the above slip relation has been rigorously derived by Maxwell (1879) from considerations of the kinetic theory of dilute, monatomic gases. Gas molecules, modeled as rigid spheres, continuously strike and reflect from a solid surface, just as they continuously collide with each other. For an idealized perfectly smooth wall (at the molecular scale), the incident angle exactly equals the reflected angle, and the molecules conserve their tangential momentum and thus exert no shear on the wall. This is termed specular reflection and results in perfect slip at the wall. For an extremely rough wall, on the other hand, the molecules reflect at some random angle uncorrelated with their entry angle. This perfectly diffuse reflection results in zero tangential-momentum for the reflected fluid molecules to be balanced by a finite slip velocity in order to account for the shear stress transmitted to the wall. A force balance near the wall leads to the following expression for the slip velocity

$$u_{\text{gas}} - u_{\text{wall}} = \mathcal{L} \frac{\partial u}{\partial y}\bigg|_w \tag{4.38}$$

where $\mathcal{L}$ is the mean free path. The right-hand side can be considered as the first term in an infinite Taylor series, sufficient if the mean free path is relatively small enough. Equation (4.38) states that significant slip occurs only if the mean velocity of the molecules varies appreciably over a distance of one mean free path. This is the case, for example, in vacuum applications and/or flow in microdevices. The number of collisions between the fluid molecules and the solid in those cases is not large enough for even an approximate flow equilibrium to be established. Furthermore, additional (nonlinear) terms in the Taylor series would be needed as L increases and the flow is further removed from the equilibrium state.

For real walls some molecules reflect diffusively and some reflect specularly. In other words, a portion of the momentum of the incident molecules is lost to the wall, and a (typically smaller) portion is retained by the reflected molecules. The tangential-momentum-accommodation coefficient σ_v is defined as the fraction of molecules reflected diffusively. This coefficient depends on the fluid, the solid, and the surface finish and has been determined experimentally to be between 0.2–0.8 [Thomas and Lord, 1974; Seidl and Steiheil, 1974; Porodnov et al., 1974; Arkilic et al., 1997b; Arkilic, 1997], the lower limit being for exceptionally smooth surfaces while the upper limit is typical of most practical surfaces. The final expression derived by Maxwell for an isothermal wall reads

$$u_{\text{gas}} - u_{\text{wall}} = \frac{2-\sigma_v}{\sigma_v}\,\mathcal{L} \frac{\partial u}{\partial y}\bigg|_w \tag{4.39}$$

For $\sigma_v = 0$ the slip velocity is unbounded, while for $\sigma_v = 1$, Equation (4.39) reverts to (4.38).

Similar arguments were made for the temperature-jump boundary condition by von Smoluchowski (1898). For an ideal gas flow in the presence of wall-normal and tangential temperature gradients, the complete (first-order) slip-flow and temperature-jump boundary conditions read

$$u_{\text{gas}} - u_{\text{wall}} = \frac{2-\sigma_v}{\sigma_v} \frac{1}{\rho\sqrt{\dfrac{2\Re T_{\text{gas}}}{\pi}}} \tau_w + \frac{3}{4} \frac{Pr\,(\gamma-1)}{\gamma\rho\Re T_{\text{gas}}}(-q_x)_w$$

$$= \frac{2-\sigma_v}{\sigma_v} \mathcal{L}\left(\frac{\partial u}{\partial y}\right)_w + \frac{3}{4}\frac{\mu}{\rho T_{\text{gas}}}\left(\frac{\partial T}{\partial x}\right)_w \tag{4.40}$$

$$T_{\text{gas}} - T_{\text{wall}} = \frac{2-\sigma_T}{\sigma_T}\left[\frac{2(\gamma-1)}{(\gamma+1)}\right] \frac{1}{\rho\Re\sqrt{\dfrac{2\Re T_{\text{gas}}}{\pi}}}(-q_y)_w$$

$$= \frac{2-\sigma_T}{\sigma_T}\left[\frac{2\gamma}{(\gamma+1)}\right]\frac{\mathcal{L}}{Pr}\left(\frac{\partial T}{\partial y}\right)_w \tag{4.41}$$

where x and y are the streamwise and normal coordinates, ρ and μ are respectively the fluid density and viscosity, $\Re$ is the gas constant, T_{gas} is the temperature of the gas adjacent to the wall, T_{wall} is the wall temperature, τ_w is the shear stress at the wall, Pr is the Prandtl number, γ is the specific heat ratio, and q_x and q_y are respectively the tangential and normal heat flux at the wall.

The tangential-momentum-accommodation coefficient σ_v and the thermal-accommodation coefficient σ_T are given by respectively

$$\sigma_v = \frac{\tau_i - \tau_r}{\tau_i - \tau_w} \tag{4.42}$$

$$\sigma_T = \frac{dE_i - dE_r}{dE_i - dE_w} \tag{4.43}$$

where the subscripts i, r, and w stand for respectively incident, reflected, and solid wall conditions, τ is a tangential momentum flux, and dE is an energy flux.

The second term in the right-hand side of Equation (4.40) is the *thermal creep*, which generates slip velocity in the fluid opposite to the direction of the tangential heat flux (i.e., flow in the direction of increasing temperature). At sufficiently high Knudsen numbers, a streamwise temperature gradient in a conduit leads to a measurable pressure gradient along the tube. This may be the case in vacuum applications and MEMS devices. Thermal creep is the basis for the so-called Knudsen pump — a device with no moving parts — in which rarefied gas is hauled from a cold chamber to a hot one.[3] Clearly, such a pump performs best at high Knudsen numbers and is typically designed to operate in the free-molecule flow regime.

In dimensionless form, Equations (4.40) and (4.41), respectively, read

$$u^*_{\text{gas}} - u^*_{\text{wall}} = \frac{2-\sigma_v}{\sigma_v} Kn\left(\frac{\partial u^*}{\partial y^*}\right)_w + \frac{3}{2\pi}\frac{(\gamma-1)}{\gamma}\frac{Kn^2 Re}{Ec}\left(\frac{\partial T^*}{\partial x^*}\right)_w \tag{4.44}$$

$$T^*_{\text{gas}} - T^*_{\text{wall}} = \frac{2-\sigma_T}{\sigma_T}\left[\frac{2\gamma}{(\gamma+1)}\right]\frac{Kn}{Pr}\left(\frac{\partial T^*}{\partial y^*}\right)_w \tag{4.45}$$

[3] The term *Knudsen pump* has been used by, for example, Vargo and Muntz (1996), but according to Loeb (1961) the original experiments demonstrating such a pump were carried out by Osborne Reynolds.

where the superscript * indicates dimensionless quantity, *Kn* is the Knudsen number, *Re* is the Reynolds number, and *Ec* is the Eckert number defined by

$$Ec = \frac{v_o^2}{c_p \Delta T} = (\gamma - 1)\,\frac{T_o}{\Delta T}\,Ma^2 \tag{4.46}$$

where v_o is a reference velocity, $\Delta T = (T_{\text{gas}} - T_o)$, and T_o is a reference temperature. Note that very low values of σ_v and σ_T lead to substantial velocity slip and temperature jump even for flows with small a Knudsen number.

The first term in the right-hand side of Equation (4.44) is first order in Knudsen number, while the thermal creep term is second order, meaning that the creep phenomenon is potentially significant at large values of the Knudsen number. Equation (4.45) is first order in *Kn*. Using Equations (4.8) and (4.46), the thermal creep term in Equation (4.44) can be rewritten in terms of ΔT and Reynolds number. Thus,

$$u^*_{\text{gas}} - u^*_{\text{wall}} = \frac{2-\sigma_v}{\sigma_v}\,Kn\left(\frac{\partial u^*}{\partial y^*}\right)_w + \frac{3}{4}\,\frac{\Delta T}{T_o}\,\frac{1}{Re}\left(\frac{\partial T^*}{\partial x^*}\right)_w \tag{4.47}$$

Large temperature changes along the surface or low Reynolds numbers clearly lead to significant thermal creep.

The continuum Navier–Stokes equations with no-slip/no-temperature jump boundary conditions are valid as long as the Knudsen number does not exceed 0.001. First-order slip/temperature-jump boundary conditions should be applied to the Navier–Stokes equations in the range of $0.001 < Kn < 0.1$. The transition regime spans the range of $0.1 < Kn < 10$, in which second-order or higher slip/temperature-jump boundary conditions are applicable. Note, however, that the Navier–Stokes equations are first-order accurate in *Kn* as will be shown later, and are themselves not valid in the transition regime. Either higher-order continuum equations (e.g., Burnett equations), should be used there, or molecular modeling should be invoked abandoning the continuum approach altogether.

For isothermal walls, Beskok (1994) derived a higher-order slip-velocity condition as follows

$$u_{\text{gas}} - u_{\text{wall}} = \frac{2-\sigma_v}{\sigma_v}\left[\mathcal{L}\left(\frac{\partial u}{\partial y}\right)_w + \frac{\mathcal{L}^2}{2!}\left(\frac{\partial^2 u}{\partial y^2}\right)_w + \frac{\mathcal{L}^3}{3!}\left(\frac{\partial^3 u}{\partial y^3}\right)_w + \cdots\right] \tag{4.48}$$

Attempts to implement the above slip condition in numerical simulations are rather difficult. Second-order and higher derivatives of velocity cannot be computed accurately near the wall. Based on asymptotic analysis, Beskok (1996) and Beskok and Karniadakis (1994, 1999) proposed the following alternative higher-order boundary condition for the tangential velocity, including the thermal creep term,

$$u^*_{\text{gas}} - u^*_{\text{wall}} = \frac{2-\sigma_v}{\sigma_v}\,\frac{Kn}{1-b\,Kn}\left(\frac{\partial u^*}{\partial y^*}\right)_w + \frac{3}{2\pi}\,\frac{(\gamma-1)}{\gamma}\,\frac{Kn^2 Re}{Ec}\left(\frac{\partial T^*}{\partial x^*}\right)_w \tag{4.49}$$

where b is a high-order slip coefficient determined from the presumably known no-slip solution, thus avoiding the computational difficulties mentioned above. If this high-order slip coefficient is chosen as $b = u''_w/u'_w$, where the prime denotes derivative with respect to y and the velocity is computed from the no-slip Navier–Stokes equations, Equation (4.49) becomes second-order accurate in Knudsen number. Beskok's procedure can be extended to third- and higher-orders for both the slip-velocity and thermal creep terms.

Similar arguments can be applied to the temperature-jump boundary condition, and the resulting Taylor series reads in dimensionless form (Beskok, 1996),

$$T^*_{\text{gas}} - T^*_{\text{wall}} = \frac{2-\sigma_T}{\sigma_T}\left[\frac{2\gamma}{(\gamma+1)}\right]\frac{1}{Pr}\left[Kn\left(\frac{\partial T^*}{\partial y^*}\right)_w + \frac{Kn^2}{2!}\left(\frac{\partial^2 T^*}{\partial y^{*2}}\right)_w + \cdots\right] \tag{4.50}$$

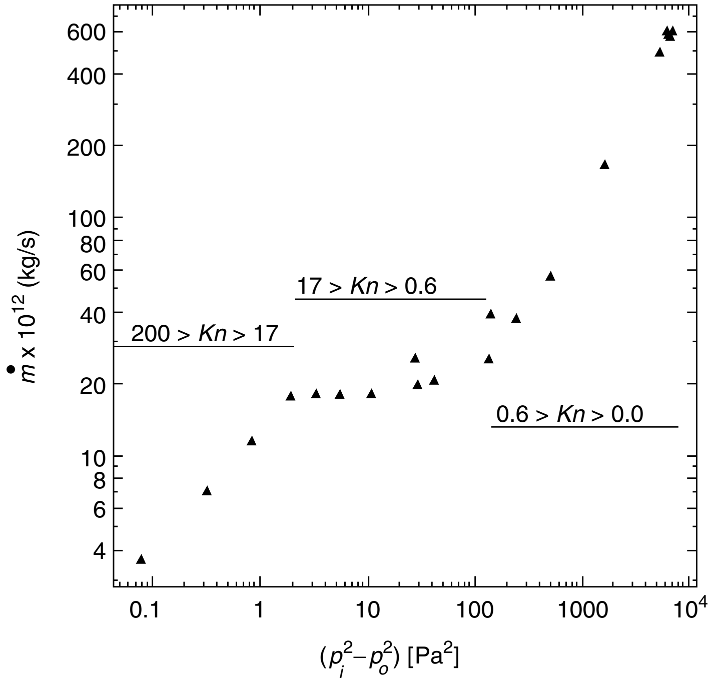


FIGURE 4.3 Variation of mass flowrate as a function of $(p_i^2 - p_o^2)$. Original data acquired by S.A. Tison and plotted by Beskok et al. (1996). (Reprinted with permission from Beskok et al. [1996] "Simulation of Heat and Momentum Transfer in Complex Micro-Geometries," *J. Thermophys. & Heat Transfer* **8**, pp. 355–70.)

Again, the difficulties associated with computing second- and higher-order derivatives of temperature are alleviated using an identical procedure to that utilized for the tangential velocity boundary condition.

Several experiments in low-pressure macroducts or in microducts confirm the necessity of applying slip boundary condition at sufficiently large Knudsen numbers. Among them are those conducted by Knudsen (1909), Pfahler, et al. (1991), Tison (1993), Liu et al. (1993, 1995), Pong et al. (1994), Arkilic et al. (1995), Harley et al. (1995), and Shi et al. (1995, 1996). The experiments are complemented by the numerical simulations carried out by Beskok (1994, 1996), Beskok and Karniadakis (1994, 1999), Beskok et al. (1996), and Karniadakis and Beskok (2002). Here we present selected examples of the experimental and numerical results.

Tison (1993) conducted pipe flow experiments at very low pressures. His pipe had a diameter of 2 mm and a length-to-diameter ratio of 200. Both inlet and outlet pressures were varied to yield Knudsen number in the range of $Kn = 0$–200. Figure 4.3 shows the variation of mass flow rate as a function of $(p_i^2 - p_o^2)$, where p_i is the inlet pressure and p_o is the outlet pressure.[4] The pressure drop in this rarefied pipe flow is nonlinear, characteristic of low-Reynolds-number compressible flows. Three distinct flow regimes are identified: (1) slip flow regime, $0 < Kn < 0.6$; (2) transition regime, $0.6 < Kn < 17$, where the mass flowrate is almost constant as the pressure changes; and (3) free-molecule flow, $Kn > 17$. Note that the demarcation between these three regimes is slightly different from that mentioned earlier. As stated, the different Knudsen number regimes are determined empirically and are therefore only approximate for a particular flow geometry.

Shih et al. (1995) conducted their experiments in a microchannel using helium as a fluid. The inlet pressure varied, but the duct exit was atmospheric. Microsensors were fabricated in situ along their MEMS channel to measure the pressure. Figure 4.4 shows their measured mass flow rate versus the inlet

[4] The original data in this figure were acquired by S.A. Tison and plotted by Beskok et al. (1996).

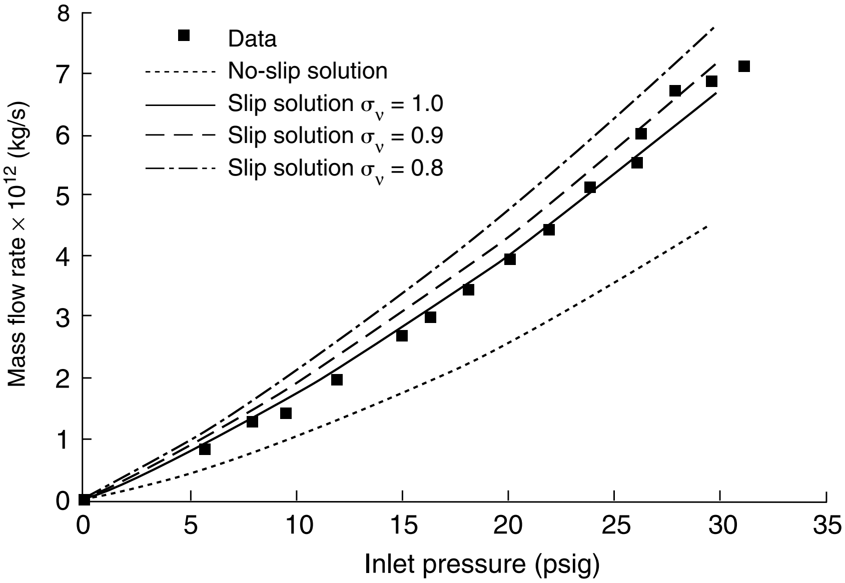


FIGURE 4.4 Mass flowrate versus inlet pressure in a microchannel. (Reprinted with permission from Shih et al. [1995] "Non-Linear Pressure Distribution in Uniform Microchannels," ASME AMD-MD-Vol. 238, New York.)

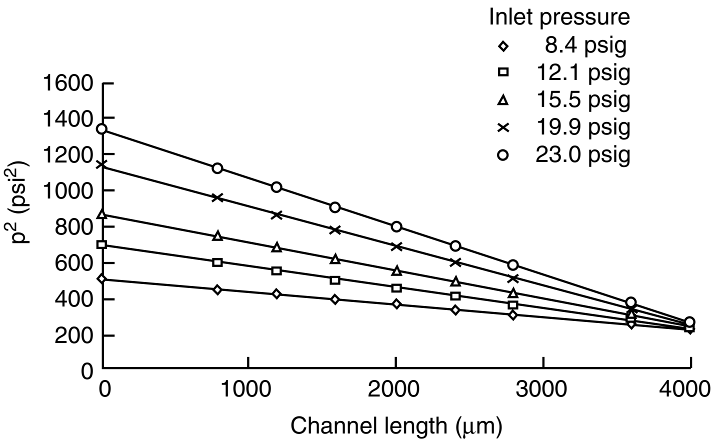


FIGURE 4.5 Pressure distribution of nitrous oxide in a microduct. Solid lines are theoretical predictions. (Reprinted with permission from Shih et al. [1996] "Monatomic and Polyatomic Gas Flow through Uniform Microchannels," in *Applications of Microfabrication to Fluid Mechanics*, K. Breuer, P. Bandyopadhyay, and M. Gad-el-Hak, eds., ASME DSC-Vol. 59, pp. 197–203, New York.)

pressure. The data are compared to the no-slip solution and the slip solution using three different values of the tangential-momentum-accommodation coefficient, 0.8, 0.9, and 1.0. The agreement is reasonable with the case $\sigma_v = 1$, indicating perhaps that the channel used by Shih et al., was quite rough on the molecular scale. In a second experiment [Shih et al., 1996], nitrous oxide was used as the fluid. The square of the pressure distribution along the channel is plotted for five different inlet pressures in Figure 4.5. The experimental data (symbols) compare well with the theoretical predictions (solid lines). Again, the nonlinear pressure drop shown indicates that the gas flow is compressible.

Arkilic (1997) provided an elegant analysis of the compressible, rarefied flow in a microchannel. The results of his theory are compared to the experiments of Pong et al., (1994) in Figure 4.6. The dotted line is the incompressible flow solution, where the pressure is predicted to drop linearly with streamwise distance.

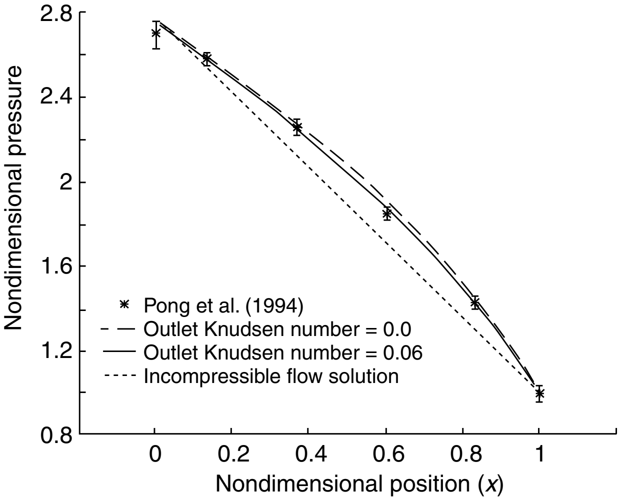


FIGURE 4.6 Pressure distribution in a long microchannel. The symbols are experimental data while the lines are different theoretical predictions. (Reprinted with permission from Arkilic [1997] Measurement of the Mass Flow and Tangential Momentum Accommodation Coefficient in Silicon Micromachined Channels, Ph.D. thesis, Massachusetts Institute of Technology.)

The dashed line is the compressible flow solution that neglects rarefaction effects (assumes $Kn = 0$). Finally, the solid line is the theoretical result that takes into account both compressibility and rarefaction via slip-flow boundary condition computed at the exit Knudsen number of $Kn = 0.06$. That theory compares most favorably with the experimental data. In the compressible flow through the constant-area duct, density decreases and thus velocity increases in the streamwise direction. As a result, the pressure distribution is nonlinear with negative curvature. A moderate Knudsen number (i.e., moderate slip) actually diminishes, albeit rather weakly, this curvature. Thus, compressibility and rarefaction effects lead to opposing trends, as pointed out by Beskok et al. (1996).

4.7 Molecular-Based Models

In the continuum models discussed thus far, the macroscopic fluid properties are the dependent variables while the independent variables are the three spatial coordinates and time. The molecular models recognize the fluid as a myriad of discrete particles: molecules, atoms, ions, and electrons. The goal here is to determine the position, velocity, and state of all particles at all times. The molecular approach is either deterministic or probabilistic (refer to Figure 4.1). Provided that there is a sufficient number of microscopic particles within the smallest significant volume of a flow, the macroscopic properties at any location in the flow can then be computed from the discrete-particle information by a suitable averaging or weighted averaging process. The present section discusses molecular-based models and their relation to the continuum models previously considered.

The most fundamental of the molecular models is deterministic. The motion of the molecules is governed by the laws of classical mechanics, although at the expense of greatly complicating the problem, the laws of quantum mechanics can also be considered in special circumstances. The modern molecular dynamics computer simulations (MD) have been pioneered by Alder and Wainwright (1957, 1958, 1970) and reviewed by Ciccotti and Hoover (1986), Allen and Tildesley (1987), Haile (1993), and Koplik and Banavar (1995). The simulation begins with a set of N molecules in a region of space, each assigned a random velocity corresponding to a Boltzmann distribution at the temperature of interest. The interaction between the particles is prescribed typically in the form of a two-body potential energy and the time evolution of the molecular positions is determined by integrating Newton's equations of motion. Because MD is based on the most basic set of equations, it is valid in principle for any flow extent and any range of parameters. The method is straightforward in principle but there are two hurdles: (1) choosing a proper and convenient potential for particular fluid and solid combinations, and (2) the colossal computer resources required to simulate a reasonable flowfield extent.

For purists, the former difficulty is a sticky one. There is no totally rational methodology by which a convenient potential can be chosen. Part of the art of MD is to pick an appropriate potential and validate the simulation results with experiments or other analytical/computational results. A commonly used potential between two molecules is the generalized Lennart-Jones 6–12 potential, to be used in the following section and further discussed in the section following that.

The second difficulty, and by far the most serious limitation of molecular dynamics simulations, is the number of molecules N that can realistically be modeled on a digital computer. Since the computation of an element of trajectory for any particular molecule requires consideration of *all* other molecules as potential collision partners, the amount of computation required by the MD method is proportional to N^2. Some savings in computer time can be achieved by cutting off the weak tail of the potential (see Figure 4.11) at, say, $r_c = 2.5\,\sigma$, and shifting the potential by a linear term in r so that the force goes smoothly to zero at the cutoff. As a result, only nearby molecules are treated as potential collision partners, and the computation time for N molecules no longer scales with N^2.

The state of the art of molecular dynamics simulations in the early 2000s is such that with a few hours of CPU time general purpose supercomputers can handle around 100,000 molecules. At enormous expense, the fastest parallel machine available can simulate around 10 million particles. Because of the extreme diminution of molecular scales, the above translates into regions of liquid flow of about 0.06 μm (600 angstroms) in linear size, over time intervals of around 0.001 μsec, enough for continuum behavior to set in for simple molecules. To simulate 1 sec of real time for complex molecular interactions (e.g., vibration modes, reorientation of polymer molecules, collision of colloidal particles, etc.) requires unrealistic CPU time measured in hundreds of years.

MD simulations are highly inefficient for dilute gases where the molecular interactions are infrequent. The simulations are more suited for dense gases and liquids. Clearly, molecular dynamics simulations are reserved for situations where the continuum approach or the statistical methods are inadequate to compute from first principles important flow quantities. Slip boundary conditions for liquid flows in extremely small devices are such a case, as will be discussed in the following section.

An alternative to the deterministic molecular dynamics is the statistical approach where the goal is to compute the probability of finding a molecule at a particular position and state. If the appropriate conservation equation can be solved for the probability distribution, important statistical properties, such as the mean number, momentum, or energy of the molecules within an element of volume, can be computed from a simple weighted averaging. In a practical problem, it is such average quantities that concern us rather than the detail for every single molecule. Clearly, however, the accuracy of computing average quantities via the statistical approach improves as the number of molecules in the sampled volume increases. The kinetic theory of dilute gases is well advanced, but that of dense gases and liquids is much less so due to the extreme complexity of having to include multiple collisions and intermolecular forces in the theoretical formulation. The statistical approach is well covered in books such as those by Kennard (1938), Hirschfelder et al. (1954), Schaaf and Chambré (1961), Vincenti and Kruger (1965), Kogan (1969), Chapman and Cowling (1970), Cercignani (1988, 2000) and Bird (1994), and review articles such as those by Kogan (1973), Muntz (1989), and Oran et al. (1998).

In the statistical approach, the fraction of molecules in a given location and state is the sole dependent variable. The independent variables for monatomic molecules are time, the three spatial coordinates, and the three components of molecular velocity. Those describe a six-dimensional phase space.[5] For diatomic or polyatomic molecules, the dimension of phase space is increased by the number of internal degrees of freedom. Orientation adds an extra dimension for molecules that are not spherically symmetric. Finally, for mixtures of gases, separate probability distribution functions are required for each species. Clearly, the

[5] The evolution equation of the probability distribution is considered, hence time is the seventh independent variable.

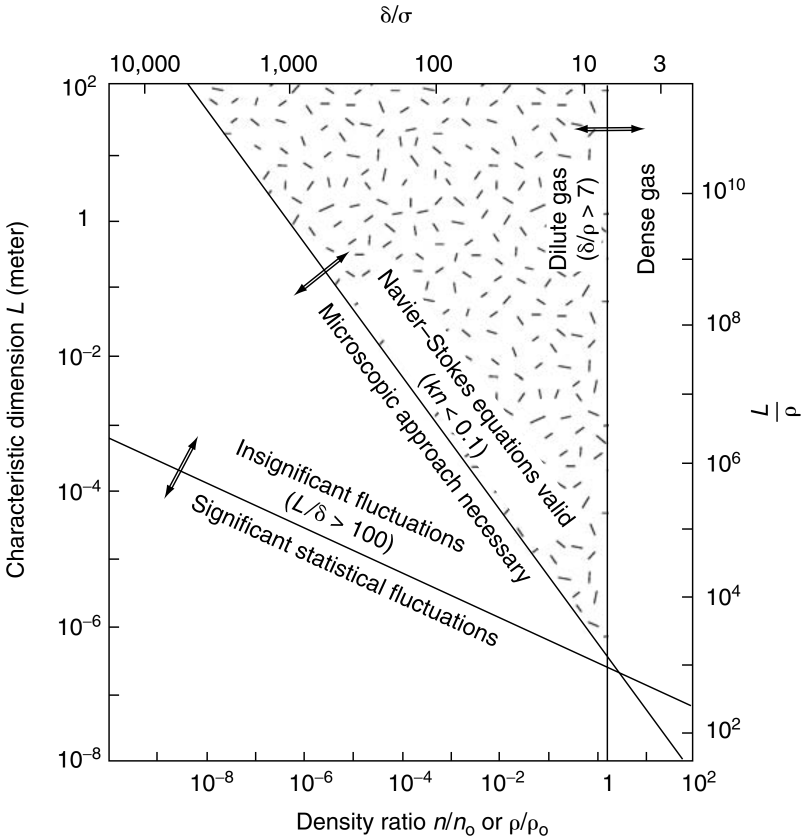


FIGURE 4.7 Effective limits of different flow models. (Reprinted with permission from Bird [1994] *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*, Clarendon Press, Oxford.)

complexity of the approach increases dramatically as the dimension of phase space increases. The simplest problems are, for example, those for steady, one-dimensional flow of a simple monatomic gas.

To simplify the problem we restrict the discussion here to monatomic gases having no internal degrees of freedom. Furthermore, the fluid is restricted to dilute gases and molecular chaos is assumed. The former restriction requires the average distance between molecules δ to be an order of magnitude larger than their diameter σ. That will almost guarantee that all collisions between molecules are binary collisions, avoiding the complexity of modeling multiple encounters.[6] The molecular chaos restriction improves the accuracy of computing the macroscopic quantities from the microscopic information. In essence, the volume over which averages are computed has to have enough molecules to reduce statistical errors. It can be shown that computing macroscopic flow properties by averaging over a number of molecules will result in statistical fluctuations with a standard deviation of approximately 0.1% if one million molecules are used and around 3% if one thousand molecules are used. The molecular chaos limit requires the length-scale L for the averaging process to be at least 100 times the average distance between molecules (i.e., typical averaging over at least one million molecules).

Figure 4.7, adapted from Bird (1994), shows the limits of validity of the dilute gas approximation ($\delta/\sigma > 7$), the continuum approach ($Kn < 0.1$, as discussed previously), and the neglect of statistical fluctuations ($L/\delta > 100$). Using a molecular diameter of $\sigma = 4 \times 10^{-10}$ m as an example, the three limits are conveniently expressed as functions of the normalized gas density ρ/ρ_o or number density n/n_o, where the reference densities ρ_o and n_o are computed at standard conditions. All three limits are straight lines in

[6] Dissociation and ionization phenomena involve triple collisions and therefore require separate treatment.

the log–log plot of L versus ρ/ρ_o, as depicted in Figure 4.7. Note the shaded triangular wedge inside which both the Boltzmann and Navier–Stokes equations are valid. Additionally, the lines describing the three limits very nearly intersect at a single point. As a consequence, the continuum breakdown limit always lies between the dilute gas limit and the limit for molecular chaos. As density or characteristic dimension is reduced in a dilute gas, the Navier–Stokes model breaks down before the level of statistical fluctuations becomes significant. In a dense gas, on the other hand, significant fluctuations may be present even when the Navier–Stokes model is still valid.

The starting point in statistical mechanics is the Liouville equation, which expresses the conservation of the N-particle distribution function in $6N$-dimensional phase space,[7] where N is the number of particles under consideration. Considering only external forces that do not depend on the velocity of the molecules,[8] the Liouville equation for a system of N mass points reads

$$\frac{\partial \mathcal{F}}{\partial t} + \sum_{k=1}^{N} \vec{\xi}k \cdot \frac{\partial \mathcal{F}}{\partial \vec{x}_k} + \sum_{k=1}^{N} \vec{F}_k \cdot \frac{\partial \mathcal{F}}{\partial \vec{\xi}_k} = 0 \tag{4.51}$$

where $\mathcal{F}$ is the probability of finding a molecule at a particular point in phase space, t is time, $\vec{\xi}_k$ is the three-dimensional velocity vector for the kth molecule, $\vec{x}_k$ is the three-dimensional position vector for the kth molecule, and $\vec{F}$ is the external force vector. Note that the dot product in Equation (4.51) is carried out over each of the three components of the vectors $\vec{\xi}$, $\vec{x}$ and $\vec{F}$ and that the summation is overall molecules. Obviously such an equation is not tractable for a realistic number of particles.

A hierarchy of reduced distribution functions may be obtained by repeated integration of the Liouville equation above. The final equation in the hierarchy is for the single particle distribution, which also involves the two-particle distribution function. Assuming molecular chaos, that final equation becomes a closed one (i.e., one equation in one unknown) and is known as the Boltzmann equation, the fundamental relation of the kinetic theory of gases. That final equation in the hierarchy is the only one that offers any hope of obtaining analytical solutions.

A simpler direct derivation of the Boltzmann equation is provided by Bird (1994). For monatomic gas molecules in binary collisions, the integro-differential Boltzmann equation reads

$$\frac{\partial(nf)}{\partial t} + \xi_j \frac{\partial(nf)}{\partial x_j} + F_j \frac{\partial(nf)}{\partial \xi_j} = J(f, f^*), \quad j = 1, 2, 3 \tag{4.52}$$

where nf is the product of the number density and the normalized velocity distribution function ($dn/n = fd\vec{\xi}$), x_j, and ξ_j are respectively the coordinates and speeds of a molecule,[9] F_j is a known external force, and $J(f, f^*)$is the nonlinear collision integral that describes the net effect of populating and depopulating collisions on the distribution function. The collision integral is the source of difficulty in obtaining analytical solutions to the Boltzmann equation and is given by

$$J(f,f^*) = \int_{-\infty}^{\infty} \int_{0}^{4\pi} n^2(f^* f_1^* - f f_1)\, \vec{\xi}_r\, \sigma d\Omega (d\vec{\xi})_1 \tag{4.53}$$

where the superscript * indicates postcollision values, f and f_1 represent two different molecules, $\vec{\xi}_r$ is the relative speed between two molecules, σ is the molecular cross-section, Ω is the solid angle, and $d\vec{\xi} = d\xi_1 d\xi_2 d\xi_3$.

Once a solution for f is obtained, macroscopic quantities, such as density, velocity, and temperature, can be computed from the appropriate weighted integral of the distribution function. For example,

$$\rho = mn = m\int (n\,f) d\vec{\xi} \tag{4.54}$$

$$u_i = \int \xi_i f d\vec{\xi} \tag{4.55}$$

[7] Three positions and three velocities for *each* molecule of a monatomic gas with no internal degrees of freedom.

[8] This excludes Lorentz forces, for example.

[9] Constituting together with time the seven independent variables of the single-dependent-variable equation.

$$\frac{3}{2}kT = \int \frac{1}{2} m\xi_i \xi_i f \mathrm{d}\vec{\xi} \qquad (4.56)$$

If the Boltzmann equation is nondimensionalized with a characteristic length L and characteristic speed $[2(k/m)T]^{1/2}$, where k is the Boltzmann constant, m is the molecular mass, and T is temperature, the inverse Knudsen number appears explicitly in the right-hand side of the equation as follows:

$$\frac{\partial \hat{f}}{\partial \hat{t}} + \hat{\xi}_j \frac{\partial \hat{f}}{\partial \hat{x}_j} + \hat{F}_j \frac{\partial \hat{f}}{\partial \hat{\xi}_j} = \frac{1}{Kn} \hat{J}(\hat{f}, \hat{f}^{\star}), \quad j = 1,2,3 \qquad (4.57)$$

where the topping symbol $\hat{}$ represents a dimensionless variable, and $\hat{f}$ is nondimensionalized using a reference number density n_o.

The five conservation equations for the transport of mass, momentum, and energy can be derived by multiplying the Boltzmann equation above by the molecular mass, momentum, and energy respectively, then integrating overall possible molecular velocities. Subject to the restrictions of dilute gas and molecular chaos stated earlier, the Boltzmann equation is valid for all ranges of Knudsen number from 0 to ∞. Analytical solutions to this equation for arbitrary geometries are difficult mostly because of the nonlinearity of the collision integral. Simple models of this integral have been proposed to facilitate analytical solutions [see, for example, Bhatnagar et al. (1954)].

There are two important asymptotes to Equation (4.57). First, as $Kn \rightarrow \infty$, molecular collisions become unimportant. This is the free-molecule flow regime depicted in Figure 4.2 for $Kn > 10$, where the only important collision is that between a gas molecule and the solid surface of an obstacle or a conduit. Analytical solutions are then possible for simple geometries, and numerical simulations for complicated geometries are straightforward once the surface-reflection characteristics are accurately modeled. Second, as $Kn \rightarrow 0$, collisions become important and the flow approaches the continuum regime of conventional fluid dynamics. The Second Law specifies a tendency for thermodynamic systems to revert to equilibrium state, smoothing any discontinuities in macroscopic flow quantities. The number of molecular collisions in the limit $Kn \rightarrow 0$ is so large that the flow approaches the equilibrium state in a time that is short compared to the macroscopic timescale. For example, for air at standard conditions ($T = 288$ K; $p = 1$ atm), each molecule experiences on average 10 collisions per nanosecond and travels 1 micron in the same time. Such a molecule has already *forgotten* its previous state after 1 nsec. In a particular flowfield, if the macroscopic quantities vary little over a distance of 1 μm or over a time interval of 1 nsec, the flow of STP air is near equilibrium.

At $Kn = 0$, the velocity distribution function is everywhere of the local equilibrium or Maxwellian form

$$\hat{f}^{(0)} = \frac{n}{n_o} \pi^{-3/2}\, exp[-(\hat{\xi} - \hat{u})^2] \qquad (4.58)$$

where $\hat{\xi}$ and $\hat{u}$ are the dimensionless speeds respectively of a molecule and of the flow. In this Knudsen number limit, the velocity distribution of each element of the fluid instantaneously adjusts to the equilibrium thermodynamic state appropriate to the local macroscopic properties as this molecule moves through the flow field. From the continuum viewpoint, the flow is isentropic, and heat conduction and viscous diffusion and dissipation vanish from the continuum conservation relations.

The Chapman–Enskog theory attempts to solve the Boltzmann equation by considering a small perturbation of $\hat{f}$ from the equilibrium Maxwellian form. For small Knudsen numbers, the distribution function can be expanded in terms of Kn in the form of a power series

$$\hat{f} = \hat{f}^{(0)} + Kn\hat{f}^{(1)} + Kn^2\hat{f}^{(2)} + \cdots \qquad (4.59)$$

By substituting the above series in the Boltzmann Equation (4.57) and equating terms of equal order, we arrive at the following recurrent set of integral equations:

$$\begin{aligned} &\hat{J}(\hat{f}^{(0)}, \hat{f}^{(0)}) = 0, \\ &\hat{J}(\hat{f}^{(0)}, \hat{f}^{(1)}) = \frac{\partial \hat{f}}{\partial \hat{t}} + \hat{\xi}_j \frac{\partial \hat{f}^{(0)}}{\partial \hat{x}_j} + \hat{F}_j \frac{\partial \hat{f}^{(0)}}{\partial \hat{\xi}_j}, \cdots \end{aligned} \qquad (4.60)$$

The first integral is nonlinear, and its solution is the local Maxwellian distribution, Equation (4.58). Each of the distribution functions $\hat{f}^{(1)}$, $\hat{f}^{(2)}$, etc., satisfies an inhomogeneous linear equation whose solution leads to the transport terms needed to close the continuum equations appropriate to the particular level of approximation. The continuum stress tensor and heat flux vector can be written in terms of the distribution function, which in turn can be specified in terms of the macroscopic velocity and temperature and their derivatives [Kogan, 1973]. The zeroth-order equation yields the Euler equations, the first-order equation results in the linear transport terms of the Navier–Stokes equations, the second-order equation gives the nonlinear transport terms of the Burnett equations, and so on. Keep in mind, however, that the Boltzmann equation as developed in this section is for a monatomic gas. This excludes the all-important air, which is composed largely of diatomic nitrogen and oxygen.

As discussed earlier, the Navier–Stokes equations can and should be used up to a Knudsen number of 0.1. Beyond that, the transition flow regime commences ($0.1 < Kn < 10$). In this flow regime, the molecular mean free path for a gas becomes significant relative to a characteristic distance for important flow-property changes to take place. The Burnett equations can be used to obtain analytical/numerical solutions for at least a portion of the transition regime for a monatomic gas, although their complexity has limited the results for realistic geometries (Agarwal et al., 1999, 2001; Lockerby and Reese, 2003). There is also a certain degree of uncertainty about the proper boundary conditions to use with the continuum Burnett equations, and experimental validation of the results has been very scarce. Additionally, as the gas flow departs farther from equilibrium, the bulk viscosity ($=\lambda + \frac{2}{3}\mu$, where λ is the second coefficient of viscosity) is no longer zero, and Stokes' hypothesis no longer holds (see Gad-el-Hak, 1995, for an interesting summary of the issue of bulk viscosity).

In the transition regime, the molecularly-based Boltzmann equation cannot easily be solved either, unless the nonlinear collision integral is simplified. So, clearly, the transition regime is in dire need of alternative solutions. MD simulations as mentioned earlier are not suited for dilute gases. The best approach for the transition regime right now is the direct simulation Monte Carlo (DSMC) method developed by Bird (1963, 1965, 1976, 1978, 1994) and briefly described below. Some recent reviews of DSMC include those by Muntz (1989), Cheng (1993), Cheng and Emmanuel (1995), and Oran et al. (1998). The mechanics as well as the history of the DSMC approach and its ancestors are well described in Bird (1994).

Unlike molecular dynamics simulations, DSMC is a statistical computational approach to solving rarefied gas problems. Both approaches treat a gas as discrete particles. Subject to the dilute gas and molecular chaos assumptions, the direct simulation Monte Carlo method is valid for all ranges of Knudsen number, although it becomes quite expensive for $Kn < 0.1$. Fortunately, this is the continuum regime where the Navier–Stokes equations can be used analytically or computationally. DSMC is therefore ideal for the transition regime ($0.1 < Kn < 10$), where the Boltzmann equation is difficult to solve. The Monte Carlo method is, like its namesake, a random-number strategy based directly on the physics of the individual molecular interactions. The idea is to track a large number of randomly selected, statistically representative particles, and to use their motions and interactions to modify their positions and states. The primary approximation of the direct simulation Monte Carlo method is to uncouple the molecular motions and the intermolecular collisions over small time intervals. A significant advantage of this approximation is that the amount of computation required is proportional to N, in contrast to N^2 for molecular dynamics simulations. In essence, particle motions are modeled deterministically, while collisions are treated probabilistically, each simulated molecule representing a large number of actual molecules. Typical computer runs of DSMC in the 1990s involved tens of millions of intermolecular collisions and fluid–solid interactions.

The DSMC computation is started from some initial condition and followed in small time steps that can be related to physical time. Colliding pairs of molecules in a small geometric cell in physical space are randomly selected after each computational time step. Complex physics such as radiation, chemical reactions, and species concentrations can be included in the simulations without the necessity of nonequilibrium thermodynamic assumptions that commonly afflict nonequilibrium continuum-flow calculations. DSMC is more computationally intensive than classical continuum simulations, and should therefore be used only when the continuum approach is not feasible.

The DSMC technique is explicit and time marching and therefore always produces unsteady flow simulations. For macroscopically steady flows, Monte Carlo simulation proceeds until a steady flow is established within a desired accuracy at sufficiently large time. The macroscopic flow quantities are then the time average of all values calculated after reaching the steady state. For macroscopically unsteady flows, ensemble averaging of many independent Monte Carlo simulations is carried out to obtain the final results within a prescribed statistical accuracy.

4.8 Liquid Flows

From the continuum point of view, liquids and gases are both fluids obeying the same equations of motion. For incompressible flows, for example, the Reynolds number is the primary dimensionless parameter that determines the nature of the flow field. True, water, for example, has density and viscosity that are respectively three orders and two orders of magnitude higher than those for air, but if the Reynolds number and geometry are matched, liquid and gas flows should be identical.[10] For MEMS applications, however, we anticipate the possibility of nonequilibrium flow conditions and the consequent invalidity of the Navier–Stokes equations and the no-slip boundary conditions. Such circumstances can best be researched using the molecular approach. We discussed this for gases earlier and will give the corresponding arguments for liquids in the present section. The literature on non-Newtonian fluids in general and polymers in particular is vast (for example, the bibliographic survey by Nadolink and Haigh, 1995, cites over 4,900 references on polymer drag reduction alone) and provides a rich source of information on the molecular approach for liquid flows.

Solids, liquids, and gases are distinguished merely by the degree of proximity and the intensity of motions of their constituent molecules. In solids, the molecules are packed closely and confined, each hemmed in by its neighbors [Chapman and Cowling, 1970]. Only rarely would one solid molecule slip from its neighbors to join a new set. As the solid is heated, molecular motion becomes more violent, and a slight thermal expansion takes place. At a certain temperature that depends on ambient pressure, sufficiently intense motion of the molecules enables them to pass freely from one set of neighbors to another. The molecules are no longer confined but are nevertheless still closely packed, and the substance is now considered a liquid. Further heating of the matter eventually releases the molecules altogether, allowing them to break the bonds of their mutual attractions. Unlike solids and liquids, the resulting gas expands to fill any available volume.

Unlike solids, neither liquids nor gases can resist finite shear force without continuous deformation; that is, the definition of a fluid medium. In contrast to the reversible, elastic, static deformation of a solid, the continuous deformation of a fluid resulting from the application of a shear stress results in an irreversible work that eventually becomes random thermal motion of the molecules — that is, viscous dissipation. There are around 25 million molecules of STP air in a 1 μm cube. The same cube would contain around 34 billion molecules of water. So liquid flows are a continuum even in extremely small devices through which gas flows would not be a continuum. The average distance between molecules in the gas example is one order of magnitude higher than the diameter of its molecules, while that for the liquid phase approaches the molecular diameter. As a result, liquids are almost incompressible. Their isothermal compressibility coefficient α and bulk expansion coefficient β are much smaller than those for gases. For water, for example, a hundredfold increase in pressure leads to a less than 0.5% decrease in volume. Sound speeds through liquids are also higher than through gases, and as a result most liquid flows are incompressible.[11] The exception is the propagation of ultra-high-frequency sound waves and cavitation phenomena.

The mechanism by which liquids transport mass, momentum, and energy must be very different from that for gases. In dilute gases, intermolecular forces play no role, and the molecules spend most of their time in free flight between brief collisions that abruptly change their direction and speed. The random molecular motions are responsible for gaseous transport processes. In liquids, on the other hand, the molecules are

[10] Barring phenomena unique to liquids such as cavitation, free surface flows, etc.

[11] Note that we distinguish between a fluid's and a flow's being compressible/incompressible. For example, the *flow* of the highly compressible air can be either compressible or incompressible.

closely packed though not fixed in one position. In essence, the liquid molecules are always in a collision state. Applying a shear force must create a velocity gradient so that the molecules move relative to one another, *ad infinitum* as long as the stress is applied. For liquids, momentum transport due to the random molecular motion is negligible compared to that due to the intermolecular forces. The straining between liquid molecules causes some to separate from their original neighbors, bringing them into the force field of new molecules. Across the plane of the shear stress, the sum of all intermolecular forces must, on average, balance the imposed shear. Liquids at rest transmit only normal force, but when a velocity gradient occurs, the net intermolecular force has a tangential component.

The incompressible Navier–Stokes equations describe liquid flows under most circumstances. Liquids, however, do not have a well-advanced molecular-based theory like that for dilute gases. The concept of mean free path is not very useful for liquids, and the conditions under which a liquid flow fails to be in quasi-equilibrium state are not well defined. There is no Knudsen number to guide us through the maze of liquid flows. We do not know from first principles the conditions under which the no-slip boundary condition becomes inaccurate or the point at which the stress–rate of strain relation or the heat flux–temperature gradient relation fails to be linear. Certain empirical observations indicate that those simple relations that we take for granted occasionally fail to accurately model liquid flows. For example, it has been shown in rheological studies (Loose and Hess, 1989) that non-Newtonian behavior commences when the strain rate approximately exceeds twice the molecular frequency-scale

$$\dot{\gamma} = \frac{\partial u}{\partial y} \geqslant 2\,\mathcal{T}^{-1} \tag{4.61}$$

where the molecular time-scale $\mathcal{T}$ is given by

$$\mathcal{T} = \left[\frac{m\sigma^2}{\varepsilon}\right]^{\frac{1}{2}} \tag{4.62}$$

where m is the molecular mass, and σ and ε are respectively the characteristic length- and energy-scale for the molecules. For ordinary liquids such as water, this time-scale is extremely small and the threshold shear rate for the onset of non-Newtonian behavior is therefore extraordinarily high. For high-molecular-weight polymers, on the other hand, m and σ are both many orders of magnitude higher than their respective values for water, and the linear stress–strain relation breaks down at realistic values of the shear rate.

The moving contact line when a liquid spreads on a solid substrate is an example where slip-flow must be allowed to avoid singular or unrealistic behavior in the Navier–Stokes solutions [Dussan and Davis, 1974; Dussan, 1976, 1979; Thompson and Robbins, 1989]. Other examples where slip-flow must be admitted include corner flows [Moffatt, 1964; Koplik and Banavar, 1995] and extrusion of polymer melts from capillary tubes [Pearson and Petrie, 1968; Richardson, 1973; Den, 1990].

Existing experimental results of liquid flow in microdevices are contradictory. This is not surprising given the difficulty of such experiments and the lack of a guiding rational theory. Pfahler et al. (1990, 1991), Pfahler (1992), and Bau (1994) summarize the relevant literature. For small-length-scale flows, a phenomenological approach for analyzing the data is to define an *apparent* viscosity μ_a calculated so that if it were used in the traditional no-slip Navier–Stokes equations instead of the fluid viscosity μ, the results would be in agreement with experimental observations. Israelachvili (1986) and Gee et al. (1990) found that $\mu_a = \mu$ for thin-film flows as long as the film thickness exceeds 10 molecular layers ($\approx$5 nm). For thinner films, μ_a depends on the number of molecular layers and can be as much as 10^5 times larger than μ. Chan and Horn's (1985) results are somewhat different: the apparent viscosity deviates from the fluid viscosity for films thinner than 50 nm.

In polar-liquid flows through capillaries, Migun and Prokhorenko (1987) report that μ_a increases for tubes smaller than 1 μm in diameter. In contrast, Debye and Cleland (1959) report μ_a smaller than μ for paraffin flow in porous glass with average pore size several times larger than the molecular length-scale. Experimenting with microchannels ranging in depths from 0.5 μm to 50 μm, Pfahler, et al. (1991) found that μ_a is consistently smaller than μ for both liquid (isopropyl alcohol, silicone oil) and gas (nitrogen, helium) flows in microchannels. For liquids, the apparent viscosity decreases with decreasing channel depth. Other researchers using small capillaries report that μ_a is about the same as μ [Anderson and Quinn, 1972; Tukermann and Pease, 1981, 1982; Tuckermann, 1984; Guvenc, 1985; Nakagawa et al., 1990].

Very recently, Sharp (2001) and Sharp et al. (2001) asserted that, despite the significant inconsistencies in the literature regarding liquid flows in microchannels, such flows are well predicted by macroscale continuum theory. A case can be made to the contrary, however, as will be seen at the end of this section, and the final verdict on this controversy is yet to come.

The above contradictory results point to the need for replacing phenomenological models with first-principles models. The lack of molecular-based theory of liquids — despite extensive research by the rheology and polymer communities — leaves molecular dynamics simulations as the nearest alternative to a first-principles model. MD simulations offer a unique approach to checking the validity of the traditional continuum assumptions. However, as was pointed out earlier, such simulations are limited to exceedingly minute flow extent.

Thompson and Troian (1997) provide molecular dynamics simulations to quantify the slip-flow boundary condition dependence on shear rate. Recall the linear Navier boundary condition introduced earlier

$$\Delta u|_w = u_{\text{fluid}} - u_{\text{wall}} = L_s \left.\frac{\partial u}{\partial y}\right|_w \qquad (4.63)$$

where L_s is the constant slip length, and

$$\left.\frac{\partial u}{\partial y}\right|_w$$

is the strain rate computed at the wall. The goal of Thompson and Troian's simulations was to determine the degree of slip at a solid–liquid interface as the interfacial parameters and the shear rate change. In their simulations, a simple liquid underwent planar shear in a Couette cell as shown in Figure 4.8. The

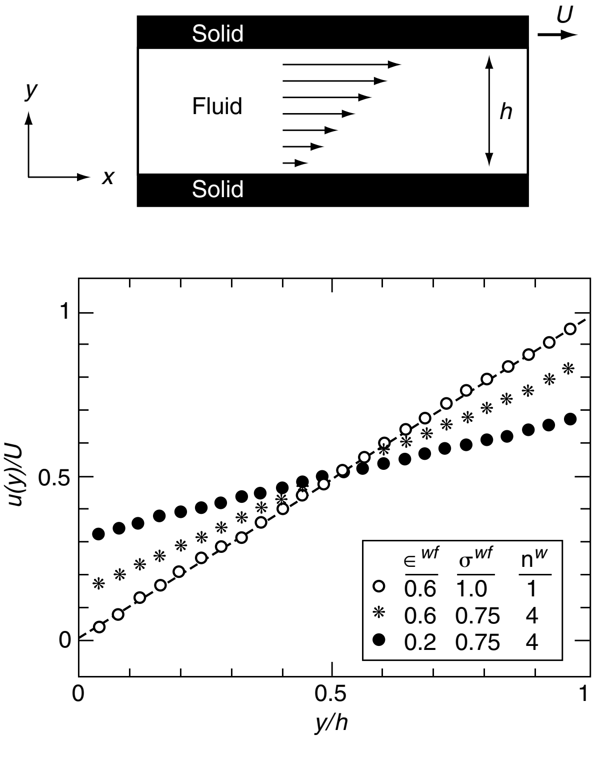


FIGURE 4.8 Velocity profiles in a Couette flow geometry at different interfacial parameters. All three profiles are for $U = \sigma\mathcal{T}^{-1}$, and $h = 24.57\sigma$. The dashed line is the no-slip Couette-flow solution. (Reprinted with permission from Thompson and Troian [1997] "A General Boundary Condition for Liquid Flow at Solid Surfaces," *Nature* **389**, pp. 360–62.)

typical cell measured 12.51 × 7.22 × h, in units of molecular length-scale σ, where the channel depth h varied in the range of 16.71σ–24.57σ, and the corresponding number of molecules simulated ranged from 1,152 to 1,728. The liquid is treated as an isothermal ensemble of spherical molecules. A shifted Lennart-Jones 6–12 potential is used to model intermolecular interactions, with energy- and length-scales ε and σ, and cut-off distance $r_c = 2.2\sigma$:

$$V(r) = 4\varepsilon\left[\left(\frac{r}{\sigma}\right)^{-12} - \left(\frac{r}{\sigma}\right)^{-6} - \left(\frac{r_c}{\sigma}\right)^{-12} + \left(\frac{r_c}{\sigma}\right)^{-6}\right] \tag{4.64}$$

The truncated potential is set to zero for $r > r_c$.

The fluid–solid interaction is also modeled with a truncated Lennart-Jones potential, with energy- and length-scales ε^{wf} and σ^{wf}, and cut-off distance r_c. The equilibrium state of the fluid is a well-defined liquid phase characterized by number density $n = 0.81\sigma^{-3}$ and temperature $T = 1.1\varepsilon/k$, where k is the Boltzmann constant.

The steady state velocity profiles resulting from Thompson and Troian's (1997) MD simulations are depicted in Figure 4.8 for different values of the interfacial parameters ε^{wf}, σ^{wf}, and n^w. Those parameters, shown in units of the corresponding fluid parameters ε, σ, and n, characterize respectively the strength of the liquid–solid coupling, the thermal roughness of the interface, and the commensurability of wall and liquid densities. The macroscopic velocity profiles recover the expected flow behavior from continuum hydrodynamics with boundary conditions involving varying degrees of slip. Note that when slip exists, the shear rate $\dot{\gamma}$ no longer equals U/h. The degree of slip increases (i.e., the amount of momentum transfer at the wall–fluid interface decreases) as the relative wall density n^w increases or the strength of the wall–fluid coupling σ^{wf} decreases — in other words, when the relative surface energy corrugation of the wall decreases. Conversely, the corrugation is maximized when the wall and fluid densities are commensurate and the strength of the wall–fluid coupling is large. In this case, the liquid *feels* the corrugations in the surface energy of the solid owing to the atomic close-packing. Consequently, there is efficient momentum transfer, and the no-slip condition applies, or in extreme cases, a "stick" boundary condition takes hold.

Variations of the slip length L_s and viscosity μ as functions of shear rate $\dot{\gamma}$ are shown in Figure 4.9 for five different sets of interfacial parameters. For Couette flow, the slip length is computed from its definition, $\Delta u|_w/\dot{\gamma} = (U/\dot{\gamma} - \mathrm{h})/2$. The slip length, viscosity, and shear rate are normalized in the figure using the respective molecular scales for length σ, viscosity $\varepsilon\mathcal{T}\sigma^{-3}$, and inverse time $\mathcal{T}^{-1}$. The viscosity of the fluid is constant over the entire range of shear rates (Figure 4.9b) indicating Newtonian behavior. As indicated earlier, non-Newtonian behavior is expected for $\dot{\gamma} \geqslant 2\mathcal{T}^{-1}$, well above the shear rates used in Thompson and Troian's simulations.

At low shear rates, the slip length behavior is consistent with the Navier model (i.e., is independent of the shear rate). Its limiting value L_s^o ranges from 0 to $\sim 17\sigma$ for the range of interfacial parameters chosen (Figure 4.9a). In general, the amount of slip increases with decreasing surface energy corrugation. Most interestingly, at high shear rates the Navier condition breaks down as the slip length increases rapidly with $\dot{\gamma}$. The critical shear-rate value for the slip length to diverge, $\dot{\gamma}_c$, decreases as the surface energy corrugation decreases. Surprisingly, the boundary condition is nonlinear even though the liquid is still Newtonian. In dilute gases, the linear slip condition and the Navier–Stokes equations, with their linear stress–strain relation, are both valid to the same order of approximation in Knudsen number. In other words, deviation from linearity is expected to take place at the same value of $Kn = 0.1$. In liquids, in contrast, the slip length appears to become nonlinear and to diverge at a critical value of shear rate well below the shear rate at which the linear stress–strain relation fails. Moreover, the boundary condition deviation from linearity is not gradual but is rather catastrophic. The critical value of shear rate $\dot{\gamma}_c$ signals the point at which the solid can no longer impart momentum to the liquid. This means that the same liquid molecules sheared against different substrates will experience varying amounts of slip and vice versa.

Based on the above results, Thompson and Troian (1997) suggest a universal boundary condition at a solid–liquid interface. Scaling the slip length L_s by its asymptotic limiting value L_s^o and the shear rate $\dot{\gamma}$ by its critical value $\dot{\gamma}_c$ collapses the data in the single curve shown in Figure 4.10. The data points are well described by the relation

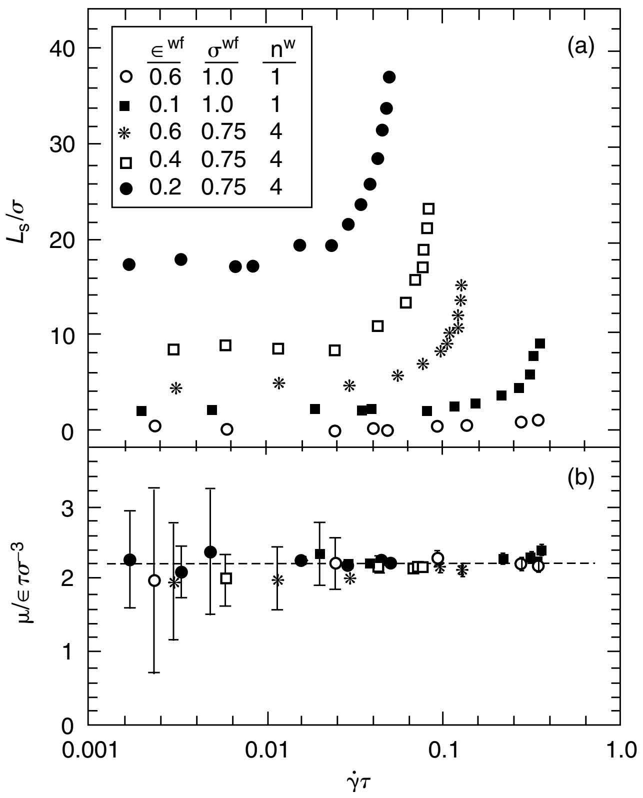


FIGURE 4.9 Variation of slip length and viscosity as functions of shear rate. (Reprinted with permission from Thompson and Troian [1997] "A General Boundary Condition for Liquid Flow at Solid Surfaces," *Nature* **389**, pp. 360–62.)

$$L_s = L_s^o\left[1 - \frac{\dot{\gamma}}{\dot{\gamma}_c}\right]^{-\frac{1}{2}} \tag{4.65}$$

The nonlinear behavior close to a critical shear rate suggests that the boundary condition can significantly affect flow behavior at macroscopic distances from the wall. Experiments with polymers confirm this observation [Atwood and Schwalter, 1989]. The rapid change in the slip length suggests that for flows in the vicinity of $\dot{\gamma}_c$, small changes in surface properties can lead to large fluctuations in the apparent boundary condition. Thompson and Troian (1997) conclude that the Navier slip condition is but the low-shear-rate limit of a more generalized universal relationship that is nonlinear and divergent. Their relation provides a mechanism for relieving the stress singularity in spreading contact lines and corner flows, as it naturally allows for varying degrees of slip on approach to regions of higher rate of strain.

To place the above results in physical terms, consider water[12] at a temperature of $T = 288$ K. The energy-scale in the Lennart-Jones potential is then $\varepsilon = 3.62 \times 10^{-21}$ J. For water, $m = 2.99 \times 10^{-26}$ kg, $\sigma = 2.89 \times 10^{-10}$ m, and at standard temperature $n = 3.35 \times 10^{28}$ molecules/m^3. The molecular time-scale can thus be computed,

$$\mathcal{T} = [m\sigma^2/\varepsilon]^{\frac{1}{2}} = 8.31 \times 10^{-13}\,\text{s}$$

[12] Water molecules are complex, forming directional, short-range covalent bonds and thus requiring a more complex potential than the Lennart-Jones to describe the intermolecular interactions. For the purpose of the qualitative example described here, however, we use the computational results of Thompson and Troian (1997), who employed the L–J potential.

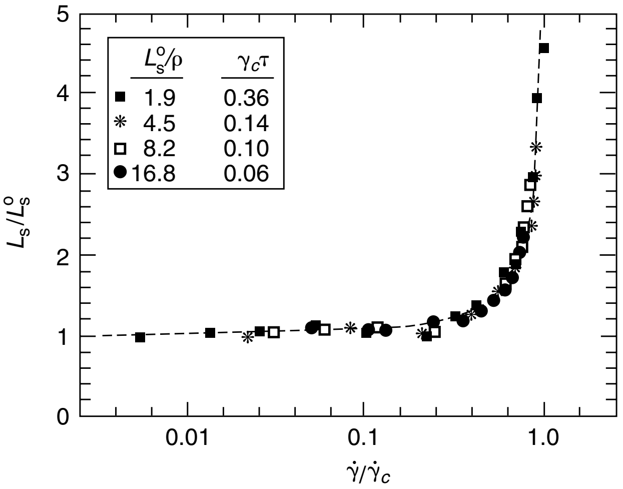


FIGURE 4.10 Universal relation of slip length as a function of shear rate. (Reprinted with permission from Thompson and Troian [1997] "A General Boundary Condition for Liquid Flow at Solid Surfaces," *Nature* **389**, pp. 360–62.)

For the third case depicted in Figure 4.10 (the open squares), $\dot{\gamma}_c\mathcal{T} = 0.1$, and the critical shear rate at which the slip condition diverges is thus $\dot{\gamma}_c = 1.2 \times 10^{11}\,\text{s}^{-1}$. Such an enormous rate of strain[13] may be found in extremely small devices having extremely high speeds. On the other hand, the conditions to achieve a measurable slip of 17σ (the solid circles in Figure 4.9) are not difficult to encounter in microdevices: density of solid that is four times that of liquid, and energy-scale for wall-fluid interaction that is one-fifth of energy-scale for liquid.

The limiting value of slip length is independent of the shear rate and can be computed for water as $L_s^o = 17\,\sigma = 4.91 \times 10^{-9}\,\text{m}$. Consider a water microbearing having a shaft diameter of 100 µm, a rotation rate of 20,000 rpm, and a minimum gap of $h = 1\,\mu\text{m}$. In this case, $U = 0.1$ m/sec, and the no-slip shear rate is $U/h = 10^5\,\text{s}^{-1}$. When slip occurs at the limiting value just computed, the shear rate and the wall slip-velocity are computed as follows

$$\dot{\gamma} = \frac{U}{h + 2L_s^o} = 9.90 \times 10^4\,\text{s}^{-1} \tag{4.66}$$

$$\Delta u|_w = \dot{\gamma}L_s = 4.87 \times 10^{-4}\,\text{m/s} \tag{4.67}$$

As a result of the Navier slip, the shear rate is reduced by 1% from its no-slip value, and the slip velocity at the wall is about 0.5% of U, small but not insignificant.

4.9 Surface Phenomena

The surface-to-volume ratio for a machine with a characteristic length of 1 m is $1\,\text{m}^{-1}$, while that for a MEMS device having a size of 1 µm is $10^6\,\text{m}^{-1}$. The millionfold increase in surface area relative to the mass of the minute device substantially affects the transport of mass, momentum, and energy through the surface. Obviously surface effects dominate in small devices. The surface boundary conditions in MEMS flows have already been discussed earlier. We have shown that in microdevices it is possible to have measurable slip-velocity and temperature jump at a solid–fluid interface. In this section, we illustrate other ramifications of the large surface-to-volume ratio unique to MEMS and provide a molecular viewpoint to surface forces.

[13] Note however that $\dot{\gamma}_c$ for high-molecular-weight polymers would be many orders of magnitude smaller than the value developed here for water.

In microdevices, both radiative and convective heat loss/gain are enhanced by the huge surface-to-volume ratio. Consider a device having a characteristic length L_s. Use of the lumped capacitance method to compute the rate of convective heat transfer, for example, is justified if the Biot number ($\equiv hL_s/\kappa_s$, where h is the convective heat transfer coefficient of the fluid and κ_s is the thermal conductivity of the solid) is less than 0.1. Small L_s implies a small Biot number and a nearly uniform temperature within the solid. Within this approximation, the rate at which heat is lost to the surrounding fluid is given by

$$\rho_s L_s^3 c_a \frac{dT}{dt} = -hL_s^2(T_s - T_\infty) \tag{4.68}$$

where ρ_s and c_s are respectively the density and specific heat of the solid, T_s is its (uniform) temperature, and T_∞ is the ambient fluid temperature. Solution of the above equation is trivial, and the temperature of a hot surface drops exponentially with time from an initial temperature T_i,

$$\frac{T_s(t) - T_\infty}{T_i - T_\infty} = \exp\left[-\frac{t}{\mathcal{T}}\right] \tag{4.69}$$

where the time constant $\mathcal{T}$ is given by

$$\mathcal{T} = \frac{\rho_s L_s^3 c_s}{hL_s^2} \tag{4.70}$$

For small devices, the time it takes the solid to cool is proportionally small. Clearly, the millionfold increase in surface-to-volume ratio implies a proportional increase in the rate at which heat escapes. Identical scaling arguments can be made regarding mass transfer.

Another effect of the diminished scale is the increased importance of surface forces and the waning importance of body forces. Based on biological studies, Went (1968) concludes that the demarcation length-scale is around 1 mm. Below that, surface forces dominate over gravitational forces. A 10 mm piece of paper will fall when gently placed on a smooth vertical wall, while a 0.1 mm piece will stick. Try it! *Stiction* is a major problem in MEMS applications. Certain structures such as long, thin polysilicon beams and large, thin comb-drives have a propensity to stick to their substrates and thus fail to perform as designed [Mastrangelo and Hsu, 1992; Tang et al., 1989].

Conventional dry friction between two solids in relative motion is proportional to the normal force that is usually a component of the moving device weight. The friction is independent of the contact-surface area because the van der Waals cohesive forces are negligible relative to the weight of the macroscopic device. In MEMS applications, the cohesive intermolecular forces between two surfaces are significant, and the stiction is independent of the device mass but is proportional to its surface area. The first micromotor did not move — despite large electric current through it — until the contact area between the 100 μm rotor and the substrate was reduced significantly by placing dimples on the rotor's surface [Fan et al., 1988, 1989; Tai and Muller, 1989].

One last example of surface effects that to my knowledge has not been investigated for microflows is the adsorbed layer in gaseous wall-bounded flows. It is well known [Brunauer, 1944; Lighthill, 1963] that when a gas flows in a duct, the gas molecules are attracted to the solid surface by the van der Waals and other forces of cohesion. The potential energy of the gas molecules drops on reaching the surface. The adsorbed layer partakes the thermal vibrations of the solid, and the gas molecules can only escape when their energy exceeds the potential energy minimum. In equilibrium, at least part of the solid would be covered by a monomolecular layer of adsorbed gas molecules. Molecular species with significant partial pressure — relative to their vapor pressure — may locally form layers two or more molecules thick. Consider, for example, the flow of a mixture of dry air and water vapor at STP. The energy of adsorption of water is much larger than that for nitrogen and oxygen, making it more difficult for water molecules to escape the potential energy trap. It follows that the life time of water molecules in the adsorbed layer significantly exceeds that for the air molecules (60,000-fold, in fact) and, as a result, the thin surface layer would be mostly water. For example, if the proportion of water vapor in the ambient air is 1:1,000 (i.e., very low humidity level), the ratio of water to air in the adsorbed layer would be 60:1. Microscopic roughness of the solid surface causes partial condensation of the water along portions having sufficiently strong concave curvature. So, surfaces exposed to nondry air flows are mainly liquid water surfaces. In most applications, this thin adsorbed layer has little effect on the flow dynamics despite the fact that the density and viscosity of liquid water are far greater than those for air. In MEMS applications, however, the layer thickness may be a significant portion of the characteristic flow dimension, and the water layer may have a measurable effect on the gas flow. A hybrid approach combining molecular dynamics and continuum flow simulations or MD–Monte Carlo simulations may be used to investigate this issue.

Majumdar and Mezic (1998, 1999) have studied the stability and rupture into droplets of thin liquid films on solid surfaces. They point out that the free energy of a liquid film consists of a surface tension component as well as highly nonlinear volumetric intermolecular forces resulting from van der Waals, electrostatic, hydration, and elastic strain interactions. For water films on hydrophilic surfaces such as silica and mica, Majumdar and Mezic (1998) estimate the equilibrium film thickness to be about 0.5 nm (2 monolayers) for a wide range of ambient-air relative humidities. The equilibrium thickness grows very sharply, however, as the relative humidity approaches 100%.

Majumdar and Mezic's (1998, 1999) results raise many questions. What are the stability characteristics of their water film in the presence of air flow above it? Would this water film affect the accommodation coefficient for microduct air flow? In a modern Winchester-type hard disk, the drive mechanism has a read/write head that floats 50 nm above the surface of the spinning platter. The head and platter together with the intervening air layer form a slider bearing. Would the computer performance be affected adversely by the high relative humidity on a particular day when the adsorbed water film is no longer "thin"? If a microduct hauls liquid water, would the water film adsorbed by the solid walls influence the effective viscosity of the water flow? Electrostatic forces can extend to almost 1 μm (the Debye length), and that length is known to be highly pH-dependent. Would the water flow be influenced by the surface and liquid chemistry? Would this explain the contradictory experimental results of liquid flows in microducts discussed earlier?

The few examples above illustrate the importance of surface effects in small devices. From the continuum viewpoint, forces at a solid–fluid interface are the limit of pressure and viscous forces acting on a parallel elementary area displaced into the fluid when the displacement distance is allowed to tend to zero. From the molecular point of view, all macroscopic surface forces are ultimately traced to intermolecular forces, which subject is extensively covered in Israelachvilli (1991) and the references therein. Here we provide a very brief introduction to the molecular viewpoint. The four forces in nature are (1) the strong and (2) the weak forces describing the interactions between neutrons, protons, electrons, etc.; (3) the electromagnetic forces between atoms and molecules; and (4) gravitational forces between masses. The range of action of the first two forces is around 10^{-5} nm, and hence neither concerns us overly in MEMS applications. The electromagnetic forces are effective over a much larger though still small distance on the order of the interatomic separations (0.1–0.2 nm). Effects over longer range — several orders of magnitude longer — can and do rise from the short-range intermolecular forces. For example, the rise of liquid columns in capillaries and the action of detergent molecules in removing oily dirt from fabric are the result of intermolecular interactions. Gravitational forces decay with the distance to the second power, while intermolecular forces decay much quicker, typically with the seventh power. Cohesive forces are therefore negligible once the distance between molecules exceeds a few molecular diameters, while massive bodies like stars and planets are still strongly interacting via gravity over astronomical distances.

Electromagnetic forces are the source of all intermolecular interactions and the cohesive forces holding atoms and molecules together in solids and liquids. They can be classified as (1) purely electrostatic forces arising from the Coulomb force between charges, interactions between charges, permanent dipoles, quadrupoles, etc.; (2) polarization forces arising from the dipole moments induced in atoms and molecules by the electric field of nearby charges and permanent dipoles; and (3) quantum mechanical forces that give rise to covalent or chemical bonding and to repulsive steric or exchange interactions that balance the attractive forces at very short distances. The Hellman–Feynman theorem of quantum mechanics states that once the spatial distribution of the electron clouds has been determined by solving the appropriate Schrödinger equation, intermolecular forces may be calculated on the basis of classical electrostatics, in effect reducing all intermolecular forces to Coulombic forces. Note however that intermolecular forces exist even when the molecules are totally neutral. Solutions of the Schrödinger equation for general atoms and molecules are not easy of course, and modeling alternatives are sought to represent intermolecular forces. The van der Waals attractive forces are usually represented with a potential that varies as the inverse-sixth power of distance, while the repulsive forces are represented with either a power or an exponential potential.

A commonly used potential between two molecules is the generalized Lennart-Jones (L–J 6–12) pair potential given by

$$V_{ij}(r) = 4\varepsilon\left[c_{ij}\left(\frac{r}{\sigma}\right)^{-12} - d_{ij}\left(\frac{r}{\sigma}\right)^{-6}\right] \tag{4.71}$$

where V_{ij} is the potential energy between two particles *i* and *j*, *r* is the distance between the two molecules, ε and σ are respectively characteristic energy- and length-scales, and c_{ij} and d_{ij} are parameters to be chosen for the particular fluid and solid combinations under consideration. The first term in the right-hand side is the strong repulsive force that is felt when two molecules are at extremely close range comparable to the molecular length-scale. That short-range repulsion prevents overlap of the molecules in physical space. The second term is the weaker van der Waals attractive force that commences when the molecules are sufficiently close (several times σ). That negative part of the potential represents the attractive polarization interaction of neutral, spherically symmetric particles. The power of 6 associated with this term is derivable from quantum mechanics considerations, while the power of the repulsive part of the potential is found empirically. The Lennart-Jones potential is zero at very large distances, has a weak negative peak at *r* slightly larger than σ, is zero at $r = \sigma$, and is infinite as $r \to 0$.

The force field resulting from this potential is given by

$$F_{ij}(r) = -\frac{\partial V_{ij}}{\partial r} = \frac{48\varepsilon}{\sigma}\left[c_{ij}\left(\frac{r}{\sigma}\right)^{-13} - \frac{d_{ij}}{2}\left(\frac{r}{\sigma}\right)^{-7}\right] \tag{4.72}$$

A typical L–J 6–12 potential and force field are shown in Figure 4.11, for $c = d = 1$. The minimum potential $V_{\min} = -\varepsilon$, corresponds to the equilibrium position (zero force) and occurs at $r = 1.12\,\sigma$. The attractive van der Waals contribution to the minimum potential is -2ε, while the repulsive energy contribution is $+\varepsilon$. Thus the inverse 12th-power repulsive force term decreases the strength of the binding energy at equilibrium by 50%.

The L–J potential is commonly used in molecular dynamics simulations to model intermolecular interactions between dense gas or liquid molecules and between fluid and solid molecules. As mentioned

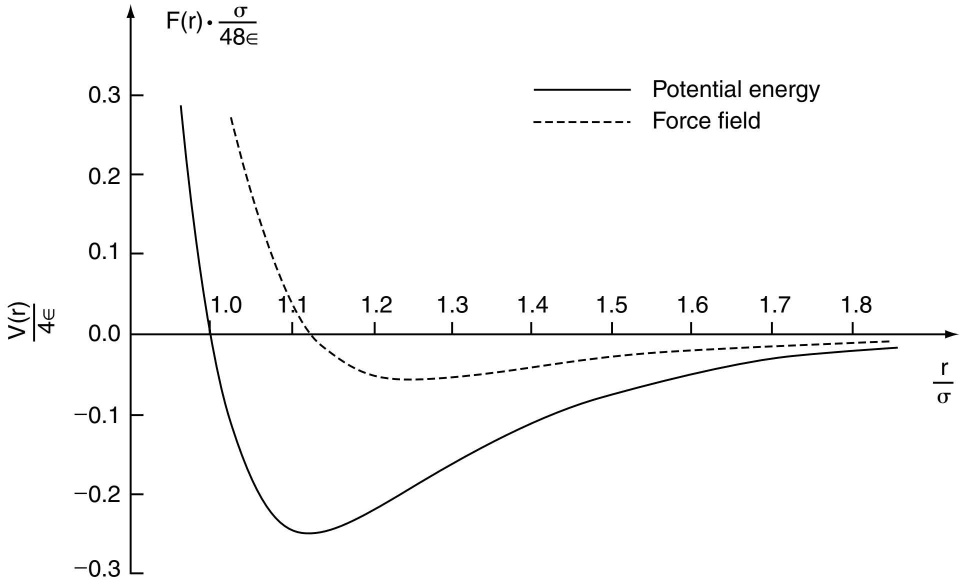


FIGURE 4.11 Typical Lennart-Jones 6–12 potential and the intermolecular force field resulting from it. Only a small portion of the potential function is shown for clarity.

earlier, such potential is not accurate for complex substances, such as water, whose molecules form directional covalent bonds. As a result, MD simulations for water are much more involved.

4.10 Parting Remarks

Richard Feynman's 40-year-old vision of building minute machines is now a reality. Microelectromechanical systems have witnessed explosive growth during the last decade and are finding increased applications in a variety of industrial and medical fields. The physics of fluid flows in microdevices has been explored in this chapter. While we now know considerably more than we did just few years ago, much physics remains to be explored so that rational tools can be developed for the design, fabrication, and operation of MEMS devices.

The traditional Navier–Stokes model of fluid flows with no-slip boundary conditions works only for a certain range of the governing parameters. This model basically demands two conditions: (1) the fluid is a continuum, which condition is almost always satisfied as there are usually more than 1 million molecules in the smallest volume in which appreciable macroscopic changes take place (this is the molecular chaos restriction); and (2) the flow is not too far from thermodynamic equilibrium, which is satisfied if there is a sufficient number of molecular encounters during a time that is small compared to the smallest time-scale for flow changes. During this time, the average molecule would have moved a distance small compared to the smallest flow length-scale.

For gases, the Knudsen number determines the degree of rarefaction and the applicability of traditional flow models. As $Kn \rightarrow 0$, the time- and length-scales of molecular encounters are vanishingly small compared to those for the flow, and the velocity distribution of each element of the fluid instantaneously adjusts to the equilibrium thermodynamic state appropriate to the local macroscopic properties as this molecule moves through the flow field. From the continuum viewpoint, the flow is isentropic, and heat conduction and viscous diffusion and dissipation vanish from the continuum conservation relations leading to the Euler equations of motion. At small but finite Kn, the Navier–Stokes equations describe near-equilibrium, continuum flows.

Slip flow must be taken into account for $Kn > 0.001$. The slip boundary condition is at first linear in Knudsen number; then nonlinear effects take over beyond a Knudsen number of 0.1. At the same transition regime (i.e., $0.1 < Kn < 10$), the linear-stress–rate-of-strain and heat-flux–temperature-gradient relations needed to close the Navier–Stokes equations also break down, and alternative continuum equations (e.g., Burnett or higher-order equations) or molecular-based models must be invoked. In the transition regime, provided that the dilute gas and molecular chaos assumptions hold, solutions to the difficult Boltzmann equation are sought, but physical simulations such as Monte Carlo methods are more readily executed in this range of Knudsen number. In the free-molecule flow regime (i.e., $Kn > 10$), the nonlinear collision integral is negligible, and the Boltzmann equation is drastically simplified. Analytical solutions are possible in this case for simple geometries and numerical integration of the Boltzmann equation is straightforward for arbitrary geometries provided that the surface-reflection characteristics are accurately modeled.

Gaseous flows are often compressible in microdevices even at low Mach numbers. Viscous effects can cause sufficient pressure drop and density changes for the flow to be treated as compressible. In a long, constant-area microduct, all Knudsen number regimes may be encountered, and the degree of rarefaction increases along the tube. The pressure drop is nonlinear and the Mach number increases downstream, limited only by choked-flow condition.

Similar deviation and breakdown of the traditional Navier–Stokes equations occur for liquids as well, but there the situation is more murky. Existing experiments are contradictory. There is no kinetic theory of liquids, and first-principles prediction methods are scarce. Molecular dynamics simulations can be used, but they are limited to extremely small flow extents. Nevertheless, measurable slip is predicted from MD simulations at realistic shear rates in microdevices.

Much nontraditional physics is still to be learned, and many exciting applications of microdevices are yet to be discovered. The future is bright for this emerging field of science and technology. Feynman was right about the possibility of building mite-size machines, but was unduly cautious in forecasting that while such machines would be fun to make, they might or might not be useful.

References

Agarwal, R., Yun, K., and Balakrishnan, R. (1999) "Beyond Navier Stokes: Burnett Equations for Flow Simulations in Continuum–Transition Regime," AIAA Paper No. 99-3580, Reston, Virginia.

Agarwal, R.K., Yun, K.Y., and Balakrishnan, R. (2003) "Beyond Navier Stokes: Burnett Equations for Flow Simulations in Continuum–Transition Regime," *Phys. Fluids* **13**, pp. 3061–85.

Alder, B.J., and Wainwright, T.E. (1957) "Studies in Molecular Dynamics," *J. Chem. Phys.* **27**, pp. 1208–9.

Alder, B.J., and Wainwright, T.E. (1958) "Molecular Dynamics by Electronic Computers," in *Transport Processes in Statistical Mechanics*, I. Prigogine, ed., pp. 97–131, Interscience, New York.

Alder, B.J., and Wainwright, T.E. (1970) "Decay of the Velocity Auto-Correlation Function," *Phy. Rev. A* **1**, pp. 18–21.

Allen, M.P., and Tildesley, D.J. (1987) *Computer Simulation of Liquids*, Clarendon Press, Oxford.

Anderson, J.L., and Quinn, J.A. (1972) "Ionic Mobility in Microcapillaries," *J. Chem. Phys.* **27**, pp. 1208–9.

Arkilic, E.B. (1997) Measurement of the Mass Flow and Tangential Momentum Accommodation Coefficient in Silicon Micromachined Channels, Ph.D. thesis, Massachusetts Institute of Technology.

Arkilic, E.B., Schmidt, M.A., and Breuer, K.S. (1995) "Slip Flow in Microchannels," in *Rarefied Gas Dynamics,* vol. 19, J. Harvey and G. Lord, eds., Oxford University Press, Oxford.

Arkilic, E.B., Schmidt, M.A., and Breuer, K.S. (1997a) "Gaseous Slip Flow in Long Microchannels," *J. MEMS* **6**, pp. 167–78.

Arkilic, E.B., Schmidt, M.A., and Breuer, K.S. (1997b) "TMAC Measurement in Silicon Micromachined Channels," in *Rarefied Gas Dynamics* vol. 20, C. Shen, ed., Beijing University Press, Beijing.

Atwood, B.T., and Schowalter, W.R. (1989) "Measurements of Slip at the Wall during Flow of High-Density Polyethylene through a Rectangular Conduit," *Rheol. Acta* **28**, pp. 134–46.

Bau, H.H. (1994) "Transport Processes Associated with Micro-Devices," *Therm. Sci. Eng.* **2**, pp. 172–78.

Beskok, A. (1994) Simulation of Heat and Momentum Transfer in Complex Micro-Geometries, M.Sc. thesis, Princeton University.

Beskok, A. (1996) Simulations and Models of Gas Flows in Microgeometries, Ph.D. thesis, Princeton University.

Beskok, A., and Karniadakis, G.E. (1994) "Simulation of Heat and Momentum Transfer in Complex Micro-Geometries," *J. Thermophys. Heat Trans.* **8**, pp. 355–70.

Beskok, A., and Karniadakis, G.E. (1999) "A Model for Flows in Channels, Pipes, and Ducts at Micro and Nano Scales," *Microscale Thermophys. Eng.* **3**, pp. 43–77.

Beskok, A., Karniadakis, G.E., and Trimmer, W. (1996) "Rarefaction and Compressibility Effects in Gas Microflows," *J. Fluids Eng.* **118**, pp. 448–56.

Bhatnagar, P.L., Gross, E.P., and Krook, M. (1954) "A Model for Collision Processes in Gases: 1. Small Amplitude Processes in Charged and Neutral One-Component Systems," *Phys. Rev.* **94**, pp. 511–24.

Bird, G.A. (1963) "Approach to Translational Equilibrium in a Rigid Sphere Gas," *Phys. Fluids* **6**, pp. 1518–19.

Bird, G.A. (1965) "The Velocity Distribution Function within a Shock Wave," *J. Fluid Mech.* **30**, pp. 479–87.

Bird, G.A. (1976*) Molecular Gas Dynamics*, Clarendon Press, Oxford.

Bird, G.A. (1978) "Monte Carlo Simulation of Gas Flows," *Annu. Rev. Fluid Mech.* **10**, pp. 11–31.

Bird, G.A. (1994) *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*, Clarendon Press, Oxford.

Brunauer, S. (1944) *Physical Adsorption of Gases and Vapours*, Oxford University Press, Oxford.

Cercignani, C. (1988) *The Boltzmann Equation and Its Applications*, Springer-Verlag, Berlin.

Cercignani, C. (2000) *Rarefied Gas Dynamics: From Basic Concepts to Actual Calculations*, Cambridge University Press, London.

Chan, D.Y.C., and Horn, R.G. (1985) "Drainage of Thin Liquid Films," *J. Chem. Phys.* **83**, pp. 5311–24.

Chapman, S., and Cowling, T.G. (1970) *The Mathematical Theory of Non-Uniform Gases*, 3rd ed., Cambridge University Press, Cambridge.

Cheng, H.K. (1993) "Perspectives on Hypersonic Viscous Flow Research," *Annu. Rev. Fluid Mech.* **25**, pp. 455–84.

Cheng, H.K., and Emmanuel, G. (1995) "Perspectives on Hypersonic Nonequilibrium Flow," *AIAA J.* **33**, pp. 385–400.

Ciccotti, G., and Hoover, W.G., eds. (1986) *Molecular Dynamics Simulation of Statistical Mechanics Systems*, North Holland, Amsterdam.

Debye, P., and Cleland, R.L. (1959) "Flow of Liquid Hydrocarbons in Porous Vycor," *J. Appl. Phys.* **30**, pp. 843–49.

Den, L.M. (1990) "Issues in Viscoelastic Fluid Mechanics," *Annu. Rev. Fluid Mech.* **22**, pp. 13–34.

Dussan, E.B. (1976) "The Moving Contact Line: The Slip Boundary Condition," *J. Fluid Mech.* **77**, pp. 665–84.

Dussan, E.B. (1979) "On the Spreading of Liquids on Solid Surfaces: Static and Dynamic Contact Lines," *Annu. Rev. Fluid Mech.* **11**, pp. 371–400.

Dussan, E.B., and Davis, S.H. (1974) "On the Motion of Fluid–Fluid Interface Along a Solid Surface," *J. Fluid Mech.* **65**, pp. 71–95.

Fan, L.-S., Tai, Y.-C., and Muller, R.S. (1988) "Integrated Movable Micromechanical Structures for Sensors and Actuators," in *IEEE Transactions on Electronic Devices*, vol. 35, pp. 724–30.

Fan, L.-S., Tai, Y.-C., and Muller, R.S. (1989) "IC-Processed Electrostatic Micromotors," *Sensor. Actuator.* **20**, pp. 41–47.

Gad-el-Hak, M. (1995) "Questions in Fluid Mechanics: Stokes' Hypothesis for a Newtonian, Isotropic Fluid," *J. Fluids Eng.* **117**, pp. 3–5.

Gad-el-Hak, M. (1999) "The Fluid Mechanics of Microdevices: The Freeman Scholar Lecture," *J. Fluids Eng.* **121**, pp. 5–33.

Gad-el-Hak, M. (2000) *Flow Control: Passive, Active, and Reactive Flow Management*, Cambridge University Press, London.

Gee, M.L., McGuiggan, P.M., Israelachvili, J.N., and Homola, A.M. (1990) "Liquid to Solidlike Transitions of Molecularly Thin Films under Shear," *J. Chem. Phys.* **93**, pp. 1895–906.

Guvenc, M.G. (1985) "V-Groove Capillary for Low Flow Control and Measurement," in *Micromachining and Micropackaging of Transducers*, C.D. Fung, P.W. Cheung, W.H. Ko, and D.G. Fleming, eds., pp. 215–23, Elsevier, Amsterdam.

Haile, J.M. (1993) *Molecular Dynamics Simulation: Elementary Methods*, Wiley, New York.

Harley, J.C., Huang, Y., Bau, H.H., and Zemel, J.N. (1995) "Gas Flow in Micro-Channels," *J. Fluid Mech.* **284**, pp. 257–74.

Hirschfelder, J.O., Curtiss, C.F., and Bird, R.B. (1954) *Molecular Theory of Gases and Liquids*, Wiley, New York.

Israelachvili, J.N. (1986) "Measurement of the Viscosity of Liquids in Very Thin Films," *J. Colloid Interface Sci.* **110**, pp. 263–71.

Israelachvili, J.N. (1991) *Intermolecular and Surface Forces*, 2nd ed., Academic Press, New York.

Karniadakis, G.E., and Beskok A. (2002) *Microflows: Fundamentals and Simulation*, Springer-Verlag, New York.

Kennard, E.H. (1938) *Kinetic Theory of Gases*, McGraw-Hill, New York.

Knight, J. (1999) "Dust Mite's Dilemma," *New Sci.* **162**, 29 May, pp. 40–43.

Knudsen, M. (1909) "Die Gesetze der Molekularströmung und der inneren Reibungsströmung der Gase durch Röhren," *Ann. Phys.* **28**, pp. 75–130.

Kogan, M.N. (1969) *Rarefied Gas Dynamics*, Nauka, Moscow. Trans. from Russian, L. Trilling, ed., Plenum, New York.

Kogan, M.N. (1973) "Molecular Gas Dynamics," *Annu. Rev. Fluid Mech.* **5**, pp. 383–404.

Koplik, J., and Banavar, J.R. (1995) "Continuum Deductions from Molecular Hydrodynamics," *Annu. Rev. Fluid Mech.* **27**, pp. 257–92.

Kovacs, G.T.A. (1998) *Micromachined Transducers Sourcebook*, McGraw-Hill, New York.

Lighthill, M.J. (1963) "Introduction: Real and Ideal Fluids," in *Laminar Boundary Layers*, L. Rosenhead, ed., pp. 1–45, Clarendon Press, Oxford.

Liu, J., Tai, Y.C., Lee, J., Pong, K.C., Zohar, Y., and Ho, C.M. (1993) "In-Situ Monitoring and Universal Modeling of Sacrificial PSG Etching Using Hydrofluoric Acid," in *Proc. IEEE Micro Electro Mechanical Systems '93*, pp. 71–76, IEEE, New York.

Liu, J., Tai, Y.C., Pong, K., and Ho, C.M. (1995) "MEMS for Pressure Distribution Studies of Gaseous Flows in Microchannels," in *Proc. IEEE Micro Electro Mechanical Systems '95*, pp. 209–15, IEEE, New York.

Lockerby, D.A., and Reese, J.M. (2003) "High-Resolution Burnett Simulations of Micro Couette Flow and Heat Transfer," *J. Comput. Phys.* **188**, pp. 333–47.

Loeb, L.B. (1961) *The Kinetic Theory of Gases*, 3rd ed., Dover, New York.

Löfdahl, L., and Gad-el-Hak, M. (1999) "MEMS Applications in Turbulence and Flow Control," *Prog. Aero. Sci.* **35**, pp. 101–203.

Loose, W., and Hess, S. (1989) "Rheology of Dense Fluids via Nonequilibrium Molecular Hydrodynamics: Shear Thinning and Ordering Transition," *Rheol. Acta* **28**, pp. 91–101.

Madou, M. (2002) *Fundamentals of Microfabrication*, 2nd ed., CRC Press, Boca Raton.

Majumdar, A., and Mezic, I. (1998) "Stability Regimes of Thin Liquid Films," *Microscale Thermophys. Eng.* **2**, pp. 203–13.

Majumdar, A., and Mezic, I. (1999) "Instability of Ultra-Thin Water Films and the Mechanism of Droplet Formation on Hydrophilic Surfaces," *J. Heat Trans.* **121**, pp. 964–971.

Mastrangelo, C., and Hsu, C.H. (1992) "A Simple Experimental Technique for the Measurement of the Work of Adhesion of Microstructures," in *Technical Digest IEEE Solid-State Sensors and Actuators Workshop*, pp. 208–12, IEEE, New York.

Maxwell, J.C. (1879) "On Stresses in Rarefied Gases Arising from Inequalities of Temperature," *Phil. Trans. R. Soc. Part 1* **170**, pp. 231–56.

Migun, N.P., and Prokhorenko, P.P. (1987) "Measurement of the Viscosity of Polar Liquids in Microcapillaries," *Colloid J. USSR* **49**, pp. 894–97.

Moffatt, H.K. (1964) "Viscous and Resistive Eddies Near a Sharp Corner," *J. Fluid Mech.* **18**, pp. 1–18.

Muntz, E.P. (1989) "Rarefied Gas Dynamics," *Annu. Rev. Fluid Mech.* **21**, pp. 387–417.

Nadolink, R.H., and Haigh, W.W. (1995) "Bibliography on Skin Friction Reduction with Polymers and Other Boundary-Layer Additives," *Appl. Mech. Rev.* **48**, pp. 351–459.

Nakagawa, S., Shoji, S., and Esashi, M. (1990) "A Micro-Chemical Analyzing System Integrated on Silicon Chip," in *Proc. IEEE: Micro Electro Mechanical Systems*, Napa Valley, California, IEEE 90CH2832-4, IEEE, New York.

Nguyen, N.-T., and Wereley, S.T. (2002) *Fundamentals and Applications of Microfluidics*, Artech House, Norwood, Massachusetts.

Oran, E.S., Oh, C.K., and Cybyk, B.Z. (1998) "Direct Simulation Monte Carlo: Recent Advances and Applications," *Annu. Rev. Fluid Mech.* **30**, pp. 403–41.

Panton, R.L. (1996) *Incompressible Flow*, 2nd ed., Wiley-Interscience, New York.

Pearson, J.R.A., and Petrie, C.J.S. (1968) "On Melt Flow Instability of Extruded Polymers," in *Polymer Systems: Deformation and Flow*, R.E. Wetton and R.W. Whorlow, eds., pp. 163–187, Macmillian, London.

Pfahler, J. (1992) Liquid Transport in Micron and Submicron Size Channels, Ph.D. thesis, University of Pennsylvania.

Pfahler, J., Harley, J., Bau, H., and Zemel, J.N. (1990) "Liquid Transport in Micron and Submicron Channels," *Sensor. Actuator. A* **21–23**, pp. 431–34.

Pfahler, J., Harley, J., Bau, H., and Zemel, J.N. (1991) "Gas and Liquid Flow in Small Channels," in *Symp. on Micromechanical Sensors, Actuators, and Systems*, D. Cho et al., eds., ASME DSC-Vol. 32, pp. 49–60, ASME, New York.

Piekos, E.S., and Breuer, K.S. (1996) "Numerical Modeling of Micromechanical Devices Using the Direct Simulation Monte Carlo Method," *J. Fluids Eng.* **118**, pp. 464–69.

Pong, K.-C., Ho, C.-M., Liu, J., and Tai, Y.-C. (1994) "Non-Linear Pressure Distribution in Uniform Microchannels," in *Application of Microfabrication to Fluid Mechanics*, P.R. Bandyopadhyay, K.S. Breuer, and C.J. Belchinger, eds., ASME FED-Vol. 197, pp. 47–52, ASME, New York.

Porodnov, B.T., Suetin, P.E., Borisov, S.F., and Akinshin, V.D. (1974) "Experimental Investigation of Rarefied Gas Flow in Different Channels," *J. Fluid Mech.* **64**, pp. 417–37.

Prud'homme, R.K., Chapman, T.W., and Bowen, J.R. (1986) "Laminar Compressible Flow in a Tube," *Appl. Sci. Res.* **43**, pp. 67–74.

Richardson, S. (1973) "On the No-Slip Boundary Condition," *J. Fluid Mech.* **59**, pp. 707–19.

Schaaf, S.A., and Chambré, P.L. (1961) *Flow of Rarefied Gases*, Princeton University Press, Princeton, New Jersey.

Seidl, M., and Steinheil, E. (1974) "Measurement of Momentum Accommodation Coefficients on Surfaces Characterized by Auger Spectroscopy, SIMS and LEED," in *Rarefied Gas Dynamics*, vol. 9, M. Becker and M. Fiebig, eds., pp. E9.1–E9.2, DFVLR-Press, Porz-Wahn, Germany.

Sharp, K.V. (2001) Experimental Investigation of Liquid and Particle-Laden Flows in Microtubes, Ph.D. thesis, University of Illinois at Urbana.

Sharp, K.V., Adrian, R.J., Santiago, J.G., and Molho, J.I. (2001) "Liquid Flow in Microchannels," in *The Handbook of MEMS*, M. Gad-el-Hak, ed., CRC Press, Boca Raton, Florida.

Shih, J.C., Ho, C.-M., Liu, J., and Tai, Y.-C. (1995) "Non-Linear Pressure Distribution in Uniform Microchannels," ASME AMD-MD-Vol. 238, New York.

Shih, J.C., Ho, C.-M., Liu, J., and Tai, Y-.C. (1996) "Monatomic and Polyatomic Gas Flow through Uniform Microchannels," in *Applications of Microfabrication to Fluid Mechanics*, K. Breuer, P. Bandyopadhyay, and M. Gad-el-Hak, eds., ASME DSC-Vol. 59, pp. 197–203, New York.

Squires, T.M., and Quake, S.R. (2005) "Microfluidics: Fluid Physics at the Nanoliter Scale," *Rev. Mod. Phys.* **77**, pp. 977–1026.

Stone, H.A., Stroock, A.D., and Ajdari, A. (2004) "Engineering Flows in Small Devices: Microfluidics Toward a Lab-on-a-Chip," *Annu. Rev. Fluid Mech.* **36**, pp. 381–411.

Tai, Y.-C., and Muller, R.S. (1989) "IC-Processed Electrostatic Synchronous Micromotors," *Sensor. Actuator.* **20**, pp. 49–55.

Tang, W.C., Nguyen, T.-C., and Howe, R.T. (1989) "Laterally Driven Polysilicon Resonant Microstructures," *Sensor. Actuator.* **20**, pp. 25–32.

Thomas, L.B., and Lord, R.G. (1974) "Comparative Measurements of Tangential Momentum and Thermal Accommodations on Polished and on Roughened Steel Spheres," in *Rarefied Gas Dynamics*, vol. 8, K. Karamcheti, ed., Academic Press, New York.

Thompson, P.A., and Robbins, M.O. (1989) "Simulations of Contact Line Motion: Slip and the Dynamic Contact Angle," *Phys. Rev. Lett.* **63**, pp. 766–769.

Thompson, P.A., and Troian, S.M. (1997) "A General Boundary Condition for Liquid Flow at Solid Surfaces," *Nature* **389**, pp. 360–62.

Tison, S.A. (1993) "Experimental Data and Theoretical Modeling of Gas Flows through Metal Capillary Leaks," *Vacuum* **44**, pp. 1171–75.

Tuckermann, D.B. (1984) Heat Transfer Microstructures for Integrated Circuits, Ph.D. thesis, Stanford University.

Tuckermann, D.B., and Pease, R.F.W. (1981) "High-Performance Heat Sinking for VLSI," *IEEE Electron Device Lett.* **EDL-2**, no. 5, May.

Tuckermann, D.B., and Pease, R.F.W. (1982) "Optimized Convective Cooling Using Micromachined Structures," *J. Electrochem. Soc.* **129**, no. 3, C98, March.

Van den Berg, H.R., Seldam, C.A., and Gulik, P.S. (1993) "Compressible Laminar Flow in a Capillary," *J. Fluid Mech.* **246**, pp. 1–20.

Vargo, S.E., and Muntz, E.P. (1996) "A Simple Micromechanical Compressor and Vacuum Pump for Flow Control and Other Distributed Applications," AIAA Paper No. 96-0310, AIAA, Washington, D.C.

Vincenti, W.G., and Kruger, C.H., Jr. (1965) *Introduction to Physical Gas Dynamics*, Wiley, New York.

Von Smoluchowski, M. (1898) "Ueber Wärmeleitung in verdünnten Gasen," *Ann. Phys. Chem.* **64**, pp. 101–30.

Went, F.W. (1968) "The Size of Man," *Am. Sci.* **56**, pp. 400–413.

5

Integrated Simulation for MEMS: Coupling Flow-Structure-Thermal-Electrical Domains

Robert M. Kirby
University of Utah

George Em Karniadakis
Brown University

Oleg Mikulchenko and Kartikeya Mayaram
Oregon State University

5.1 Introduction

5.1.1 Full-System Simulation

Microelectromechanical systems (MEMS) involve complex functions governed by diverse transient physical and electrical processes for each of their many components. The design complexity and functionality complexity of MEMS exceeds by far the complexity of Very Large Scale Integration (VLSI) systems. Today, however, VLSI systems are simulated routinely, thanks to the many advances in computer assisted design (CAD) and simulation tools achieved over the last two decades. It is clear that similar and even greater advances are required in the MEMS field in order to make *full-system simulation* of MEMS a reality in the near future. This will enable the MEMS community to explore new pathways of discovery and expand the role and influence of MEMS at a rapid rate.

In order to develop such a systems-level simulation framework that is sufficiently accurate and robust, *all processes* involved need to be simulated at a comparable degree of accuracy and integrated seamlessly. That is, circuits, semiconductors, springs and masses, beams and membranes, as well as the flow field need to be simulated in a consistent and compatible way and in reasonable computational time. This coupling of diverse domains has already been addressed by the electrical engineering community, primarily for mixed-circuit-device simulation.

The combination of circuits and devices necessitates the use of different levels of description. At a first level for analog circuits represented by lumped continuum models, the use of ordinary differential equations (ODEs) and algebraic equations (AEs) is sufficient. However, some other devices and circuits can be described as digital automata, and thus boolean equations of mathematical logic should be employed in the description; these equations correspond to digital circuit simulation on the digital level. Finally, some semiconductor devices of the kind encountered in MEMS have to be described as linear and nonlinear partial differential equations (PDEs), and they are usually employed on the device-simulation level. Mixed-level simulation is implemented for digital-analog (or analog-mixed) circuit simulation and for analog-circuit-device simulation. In the following paragraphs, we briefly review the common practice in simulating circuits with some nonfluidic devices.

The code SPICE, which is an acronym for Simulation Program with Integrated Circuit Emphasis, was developed in the 1970s at UC Berkeley [Nagel and Pederson, 1973] and since then it has become the unofficial industrial standard by integrated circuit (IC) designers. SPICE is a general-purpose simulation program for circuits that may contain resistors, capacitors, inductors, switches, transmission lines, etc., as well as the five most common semiconductor devices: diodes, Bipolar Junction Transistor (BJTs), Junction Field Effect (JFETs), Metal Semiconductor Field Effect Transistor (MESFETs), and Metal Oxide Silicon Field Effect Transistor (MOSFETs). SPICE has built-in models for the semiconductor devices, and the user specifies only the pertinent model parameter values. However, these devices are typically simple and can be described by lumped models; that is, combinations of ordinary differential equations and algebraic equations (ODEs/AEs). In some cases, such as in submicron devices, even for usual semiconductor devices (i.e., MOSFET), simple modeling is not straightforward, and it is rather art than science to transfer from basic PDEs to approximated ODEs and algebraic equations. Mechanical systems are recast into electrical systems, which can be handled by SPICE. This can be understood more clearly by considering the analogy of a mass-spring-damper system driven by an external force with a parallel-connected RLC circuit with a current source. In this example, mass corresponds to capacitance, dampers to resistors, springs to inductive elements, and forces to currents.

Other devices cannot be represented by lumped models, and such an analogy may not be valid. While SPICE is essentially an ODE solver — that is, an analog circuit simulator only — another successful code, CODECS (acronym for Coupled Device and Circuit Simulator) provides a truly mixed-level description of both circuits and devices. This code too was developed at UC Berkeley [Mayaram and Pederson, 1987] and employs combinations of both ODEs and PDEs with algebraic equations. CODECS incorporates SPICE3, the latest version of SPICE written in C [Quarles, 1989], for the circuit simulation capability. The multirate dynamics introduced by combinations of devices and circuits is handled efficiently by a multilevel Newton method or a full-Newton method for transient analysis, unlike the standard Newton method employed in SPICE. CODECS is appropriate for one-dimensional and two-dimensional devices, but recent developments have produced efficient algorithms for three-dimensional devices as well [Mayaram et al., 1993].

The aforementioned simulation tools for IC design can be used for MEMS simulations, and in fact SPICE has been used to model electrostatic lateral resonators [Lo et al., 1996]. The assumption here is that all device components can be recast as equivalent analog circuit elements that SPICE recognizes. Clearly, this approach can be used in some well-studied structures, such as membranes or simple microbeams, but very rarely for microflows. However, in the last decade there has been an intense effort to produce such models and corresponding software, such as MEMCAD [Senturia et al., 1992], which has become a commercial package [Gilbert et al., 1993] for electrostatic and mechanical analysis of microstructures. Other such packages are the SOLIDIS and IntelliCAD (IntelliSense and ISE). In these simulation approaches, the flow field is not simulated, but its effect is typically expressed by the equivalent of a drag coefficient that provides damping. In some cases, as in the squeezed gas film in silicon accelerometers, an equivalent RLC circuit can also be obtained [Veijola et al., 1995]; however, this is the exception rather than the rule. Even the structural components are often modeled analytically, and significant effort has been devoted to constructing reduced-order macromodels [Hung et al., 1997; Gabbay, 1998]. These are typically nonlinear low-dimensional models obtained from projections of full three-dimensional simulations to a few representative modal shapes. Nonlinear function fitting is then employed so that analytical forms can be written, and these structural models are then imported to SPICE as analog circuit equivalent elements.

This reduced-order macromodeling approach has been used with success in a variety of applications including, for example, the electrostatic actuation of a suspended beam and elastically suspended plates [Gabbay, 1998]. Their great advantage is computational speed, but they are limited to small displacements and small deformations, mostly in the linear regime, and are appropriate for familiar designs only. Unfortunately, most of the MEMS devices are operating in nonlinear regimes including electrostatic actuators, flow fields, and structures. More importantly, the real impact and anticipated benefits of MEMS will come from new designs, yet unknown, that hopefully will be pretested using full simulations where all processes are simulated accurately without sacrificing important details of the physics. MEMS simulation based on full-physics models may be then more appropriate for exploring new concepts, whereas macromodeling may be employed efficiently for familiar designs and in known operating regimes.

In the following section, we address some of the specific issues encountered in each of the coupled domains, (i.e., fluid, electric, structure, thermal), and we analyze their corresponding computational complexity and proposed algorithms for their integration.

5.1.2 Computational Complexity of MEMS Flows

Liquid and gas flows in microdevices are characterized by low Reynolds number, typically of order one or less in channels with heights in the submillimeter range [Ho and Tai, 1998; Gad-el-Hak, 1999]. They are unsteady due to external excitation from a moving boundary or an electric field, often driven by high-frequency (e.g., 50 kHz) oscillators, as in the example of the MIT electrostatic comb-drive [Freeman et al., 1998]. The domain of microflows is three-dimensional and geometrically complex, consisting of large-aspect ratio components, abrupt expansions, and rough boundaries. In addition, microdevices interact with larger devices resulting in fluid flow going through disparate regimes.

Accurate and efficient simulation of microflows should take into account the above factors. For example, the significant geometric complexity of MEMS flows suggests that finite elements and boundary elements are more suitable than finite differences for efficient discretization [Ye, Kanapka, and White, 1999]. However, because of the nonlinear effects, either through the convection or boundary conditions, boundary element methods are also limited in their application range despite their efficiency for linear flows [Aluru and White, 1996]. A particularly promising approach developed recently for MEMS flows makes use of meshless and mesh-free approaches [Aluru, 1999], where particles are “sprinkled” almost randomly into the flow and boundary. This approach effectively handles the geometric complexity of MEMS flows, but the issues of accuracy and efficiency have not been fully resolved yet. As regards nonlinearities, one may argue that at such low Reynolds numbers the convection effects should be neglected, but in complex geometries with abrupt turns, the convective acceleration terms may be substantial, and thus they need to be taken into account.

The computational difficulties for liquid and gas flows are of a different type. Gas microflows are compressible and can experience large density variations. In addition, for channels of a size below 10 microns or at subatmospheric conditions, serious *rarefaction effects* may be present, (see [Beskok, Karniadakis, and Trimmer 1996] and also the chapter by A. Beskok in this volume). In this case, either modified Navier–Stokes equations with appropriate *slip boundary conditions* or higher-order approximations are necessary to describe the governing flow dynamics. To this end, a nondimensional number, the *Knudsen number* defined as the ratio of the mean-free-path to the characteristic domain size, defines which model and correspondingly which numerical method is appropriate for simulating gas microflows [Bird, 1994]. For submicron devices, atomistic or molecular simulations are necessary as the familiar concept of continuum description breaks down. The direct simulation Monte Carlo (DSMC) method, described in the article by Beskok in this volume, is one efficient method of simulating highly rarefied flows.

On the other hand, liquid flows in microscales are "granular"; that is, they form a layering structure very close to the wall over a distance of a few molecule diameters [Koplik and Banavar, 1995]. This is accompanied by large density fluctuations very close to the wall leading to anomalous heat and momentum transport. Liquid flows, in particular, are very sensitive to the wall type, and although such an issue may not be important for averaged heat and momentum transport rates in flow domains of 100 microns or greater, it is extremely important in smaller domains. This distinction suggests two possible approaches in simulating liquid flows in microscales: a phenomenological approach using the Navier–Stokes similar to macrodomain flows, and a molecular approach based on the molecular dynamics (MD) approach [Koplik and Banavar, 1995; Allen and Tildesley, 1994]. The MD approach is deterministic following the trajectories of all molecules involved, unlike the DSMC approach, which is stochastic representing collisions as a random process. The drawback of the Navier–Stokes approach is that events at the molecular level are modeled via continuum-like parameters. For example, consider the problem of routing microdroplets on a silicon surface, effectively altering dynamically the contact line of the microdrop. This is a molecular level process, but in the continuum approach it is determined via a macro-domain-type formulation (e.g., via gradients), which may lead to erroneous results. Accurate MD modeling of the contact line will be truly predictive as it will take into account the wall–fluid interaction at the molecular level. The wall type and the specific fluid type are taken into account by different potentials that describe intermolecular structure and force. However, such a detailed simulation requires an enormous number of molecules (e.g., hundreds of millions of molecules), and thus it is limited to a very small region, probably around the contact line region only. It is therefore important to develop new hybrid approaches that combine the best features of both methods [Hadjiconstantinou, 1999].

In summary, geometry and surface effects, compressibility and rarefaction, unsteadiness and unfamiliar physics make simulation of microflows a challenging task. The true challenge, however, comes from the interaction of the fluidic system with other system components, such as the structure, the electric field, and the thermal field. In the following sections, we discuss this interaction.

5.1.3 Coupled-Domain Problems

In coupled-domain problems, such as flow-structure, structure-electric, or a combination of both, there are significant disparities in temporal and spatial scales. This, in turn, implies that multiple grids and heterogeneous time-stepping algorithms may be needed for discretization, leading to very complicated and consequently computational prohibitive simulation algorithms. Simplifications are typically made with one of the fields represented at a reduced resolution level or by low-dimensional systems or even by equivalent lumped dynamical models. For example, consider the electric activation of a cantilever microbeam made of piezoelectric material. The emphasis may be on modeling the electronic circuit and the motion, and thus a simple model for the motion-induced hydrodynamic damping may be constructed avoiding full simulation of the flow around the beam.

A possible method of constructing low-order dynamical models is by projecting the results of detailed numerical simulations onto spaces spanned by a very small number of degrees of freedom — the so-called *nonlinear macromodeling* approach (see [Gabbay, 1998] and [Senturia, Aluru, and White, 1997]). To clarify the concept of a macromodel, we give a specific example (see [Senturia, Aluru, and White, 1997]) for a suspended membrane of thickness *t* deflected at its center by an amplitude *d* under the action of uniform pressure force *P*. Let us also denote by 2*a* the length of the membrane, by *E* the Young's module, by ν the Poisson ratio, and by σ the residual stress. One can use analytical methods to obtain the resulting form of the pressure-deflection relation (e.g., power series assuming a circular thin membrane). This can be extended to more general shapes and nonlinear responses, for example:

$$P = \frac{C_1 t}{a^2} + \frac{C_2 f(\nu)}{a^4} \frac{E}{1 - \nu} d^3 \tag{5.1}$$

where C_1 and C_2 are dimensionless constants that depend on the shape of the membrane, and $f(\nu)$ is a slowly varying function of the Poisson ratio. This function is determined from detailed finite element simulations over a range of length a, thickness t, and material properties ν and E. Such "best-fits" are tabulated and are used in the simulation according to the specific structure considered without the need for solving the partial differential equations governing the dynamics of the structure. They can also be built automatically as has been demonstrated in [Gabbay, 1998]. Another type of a macromodel based on neural networks training will be presented later for a flow sensor.

Unfortunately, construction of such macromodels is not always possible, and this lack of simplified models for the many and diverse components of microsystems makes system-level simulation a challenging task. On the other hand, model development for electronic components (transistors, resistors, capacitors, etc.) has reached a state of maturity. Therefore, considerable attention should be focused on models for the nonelectronic components. This is necessary for the design and verification of complete microsystems. In the remainder of this chapter, we describe an integrated approach for simulation of microsystems with the emphasis being on microfluidic systems. To this end, we resort to full simulation of the fluidic system, which involves also interactions with moving structures. To illustrate the formulation more clearly, we present next a target simulation problem that represents the aforementioned challenges.

5.1.4 A Prototype Problem

An example of a microfluidic system is a microliquid dosing system shown schematically in Figure 5.1. This system is made up of a micropump, a microflow sensor, and an electronic control circuit. The electronic circuit adjusts the pump flow rate so that a constant flow is maintained in the microchannel. A realization of this system is shown in Figure 5.2, along with the details of the control circuit. The simulation of the complete system requires models for the micropump, the microflow sensor, and the electronic components shown in Figure 5.2. When low-order full-physics models are available for all components including the fluid flow, the complete system can be simulated using a standard circuit simulator such as SPICE [Nagel, 1975; Quarles, 1989].

In the absence of macromodels for the micropump and the microflow sensor, the typical approach for microsystem simulation makes use of lumped-element equivalent circuit descriptions for these devices [Tilmans, 1996]. However, such an approach has two main limitations:

- It is suitable only for open-loop systems, where there is no feedback from the output to the input
- It is applicable only for small-signal conditions

These two limitations arise in the model development process where several assumptions are made in order to construct the lumped-element equivalent circuits. Therefore, this approach would not be suitable when the large-signal behavior of a closed-loop system is of interest.

To address the above problem, we present a coupled circuit/microfluidic device simulator that efficiently couples the discretized Navier–Stokes equations describing a microfluidic device (numerical model) to the solution of circuit equations. Such a capability is unique in that it allows direct and efficient simulation of microfluidic systems without the need for mapping finite element descriptions into

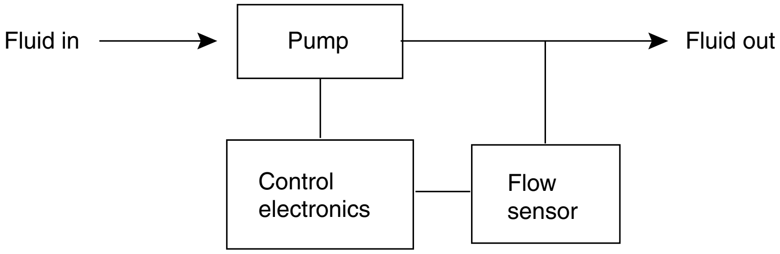


FIGURE 5.1 Block diagram of a generic microfluidic system. The flow sensor senses the flow rate, which is controlled by the electronic circuit controlling the pump.

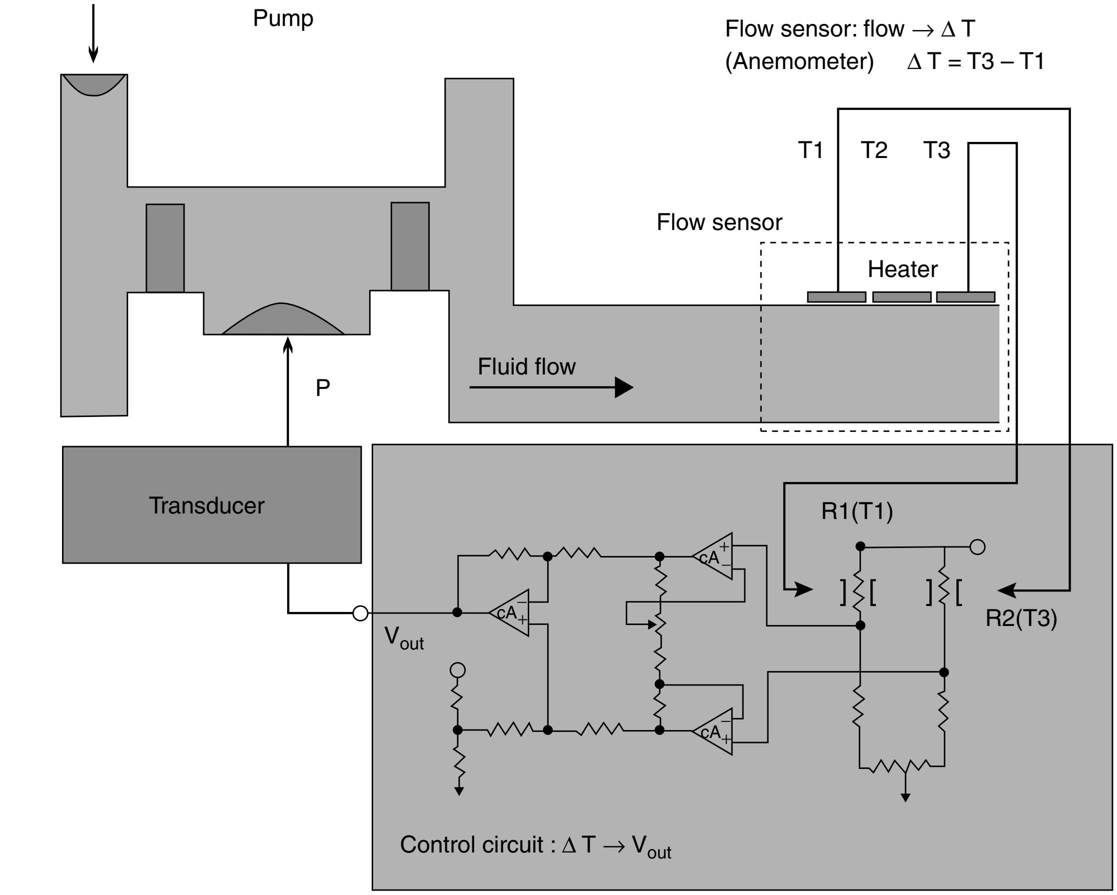


FIGURE 5.2 Realization of the microfluidic system showing the electronic control circuit. The fluid flow determines the temperature ΔT of the flow sensor. This temperature is transformed by the control electronics into the voltage *Vout,* which in turn controls the pump pressure P by a transformation of the voltage to a proportional pressure.

equivalent networks [Tilmans, 1996] or analog hardware description languages (AHDLs) [Bielefeld, Pelz, and Zimmer, 1997].

The rest of this chapter is organized as follows: an overview of coupled circuit and device simulation is given in section 2, followed by a description of the circuit and fluidic simulators in section 3. The details of the coupled circuit/fluidic simulator are presented in section 4, and an illustrative example is described in section 5. Conclusions are provided in section 6.

5.2 Coupled Circuit-Device Simulation

Coupled simulation techniques have previously been used for the simulation of a sensor system [Schroth, Blochwitz, and Gerlach, 1995]. In this approach, the finite-element program ANSYS [Moaveni, 1999] is coupled to an electrical simulator PSPICE [Keown, 1997]. Although such an approach has been demonstrated to work for system simulations, the coupling is not efficient. Special coupling algorithms and time-stepping schemes are required to enable fast simulation of microsystems. Therefore, a tight coupling between the circuit and device simulators is necessary for simulation efficiency [Mayaram and Pederson, 1992; Mayaram, Chern, and Yang, 1993].

The coupled circuit-device simulator allows verification of microfluidic systems. It provides accurate large- and small-signal simulation of systems even in the absence of proper macromodels for the microfluidic devices. On the other hand, the coupled simulator is important for constructing and validating

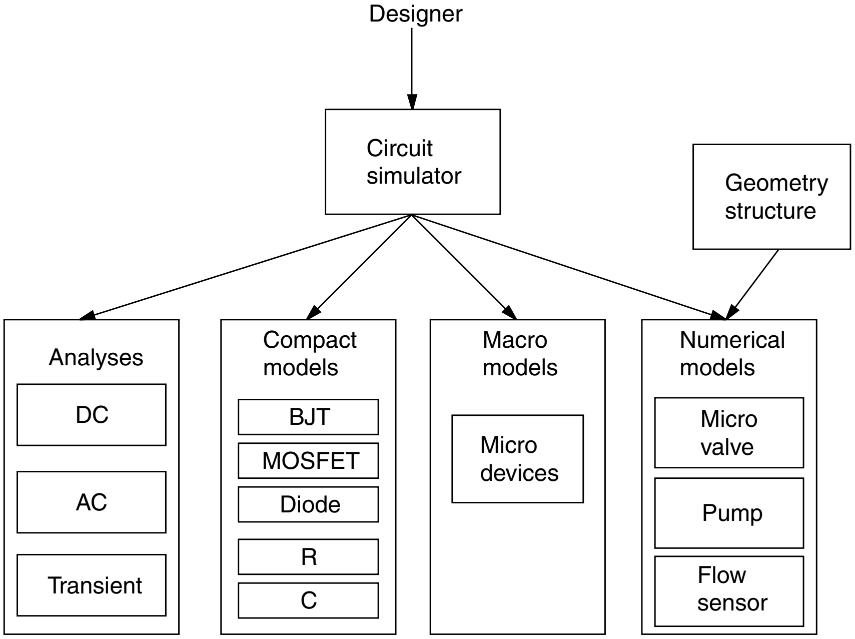


FIGURE 5.3 The coupled circuit-fluidic device simulator. Microfluidic systems including the control electronics can be simulated using accurate numerical models for all components.

macromodels. As important effects (such as highly nonlinear or distributed behavior, compressibility, or slip-flow) are identified, they can be implemented in the macromodels and verified for system simulation using the coupled simulator. Furthermore, critical devices can be simulated using the full physics-based numerical models when there are stringent accuracy requirements on the simulated results.

The concept of a coupled circuit and device simulator has proved to be extremely beneficial in the domain of integrated circuits. Since the first of such simulators, MEDUSA [Engl, Laur, and Dirks, 1982], became available in the early 1980s, there has been significant work addressing coupled simulation. These activities have focused on improved algorithms, faster execution speeds, and applications. Commercial Technology Computer Aided Design (TCAD) vendors also support a mixed circuit-device simulation capability [Technology Modeling Associates, 1997; Silvaco International, 1995]. Since the computational costs of these simulators are high, they are not used on a routine basis. However, there are several critical applications in which these simulators are extremely valuable. These include simulation of Radio Frequency (RF) circuits [Rotella et al., 1997], single-event-upset simulation of memories [Woodruff and Rudeck, 1993], simulation of power devices [Ravanelli and Hu, 1991], and validation of nonquasistatic MOSFET models [Park, Ko, and Hu, 1991].

The coupled circuit-device simulator for microfluidic applications is illustrated in Figure 5.3. This simulator supports compact models for the electronic components and available macromodels for microfluidic devices. In addition, numerical models are available for the microfluidic components that can be utilized when detailed and accurate modeling is required. As an example, specific components such as microvalves, micropumps, and micro-flow-sensors are shown in Figure 5.3. The coupling of the circuit and microfluidic components is handled by imposing suitable boundary conditions on the fluid solver. This simulator allows the simulation of a complete microfluidic system including the associated control electronics. The details of the various simulators and coupling methods are described in the sections below.

One of the biggest disadvantages of such an approach is the high computational cost involved. The main cost comes from solving the three-dimensional time-dependent Navier–Stokes equations in complex geometric domains. Thus, efficient flow solvers are critical to the success of a coupled circuit-micro-fluidic device simulator. Any performance improvements in the solution of the Navier–Stokes equations directly translate into a significant performance gain for the coupled simulator.

5.3 Overview of Simulators

The circuit simulator employed here is based on the circuit simulator SPICE3f5 [Quarles, 1989] and the microfluidic simulator on the code *NεκTαr* [Karniadakis and Sherwin, 1999; Kirby et al., 1999]. A brief description of the algorithms and software structure of each of these simulators is provided in this section.

5.3.1 The Circuit Simulator: SPICE3

Electrical circuits consist of many components (resistors, capacitors, inductors, transistors, diodes, and independent sources) that are described by algebraic and/or differential relations among the components' currents and voltages. These relationships are called the *branch constitutive relations* [Sangiovanni-Vincentelli, 1981]. The circuits also satisfy conservation laws known as the Kirchhoff's laws; these laws result in algebraic equations. Therefore, a circuit is described by a set of coupled nonlinear differential algebraic equations that are both highly nonlinear and stiff, and this imposes certain limitations on the solution methods. One of the most commonly used analyses is the *time-domain transient analysis*. We briefly describe below the solution approach used for this analysis.

Time discretization: At each time-step of the transient analysis, the time derivatives are replaced by an algebraic equation using an integration method. Typically, an implicit linear multistep method of the backward-differentiation type suitable for stiff ODEs is used [Sangiovanni-Vincentelli, 1981]:

$$v \approx \alpha_0 v_{t_n} + \sum_{k=1}^{n} \alpha_k v_{t_{n-k}} \tag{5.2}$$

Linearization: Time discretization yields a system of nonlinear algebraic equations, which are typically solved by a Newton–Raphson method. The nonlinear components are replaced by linear equivalent models for each iteration of the Newton's method

$$f(v_{t_n}^{j+1} \approx f(v_{t_n}^{j}) + \partial f(v)/\partial v|_{v_{t_n}^{j}} \cdot (v_{t_n}^{j+1} - v_{t_n}^{j}) \tag{5.3}$$

Equation solution: After time discretization and application of Newton's method a linear system of equations is obtained at each iteration of the Newton method. These equations are described by

$$\mathbf{A}\mathbf{v}^{j+1} = \mathbf{b} \tag{5.4}$$

where $\mathbf{A} \in \Re^{n \times n}$, $\mathbf{v}^{j+1} \in \Re^{n}$, $\mathbf{b} \in \Re^{n}$, and can be solved by sparse matrix techniques [Kundert, 1990].

The time-domain simulation algorithm can be summarized in the following steps [Sangiovanni-Vincentelli, 1981]:

1. Read circuit description and initialize data structures.
2. Increment time $t_n = t_{n-1} + h$.
3. Update values of independent sources at t_n.
4. Predict values of unknown variables at t_n.
5. Apply integration formula (1) to capacitors and inductors.
6. Apply linearization (2) to nonlinear circuit elements.
7. Assemble linear circuit equations.
8. Solve linear circuit equations.
9. Check convergence. If not converged go to step 6.
10. Estimate local truncation error.
11. Select new time step h; rollback time if truncation error is unacceptable.
12. If $t_n < t_{stop}$ go to step 3.

5.3.2 The Fluid Simulator: *NεκTαr*

The flow solver corresponds to a particular version of the code *NεκTαr*, which is a general purpose Computational Fluid Dynamics (CFD) code for simulating incompressible, compressible, and plasma

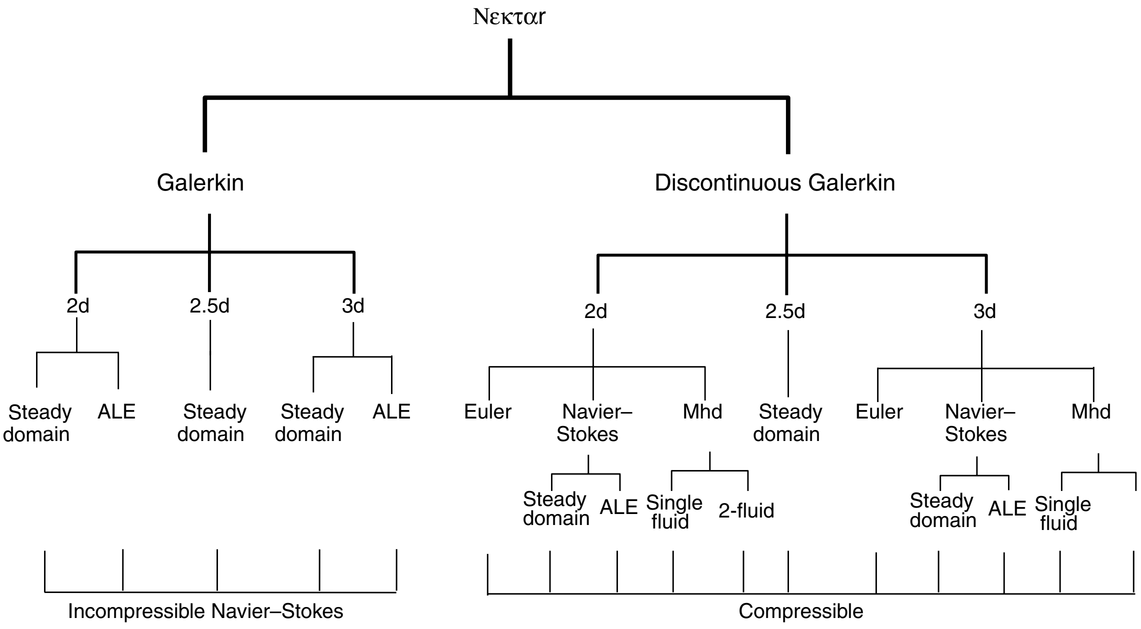


FIGURE 5.4 Hierarchy of the *NεκTαr* code. Note that "2.5*d*" refers to a three-dimensional capability with one of the directions being homogeneous in the geometry. Also, ALE refers to moving computational domains required in dynamic flow–structure interactions. Gaseous microflows can be simulated by either the compressible or incompressible version depending on the pressure/density variations.

flows in unsteady three-dimensional geometries. The major algorithmic developments are described in [Sherwin, 1995] and [Warburton, 1999], and the capabilities are summarized in Figure 5.4. The code uses meshes similar to standard finite-element and finite-volume meshes consisting of structured or unstructured grids or a combination of both. The formulation is also similar to those methods, corresponding to Galerkin and discontinuous Galerkin projections for the incompressible and compressible Navier–Stokes equations, respectively. Field variables, data, and geometry are represented in terms of hierarchical (Jacobi) polynomial expansions [Karniadakis and Sherwin, 1999]; both isoparametric and superparametric representations are employed. These expansions are ordered in vertex, edge, face, and interior (or bubble) modes. For the Galerkin formulation, the required C^0 continuity across elements is imposed by choosing appropriately the edge (and face in 3D) modes; at low-order expansions this formulation reduces to the standard finite element formulation. The discontinuous Galerkin is a flux-based formulation, and all field variables have L^2 continuity; at low order this formulation reduces to the standard finite-volume formulation.

This new generation of Galerkin and discontinuous Galerkin spectral/hp element methods implemented in the code *NεκTαr* does not replace but rather extends the classical finite element and finite volumes that the CFD practitioners are familiar with [Karniadakis and Sherwin, 1999]. The additional advantages are that convergence of the discretization and thus solution verification can be obtained without remeshing (h-refinement) and that the quality of the solution does not depend on the quality of the original discretization. In Figure 5.4 we summarize the major current capabilities of the general code *NεκTαr* for incompressible, compressible, and even plasma flows. In particular, for microflows both the compressible and incompressible versions are used. For gas microflows we account for rarefaction by using velocity-slip and temperature-jump boundary conditions as described in this volume in the chapter by Beskok (see also [Beskok, Karniadakis, and Trimmer, 1996; Beskok and Karniadakis, 1999]). An extension of the classical Maxwell's boundary condition is employed in the code in the form

$$U_g - U_w = \frac{Kn}{1 - bKn}(\nabla U)_w \cdot \hat{n} \qquad (5.5)$$

Here we define the Knudsen number $Kn = \lambda/L$ with λ the mean free path of the gas molecules and L the characteristic length scale in the flow. Also, U_g is the velocity (tangential component) of the gas at the wall, U_w is the wall velocity, and n is the unit normal vector. The constant b is adjusted to reflect the physics of the problem as we go from the slightly rarefied regime (*slip flow*) to the transition regime ($Kn \approx 1$) or free molecular regime ($Kn > 5$–10). For $b = 0$, we recover the classical linear relationship between velocity-slip and shear stress first proposed by Maxwell. However, for $b = -1$ we obtain a second-order accuracy [Beskok and Karniadakis, 1999], and in general for $b \neq 0$ Equation (5.5) leads to finite slip at the wall unlike the linear boundary condition (for $b = 0$) used in most codes. The boundary condition in Equation (5.5) has been used with success in the entire Knudsen number regime, $Kn \approx 0$–200, [see several examples in Beskok and Karniadakis (1999)].

One of the key points in obtaining *efficiency* in simulations of moving domains is the type of discretization employed in the flow solver. In *NεκTαr* we employ the so-called h-p version of the finite-element method with spectral Jacobi polynomials as basis functions. Convergence is obtained via a dual path in this approach, either by increasing the number of elements (h-refinement) or by increasing the order of the spectral polynomial (p-refinement). In the latter case a faster convergence is obtained without the need for remeshing. Instead, the number of degrees of freedom is increased in the *modal space* by increasing the polynomial order (p) while keeping the mesh unchanged. It is, of course, the cost of reconstructing the mesh that is orders of magnitude higher in time-dependent simulations both in terms of computer and human time.

Regarding the type of elements (subdomains), *NεκTαr* uses hybrid meshes (i.e., both structured and unstructured meshes). For example, in three-dimensional simulations a hybrid grid may consist of tetrahedra, hexahedra, triangular prisms, and even pyramids. In Figure 5.5 we plot the mesh used in the simulation of the pump, and in Figure 5.6 we plot the flow field at three different time instances.

In the following section, we briefly describe how we formulate the algorithm for a compatible and efficient flow–structure coupling.

5.3.2.1 Formulation for Flow–Structure Interactions

We consider the incompressible Navier–Stokes equations in a time-dependent domain $\Omega(t)$

$$u_{i,t} + u_j u_{i,j} = -(p\delta_{ij})_j + \nu u_{i,jj} + f_i \text{ in } \Omega(t) \tag{5.6}$$

$$u_{j,j} = 0 \text{ in } \Omega(t), \tag{5.7}$$

where ν is the viscosity and J_i is a body force. We assume for clarity homogeneous boundary conditions; velocity-slip boundary conditions can be included relatively easily in the Galerkin framework as mixed

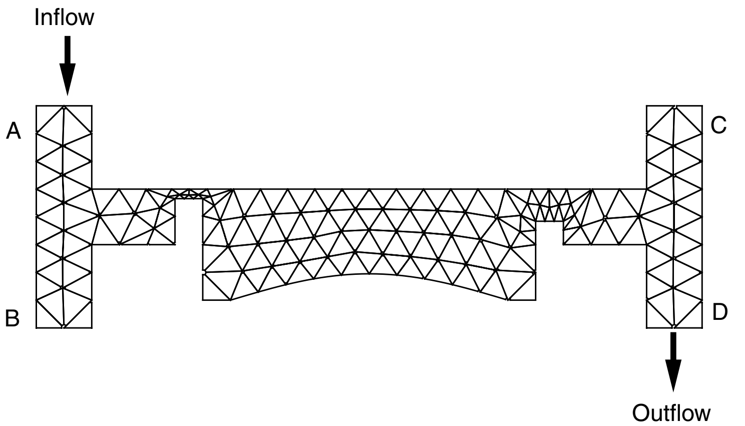


FIGURE 5.5 Mesh of the pump used in the flow simulator *NεκTαr*. This device was first introduced by [Beskok and Warburton, 1998] as a mixing device between two microchannels. Here B and C are blocked so the device is operating as a pump from A to D.

(Robin) boundary conditions. Multiplying Equation (5.6) by test functions and integrating by parts we obtain

$$\int_{\Omega(t)} v_i(u_{i,t} + u_j u_{i,j})dx = \int_{\Omega(t)} v_{i,j}(p\delta_{ij} - \nu u_{i,j} + v_i f_i)dx \tag{5.8}$$

The next step is to define the reference system on which time differentiation takes place. This was accomplished in [Ho, 1989] by use of the Reynolds transport theorem and by using the fact that the test function v_i is following the material points. Therefore, its time-derivative in that reference frame is zero,

$$\frac{dv_i}{dt} \mid x_p = v_{i,t} + w_j v_{i,j} = 0,$$

where w_j is a velocity that describes the motion of the time-dependent domain $\Omega(t)$; x_p denotes the material point. The final variational statement then becomes

$$\frac{d}{dt}\int_{\Omega(t)} v_i u_i\, dx + \int_{\Omega(t)} [v_i(u_j - w_j)u_{i,j} - v_i u_i w_{j,j}]dx = \int_{\Omega(t)} [v_{i,j} p\delta_{ij} - \nu\, v_{i,j} u_{i,u} + v_i f_i]dx \tag{5.9}$$

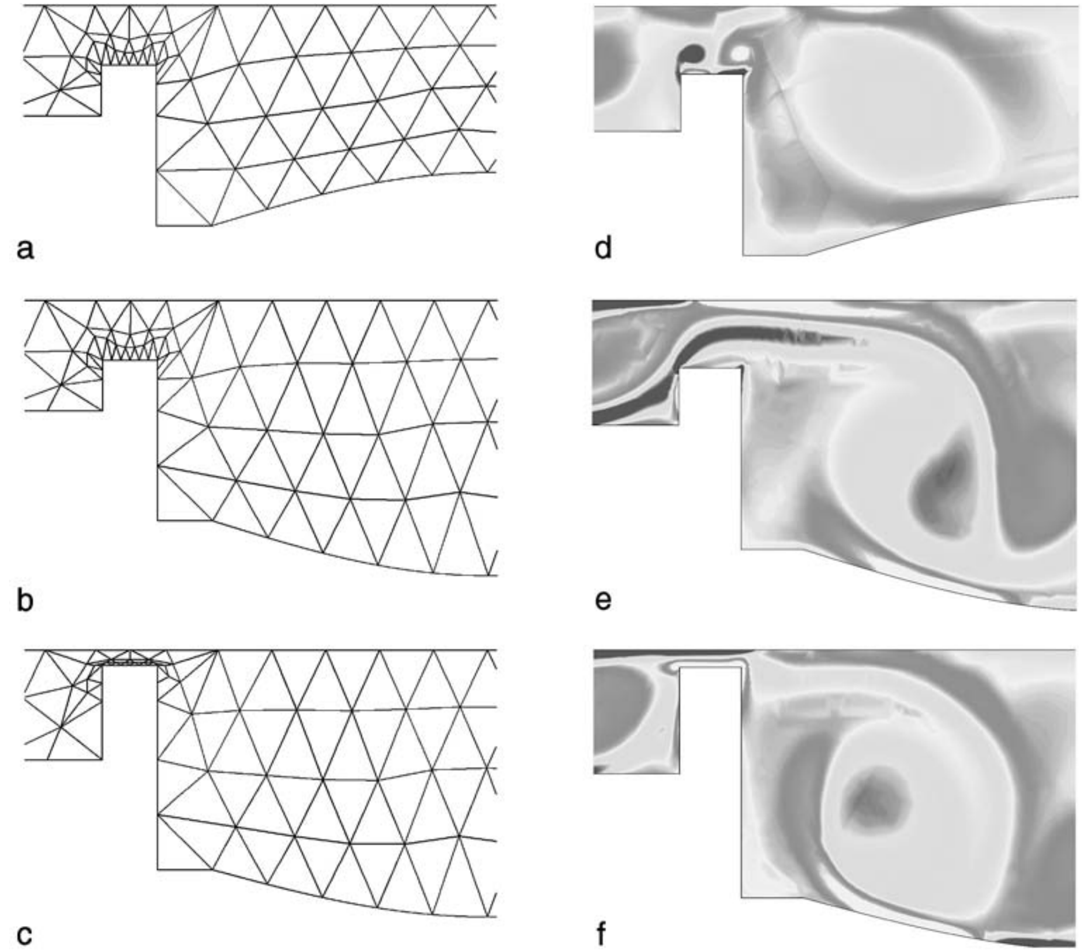


FIGURE 5.6 Close-up of the vorticity contours for Re = 30 simulation at the left valve (meshes shown on right side). Top: $\tau\omega = 0.28$ corresponds to the beginning of the suction stage. Start-up vortices due to the motion of the inlet valve can be identified. Middle: $\tau\omega = 0.72$, corresponding to the end of the suction stage. A vortex jet pair is visible in the pump cavity. Bottom: $\tau\omega = 0.84$, corresponding to early ejection stage. Further evolution of the vortex jet and the start-up vortex of the exit valve can be identified. (Reprinted with permission from A. Beskok).

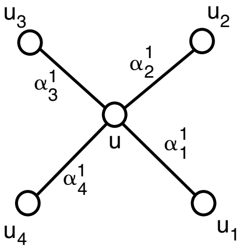


FIGURE 5.7 Graph showing vertices with associated velocities and edges with associated weights.

This is the ALE formulation of the momentum equation. It reduces to the familiar Eulerian and Lagrangian form by setting $w_j = 0$ and $w_j = u_j$ respectively. However, w_j can be chosen arbitrarily to minimize the mesh deformation. We discuss this algorithm next.

5.3.2.2 Grid Velocity Algorithm

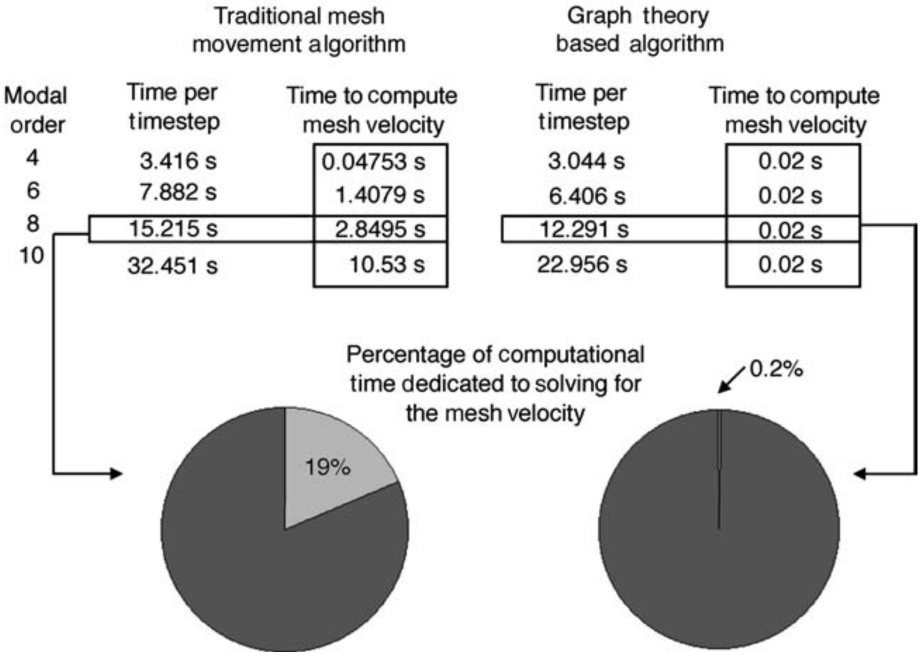
The grid velocity is arbitrary in the ALE formulation, and therefore great latitude exists in the choice of technique for updating it. Mesh constraints such as smoothness, consistency, and lack of edge crossover, combined with computational constraints such as memory use and efficiency dictate the update algorithm used. In the current work, we address the problem of solving for the mesh velocity in terms of its graph theory equivalent problem. Mesh positions are obtained using methods based on a graph theory analogy to the spring problem. Vertices are treated as *nodes*, while edges are treated as *springs* of varying length and tension. At each time step, the mesh coordinate positions are updated by equilibration of the spring network. Once the new vertex positions are calculated, the mesh velocity is obtained through differences between the original and equilibrated mesh vertex positions.

Specifically, we incorporate the idea of variable diffusivity while maintaining computational efficiency by avoiding solving full Laplacian equations. The method we use for updating the mesh velocity is a variation of the barycenter method [Battista, Eades, Tamassia, and Tollis, 1998] and relies on graph theory. Given the graph $G = (V,E)$ of element vertices V and connecting edges E, we define a partition $V = V_0 \cup V_1 \cup V_2$ of V such that V_0 contains all vertices affixed to the moving boundary, V_1 contains all vertices on the outer boundary of the computational domain, and V_2 contains all remaining interior vertices. To create the effect of variable diffusivity, we use the *concept of layers*. As is pointed in [Lohner and Yang, 1996], it is desirable for the vertices very close to the moving boundary to have a grid velocity almost equivalent to that of the boundary. Hence, locally the mesh appears to move with solid movement, whereas far away from the moving boundary the velocity must gradually go to zero. To accomplish this in our formulation, we use the concept of *local tension* within layers to allow us to prescribe the rigidity of our system. Each vertex is assigned to a layer value that heuristically denotes its distance from the moving boundary. Weights are chosen such that vertices closer to the moving boundary have a higher influence on the updated velocity value. To find the updated grid velocity u^g at a vertex $v \in V_2$, we use a force-directed method. Given a configuration as in Figure 5.7, the grid velocity at the center vertex is given by:

$$u^g = \sum_{i=1}^{\deg(v)} \alpha_i^l u_i, \sum_{i=1}^{\deg(v)} \alpha_i^l = 1,$$

where $\deg(v)$ is the number of edges meeting at the vertex v and α_i^l is the lth layer weight associated with the i-th edge. This is subjected to the following constraints: $u^g = 0(\forall v \in V_1)$, and $u^g(\forall v \in V_0)$ is prescribed to be the wall velocity. This procedure is repeated for a few cycles following an incomplete iteration algorithm, over all $v \in V_2$. (Here by incomplete we mean that only a few sweeps are performed and not full convergence is sought.) Once the grid velocity is known at every vertex, the updated vertex positions are determined using explicit time-integration of the newly found grid velocities.

An example of the relative speed-up gained following the graph-theory approach versus the classical approach of employing Poisson solvers to update the grid velocity is shown in Figure 5.8. We have computed the portion of CPU time devoted exclusively to the solver as a function of the spectral order



	Traditional mesh movement algorithm		Graph theory based algorithm	
Modal order	Time per timestep	Time to compute mesh velocity	Time per timestep	Time to compute mesh velocity
4	3.416 s	0.04753 s	3.044 s	0.02 s
6	7.882 s	1.4079 s	6.406 s	0.02 s
8	15.215 s	2.8495 s	12.291 s	0.02 s
10	32.451 s	10.53 s	22.956 s	0.02 s

FIGURE 5.8 Comparison of CPU time for the grid velocity algorithm between the classical approach (Poisson solver) and the new approach (graph algorithm). In the leftmost column is the order of spectral polynomial approximation.

employed in the discretization. The problem we considered involved the motion of a piezoelectric membrane induced by vortex shedding caused by a bluff body in front of the membrane. We see that a two- to three-orders of magnitude speed-up can be obtained using the graph-based algorithm.

5.3.3 The Structural Simulator

The membrane of the micropump is modeled using the linear string-beam equation as given by the following equation:

$$\frac{d^2y}{dt^2} + \frac{R}{m}\frac{dy}{dt} + \frac{EI}{m}\frac{d^4y}{dx^4} - \frac{T}{m}\frac{d^2y}{dx^2} = \frac{F}{m} \tag{5.10}$$

where E is the Young's modulus of elasticity, I is the second moment of inertia, T is the axial tension, F is the hydrodynamic forcing, R is the coefficient of structural damping, and m is the structural mass per unit length. In this model, the coefficients are given by the physical parameters of the membrane used within the pump, and the hydrodynamic forcing on the membrane is provided by $N\varepsilon\kappa T\alpha r$.

Assume that the membrane lies in the interval $[0,L]$. For the micropump configuration, we have chosen the boundary conditions $y(0) = y(L) = 0$, $y''(0) = y''(L) = 0$, which correspond to a fixed-hinged membrane. Equation (5.10) combined with these boundary conditions lends itself to the use of eigenfunction decomposition for the efficient solution of the membrane motion. We begin by transforming the problem to lie on the interval $[0,1]$ using the linear mapping $x = L\xi$, $\xi \in [0,1]$. The eigenfunctions of this system are given by

$$\phi_n = \frac{1}{2}\sin\sqrt{\lambda_n}\,\xi;\ \sqrt{\lambda_n} = (n-1)\pi n = 1, 2, \ldots, \infty$$

If we assume a solution of the form

$$y(\xi,t) = \sum_{n=1}^{N} A_n(t)\phi_n(\xi),$$

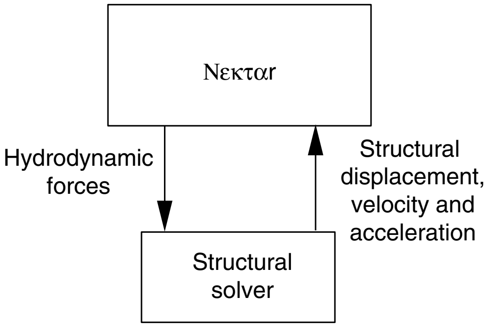


FIGURE 5.9 Coupling between *NεκTαr* and the structural solver. *NεκTαr* provides the hydrodynamic force information on the membrane. With this information the structural solver calculates the membrane's response. Structural displacement, velocity and acceleration are then returned to *NεκTαr* for determining the influence of the structure's motion on the fluid.

then by employing the Galerkin method we obtain the following evolution equation for the coefficients $A_n(t)$:

$$\frac{d^2A_n}{dt^2} + \frac{R}{m}\frac{dA_n}{dt} + \left(\frac{EI}{mL^4}\lambda_n - \frac{T}{m^2}\right)\lambda_n A_n = \frac{1}{m}\int_0^1 F d\xi \tag{5.11}$$

We then solve this evolution equation using the Newmark scheme [Hughes, 1987], which returns the coefficients for the displacement, velocity, and acceleration of the membrane. This information is then returned to *NεκTαr* as demonstrated in Figure 5.9.

5.3.4 Differences among Circuit, Fluid, and Solid Simulators

The above descriptions suggest some differences between the various simulators. The key distinguishing features are:

- The fluid simulator is computationally more expensive than the structure and circuit simulators.
- SPICE3 has a reliable error estimation for time discretization. Therefore, a rollback in time can be done if the truncation error is unacceptable. As a result, SPICE3 automatically controls the simulation time step to ensure an acceptable user-specified error. *NεκTαr* is a much more complex code and does not have an automatic time-step control scheme for coupled fluid–structure simulation.
- SPICE3 uses implicit numerical integration methods for time-domain simulation. These methods are efficient for circuit simulation because the circuit equations are stiff. For the fluid solver, however, explicit methods are simpler to implement and reasonably efficient. For this reason, *NεκTαr* uses semiimplicit methods for the time domain integration (explicit for the advection terms and implicit for the diffusion terms of the Navier–Stokes equations), which suffer from the standard CFL (Courant–Friedrichs–Levy condition for the time step) restrictions. However, the flow time step is much higher than the electronics time step due to the relevant physical time scales. Also, the Newmark scheme for the structure is unconditionally stable.

5.4 Circuit-Micro-Fluidic Device Simulation

For coupled circuit-micro-fluidic device simulation, four different physical domains (electrical, structure mechanical, fluid mechanical, and thermal) must be considered, as shown in Figure 5.10. These domains are coupled to one another as described below.

In Figure 5.2 four types of coupling can be identified. These are

- Electromechanical coupling for a piezoelectric actuation of the pump membrane
- Fluid–structure coupling due to volume displacement of the pump membrane

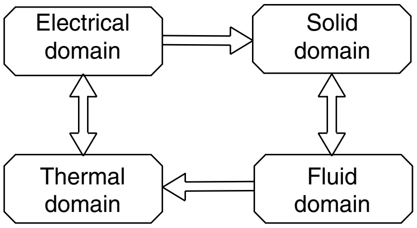


FIGURE 5.10 Coupling between the various physical domains.

- Fluid-thermal coupling because of the thermoresistor cooling in the fluid when an anemometer type of microflow sensor is used
- Electrothermal thermoresistor heating due to current flow in the microflow sensor

The overall system can be simulated using different approaches. One approach is a detailed physical simulation for each coupled domain. Another is the use of lumped-element equivalent circuits, compact, or macromodels, and/or analog hardware description languages. A third approach is to use a combination of coupled solvers, compact models, and lumped elements. In this work, we will demonstrate this third approach.

5.4.1 Software Integration

The interaction of the full system is based on different abstraction levels, using lumped circuit elements, compact/macromodels, and a direct interconnection of solvers for various domains. The circuit simulator SPICE3 is chosen as the *controlling solver* for the following reasons:

- SPICE3 has advanced time-step control.
- Models for different abstraction levels can be easily implemented in SPICE3.
- Lumped-element equivalent circuits can be readily simulated.

Relatively simple elements are implemented as lumped elements or compact models. These elements are electromechanical transducers (piezoelectric actuator) and thermoresistors. Flow sensors are much more complicated but often the fluid flow around sensors is relatively simple. For example, if the fluid flow in a channel is fully developed then it has a parabolic profile for the velocity, and thus this profile (compact model) can be used for the flow sensors as well. It is important to note that these compact models are parameterized and can be highly nonlinear. These models are obtained by insight gained from detailed physical level simulations, such as Navier–Stokes simulations, DSMC, and linearized solutions of the Boltzmann equation [Beskok and Karniadakis, 1999]. The pump can also be described as a lumped element [Klein, Matsumoto, and Gerlach, 1998]. However, these lumped-element descriptions are applicable only for small variations in the fluid flow. Usually pumps operate in a nonlinear and nonsmooth mode of fluid flow with a strong fluid–structure interaction. Therefore, a detailed physical level simulation of the pump is required. A simplification can be made by employing a macromodel of the form described in Equation (5.1), but here we employ full Navier–Stokes simulations with full dynamics.

For this reason, the following options are used:

- Electromechanical actuators, thermoresistors, and flow sensors are described as lumped elements and/or compact models.
- The pump is modeled at the detailed physical level.
- All lumped elements and models are implemented in SPICE3.
- The pump is implemented as a direct SPICE3-*NεκTαr* interconnection (Figure 5.11). SPICE3 transfers the time t_{spice} and pressure P for the membrane activation to *NεκTαr* and receives the flow rate Q and the time t_{call} for the next call to *NεκTαr*.

A detailed description of this coupling is provided later.

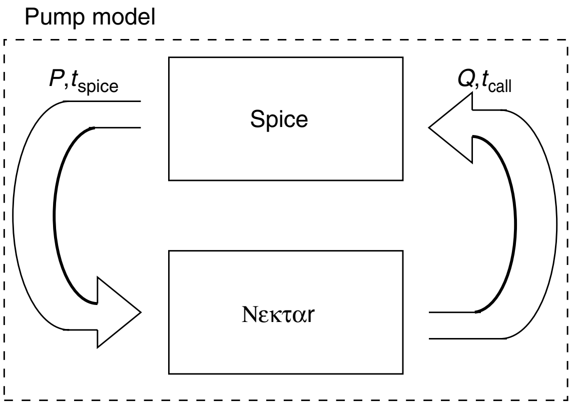


FIGURE 5.11 The SPICE3–*NεκTαr* interaction for the pump microsystem of Figure 5.2. SPICE3 provides the time t_{spice} and pressure *P* for the membrane actuation to *NεκTαr*. *NεκTαr* transfers the flow rate *Q* at time t_{call} for the next call of *NεκTαr* by SPICE3.

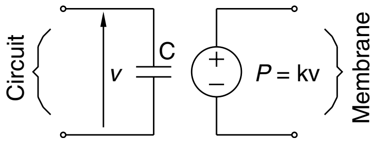


FIGURE 5.12 Lumped model for piezoelectric actuation. The voltage *V* is transformed into a pressure *P* that is used to activate the membrane of the pump.

5.4.2 Lumped-Element and Compact Models for Devices

5.4.2.1 Model for Piezoelectric Transducers

The model for electromechanical coupling with a piezoelectric actuation of the membrane is shown in Figure 5.12. This model forms the interface between the electrical and mechanical networks. The electrical characteristics of the piezoelectric actuator are described by the capacitor C. The input voltage V translates into an output pressure P by virtue of the piezoelectric effect with coefficient k. This pressure is an input argument to *NεκTαr*. The mechanical characteristics of the piezoelectric actuator are coupled with the mechanical characteristics of the substrate [Klein, 1997; Timoshenko and Woinowsky-Krieger, 1970].

5.4.2.2 Compact Model for Flow Sensor

For an anemometer type flow sensor [Rasmussen and Zaghloul, 1999] shown in Figure 5.13, a macromodel has been developed in [Mikulchenko, Rasmussen, and Mayaram, 2000]. This macromodel (Figure 5.14) is based on neural networks trained using data from detailed physical simulations.

The inputs to the neural network are the flow velocity *U* and the vector of geometrical and physical parameters Θ. The results from this model are in good agreement with the simulated data for a large range of parameters [Mikulchenko, Rasmussen, and Mayaram, 2000].

The dynamic macromodel is incorporated in SPICE3 by coupling it with a sensor circuit and a model for thermoresistors for the heater and sensors as shown in Figure 5.15. Based on the fluid flow rate the thermoresistor temperatures T1, T2, and T3 change, which in turn alters the resistance values and the sensing-circuit currents and voltages.

5.4.3 Effective Time-Stepping Algorithms

In general, the flow solver can be *NεκTαr* implemented as one big model in SPICE3. This is accomplished by *NεκTαr* from SPICE3 for each Newton iteration. However, such a coupling is extremely inefficient

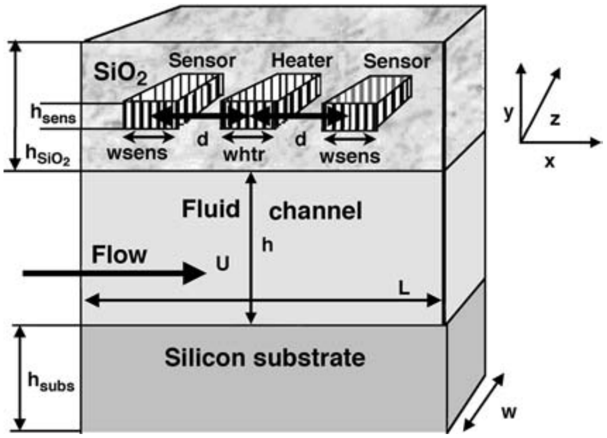


FIGURE 5.13 Structure of an anemometer-type flow sensor (thermocouple). This sensor is made up of a heating element and two sensing elements. The temperature difference between the sensors is used to measure the flow.

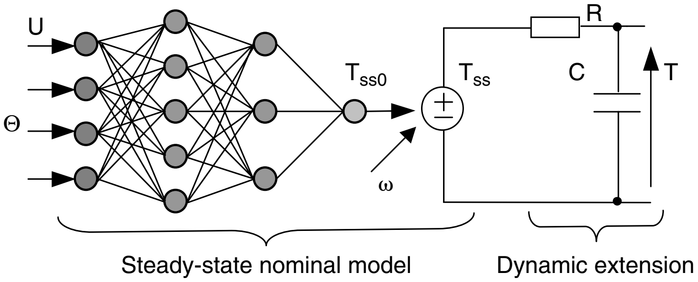


FIGURE 5.14 Dynamic macromodel for the flow sensor. The steady-state solution T_{SS0} corresponds to a nominal power for the heat source χ. The neural network output T_{SS0} is a multivariate function of the flow velocity U and the vector of geometrical and physical parameters Θ. T_{SS} is a linear function of the heat source χ and T_{SS0}.

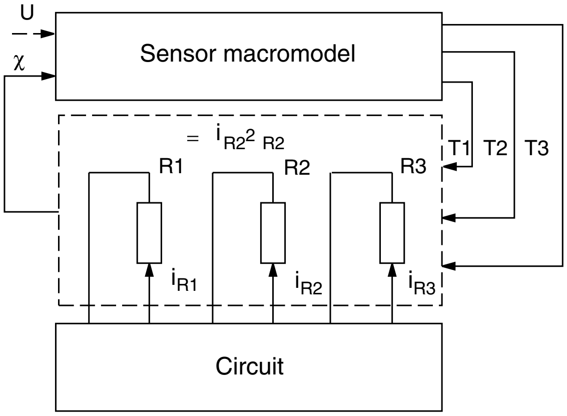


FIGURE 5.15 Macromodel implementation in SPICE3. Based on the fluid flow rate the thermoresistor temperatures T1, T2, and T3 change, which in turn alters the resistance values and the sensing-circuit currents and voltages.

because a call to *NεκTαr* is computationally very expensive. Furthermore, the time scales and nonlinearities are extremely different for the circuit and fluidic devices. If one considers only the circuit element, then a SPICE3 simulation results in nonuniform time steps and several Newton iterations for each time step. Typical time constants for circuits are of the order of $10^{-12} \ldots 10^{-6}$ seconds. On the other hand, fluidic devices have a typical time constant of the order $10^{-4} \ldots 10^{-1}$ seconds.

This property can be exploited to improve simulation performance by calling *NεκTαr* only at some of the circuit time points following *a subcycling type algorithm*. Between these time points, the *NεκTαr* outputs can be modeled as constant values. Further improvement in performance is possible by taking into account the usage of semiexplicit methods for fluid simulation. In this case, the flow rate Q_n for time point t_n is calculated by the explicit scheme: $Q_n = F(\mathbf{P}_{n-1}, \mathbf{V}_{n-1}, t_n)$, where P is the vector of the pressure at mesh points, and V is the vector of velocities at mesh points. For the SPICE3 *NεκTαr* interaction described earlier, the important quantities are the distributed pressure P for the pump membrane and the flow rate Q_n. This functional relationship can be expressed as follows: $Q_n = f(P_{n-1}, Q_{n-1}, t_n)$.

Based on this observation, an efficient time-stepping scheme is obtained as shown in Figure 5.16. Here, time is plotted on the horizontal axis, and the SPICE3 iterations are plotted on the vertical axis; $t_{S,k}$ and $t_{N,k}$ are the SPICE3 and *NεκTαr* time points, respectively. *NεκTαr* selects a time step $h_{N,i} = t_{N,i} - t_{N,i-1}$ independent of SPICE3, based on the Courant number (CFL) constraint for convection. The *NεκTαr* time points $t_{N,i}$ are used as synchronization time points with SPICE3, whereby $t_{N,i} = t_{S,k}$. The flow rate Q has a constant value between these synchronization time points. The membrane pressure $p_{j,k}$ is calculated as a function of the circuit behavior for each SPICE3 call at time $t_{S,k}$ and iteration j. The pressure $P_i = p_{M,k}$ at the final SPICE3 iteration M, for a synchronization time point $t_{S,k} = t_{N,i}$, is an input to *NεκTαr*. A *NεκTαr* call is made at $t_{N,i}$ and a new value of Q is computed using the relation $Q_{i+1} = f(P_i, Q_i, t_{N,i+1})$. This value is then used for the next *NεκTαr* time point, $t_{N,i+1}$.

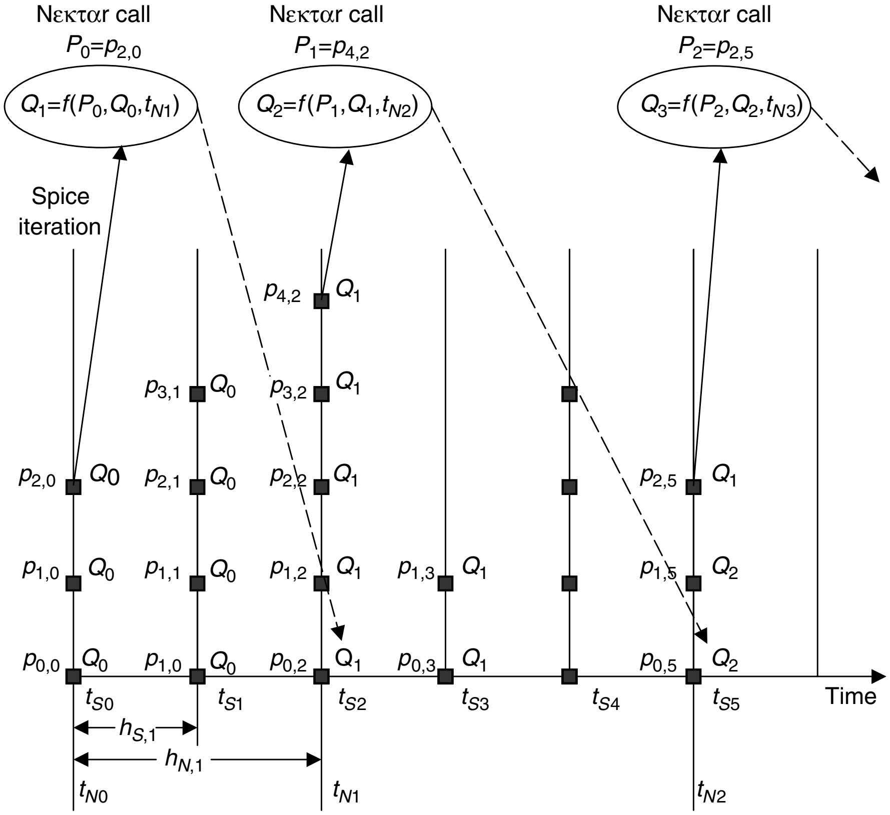


FIGURE 5.16 The time-stepping scheme for SPICE3 *NεκTαr* coupling. $T_{s,k}$ and $t_{N,k}$ are the SPICE3 and *NεκTαr* time points respectively. Q_i is a constant value for each SPICE3 iteration and at each SPICE3 time point between the *NεκTαr* time points $t_{N,i}$ and $t_{N,i+1}$. The membrane pressure $p_{j,k}$ is calculated as a function of the circuit behavior for each SPICE3 call at time $T_{s,k}$ and iteration j. SPICE3 selects time points based on a local truncation error estimate and synchronizes with *NεκTαr* at all *NεκTαr* time points. The pressure p_i for the final SPICE3 iteration at the synchronization time point $T_{s,k} = t_{N,i}$ is used as an input to *NεκTαr*. *NεTαr* call is made at $t_{N,i}$, and a new value of Q is computed for the next *NεκTαr* time point.

The main features of this time stepping scheme can be summarized as follows:

- *NεκTαr* is called from SPICE3.
- The timestep for SPICE3 is much smaller than the timestep for *NεκTαr*.
- *NεκTαr* specifies the next synchronization time point.

From this, it can be concluded that the number of *NεκTαr* calls are the same as that of stand-alone *NεκTαr*. This is the best possible situation in terms of efficiency for the coupled SPICE3–*NεκTαr* simulation.

5.5 Demonstrations of the Integrated Simulation Approach

5.5.1 Microfluidic System Description

A microliquid dosing system is used as an illustrative example. This system is made up of a micropump, a flow sensor and an electronic control circuit. The electronic circuit adjusts the pump flow rate. A simplified simulation circuit is shown in Figure 5.17.

In this system, the flow rate *Q* determines the flow sensor velocity *U* for a given set of geometry parameters (h, d, wsens). Based on the fluid flow rate, the thermoresistor temperatures T1, T2, and T3 change, which in turn alters the resistance values R1(T1), R2(T2), and R3(T3). The resistances R1(T1) and R3(T3) are included in a Wheatstone-bridge arrangement with two fixed resistors R4 and R5. The voltage difference $V_{R3(T3)} - V_{R1(T1)}$ is directly proportional to the temperature difference T3 − T1. This voltage difference is linearly transformed to the output voltage Vout by an operational amplifier with a controlled gain. This output voltage determines the pressure *P*, which activates the pump membrane and changes the flow rate *Q*. The thermoresistor of the heater (R2) is activated by the control electronics that maintain a constant heater temperature.

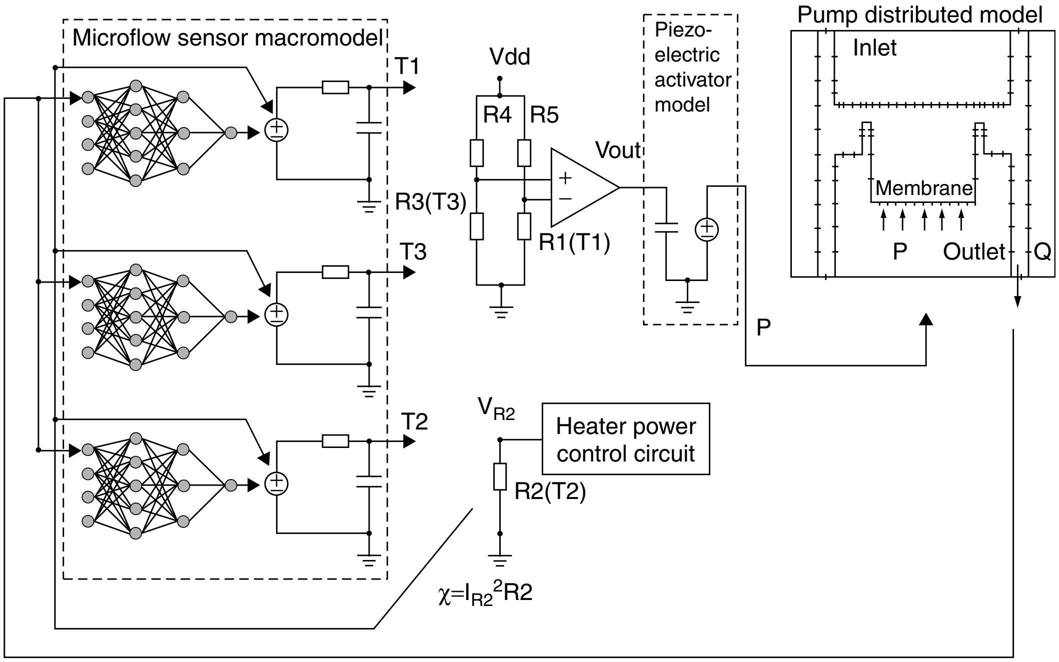


FIGURE 5.17 Description of the complete system for simulation. The pump flow rate *Q* determines the flow sensor velocity U. This yields the temperatures for the sensor thermoresistors. The difference between the resistance values R1(T1) and R3(T3) is transformed into the voltage V_{out} by the control electronics, which are used to control the pressure P for the pump membrane. This, in turn, determines the flow rate *Q*.

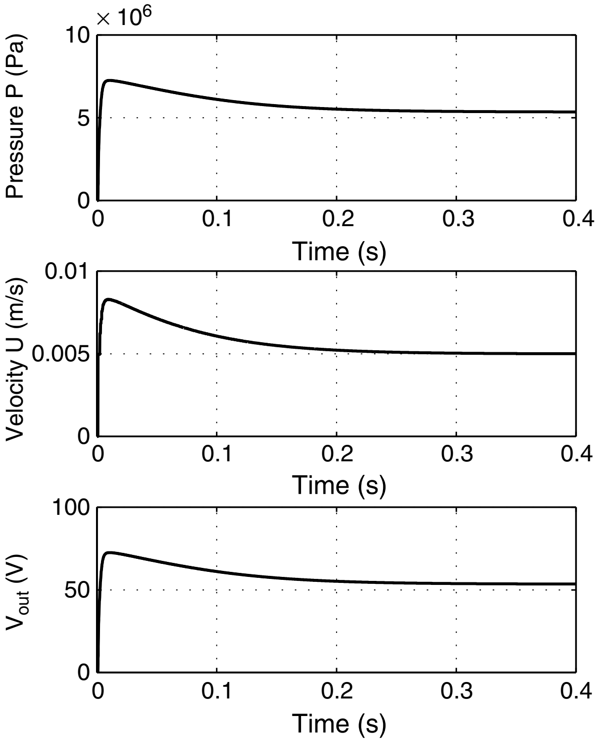


FIGURE 5.18 External pressure for the pump membrane, inlet velocity for the microflow sensor, and the amplifier output voltage for the simulation of the microfluidic system as a function of time.

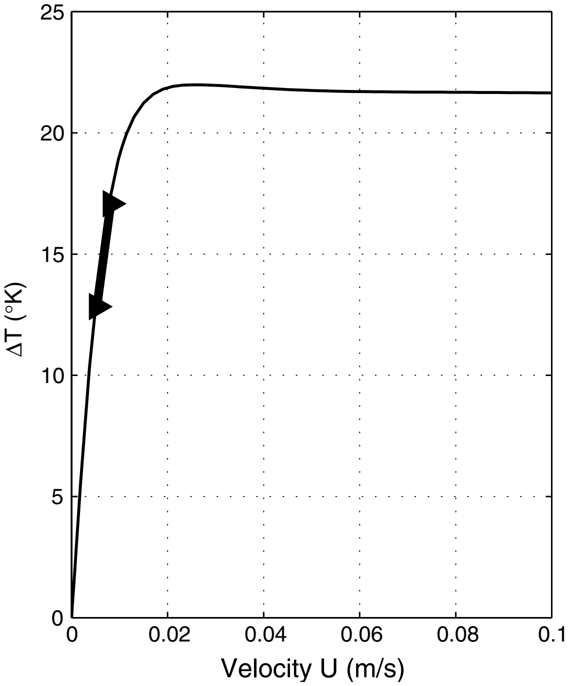


FIGURE 5.19 Flow sensor characteristics and its region of operation. A small change in velocity results in a large change in ΔT, the difference of the upstream and downstream sensor temperatures.

5.5.2 SPICE3–*NεκTαr* Integration

As mentioned earlier, *NεκTαr* is embedded as a subroutine in SPICE3. The interaction with SPICE3 is by means of the model code and the simulation engine. Synchronization time points are determined by *NεκTαr* and used by the SPICE3 transient analysis engine. The pump is modeled as a SPICE3 element with *NεκTαr* being the underlying simulation engine. The other elements in the circuit are described by lumped element descriptions and/or compact models.

5.5.3 Simulation Results

The simulation results from the coupled simulator are presented in Figure 5.18. In this simulation, one can determine the pressure on the pump membrane, the flow velocity, and the output control voltage as a function of time for various component parameters. As an example, consider the microflow sensor whose characteristics are shown in Figure 5.19. For the given range of flow velocity, the temperature difference between the upstream and downstream sensor temperatures is in the range 12–17°K. This simulation required approximately 5 minutes of CPU time on a 300 MHz Pentium II processor. Thus, the coupled simulator is reasonably efficient and provides valuable information to the system for device developers.

5.6 Summary and Discussion

Coupled-domain simulation is necessary in MEMS applications as many different physical phenomena are present and different processes are taking place simultaneously. Depending on the specific application (e.g., a microsensor versus a microactuator or a more complex system), some aspects of the device need to be simulated in detail at high resolution while others need to be accounted for by a low-dimensional description. Nonlinear macromodels are a possibility, but they are inadequate for the microfluidic system, which is typically highly unsteady and nonlinear. In addition, in the microdomain certain nonstandard flow features have to be modeled accurately, such as velocity-slip or temperature-jump in gas flows, viscous electrokinetic effects in liquid flows, and particle trajectories in particulate flows. To this end, we have developed the code that can simulate flows in the microdomains and macrodomains both for liquids and for gases. In addition, it includes a library of linear and nonlinear structures, such as beams, membranes, and cables.

For the coupled-domain simulation, the main driver program is SPICE3, a popular code for circuit simulation. In this paper, a coupled circuit and microfluidic device simulator was presented. The resulting simulator allows simulation of a complete microfluidic system in which thermal, flow, structural, and electrical domains are integrated. The coupling of these simulators was described and demonstrated for a microliquid dosing system. The integrated simulator can be utilized for parametric studies and optimal design of microfluidic systems.

The integration of different simulators required for complete MEMS simulations is a difficult problem with challenges well beyond software integration. It involves disparate temporal and spatial scales leading to great stiffness and inefficiencies, new physical assumptions and approximations for some of the components, issues of numerical stability, staggered time-marching procedures, new fast solvers for coupled problems, and optimization and control algorithms. Most of the mature algorithms from single disciplines are inefficient in this context, so new methods are required in order to produce a new generation of simulation algorithms for MEMS devices. In this chapter, we have demonstrated that this is possible by coupling two accurate codes and resolving at least at some level some of these coupling issues. However, significant improvements can be made for specific devices. For example, for the membrane-driven micropump presented here, convergence of the coupling algorithm could be accelerated by inspecting the time-dependent mass-conservation equation every SPICE time step and obtaining a new estimate of flow from

$$Q_{new} = Q_{old} + \frac{\Delta V}{\Delta t}$$

where ΔV is the change in volume due to the change in the membrane position, and Δt is the time between two consecutive SPICE calls. This requires solving for the structure only but not necessarily for the entire flow field, which is the most computationally intensive task. The structure solver is very fast and can be called as often as necessary without a serious computational overhead.

Acknowledgments

This work is supported in part by DARPA under agreement number F30602-98-2-0178. We would like to thank Prof. A. Beskok for many useful suggestions regarding this work.

References

Allen, M., and Tildesley, D. (1994) *Computer Simulation of Liquids*, Clarendon Press, Oxford.

Aluru, N. (1999) "A Reproducing Kernel Particle Method for Meshless Analysis," *Comput. Mech.* **23**, pp. 324–38.

Aluru, N., and White, J. (1996) "Direct-Newton Finite-Element/Boundary-Element Technique for Micro-Electro-Mechanical Analysis," in *Tech. Digest: Solid-State Sensor and Actuator Workshop*, *Hilton Head Is. SC*, pp. 54–57.

Aluru, N.R., and White, J. (1997) "An efficient numerical technique for electromechanical simulation of complicated microelectromechanical structures," *Sensors and Actuators*, **A-58**, pp. 1–11.

Battista, G.D., Eades, P., Tamassia, R., and Tollis, I. (1998) *Graph Drawing*, Prentice Hall, Englewood Cliffs, NJ.

Beskok, A., and Karniadakis, G.E. (1999) "A Model for Flows in Channels, Pipes and Ducts at Micro- and Nano-Scales," *J. Microscale Thermophys. Eng.* **3**, pp. 43–77.

Beskok, A., Karniadakis, G.E., and Trimmer, W. (1996) "Rarefaction and Compressibility Effects in Gas Microflows," *J. Fluids Eng.* **118**, p. 448.

Bielefeld, J., Pelz, G., and Zimmer, G. (1997) "AHDL-Model of a 2D Mechanical Finite-Element Usable for Microelectro-Mechanical Systems," in *BMAS'97*, pp. 177–81, IEEE, Washington, D.C.

Bird, G. (1994) *Molecular Gas Dynamic and the Direct Simulation of Gas Flows*, Oxford Science Publications, Oxford.

Engl, W.L., Laur, R., and Dirks, H.K. (1982) "MEDUSA-A Simulator for Modular Circuits," *IEEE Trans. Comput. Aid. Design* **1**, April, pp. 85–93.

Freeman, D., Aranyosi, A., Gordon, M., and Hong, S. (1998) "Multidimensional Motion Analysis of MEMS Using Computer Microvision," in *Solid-State Sensor and Actuator Workshop*, *Hilton Head Is., S.C.*, June 1998, pp. 150–55.

Gabbay, L.D. (1998) Computer Aided Macromodeling for MEMS, Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Gad-El-Hak, M. (1999) "The Fluid Mechanics of Microdevices," *J. Fluids Eng.* **12**(1), pp. 5–33.

Gilbert, J.R. et al. (1993) "Implementation of a MEMCAD System for Electrostatic and Mechanical Analysis of Complex Structures from Mask Descriptions," *Proc. IEEE Workshop on Microelectromechanical Systems, MEMS'93*, Ft. Lauderdale, February 1993, pp. 207–12.

Hadjiconstantinou, N. (1999) "Hybrid Atomistic-Continuum Formulations and the Moving Contact-Line Problem," *J. Comp. Phys.* **154**, pp. 245–65.

Ho, C.-M., and Tai, Y.-C. (1998) "Micro-Electro-Mechanical Systems (MEMS) and Fluid Flows," *Annu. Rev. Fluid Mech.* **30**, pp. 579–612.

Ho, L.-W. (1989) A Legendre Spectral Element Method for Simulation of Incompressible Unsteady Free-Surface Flows, Ph.D. thesis, Department of Mechanical Engineering, Massachusetts Institute of Technology.

Hughes, T.J.R. (1987) *The Finite Element Method*, Prentice-Hall, Englewood Cliffs, New Jersey.

Hung, E.S., Yang, Y.-J., and Senturia, S. (1997) "Low-Order Models for Fast Dynamical Simulations of MEMS Microstructures," in *Transducers'97*, June 1997, pp. 1101–4, IEEE, Chicago.

Karniadakis, G.E., and Sherwin, S.J. (1999) "*Spectral/hp Element Methods for CFD*," Oxford University Press, Oxford.

Keown, J. (1997) *Microsim PSPCE and Circuit Analysis*, 3rd ed., Prentice Hall, London.

Kirby, R.M., Warburton, T.C., Sherwin, S.J., Beskok, A., and Karniadakis, G.E. (1999) "The *NεκTαr* Code: Dynamic Simulations with Remeshing," *Second International Conference on Computational Technologies for Fluid/Thermal/Chemical Systems with Industrial Applications*, August 1–5.

Klein, A., and Gerlach, G. (1997) "Modelling of Piezoelectric Bimorph Structures Using an Analog Hardware Description Language," in *Second International Conference on the Simulation and Design of Microsystems and Microstructures (MICROSIM'97), Lausanne (Switzerland)*, September in Adey and Renaud "Microsim II-Simulation and Design of Microsystems and Microstructures", Computational Mechanics Publications, Southampton, UK/Boston, USA, 1998, pp. 229–38.

Klein, A., Matsumoto, S., and Gerlach, G. (1998) "Modeling and Design Optimization of a Novel Micropump," in *First International Conference on Modeling and Simulation of Microsystems, Semiconductors, Sensors and Actuators*, 6–8 April, Santa Clara, California, pp. 506–11.

Koplik, J., and Banavar, J. (1995) "Continuum Deductions from Molecular Hydrodynamics," *Annu. Rev. Fluid Mech.* **27**, pp. 257–92.

Kundert, K.S. (1990) "Sparse-Matrix Techniques and Their Application to Circuit Simulation," in *Circuit Analysis, Simulation and Design*, North-Holland Co, New York, NY.

Lo, N.R. et al. (1996) "Parameterized Layout, Synthesis, Extraction, and SPICE Simulation for MEMS," *ICASE '96*, pp. 481–84.

Lohner, R., and Yang, C. (1996) "Improved ALE Mesh Velocities for Moving Bodies," *Comm. Num. Meth. Eng., Phys.* **12**, pp. 599–608.

Mayaram, K., Chern, J., and Yang, P. (1993) "Algorithms for Transient Three-Dimensional Mixed-Level Circuit and Device Simulation," *IEEE Trans. Comput. Aid. Design* **12**, pp. 1726–33.

Mayaram, K., and Pederson, D.O. (1987) "Analysis of MOS Transformer-Coupled Oscillators", *IEEE J. Solid-State Circuits*, **22**, pp. 1155–62.

Mayaram, K., and Pederson, D.O. (1992) "Coupling Algorithms for Mixed-Level Circuit and Device Simulation," *IEEE Trans. Comput. Aid. Design* **11**, pp. 1003–12.

Mikulchenko, O., Rasmussen, A., and Mayaram, K. (2000) "A Neural Network Based Macromodel of Microflow Sensors," *Proc. of MSM2000*, NSTI (Nano Science and Technology Institute), pp. 540–43, Cambridge, Massachusetts.

Moaveni, S. (1999) *Finite Element Analysis: Theory and Application with ANSYS*, Prentice Hall, New Jersey.

Nagel, L.W. (1975) "SPICE2: A Computer Program to Simulate Semiconductor Circuits," Tech. Rep. No. UCB/ERL M520, Electronics Research Lab., Univ. of California, Berkeley.

Park, H.J., Ko, P.K., and Hu, C. (1991) "A Charge Conserving Non-Quasi-Static (NQS) Model for SPICE Transient Analysis," *IEEE Trans. Comput. Aid. Design* **10**, pp. 629–42.

Quarles, T.L. (1989) "The SPICE3 Implementation Guide," Tech. Rep. No. UCB/ERL M89/44, Electronics Research Lab., Univ. of California, Berkeley.

Rasmussen, A., and Zaghloul, M.E. (1999) "The Design and Fabrication of Microfluidic Flow Sensors," in *Proc. ISCAS–99*, pp. 136–39, IEEE, Orlando, FL.

Ravanelli, E., and Hu, C. (1991) "Device-Circuit Mixed Simulation of VDMOS Charge Transients," *Solid State Electron.* **34**, pp. 1353–60.

Rotella, F.M., Troyanovsky, B., Yu, Z., Dutton, R., and Ma, G. (1997) "Harmonic Balance Device Analysis of an LDMOS RF Power Amplifier with Parasitics and Matching Network," in *SISPAD–97*, pp. 157–59, IEEE, Picataway, NJ.

Sangiovanni-Vincentelli, A.L. (1981) "Circuit Simulation," in *Computer Design Aids for VLSI Circuits*, Sijthoff and Noordhoff, pp. 19–113, Boston, MA.

Schroth, A., Blochwitz, T., and Gerlach, G. (1995) "Simulation of a Complex Sensor System Using Coupled Simulation Programs," in *Transducers '95*, pp. 33–35, IEEE, Stockholm, Sweden.

Senturia, S., Aluru, N., and White, J. (1997) "Simulating the Behavior of MEMS Devices: Computational Methods and Needs," *IEEE Computational Science & Engineering*, vol. **4**(1) January–March, pp. 30–43.

Senturia, S. et al. (1992) "A Computer-Aided Design System for Microelectromechanical Systems (MEMCAD)," *J. Microelectromech. Syst.* **1**, pp. 3–13.

Sherwin, S.J. (1995*)* Triangular and Tetrahedral Spectral/hp Finite Element Methods for Fluid Dynamics, Ph.D. thesis, Princeton University.

Silvaco International (1995) *ATLAS User's Manual: Mixed Mode*, June, Santa Clara, CA.

Technology Modeling Associates (1997) *MEDICI User's Manual: Circuit Analysis*, February.

Tilmans, H.A.C. (1996) "Equivalent Circuit Representation of Electromechanical Transducers: 1. Lumped-Parameter Systems," *J. Micromech. Microeng.* **6**, pp. 157–76, Synopsis, Inc.

Timoshenko, S.P., and Woinowsky-Krieger, S. (1970) *Theory of Plates and Shells*, 2nd ed., McGraw-Hill, New York.

Veijola, T. et al. (1995) "Equivalent-Circuit Model of the Squeezed Gas Film in a Silicon Accelerometer," *Sensor. and Actuator. A*, **48**, pp. 239–48.

Warburton, T.C. (1999*)* Spectral/hp Methods on Polymorphic Multi-Domains: Algorithms and Applications, Ph.D. thesis, Division of Applied Mathematics, Brown University.

Woodruff, R.L., and Rudeck, P.J. (1993) "Three-Dimensional Numerical Simulation of Single Event Upset of an SRAM Cell," *IEEE Trans. Nucl. Sci.* **40**, pp. 1795–1803.

Ye, W., Kanapka, J., and White, J. (1999) "A Fast 3D Solver for Unsteady Stokes Flow with Applications to Micro-Electro-Mechanical Systems," in *Proceedings of the Second International Conference on Modeling and Simulation of Microsystems*, San Juan, Puerto Rico, April 19–21, pp. 518–21.

6

Molecular-Based Microfluidic Simulation Models

Ali Beskok
Texas A&M University

6.1 Introduction

Simulation of microscale thermal fluidic transport is gaining importance due to the emerging technologies of the 21st century, such as microelectromechanical systems (MEMS) and nanotechnologies. Miniaturization of device scales has made possible for the first time the integration of sensing, computation, actuation, control, communication, and power generation within the same microchip. The small size, light weight, and high-durability of MEMS, combined with their mass fabrication, result in low-cost systems with a wide variety of applications from control systems to advanced energy systems to biological, medical, and chemical uses. Despite the diverse prospects and fast growth of MEMS, further miniaturization of device scales presents the challenge of better understanding micron and submicron scale physics.

The microscale thermal/fluidic transport phenomenon differs from its larger scale counterparts mainly due to the size, surface, and interface effects [Ho and Tai, 1998; Gad-el-Hak, 1999]. Reduction of the characteristic device dimensions to micrometer scale drastically decreases the volume-to-surface area ratio. Hence, the surface forces are more dominant than the body forces in such small scales. The origin of the surface forces is atomistic and based on the short-ranged van der Waals forces and longer-ranged electrostatic, or Coulombic, forces. Although a molecular-simulation-based approach for understanding fluid forces on surfaces is fundamental in nature, it is very difficult to apply to engineering problems due to the vast number of molecules involved in the analysis; however, direct application of the well-known continuum equations is not appropriate, either. For example, the Navier–Stokes level of constitutive relations that model the shear stress, being linearly proportional to the strain rate, is not valid for gases when the Knudsen number $Kn > 0.1$, or for liquids when the strain rate exceeds twice the molecular frequency scale [Gad-el-Hak, 1999]. Significant differences between the thermal/fluidic transport of gas and liquid states also exist. For example, dilute gases spend most of their time in free flight with abrupt changes in their direction and speed caused by binary intermolecular collisions. The liquid molecules are closely packed, however, and they experience multiple collisions with large intermolecular forces. The fundamental simulation approaches for liquid and gas flows differ from a microscopic point of view. In this chapter, we will address separately the numerical simulation methods relevant for dilute gases and liquids. However, the main emphasis of the chapter is microscale gas transport modeling with the direct simulation Monte Carlo (DSMC) algorithm. Other microscopic simulation methods, such as the Boltzmann equation approach, lattice Boltzmann method, and molecular dynamics (MD), are briefly introduced to guide the reader to the appropriate resources in these areas.

6.2 Gas Flows

The ratio of the gas mean free path l to a characteristic microfluidic length scale h is known as the Knudsen number, $Kn = \lambda/h$. Because the momentum and energy transfers happen with intermolecular and gas/wall collisions, the mean free path indicates an intrinsic length scale of thermal/fluidic transport for gases. In standard pressure and temperature (STP), the mean free path for air is about 65 nm. For macroscopic devices, the Knudsen number is very small, so the surrounding air can be treated as a continuous medium. However, in microscales, the Knudsen number can be fairly large due to the small length scales. Momentum and energy transport in micron and submicron scales show significant deviations from their larger scale counterparts. For example, recent microchannel experiments show increased mass flow rates compared to the Navier–Stokes-based continuum estimates [Arkilic et al., 1997; Harley et al., 1995; Liu et al., 1993; Pong et al., 1994]. Similarly, in the case of magnetic disk storage units, the head floating about 50 nm above the media exhibits an order of magnitude reduction of load capacity compared to predictions by the continuum Reynolds equations [Fukui and Kaneko, 1990]. These deviations are explained as a function of the Knudsen number by dividing the flow into four regimes: continuum ($Kn \leqslant 0.01$), slip ($0.01 \leqslant Kn \leqslant 0.1$), transitional ($0.1 \leqslant Kn \leqslant 10$), and free-molecular ($Kn > 10$). Operation regimes of typical MEMS devices at standard temperature and pressure are shown in Figure 6.1. MEMS operate in a wide variety of flow regimes covering the continuum, slip, and early transitional flow regimes. Further miniaturization of MEMS device components and nanotechnology applications [Drexler, 1990] corresponds to higher Knudsen numbers, making it necessary to study the mass, momentum, and energy transport in the entire Knudsen regime.

It may be misleading to identify the flow regimes as slip and continuum. Within this text and in most of the microscale transport literature, *continuum* refers to the Navier–Stokes equations subject to the no-slip-boundary conditions. This identification leads to two common misconceptions. First, if the Navier–Stokes equations cannot be applied, then the continuum approximation should break down. This is misleading, for we will see shortly that it is possible to derive conservation equations with more advanced constitutive laws than the Navier–Stokes equations. One example of this is the Burnett equations. The second misconception is that in the slip flow regime the boundary conditions suddenly change from no-slip to slip. This is also misleading, as the no-slip-boundary condition is just an empirical finding and the Navier–Stokes equations are valid both for slip and continuum flow regimes. Hence, the slip effects become important gradually with increased Kn. Nevertheless, the identification of flow regimes was made for rarefied gas flows almost a century ago. For $Kn \leqslant 0.1$ flows, the Navier–Stokes equations subject to the velocity-slip and temperature-jump boundary conditions should be used. The slip conditions are [Kennard, 1938; Schaaf and Chambre, 1961]:

$$u_s - u_w = \frac{2-\sigma_v}{\sigma_v}\frac{1}{\rho(2RT_w/\pi)^{1/2}}\tau_s + \frac{3}{4}\frac{Pr(\gamma-1)}{\gamma\rho RT_w}(-q_s) \tag{6.1}$$

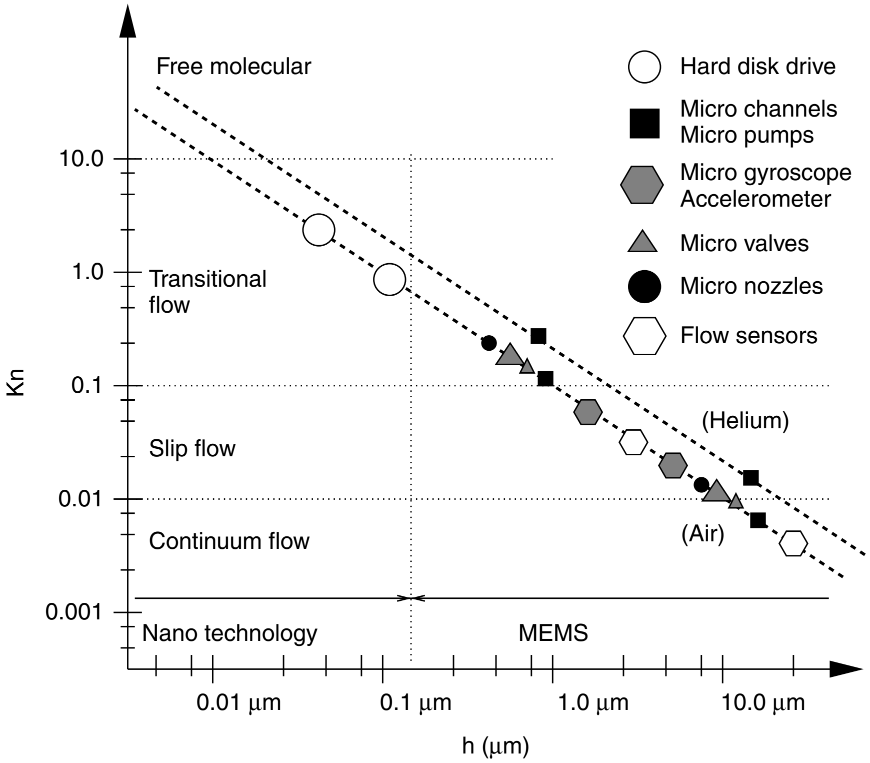


FIGURE 6.1 The operation range for typical MEMS and nanotechnology applications under standard conditions spans the entire Knudsen regime (continuum, slip, transition, and free molecular flow regimes).

$$T_s - T_w = \frac{2-\sigma_T}{\sigma_T}\left[\frac{2(\gamma-1)}{\gamma+1}\right]\frac{1}{R\rho(2RT_w/\pi)^{1/2}}(-q_n) \tag{6.2}$$

where q_n and q_s are the normal and tangential heat-flux components, τ_s is the viscous stress component corresponding to the skin friction, R is the specific gas constant, g is the ratio of specific heats, r is the density, Pr is the Prandtl number, and T_w and u_w are the wall temperature and velocity respectively. The gas slip velocity and temperature near the wall (jump) are given by u_s and T_s respectively. The term in the above equation proportional to $(-q_s)$ is associated with the phenomenon of thermal creep, which can cause variations of pressure along tubes in the presence of tangential temperature gradients [Beskok et al., 1995; Sone, 2000; Vargo and Muntz, 1996; Vargo et al., 1998].

In a recent work, a Padé approximation of 1 was developed resulting in a velocity slip condition valid in the entire Knudsen regime. Excluding the thermal creep terms, this new slip condition is given in the following form [Beskok et al., 1996; Beskok and Karniadakis, 1999]:

$$U_s - U_w = \frac{2-\sigma_v}{\sigma_v}\left[\frac{Kn}{1-bKn}\right]\frac{\partial U}{\partial n} \tag{6.3}$$

where U_w and U_s are the wall and gas-slip velocity nondimensionalized with a reference velocity respectively. Here, b is the general slip coefficient determined by the following procedures:

- A perturbation expansion in Kn for $Kn < 1$, such that Equation (6.3) is equivalent to a second-order slip condition [Beskok et al., 1996].
- Matching the velocity profiles with the direct simulation Monte Carlo (DSMC) results in the transitional and free molecular flow regimes [Beskok and Karniadakis, 1999].

Hence the value of b is defined analytically in the slip and early transition flow regime, but in the transitional and free molecular flow regimes it is an empirical parameter.

In Equations (6.1) and (6.3), σ_v, and σ_T are the tangential momentum and thermal accommodation coefficients respectively. The accommodation coefficients model the momentum and energy exchange of gas molecules impinging on the walls. Hence, they characterize the surface effects. For example, $\sigma_v = 0.2$ enhances the apparent slip by almost an order of magnitude. The accommodation coefficients are usually determined experimentally. Due to the difficulties of experimentation in microscales, the accommodation coefficients are obtained by assuming slip flow and matching the value of the accommodation coefficients to maintain the measured mass flow rate. This has resulted in $\sigma_v = 0.80$ for nitrogen, argon, and carbon dioxide in contact with prime silicon crystal [Arkilic et al., 1997]. Lower accommodation coefficients are expected due to the low surface roughness of the prime silicon crystal. However, for a general micromachined surface and gas pair, the values of the accommodation coefficients are not known a priori. For low-pressure, rarefied gas flows, the values of the accommodation coefficients are tabulated as a function of the specific gas and surface quality [Seidl and Steinheil, 1974]; under laboratory conditions, values as low as 0.2 have been observed [Lord, 1976]. Very low values of sv will increase the slip on the walls considerably, even for small Knudsen number flows.

In the transitional flow regime, the constitutive laws defining the stress tensor and the heat-flux vector must be updated for increased rarefaction effects resulting in Wood's, Burnett's, or Grad's equations. It is also possible to use the Boltzmann transport equation in this regime (see section 6.2.7). In a recent work, Myong has developed thermodynamically consistent hydrodynamic computational models for high-Knudsen-number gas flows, uniformly valid in all Mach numbers and satisfying the second law of thermodynamics [Myong, 1999].

In the free molecular flow regime ($Kn \geqslant 10$), the molecule–wall interactions dominate the transport with significantly reduced intermolecular collisions. Hence, the collisionless Boltzmann equation is commonly used in this flow regime.

6.2.1 Molecular Magnitudes

Before studying the molecular-based numerical simulation algorithms, it is crucial to understand the complexity of the molecular simulation problem. In this section, we present relationships for the number density of molecules n, mean molecular spacing d, molecular diameter d_m, mean free path λ, mean collision time t_c, and mean square molecular speed $\sqrt{\overline{C}^2}$.

The number of molecules in one mole of gas is a constant known as the Avogadro's number, 6.02252×10^{23}/mole, and the volume occupied by one mole of gas at a given temperature and pressure is a constant, regardless of the composition of the gas [Vincenti and Kruger, 1977]. This leads to the perfect gas relationship given by:

$$P = nk_bT \tag{6.4}$$

where P is the pressure, T is the temperature, n is the number density of the gas, and k_b is the Boltzmann constant ($k_b = 1.3805 \times 10^{-23}$ J/K). This ideal gas law is valid for dilute gases at any pressure (above the saturation pressure). Hence, for most microscale gas flow applications we can predict the number density of the molecules at a given temperature and pressure using Equation (6.4). At atmospheric pressure and 0°C (standard conditions) the number density is about $n \cong 2.69 \times 10^{25}\,\text{m}^{-3}$. If all of these molecules are placed in a 1-m cube in an equidistant fashion, the mean molecular spacing will be

$$\delta = n^{-1/3} \tag{6.5}$$

Under standard conditions the mean molecular spacing is $\delta \cong 3.3 \times 10^{-9}$ m.

The mean molecular diameter (d_m) of typical gases, based on the measured coefficient of viscosity and the Chapman–Enskog theory of transport properties for hard-sphere molecules, is on the order of 10^{-10} m. For air under standard conditions, $d_m \cong 3.7 \times 10^{-10}$ m, as tabulated in Bird (1994). Comparison of the mean molecular spacing δ and the typical molecular diameter d_m shows an order of magnitude difference. This leads to the concept of "dilute gas," where $\delta/d_m \gg 1$. For dilute gases, binary intermolecular collisions are more likely than the simultaneous multiple collisions. On the other hand, dense gases and liquids go through multiple collisions at a given instant, making the treatment of the intermolecular collision process more difficult. The dilute gas approximations, along with molecular chaos and equipartition of energy principles, lead us to the well established kinetic theory of gases and formulation of the Boltzmann transport equation starting from the Liouville equation. The assumptions and simplifications of this derivation are given in Vincenti and Kruger (1977) and Bird (1994).

Momentum and energy transport in the bulk of the fluid happen with intermolecular collisions, as does settling to a thermodynamic equilibrium state. Hence, the time and length scales associated with the intermolecular collisions are important parameters for many applications. The distance traveled by the molecules between the intermolecular collisions is known as the mean free path. For a simple gas of hard-sphere molecules in thermodynamic equilibrium, the mean free path is given in the following form [Bird, 1994]:

$$\lambda = (2^{1/2}\pi d_m^2 n)^{-1} \tag{6.6}$$

The gas molecules are traveling with high speeds proportional to the speed of sound. By simple considerations, the mean-square molecular speed of the gas molecules is given by [Vincenti and Kruger, 1977]:

$$\sqrt{\overline{C}^2} = \sqrt{\frac{3P}{\rho}} = \sqrt{3RT} \tag{6.7}$$

where R is the specific gas constant. For air under standard conditions, this corresponds to 486 m/sec. This value is about 3 to 5 orders of magnitude larger than the typical average speeds obtained in gas microflows. (The importance of this discrepancy will be discussed in Section 6.2.3.) In regard to the time scales of intermolecular collisions, we can obtain an average value by taking the ratio of the mean free path to the mean-square molecular speed. This results in $t_c \cong 10^{-10}$ for air under standard conditions. This time scale should be compared to a typical microscale process time scale to determine the validity of the thermodynamic equilibrium assumption.

So far we have identified the vast number of molecules and the associated time and length scales for gas flows. That it is possible to lump all of the microscopic quantities into time- and/or space-averaged macroscopic quantities, such as fluid density, temperature, and velocity. It is crucial to determine the limitations of these continuum-based descriptions; in other words:

- How small should a sample size be so that we can still talk about the macroscopic properties and their spatial variations?
- At what length scales do the statistical fluctuations become significant?

It turns out that a sampling volume that contains 10,000 molecules typically results in 1% statistical fluctuations in the averaged quantities [Bird, 1994]. This corresponds to a volume of 3.7×10^{-22} m^3 for air at standard conditions. If we try to measure an "instantaneous" macroscopic quantity such as velocity in a three-dimensional space, one side of our sampling cube will typically be about 72 nm. This length scale is slightly larger than the mean free path of air λ under standard conditions. Therefore, in complex micro-geometries where three-dimensional spatial gradients are expected, the definition of instantaneous macroscopic values may become problematic for $Kn > 1$. If we would like to subdivide this domain further to obtain an instantaneous velocity distribution, the statistical fluctuations will be increased significantly as the sample volume is decreased. Hence, we may not be able to define instantaneous velocity distribution in a 72 nm^3 volume. On the other hand, it is always possible to perform time or ensemble averaging of the data at such small scales. Hence, we can still talk about a velocity profile in an averaged sense.

To describe the statistical fluctuation issues further, we present in Figure 6.2 the flow regimes and the limit of the onset of statistical fluctuations as a function of the characteristic dimension L and the normalized number density n/n_o. The 1% statistical scatterline is defined in a cubic volume of side L, which contains approximately 10,000 molecules. Using Equation (6.5), we find that $L/\delta \approx 20$ satisfies this condition approximately, and the 1% fluctuation line varies as $(n/n_o)^{-1/3}$. Under standard conditions, 1% fluctuation is observed at $L = 72$ nm, and the Knudsen number based on this value is $Kn \approx 1$. Figure 6.2 also shows the continuum, slip, transitional, and free molecular flow regimes for air at

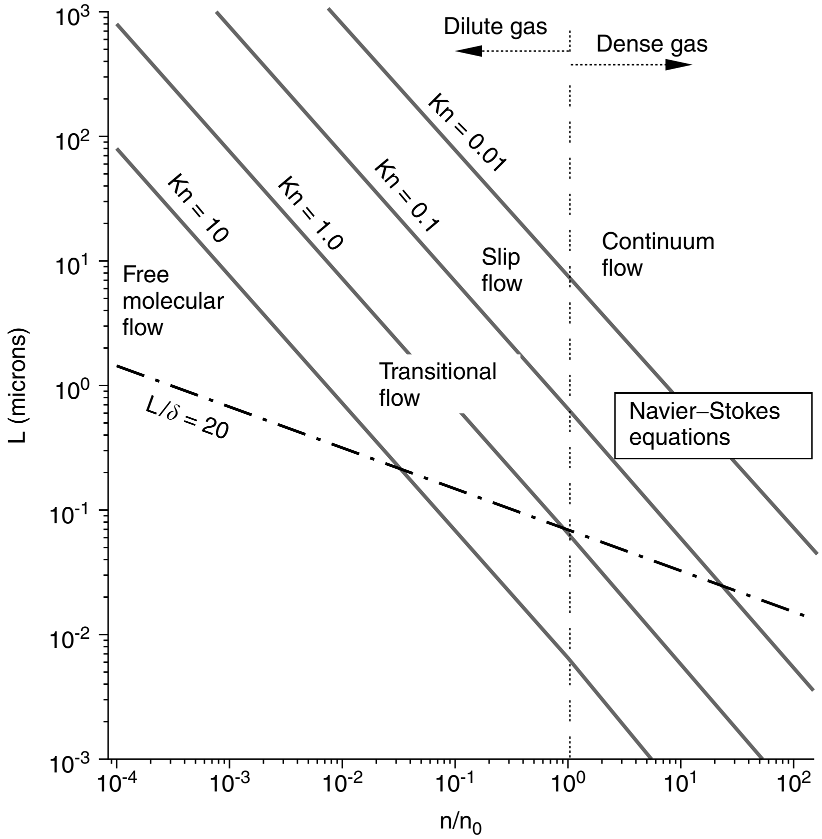


FIGURE 6.2 Limit of approximations in modeling gas microflows. L is the characteristic length, n/n_o is the number density normalized with the corresponding standard conditions. The lines that define the various Knudsen regimes are based on air at isothermal conditions ($T = 273$ K). The $L/\delta = 20$ line corresponds to the 1% statistical scatter in the macroscopic properties. The area below this line experiences increased statistical fluctuations.

273 K and at various pressures. The mean free path varies inversely with the pressure. Hence, at isothermal conditions, the Knudsen number varies as $(n/n_o)^{-1}$. The fundamental question of dynamic similarity of low-pressure gas flows to gas microflows under geometrically similar and identical Knudsen, Mach, and Reynolds number conditions can be answered to some degree by Figure 6.2. Provided that there are no unforeseen microscale-specific effects, the two flow cases should be dynamically similar. However, a distinction between the low-pressure and gas microflows is the difference in the length scales for which the statistical fluctuations become important.

It is interesting to note that for low-pressure rarefied gas flows the length scales for the onset of significant statistical scatter correspond to much larger Knudsen values than do the gas microflows. For example, $Kn = 1.0$ flow obtained at standard conditions in a 72 nm cube volume permits us to perform one instantaneous measurement in the entire volume with 1% scatter. However, at 100 pascal pressure and 273 K temperature, $Kn = 1.0$ flow corresponds to a length scale of 65 mm. For this case, 1% statistical scatter in the macroscopic quantities is observed in a cubic volume of side 0.72 μm, allowing about 90 pointwise instantaneous measurements. This is valid for instantaneous measurements of macroscopic properties in complex three-dimensional conduits. In large-aspect-ratio microdevices, one can always perform spanwise averaging to define an averaged velocity profile. Also, for practical reasons one can also define averaged macroscopic properties either by time or ensemble averaging (such examples are presented in Section 6.2.4).

6.2.2 An Overview of the Direct Simulation Monte Carlo Method

In this section, we present the algorithmic details, advantages, and disadvantages of using the direct simulation Monte Carlo algorithm for microfluidic applications. The DSMC method was invented by Graeme

A. Bird (1976, 1994). Several review articles about the DSMC method are currently available [Bird 1978, 1998; Muntz, 1989; Oran et al., 1998]. Most of these articles present an extended review of the DSMC method for low-pressure rarefied gas flow applications, with the exception of Oran et al. (1998), who also address microfluidic applications.

The previous section describes molecular magnitudes and associated time and length scales. Under standard conditions in a volume of 10 μm^3, there are about 2.69×10^{10} molecules. A molecular-based simulation model that can compute the motion and interactions of all these molecules is not possible. The typical DSMC method uses hundreds of thousands or even millions of simulated molecules or particles that mimic the motion of real molecules.

The DSMC method is based on splitting the molecular motion and intermolecular collisions by choosing a time step less than the mean collision time and tracking the evolution of this molecular process in space and time. For efficient numerical implementation, the space is divided into cells similar to the finite-volume method. The DSMC cells are chosen proportional to the mean free path λ. In order to resolve large gradients in flow with realistic (physical) viscosity values, the average cell size should be $\Delta x_c \cong \lambda/3$ [Oran et al., 1998]. The time- and cell-averaged molecular quantities are obtained as the macroscopic values at the cell centers. The DSMC involves four main processes: motion of the particles, indexing and cross-referencing of particles, simulation of collisions, and sampling of the flow field. The basic steps of a DSMC algorithm are given in Figure 6.3.

The first process involves motion of the simulated molecules during a time interval of Δt. Because the molecules will go through intermolecular collisions, the time step for simulation chosen is smaller than the mean collision time Δt_c. Once the molecules are advanced in space, some of them will have gone through wall collisions or will have left the computational domain through the inflow–outflow boundaries. Hence, the boundary conditions must be enforced at this level, and the macroscopic properties along the solid surfaces must be sampled. This is done by modeling the surface molecule interactions by applying the conservation laws on individual molecules rather than using a velocity distribution function that is commonly utilized in the Boltzmann algorithms. This approach allows inclusion of many other physical processes, such as chemical reactions, radiation effects, three-body collisions, and ionized flow effects, without major modifications to the basic DSMC procedure [Oran et al., 1998]. However, a priori knowledge of the accommodation coefficients must be used in this process. Hence, this constitutes a weakness of the DSMC method similar to the Navier–Stokes-based slip and even Boltzmann equation-based simulation models. The following section discusses this issue in detail.

The second process is the indexing and tracking of the particles. This is necessary because the molecules might have moved to new cell locations during the first stage. The new cell location of the molecules is indexed, and thus the intermolecular collisions and flow field sampling can be handled accurately. This is a crucial step for an efficient DSMC algorithm. The indexing, molecule tracking, and data structuring algorithms should be carefully designed for the specific computing platforms, such as vector super computers and workstation architectures.

The third step is simulation of collisions via a probabilistic process. Because only a small portion of the molecules is simulated and the motion and collision processes are decoupled, probabilistic treatment becomes necessary. A common collision model is the no-time-counter technique of Bird (1994) that is used in conjunction with the subcell technique where the collision rates are calculated within the cells and the collision pairs are selected within the subcells. This improves the accuracy of the method by maintaining the collisions of the molecules with their closest neighbors [Oran et al., 1998].

The last process is the calculation of appropriate macroscopic properties by the sampling of molecular (microscopic) properties within a cell. The macroscopic properties for unsteady flow conditions are obtained by the ensemble average of many independent calculations. For steady flows, time averaging is also used.

6.2.3 Limitations, Error Sources, and Disadvantages of the DSMC Approach

Following the work of Oran et al., (1998), we identify several possible limitations and error sources within a DSMC method.

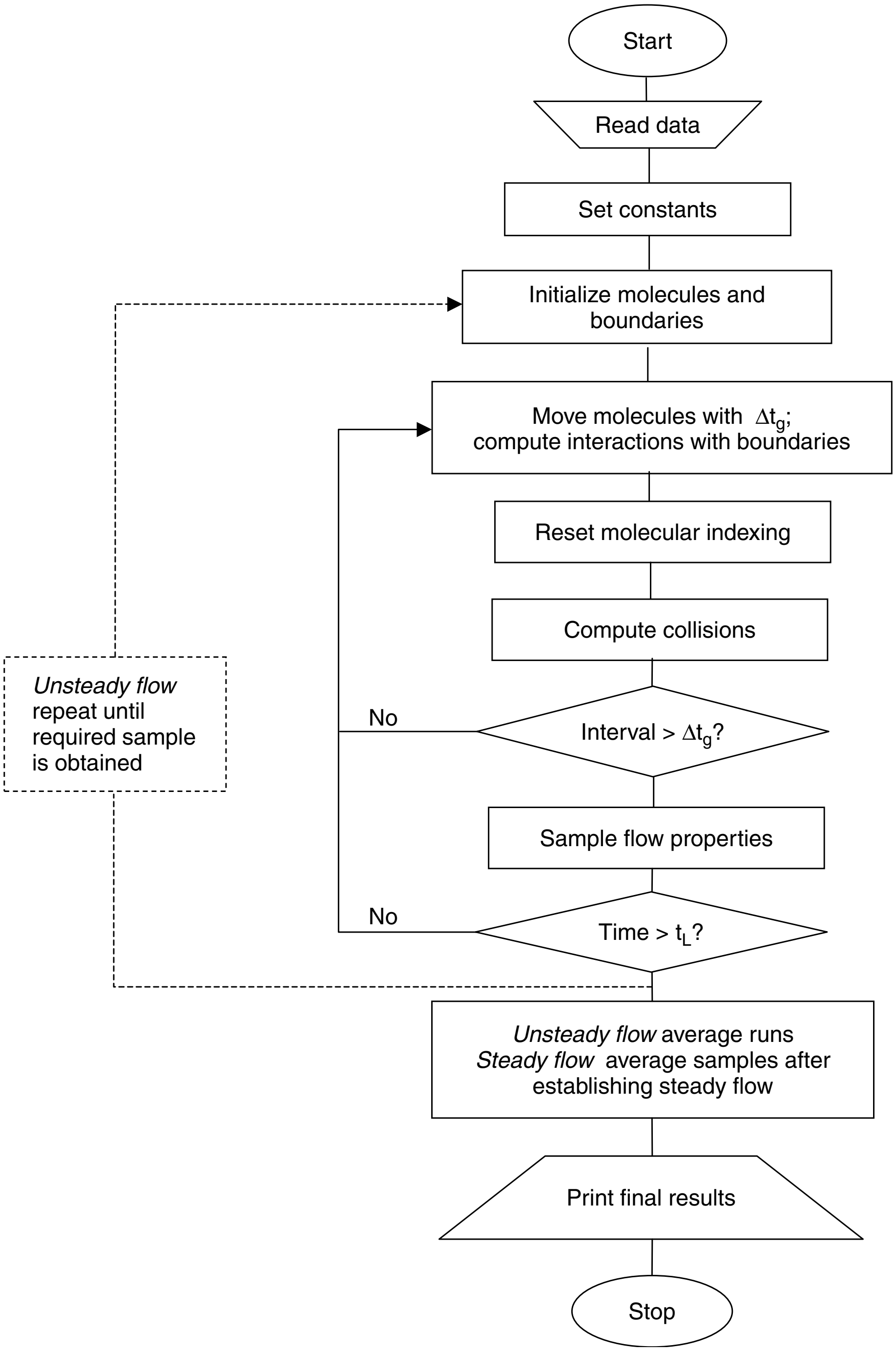

Start
Read data
Set constants
Initialize molecules and boundaries
Move molecules with Δt_g; compute interactions with boundaries
Reset molecular indexing
Compute collisions
Unsteady flow repeat until required sample is obtained
No
Interval > Δt_g?
Sample flow properties
No
Time > t_L?
Unsteady flow average runs
Steady flow average samples after establishing steady flow
Print final results
Stop

FIGURE 6.3 Typical steps for a DSMC method. (Reprinted with permission from Oran, E.S. et al. [1998] "Direct Simulation Monte Carlo: Recent Advances and Applications," *Ann. Rev. Fluid Mech.* **30**, 403–441.)

1. Finite cell size: the typical DSMC cell should be about one-third of the local mean free path. Values of cell sizes larger than this may result in enhanced diffusion coefficients. In DSMC, one cannot directly specify the dynamic viscosity and thermal conductivity of the fluid. The dynamic viscosity is calculated via diffusion of linear momentum. Breuer et al. (1995) performed one-dimensional Rayleigh flow problems in the continuum flow regime and showed that for cell sizes larger than one mean free path the apparent viscosity of the fluid was increased. Some numerical experimentation details for this finding are also given in Beskok (1996). More recently, the viscosity and thermal

conductivity dependence on cell size have been obtained more systematically by using the Green–Kubo theory [Alexander et al., 1998; Hadjiconstaninou, 2000].

2. Finite time step: due to the time splitting of the molecular motion and collisions, the maximum allowable time step is smaller than the local collision time t_c. Values of time steps larger than t_c result in molecules traveling through several cells prior to a cell-based collision calculation.

   The time-step and cell-size restrictions presented in items 1 and 2 are not a Courant–Friederichs–Lewy (CFL) stability condition. The DSMC method is always stable. However, overlooking the physical restrictions stated in items 1 and 2 may result in highly diffusive numerical results.
3. Ratio of the simulated particles to the real molecules: due to the vast number of molecules and limited computational resources, one always has to choose a sample of molecules to simulate. If the ratio of the actual molecules to the simulated molecules gets too high, the statistical scatter of the solution is increased. The details for the statistical error sources and the corresponding remedies can be found in Oran et al. (1998), Bird (1994) and Chen and Boyd (1996). A relatively well-resolved DSMC calculation requires a minimum of 20 simulated particles per cell.
4. Boundary condition treatment: the inflow–outflow boundary conditions can become particularly important in a microfluidic simulation. A subsonic microchannel flow simulation may require specification of inlet and exit pressures. The flow will develop under this pressure gradient and result in a certain mass flow rate. During such simulations, specification of back pressure for subsonic flows is challenging. In the DSMC studies, one can simulate the entry problem to the channels by specifying the number density, temperature, and average macroscopic velocity of the molecules at the inlet of the channel. At the outflow, the number density and temperature corresponding to the desired back pressure can be specified. This and similar treatments facilitate significantly reducing the spurious numerical boundary layers at inflow and outflow regions. For high Knudsen number flows (i.e., $Kn > 1$) in a channel with blockage (such as a sphere in a pipe), the location of the inflow and outflow boundaries is important. For example, the molecules reflected from the front of the body may reach the inflow region with very few intermolecular collisions, creating a diffusing flow at the front of the bluff body [Liu et al., 1998]. (The details of this case are presented in Section 6.2.4.)
5. Uncertainties in the physical input parameters: these typically include the input for molecular collision cross-section models, such as the hard sphere (HS), variable hard sphere (VHS), and variable soft sphere (VSS) models [Oran et al., 1998; Vijayakumar et al., 1999]. The HS model is usually sufficient for monatomic gases or for cases with negligible vibrational and rotational nonequilibrium effects, such as in the case of nearly isothermal flow conditions.

Along with these possible error sources and limitations, some particular disadvantages of the DSMC method for simulation of gas microflows are:

1. Slow convergence: the error in the DSMC method is inversely proportional to the square root of the number of simulated molecules. Reducing the error by a factor of two requires increasing the number of simulated molecules by a factor of four. This is a very slow convergence rate compared to the continuum-based simulations with spatial accuracy of second or higher order. Therefore, continuum-based simulation models should be preferred over the DSMC method whenever possible.
2. Large statistical noise: gas microflows are usually low subsonic flows with typical speeds of 1 mm/sec to 1 m/sec (exceptions to this are the micronozzles utilized in synthetic jets and satellite thruster control applications). The macroscopic fluid velocity is obtained by time or ensemble averaging of the molecular velocities. This difference of five to two orders of magnitude between the molecular and average speeds results in large statistical noise and requires a very long time averaging for gas microflow simulations. The statistical fluctuations decrease with the square root of the sample size. Time or ensemble averages of low-speed microflows on the order of 0.1 m/sec require about 108 samples in order to distinguish such small macroscopic velocities. Fan and Shen (1999) introduced the information preservation (IP) technique for the DSMC method, which enables efficient DSMC simulations for low-speed flows (the IP scheme is briefly covered in Section 6.2.5).

3. Extensive number of simulated molecules: if we discretize a rectangular domain of 1 mm × 100 µm × 1 µm under standard conditions for $Kn = 0.065$ flow, we will need at least 20 cells per micrometer length scale. This results in a total of 8×10^8 cells. Each of these cells should contain at least 20 simulated molecules, resulting in a total of 1.6×10^{10} particles. Combined with the number of time-step restrictions, simulation of low-speed microflows with DSMC easily exceeds the capabilities of current computers. An alternative treatment to overcome the extensive number of simulated molecules and long integration times is utilization of the dynamic similarity of low-pressure rarefied gas flows to gas microflows under atmospheric conditions. The key parameters for the dynamic similarity are the geometric similarity and matching of the flow Knudsen, Mach, and Reynolds numbers. Performing actual experiments under dynamically similar conditions may be very difficult; however, parametric studies via numerical simulations are possible. The fundamental question to answer for such an approach is whether or not a specific, unforeseen microscale phenomenon is missed with the dynamic similarity approach. In response to this question, all numerical simulations are inherently model based. Unless microscale-specific models are implemented in the algorithm, we will not be able to obtain more physical information from a microscopic simulation than from a dynamically similar low-pressure simulation. One limitation of the dynamic similarity concept is the onset of statistical scatter in the instantaneous macroscopic flow quantities for gas microflows for $Kn > 1$ (see section 6.2.1 and Figure 6.2 for details). Here, we must also remember that DSMC utilizes time or ensemble averages to sample the macroscopic properties from the microscopic variables. Hence, DSMC already determines the macroscopic properties in an averaged sense.
4. Lack of deterministic surface effects: Molecule wall interactions are specified by the accommodation coefficients σ_v, σ_T. For diffuse reflection $\sigma = 1$, and the reflected molecules lose their incoming tangential velocity while being reflected with the tangential wall velocity. For $\sigma = 0$ the tangential velocity of the impinging molecules is not changed during the molecule/wall collisions. For any other value of σ, a combination of these procedures can be applied. The molecule–wall interaction treatment implemented in DSMC is more flexible than the slip conditions given by Equations (6.1) and (6.2). However, it still requires specification of the accommodation coefficients, which are not known for any gas surface pair with a specified surface root mean square (rms) roughness. The tangential momentum accommodation coefficients for helium, nitrogen, argon and carbon dioxide on single-crystal silicon were measured by careful microchannel experiments [Arkilic, 1997].

6.2.4 Some DSMC-Based Gas Microflow Results

This section presents some DSMC results applied to gas microflows.

6.2.4.1 Microchannel Flows

The DSMC simulation results for subsonic gas flows in microchannels are presented in this section. Due to the computational difficulties explained in the previous sections, a low-aspect-ratio, two-dimensional channel with relatively high inlet velocities is studied. The results presented in the figures are for microchannels with a length-to-height ratio (L/h) of 20 under various inlet-to-exit-pressure ratios. The DSMC results are performed with 24,000 cells, of which 400 cells were in the flow direction and 60 cells were across the channel. A total of 480,000 molecules are simulated. The results are sampled (time averaged) for 10^5 times, and the sampling is performed every ten time steps.

In the following simulations, diffuse reflection ($\sigma_v = 1.0$) is assumed for interaction of gas molecules with the surfaces. Because the slip amount can be affected significantly by small variations in σ_v (Equation [6.1]), the apparent value of the accommodation coefficient σ_v is monitored throughout the simulations by recording the tangential momentum of the impinging (τ_i) and reflected (τ_r) gas molecules. Based on these values, the apparent tangential momentum accommodation coefficient, $\sigma_v = (\tau_i - \tau_r)/(\tau_i - \tau_w) = 0.99912$ with standard deviation of $\sigma_{rms} = 0.01603$, is obtained.

The velocity profiles normalized with the corresponding average speed are presented in Figure 6.4 for pressure-driven microchannel flows at $Kn = 0.1$ and $Kn = 2.0$. The figure also presents the molecule/cell

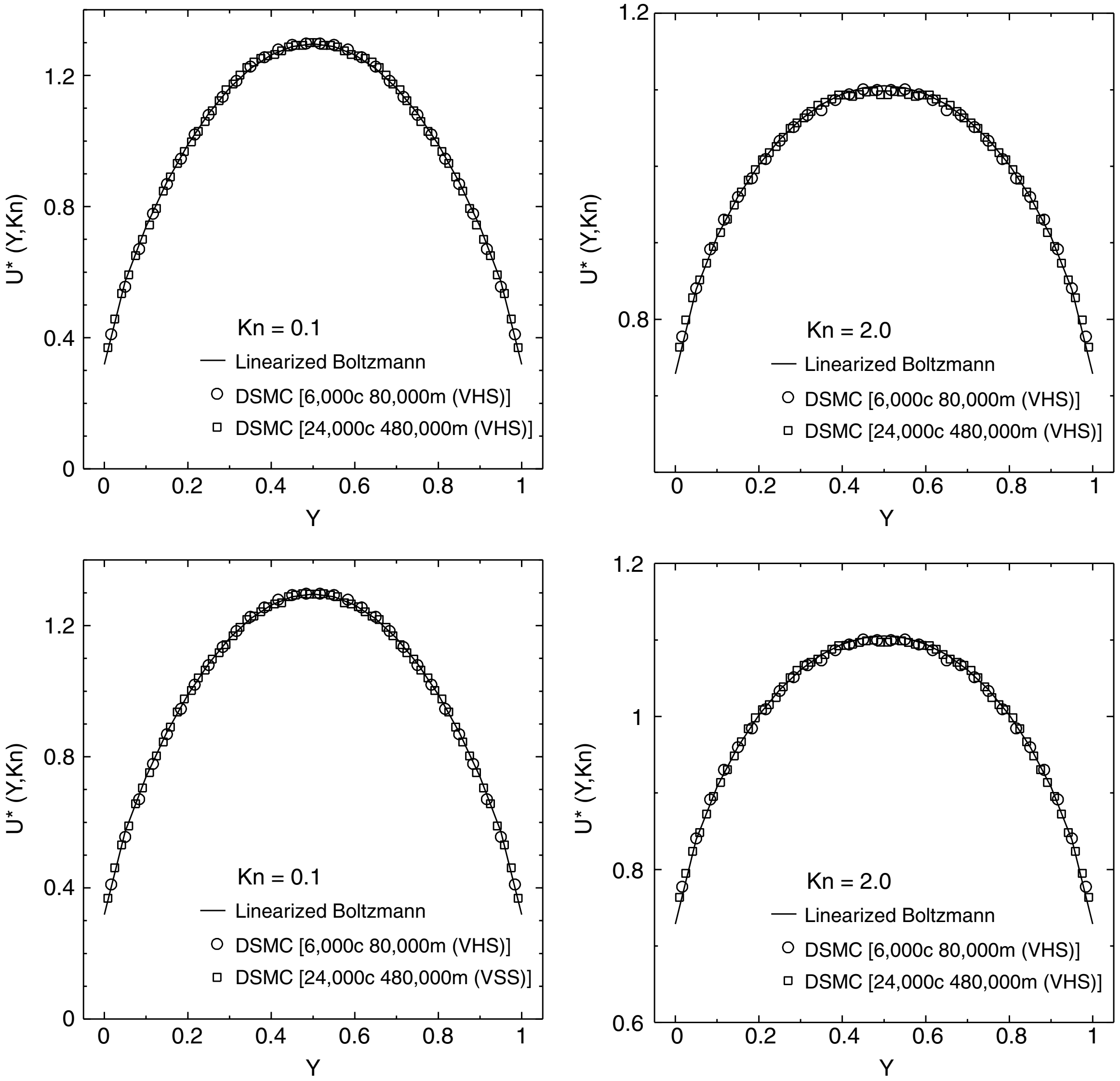


FIGURE 6.4 Velocity profiles normalized with the local average velocity in the slip and transitional flow regimes. The DSMC predictions with the VHS and VSS models agree well with the linearized Boltzmann solutions of Ohwada et al. (1989). The number of cells and simulated molecules are identified on each figure.

refinement studies as well as predictions of the VHS and VSS models. The DSMC results are compared against the linearized Boltzmann solutions [Ohwada et al., 1989], and excellent agreements of the VHS and VSS models with the linearized Boltzmann solutions are observed for these nearly isothermal flows. In regard to the molecule/cell refinement study, the number of cells and the number of simulated molecules are identified for each case. The first VHS case utilized only 6000 cells with 80,000 simulated molecules, and the results are sampled about 5×10^5 times. Sampling is performed every 20 time steps. The refined VHS and VSS cases utilized 24,000 cells and a total of 480,000 molecules. The results for these are sampled 10^5 times, every ten time steps. Although the velocity profiles for the low-resolution case (6000 cells) seem acceptable, the density and pressure profiles show large fluctuations.

The DSMC and μFlow (spectral-element-based, continuum computational fluid dynamics [CFD] solver) predictions of density and pressure variations along a pressure-driven microchannel flow are shown in Figure 6.5. For this case, the ratio of inlet to exit pressure is $\Pi = 2.28$, and the Knudsen number at the channel outlet is 0.2. Deviations of the slip flow pressure distribution from the no-slip solution are also presented in the figure. Good agreements between the DSMC and μFlow simulations are achieved.

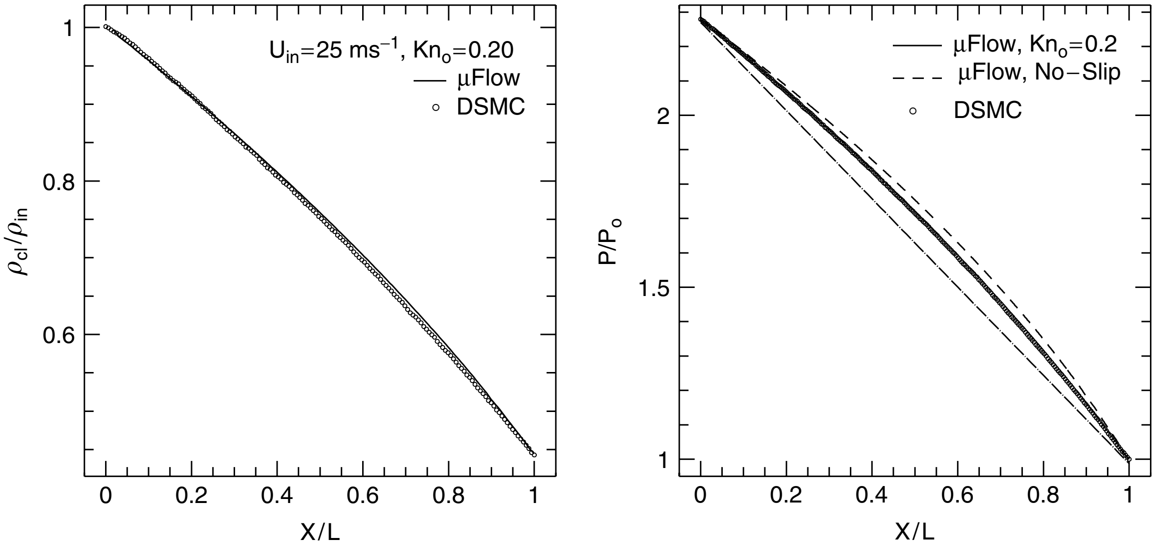


FIGURE 6.5 Density (left) and pressure (right) variation along a microchannel. Comparisons of the Navier–Stokes and DSMC predictions for ratio of inlet to exit pressure of $\Pi = 2.28$ and $Kn_o = 0.20$. (Reprinted with permission from Beskok, A. [1996] Ph.D. thesis, Princeton University.)

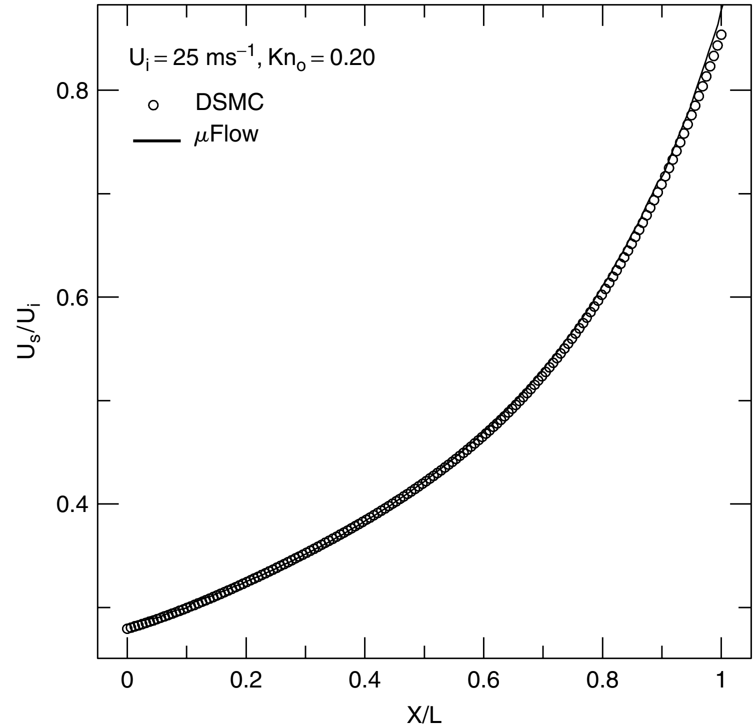


FIGURE 6.6 Wall slip velocity variation along a microchannel predicted by Navier–Stokes and DSMC simulations. (Reprinted with permission from Beskok, A. [1996] Simulations and Models for Gas Flows in Microgeometries, Ph.D. thesis, Princeton University.)

The curvature in the pressure distribution is due to the compressibility effect, and the rarefaction negates this curvature, as seen in Figure 6.5. The slip velocity variation on the channel wall is shown in Figure 6.6. Overall good agreements between both methods are observed. Pan et al. (1999) used the DSMC simulations to determine the slip distance as a function of various physical conditions such as the number density,

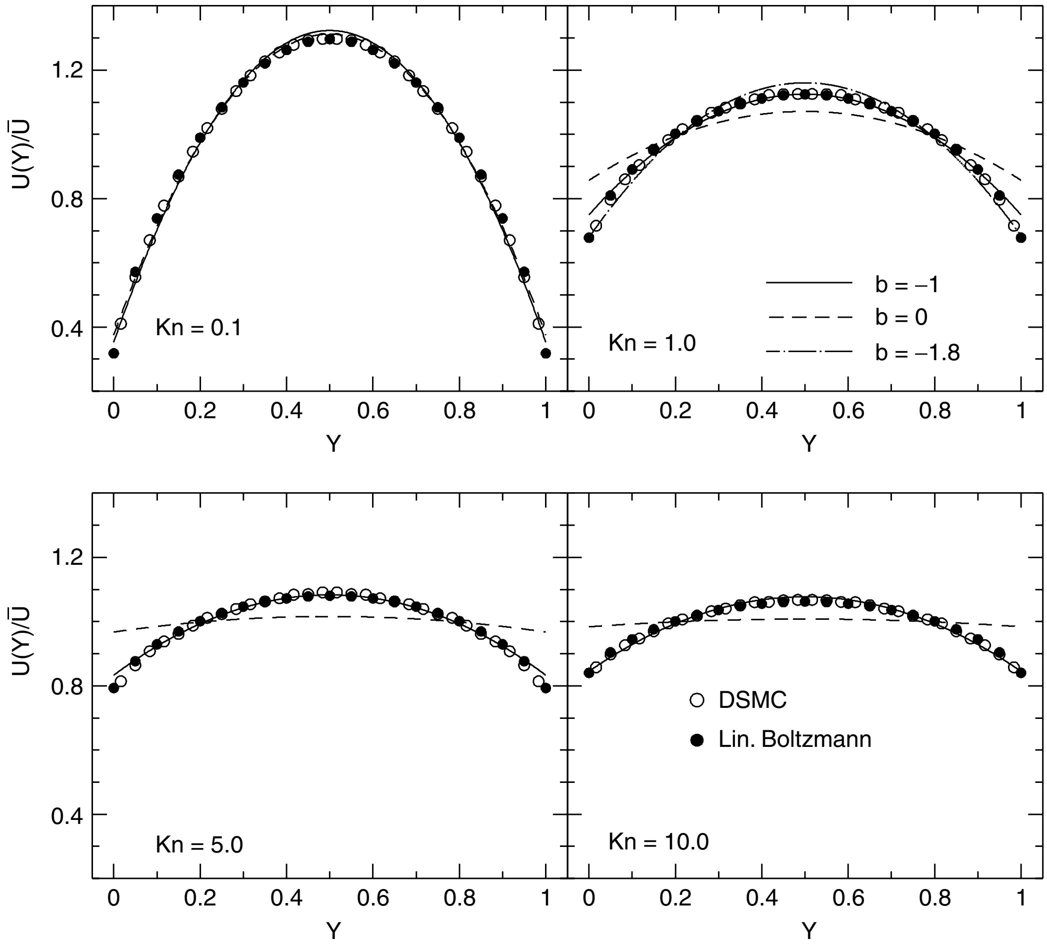


FIGURE 6.7 Comparison of the velocity profiles obtained by new slip model Equations (6.3) and (6.8) with DSMC and linearized Boltzmann solutions [Ohwada et al., 1989]. Maxwell's first-order boundary condition is shown by the dashed lines ($b = 0$), and the general slip boundary condition ($b = -1$) is shown by solid lines. (Reprinted with permission from Beskok, A., and Karniadakis, G.E. [1999] "A Model for Flows in Channels, Pipes, and Ducts at Micro and Nano Scales," *Microscale Thermophys. Eng.* 3, pp. 43–77. Reproduced with permission of Taylor & Francis, Inc.)

wall temperature, and the gas mass. They determined that an appropriate slip distance is 1.125 λ_{gw}, where the subscript *gw* indicates the gas-wall conditions [Pan et al., 1999].

In the transitional flow regime, Beskok and Karniadakis (1999) studied the Burnett equations for low-speed isothermal flows. This analysis has shown that the velocity profiles remain parabolic even for large *Kn* flows. To verify this hypothesis, they performed several DSMC simulations; the velocity distribution nondimensionalized with the local average speed is shown in Figure 6.7. They also obtained an approximation to this nondimensionalized velocity distribution in the following form:

$$U^* \,(y,Kn) \equiv U(x,y)/\overline{U}(x) = \left[\frac{-\left(\dfrac{y}{h}\right)^2 + \dfrac{y}{h} + \dfrac{Kn}{1-bKn}}{\dfrac{1}{6} + \dfrac{Kn}{1-bKn}} \right] \tag{6.8}$$

where the extended slip condition given in Equation (6.3) is used. In the above relation, the value of $b = -1$ is determined analytically for channel and pipe flows [Beskok and Karniadakis, 1999]. In Figure 6.7, the nondimensional velocity variation obtained in a series of DSMC simulations for $Kn = 0.1$, $Kn = 1.0$, $Kn = 5.0$, and $Kn = 10.0$ flows are presented along with the corresponding linearized Boltzmann solutions [Ohwada et al., 1989]. The DSMC velocity distribution and the linearized Boltzmann solutions agree quite well. One can use Equation (6.8) to compare the results with the DSMC/linearized Boltzmann data by varying the parameter b. The case $b = 0$ corresponds to Maxwell's first-order slip model, and $b = -1$ corresponds to Beskok's second-order slip boundary condition. It is clear from Figure 6.7 that Equation (6.8) with $b = -1$ results in a uniformly valid representation of the velocity distribution in the entire Knudsen regime.

The nondimensionalized centerline and wall velocities for $0.01 \leqslant Kn \leqslant 30$ flows are shown in Figure 6.8. The figure includes the data for the slip velocity and centerline velocity from 20 different DSMC runs, of which 15 are for nitrogen (diatomic molecules) and 5 for helium (monatomic molecules). The differences between the nitrogen and helium simulations are negligible; thus, this velocity scaling is independent of the gas type. The linearized Boltzmann solutions [Ohwada et al., 1989] for a monatomic gas are also indicated by triangles in Figure 6.8. The Boltzmann solutions closely match the DSMC predictions. Maxwell's first-order boundary condition $b = 0$ (shown by solid lines) predicts, erroneously, a uniform nondimensional velocity profile for large Knudsen numbers. The breakdown of the slip flow theory based on the first-order slip-boundary conditions is realized around $Kn = 0.1$ and $Kn = 0.4$ for wall and centerline velocities respectively. This finding is consistent with the commonly accepted limits of the slip flow regime [Schaaf and Chambre, 1961]. The prediction using $b = -1$ is shown by small dashed lines. The corresponding centerline velocity closely follows the DSMC results, while the slip velocity of the model with $b = -1$ deviates from DSMC in the intermediate range for $0.1 < Kn < 5$. One possible reason for this is the effect of the Knudsen layer. For small Kn flows, the Knudsen layer is thin and does not affect the overall velocity distribution too much. For very large Kn flows, the Knudsen layer covers the channel entirely. For intermediate Kn values, however, both the fully developed viscous flow and the Knudsen layer coexist in the channel. At this intermediate range, approximating the velocity profile as parabolic ignores the Knudsen layers. For this reason, the model with $b = -1$ results in 10% error in the slip velocity

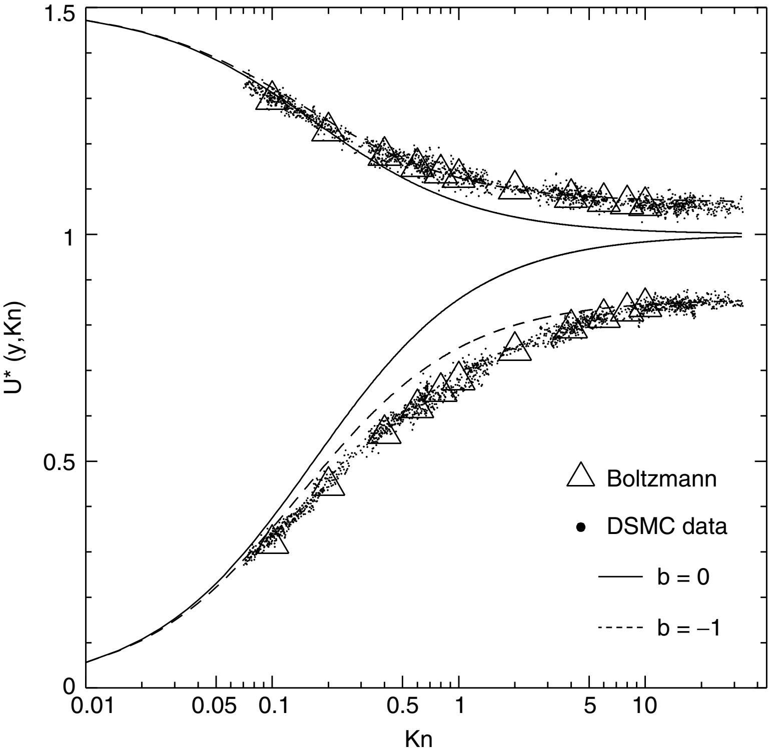


FIGURE 6.8 Centerline and wall slip velocity variations in the entire Knudsen regime. The linearized Boltzmann solutions of Ohwada et al. (1989) are shown by triangles, and the DSMC simulations are shown by closed circles. Theoretical predictions of the velocity scaling obtained by Equation (6.8) are shown for different values of b. The $b = 0$ case corresponds to Maxwell's first-order boundary condition, and $b = -1$ corresponds to the general slip-boundary condition.

at $Kn = 1$. However, the velocity distribution in the rest of the channel is described accurately for the entire flow regime. Based on these results, Beskok and Karniadakis (1999) developed a unified flow model that can predict the velocity profiles, pressure distribution, and mass flow rate in channels, pipes, and arbitrary aspect-ratio rectangular ducts in the entire Knudsen regime, including Knudsen's minimum effects [Beskok and Karniadakis, 1999; Kennard, 1938; Tison, 1993].

6.2.4.2 Separated Rarefied Gas Flows

Gas flows through complex microgeometries are prone to flow separation and recirculation. Most of the DSMC-based microflow analyses were performed in straight channels [Mavriplis et al., 1997; Oh et al., 1997] and for smooth microdiffusers [Piekos and Breuer, 1996]. Nance et al. (1997) discuss the Monte Carlo simulation for MEMS devices. The mainstream approach for gas flow modeling in MEMS is solution of the Navier–Stokes equations with slip models. This is more practical and numerically efficient than utilization of the DSMC method. However, rarefied separated gas flows are not studied extensively. To investigate the validity of slip-boundary conditions under severe adverse pressure gradients and separation, Beskok and Karniadakis (1997) performed a series of numerical simulations using the classical backward-facing step geometry with a step-to-channel-height ratio of 0.467. The variations of pressure and streamwise velocity along a step microchannel, obtained at five different cross-flow locations (y/h), are presented in Figure 6.9. The values of pressure and velocity are nondimensionalized with the corresponding freestream dynamic head and the local sound speed respectively. The specific y/h locations are selected to coincide with the DSMC cell centers to avoid interpolations or extrapolations of the DSMC method. The results show reasonable agreements of the slip-based Navier–Stokes simulations with the DSMC data.

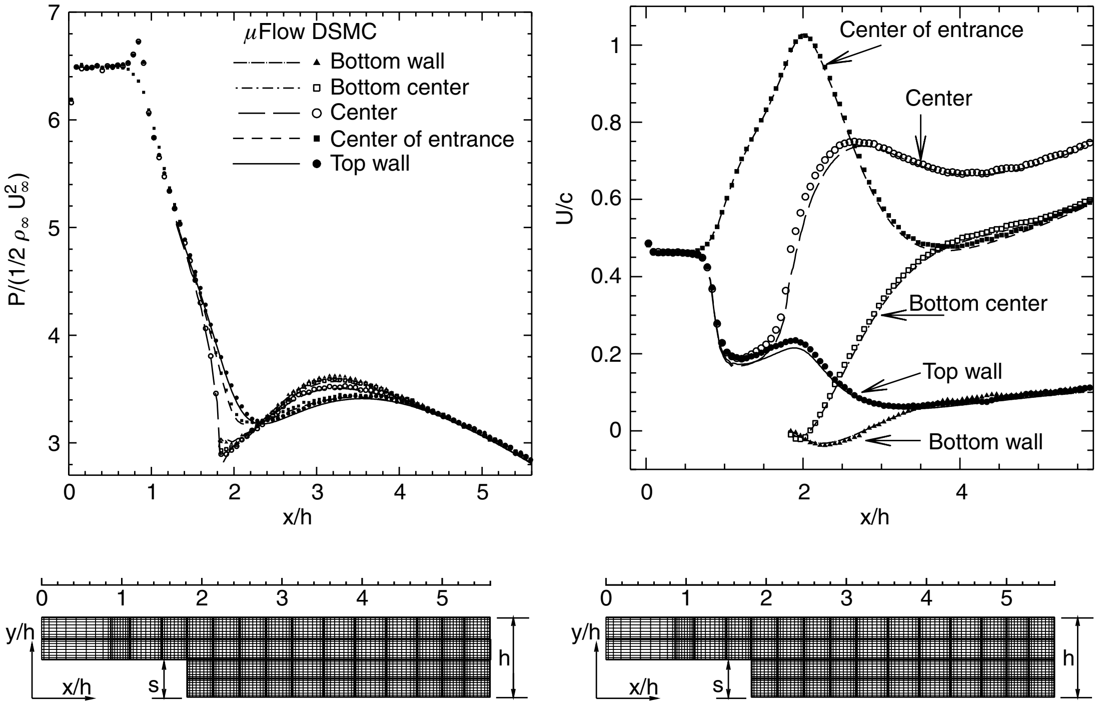


FIGURE 6.9 Pressure (left) and streamwise velocity (right) distribution along a backward-facing step channel at five selected locations. Predictions of both DSMC (symbols) and continuum-based spectral element CFD code (lines) are presented. A tenth-order spectral element grid is also shown. The top wall is at $y/h = 0.98325$; the center of the entrance is at $y/h = 0.75$; the center is at $y/h = 0.48$; the bottom center is at $y/h = 0.25$; and the bottom wall is at $y/h = 0.017$. The simulation conditions are for $Re = 80$, $Kn_{\text{out}} = 0.04$, $M_{\text{in}} = 0.45$ and $Pr = 0.7$. (Reprinted with permission from Beskok, A., and Karniadakis, G.E. (1997) "Modeling Separation in Rarefied Gas Flows," 28th AIAA Fluid Dynamics Conf., AIAA 97-1883, June 29–July 2. Copyright © 1997 by the American Institute of Aeronautics and Astronautics, Inc.)

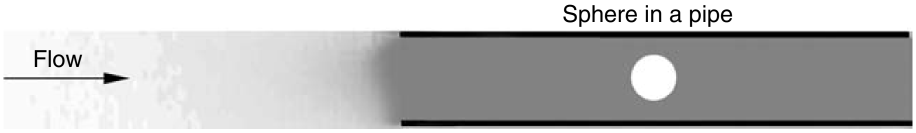


FIGURE 6.10 Velocity contours for a sphere in a pipe in the transitional flow regime ($Kn = 3.5$). Molecules reflected from the sphere create a diffusive layer at the entrance of the pipe [Liu et al., 1998].

The flow recirculation and reattachment location at the bottom wall are predicted equally well with both methods. The DSMC simulations utilized 28,000 cells (700×40) with 420,000 simulated molecules. The solution is sampled 10^5 times. The continuum-based simulations are performed by 52 spectral elements with tenth-order polynomial expansions for each direction.

6.2.4.3 Transitional Flow Past a Sphere in a Pipe

The DSMC simulations of high *Kn* rarefied flows at the entry of channels or pipes show diffusion of the molecules from the entry toward the free-stream region. To demonstrate this counterintuitive effect, Liu et al. (1998) simulated flow past a sphere in a pipe with diffuse reflection from the surfaces. To incorporate the molecules diffusing out from the entry of the pipe, the computational domain for the free-stream region had to be extended more than expected. In high Knudsen number subsonic flows, the molecules reflected from the sphere can travel toward the pipe inlet with very few intermolecular collisions and then diffuse out. Figure 6.10 presents the velocity contours for $Kn = 3.5$ flow. Diffusion of the molecules toward the inflow can be identified easily from the velocity contours. This effect was studied earlier by Kannenberg and Boyd (1996) for transitional flow entering a channel. For $Kn = 3.5$ results presented in Figure 6.10, the length of the free-stream region is equal to the length of the pipe; hence, the computational cost is increased significantly.

6.2.5 Recent Advances in the DSMC Method

This section presents recent developments in the application and implementation of the DSMC method.

6.2.5.1 Information Preservation DSMC Scheme

Fan and Shen (1999) developed an information preservation (IP) DSMC scheme for low-speed rarefied gas flows. Their method uses the molecular velocities of the DSMC method as well as an information velocity that records the collective velocity of an enormous number of molecules that a simulated particle represents. The information velocity evolves with inelastic molecular collisions, and the results presented for Couette, Poiseuille, and Rayleigh flows in the slip, transition, and free molecular regimes show very good agreements with the corresponding analytical solutions. This approach seems to decrease the sample size and correspondingly the CPU time required by a regular DSMC method for low-speed flows by orders of magnitude. This is a tremendous gain in computational time that can lead to the effective use of IP DSMC schemes for microfluidic and MEMS simulations. The IP DSMC schemes are being validated in two-dimensional, complex-geometry flows, and extensions of the IP technique for three-dimensional flows are also being developed [Cai et al., 2000].

6.2.5.2 DSMC with Moving Boundaries

Some microflow applications require numerical simulation of moving surfaces. In continuum-based approaches, arbitrary Lagrangian Eulerian (ALE) algorithms are successfully utilized for such applications [Beskok and Warburton, 2000a, 2000b]. A similar effort to expand the DSMC method for grid adaptation, including the moving external and internal boundaries, combined the DSMC method with a monotonic Lagrangian grid (MLG) method [Cybyk et al., 1995; Oran et al., 1998].

6.2.5.3 Parallel DSMC Algorithms

Because the DSMC calculations involve vast numbers of molecules, using parallel algorithms with efficient interprocessor communications and load balancing can have a significant impact on the effectiveness of

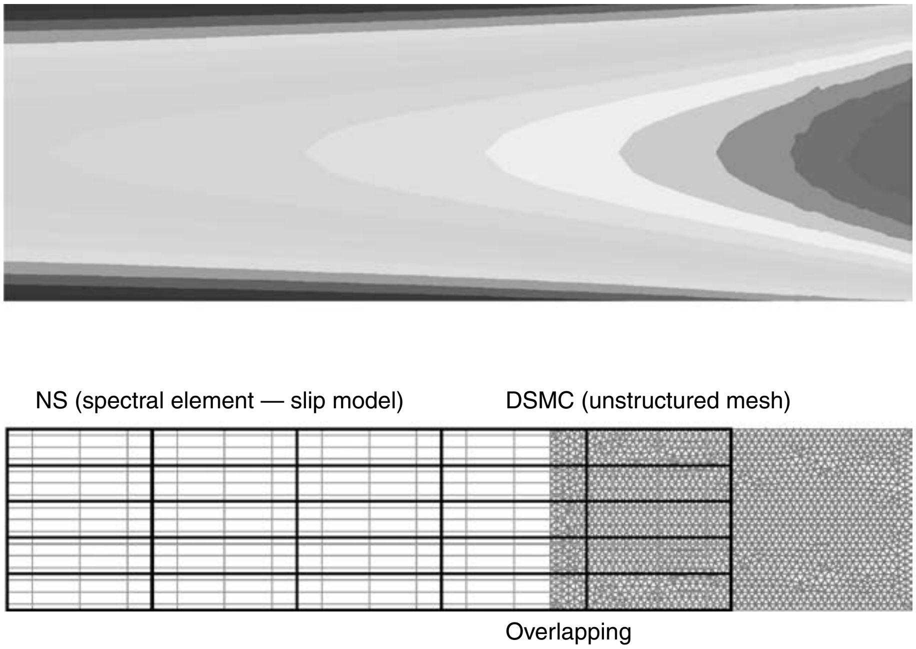


FIGURE 6.11 Streamwise velocity contours for rarefied gas flow in mixed slip/transitional regimes, obtained by a coupled DSMC/continuum solution method. Most of the channel is in the slip flow regime, and a spectral element method μFlow is utilized to solve the compressible Navier–Stokes equations with slip. The rest of the channel is in the transitional flow regime, where a DSMC method with unstructured cells is utilized. (Reprinted with permission from Liu, H.F. [2001] 2D and 3D Unstructured Grid Simulation and Coupling Techniques for Micro-Geometries and Rarefied Gas Flows, Ph.D. thesis, Brown University.)

simulations. Developments in parallel DSMC algorithms are addressed by Oran et al. (1998). For example, Dietrich and Boyd (1996) were able to obtain 90% parallel efficiency with 400 processors, simulating over 100 million molecules on a 400-node IBM SP2 computer. The computing power of their code is comparable to 75 single-processor Cray C90 vector computers. Good parallel efficiencies for DSMC algorithms can be achieved with effective load-balancing methods based on the number of molecules. This is because the computational work of the DSMC method is proportional to the number of simulated molecules.

6.2.6 DSMC Coupling with Continuum Equations

This section provides an overview of the DSMC/Navier–Stokes and DSMC/Euler equation coupling strategies. These equations are particularly important for simulation of gas flows in MEMS components. If we consider a micromotor or a micro-comb-drive mechanism, gas flow in most of the device can be simulated using the slip-based continuum models. The DSMC method should be utilized only when the gap between the surfaces becomes submicron or when the local $Kn > 0.1$. Hence, it is necessary to implement multidomain DSMC/continuum solvers. Depending on the specific application, hybrid Euler/DSMC [Roveda et al., 1998] or DSMC/Navier–Stokes algorithms [Hash and Hassan, 1995] can be used. Such hybrid methods require compatible kinetic-split fluxes for the Navier–Stokes portion of the scheme [Chou and Baganoff, 1997; Lou et al., 1998] to achieve an efficient coupling. An adaptive mesh and algorithm refinement (AMAR) procedure that embeds a DSMC-based particle method within a continuum grid has been developed; it enables molecular-based treatments even within the continuum region [Garcia et al., 1999]. Hence, the AMAR procedure can be used to deliver microscopic and macroscopic information within the same flow region.

Simulation results for a Navier–Stokes/DSMC coupling procedure obtained by Liu (2001) are shown in Figure 6.11. A structured spectral element algorithm, μFlow [Beskok, 1996], is coupled with an unstructured DSMC method, UDSMC 2-D, with an overlapping domain. Both the grid and the streamwise velocity contours are shown in the figure; smooth transition of the velocity contours from the continuum-based slip region to the DSMC region can be observed. The details of the coupling procedure are given in Liu (2001).

6.2.7 Boltzmann Equation Research

Microscale thermal/fluidic transport in the entire Knudsen regime ($0 \leqslant Kn < \infty$) is governed by the Boltzmann equation (BE). The Boltzmann equation describes the evolution of a velocity distribution function by molecular transport and binary intermolecular collisions. The assumption of binary intermolecular collisions is a key limitation in the Boltzmann formulation making it applicable for dilute gases only. The Boltzmann equation for a simple dilute gas is given in the following form [Bird, 1994]:

$$\frac{\partial nf}{\partial t} + \vec{c} \cdot \frac{\partial nf}{\partial \vec{x}} + \vec{F} \cdot \frac{\partial nf}{\partial \vec{c}} = \int_{-\infty}^{\infty} \int_{0}^{4\pi} n^2 (f^* f_1^* - ff_1) c_r \sigma \, d\Omega \, d\vec{c}_1 \tag{6.9}$$

where f is the velocity distribution function, n is the number density, $\vec{c}$ is the molecular velocity, $\vec{F}$ is the external force per unit mass, c_r is the relative speed of class molecules with respect to class $\vec{c}_1$ molecules, and σ is the differential collision cross-section. The definitions of terms in Equation (6.9) follow. The first term is the rate of change of the number of class $\vec{c}_1$ molecules in the phase space. The second term shows convection of molecules across a fluid volume by molecular velocity $\vec{c}$. The third term is convection of molecules across the velocity space as a result of the external force $\vec{F}$. The fourth term is the binary collision integral. The term $(-ff_1)$ describes the collision of molecules of class $\vec{c}$ with molecules of class $\vec{c}_1$ (resulting in creation of molecules of class $\vec{c}^*$ and $\vec{c}_1^*$, respectively), and it is known as the loss term. Similarly, in inverse collisions class $\vec{c}^*$ molecules collide with class $\vec{c}_1^*$ molecules creating class $\vec{c}$ and $\vec{c}_1$ molecules. This is shown by $f^* f_1^*$, known as the gain term. Assuming binary elastic collisions enables us to determine class $\vec{c}^*$ and $\vec{c}_1^*$ conditions [Bird, 1994]. The difficulty of the Boltzmann equation arises due to the nonlinearity and complexity of the collision integral terms and the multidimensionality of the equation (x, c, t).

Current numerical methods, which are usually very expensive, are applied for simple geometries, such as pipes and channels. In particular, a number of investigators have considered semianalytical and numerical solutions of the linearized Boltzmann equation to be valid for flows with small pressure and temperature gradients [Aoki, 1989; Huang et al., 1966; Loyalka and Hamoodi, 1990; Ohwada et al., 1989; Sone, 1989]. These studies used HS and Maxwellian molecular models. Simplifications for the collision integral based on the BGK model [Bhatnagar et al., 1954] are used in the Boltzmann equation studies. The BGK model for a rarefied gas with no external forcing is given as:

$$\frac{\partial nf}{\partial t} + \vec{c} \cdot \frac{\partial nf}{\partial \vec{x}} = \nu n (f_o - f) \tag{6.10}$$

where ν is the collision frequency and f_o is the local Maxwellian (equilibrium) distribution. The right-hand side of Equation (6.10) becomes zero when the flow is in local equilibrium (continuum flow) or when the collision frequency goes to zero (corresponding to the free molecular flow). The BGK model captures both limits correctly. However, there are justified concerns about the validity of the BGK model in the transition flow regime. A model's ability to capture the two asymptotic limits ($Kn \rightarrow 0$ and $Kn \rightarrow \infty$) is not necessarily sufficient for its accuracy in the intermediate regimes [Bird, 1994].

Veijola et al. (1995, 1998) presented a Boltzmann equation analysis of silicon accelerometer motion and squeeze-film damping as a function of the Knudsen number and the time-periodic motion of the surfaces. Although the mixed compressibility and rarefaction effects make the squeeze-film damping analysis very challenging, it has many practical applications including computer disk hard drives, microaccelerometers, and noncontact gas buffer seals [Fukui and Kaneko, 1988, 1990]. Saripov and Seleznev (1998) give a comprehensive theory of internal rarefied gas flows including the numerical simulation data. See this article for further theoretical and numerical details on the Boltzmann equation.

The wall-boundary conditions for Boltzmann solutions typically use diffuse and mixed diffuse/specular reflections. For diffuse reflection, the molecules reflected from a solid surface are assumed to have reached thermodynamic equilibrium with the surface. Thus, they are reflected with a Maxwellian distribution corresponding to the temperature and velocity of the surface.

6.2.8 Hybrid Boltzmann/Continuum Simulation Methods

Solution of the Navier–Stokes equation is numerically more efficient than solution of the Boltzmann equation; therefore, it is desirable to develop coupled multidomain Boltzmann/Navier–Stokes models for simulation of mixed regime flows in MEMS and microfluidic applications. Because the typical DSMC method for this coupling results in large statistical noise, solution of the Boltzmann equation may be preferred. The hybrid Boltzmann/Navier–Stokes simulation approach can be achieved by calculating the macroscopic fluid properties from the Boltzmann solutions by moment methods [Bird, 1994], and using the kinetic flux-vector splitting procedure of Chou and Baganoff (1997). Another continuum to Boltzmann coupling can be obtained by using local Chapman–Enskog expansions to the BGK equation [Chapman and Cowling, 1970] and evaluating the distribution function for the kinetic region [Jamamato and Sanryo, 1990].

6.2.9 Lattice Boltzmann Methods

Another approach for simulating flows in microscales is the lattice Boltzmann method (LBM), which is based on the solution of the Boltzmann equation on a previously defined lattice structure with simplistic molecular collision rules. Details of the lattice Boltzmann method are given in a review article by Chen and Doolen (1998). The LBM can be viewed as a special finite differencing scheme for the kinetic equation of the discrete-velocity distribution function, and it is possible to recover the Navier–Stokes equations from the discrete lattice Boltzmann equation with sufficient lattice symmetry [Frisch et al., 1986].

The main advantages of the LBM compared to other continuum-based numerical methods include [Chen and Doolen, 1998]:

- The convection operator is linear in the phase space.
- The LBM is able to obtain both compressible and incompressible Navier–Stokes limits.
- The LBM utilizes a minimal set of velocities in the phase space compared to the continuous velocity distribution function of the Boltzmann algorithms.

With these advantages, the LBM has developed significantly within the last decade. The molecular motions for LBM are allowed on a previously defined lattice structure with restriction on molecular velocities to a few values. Particles move to a neighboring lattice location in every time step. Rules of molecular interactions conserve mass and momentum. Successful thermal and hydrodynamic analysis of multiphase flows including real gas effects can also be obtained [He et al., 1998; Luo, 1998; Qian, 1993; Shan and Chen, 1994]. Another useful application of the LBM is for granular flows, which can be expanded to include flow-through microfiltering systems [Angelopoulos et al., 1998; Bernsdorf et al., 1999; Michael et al., 1997; Spaid and Phelan, 1997; Vangenabeek and Rothman, 1996].

Lattice Boltzmann methods have relatively simple algorithms, and they are introduced as an alternative to the solution of the Navier–Stokes equations [Frisch et al., 1986; Qian et al., 1992]. In contrast to the continuum algorithms, which have difficulties in simulating rarefied flows with consistent slip-boundary conditions, the lattice Boltzmann method initially had difficulties in imposing the no-slip-boundary condition accurately. However, this problem has been successfully resolved [Inamuro et al., 1997; Lavallée et al., 1991; Noble et al., 1995; Zou and He, 1997].

Rapid development of the lattice Boltzmann method with relatively simpler algorithms that can handle both the rarefied and continuum gas flows from a kinetic theory point of view and the ability of the method to capture the incompressible flow limit make the LBM a great candidate for microfluidic simulations. The author is not familiar with applications of the lattice Boltzmann method specifically for microfluidic problems.

6.3 Liquid and Dense Gas Flows

Liquids do not have a well-advanced molecular-based theory. Similar limitations also exist for dense gases where simultaneous intermolecular collisions can exist. The stand-alone Navier–Stokes simulations cannot describe the liquid and dense gas flows in submicron-scale conduits. The effects of van der Waals forces between the fluid and the wall molecules and the presence of longer range Coulombic forces and an electrical double layer (EDL) can significantly affect the microscale transport [Ho and Tai, 1998; Gad-el-Hak, 1999]. For example, the streaming potential effect present in pressure-driven flows under the influence of EDL can explain deviations in the Poiseuille number reported in the seminal liquid microflow experiments of Pfahler et al. (1991).

In recent years, there has been increased interest in the development of micropumps and valves with nonmoving components for medical, pharmaceutical, defense, and environmental-monitoring applications. Electrokinetic body forces can be used to develop microfluidic flow control and manipulation systems with nonmoving components. This section briefly reviews continuum equations for electrokinetic phenomena and the electric double layer.

6.3.1 Electric Double Layer and Electrokinetic Effects

The electric double layer is formed due to the presence of static charges on surfaces. Generally, a dielectric surface acquires charges when it is in contact with a polar medium or by chemical reaction, ionization, or ion absorption. For example, when a glass surface is immersed in water, it undergoes a chemical reaction that results in a net negative surface potential [Cummings et al., 1999; Probstein, 1994]. This influences the distribution of ions in the buffer solution. Figure 6.12 shows the schematic view of a solid surface in contact with a polar medium. Here a net negative electric potential is generated on the surface. Due to this surface electric potential, positive ions in the liquid are attracted to the wall; on the other hand, the negative ions are repelled from the wall. This results in redistribution of the ions close to the wall, keeping the bulk of the liquid far away from the wall electrically neutral. The distance from the wall at which the electric potential energy is equal to the thermal energy is known as the Debye length (λ), and this zone is known as the electric double layer. The electric potential distribution within the fluid is described by the Poisson–Boltzmann equation:

$$\nabla^2 (\psi/\zeta) = \frac{-4\pi h^2 \rho_e}{D\zeta} = \beta \sin h(\alpha\psi/\zeta) \tag{6.11}$$

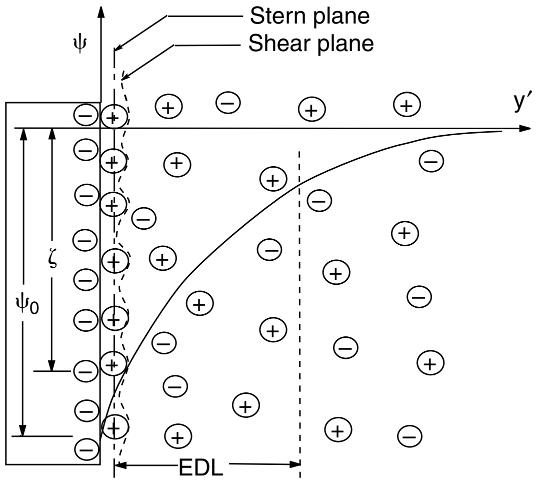


FIGURE 6.12 Schematic diagram of the electric double layer next to a negatively charged solid surface. Here, ψ is the electric potential, ψ_s is the surface electric potential, ζ is the zeta potential, and y' is the distance measured from the wall.

where ψ is the electric potential field, ζ is the zeta potential, ρ_e is the net charge density, D is the dielectric constant, and α is the ionic energy parameter given as:

$$\alpha = ez\zeta/k_b\mathrm{T} \tag{6.12}$$

where e is the electron charge, z is the valence, k_b is the Boltzmann constant, and T is the temperature. In Equation (6.11), the spatial gradients are nondimensionalized with a characteristic length h. Parameter β relates the ionic energy parameter α and the characteristic length h to the Debye–Hückel parameter $\omega = 1/\lambda$ as shown below:

$$\beta = (\omega h)^2/\alpha$$

where

$$\omega = \frac{1}{\lambda} = \sqrt{\frac{2n_0e^2z^2}{\varepsilon k_bT}}$$

The electrokinetic phenomenon can be divided into four parts [Probstein, 1994]:

1. Electro-osmosis: motion of ionized liquid relative to the stationary charged surface by an applied electric field
2. Streaming potential: electric field created by the motion of ionized fluid along stationary charged surfaces (opposite of electro-osmosis)
3. Electrophoresis: motion of the charged surface relative to the stationary liquid, by an applied electric field
4. Sedimentation potential: electric field created by the motion of charged particles relative to a stationary liquid (opposite of electrophoresis).

6.3.2 The Electro-Osmotic Flow

The electro-osmotic flow is created by applying an electric field in the streamwise direction, where this electric field ($\vec{\mathrm{E}}$) interacts with the electric charge distribution in the channel (ρ_e) and creates an electrokinetic body force on the fluid. The ionized incompressible fluid flow with electro-osmotic body forces is governed by the incompressible Navier–Stokes equation:

$$\rho_f\left(\frac{\partial \vec{V}}{\partial t} + (\vec{V}\cdot\nabla)\vec{V}\right) = -\nabla P + \mu\nabla^2\vec{V} + \rho_e\vec{E} \tag{6.13}$$

The main simplifying assumptions and approximations are

- The fluid viscosity is independent of the shear rate; hence, the Newtonian fluid is assumed.
- The fluid viscosity is independent of the local electric field strength. This condition is an approximation. Because the ion concentration within the EDL is increased, the viscosity of the fluid may be affected; however, such effects are usually neglected for dilute solutions.
- The Poisson–Boltzmann equation, Equation (6.11), is valid; hence, the ion convection effects are negligible.
- The solvent is continuous and its permittivity is not affected by the overall and local electric field strength.

Based on these continuum-based equations, various researchers have developed numerical models to simulate electrokinetic effects in microdevices [Yang et al., 1998; Ermakov et al., 1998; Dutta et al., 1999, 2000]. The EDL thickness can be as small as a few nanometers, and in such small scales the continuum description given by the Poisson–Boltzmann equation may break down [Dutta and Beskok, 2001]. Hence, the molecular dynamics method can be used to study the EDL effects in such small scales [Lyklema et al., 1998].

6.3.3 Molecular Dynamics Method

The molecular dynamics method requires simulation of motion and interactions of all molecules in a given volume. The intermolecular interactions are described by a potential energy function, typically the Lennart-Jones 12–6 potential given as [Allen and Tildesley, 1994]:

$$V^{LJ}(r) = 4\varepsilon\left[c_{ij}\left(\frac{\sigma}{r}\right)^{12} - d_{ij}\left(\frac{\sigma}{r}\right)^{6}\right] \tag{6.14}$$

where r is the molecular separation and σ and ε are the length and energy parameters in the pair potential respectively. The coefficients c_{ij} and d_{ij} are parameters chosen for particular fluid–fluid and fluid–surface combinations [see Allen and Tildesley (1994) for details]. The first term on the right-hand side shows the strong repulsive force felt when the two molecules are extremely close to each other, and the second term represents the van der Waals forces. The force field is found by differentiation of this potential for each molecule pair, and the molecular motions are obtained by numerical integration of Newton's equations of motion. Because the motion and interactions of all molecules are simulated, MD simulations are expensive. The computational work scales like the square of the number of the simulated molecules O (N^2). Reduction in the computational intensity can be achieved by fast multipole algorithms or by implementation of a simple cutoff distance [Gad-el-Hak, 1999]. The MD simulations are usually performed for simple liquid molecules and for dense gases. Potential functions other than the Lennart-Jones 12–6 potential are also available. In addition to the prohibitively large number of molecules involved in the simulation, however, an intrinsic limitation of the molecular dynamics method is the insight required to select the appropriate potential functions. For example, the electrokinetic transport simulations require inclusion of electrostatic forces in the potential function.

6.3.4 Treatment of Surfaces

In the molecular dynamics method, the fluid–surface interactions can be handled more realistically by including solid atoms attached to the lattice sites via a confining potential and letting them interact with gas–liquid molecules through a Lennart-Jones potential, Equation (6.14). The solid atoms exhibit random thermal motions corresponding to the surface temperature T_{wall}. The desired temperature of the simulation is maintained by controlling the outer parts of the solid-lattice structure [Koplik and Banavar, 1995a]. Using realistic atomistic surface discretization increases the number of molecules even further, but this may become necessary to determine the surface roughness effects in microtransport. Also considering that the microfabricated surfaces can have rms surface roughness on the order of a few nanometers (depending on the fabrication process), realistic molecular-based surface treatments for liquid flows in nanoscales can be achieved using the molecular dynamics method [Tehver et al., 1998].

The molecular dynamics method is used to determine the validity range of the Newtonian fluid approximation and the no-slip-boundary conditions for simple liquids in submicron and nanoscale channels. Koplik et al. (1989) investigated dense gas and liquid Poiseuille flows with MD simulations. The molecular structure of the wall is also included in these simulations, resulting in fluid–wall interactions for smooth surfaces. Their findings for liquid flows show insignificant velocity slip effects. However, considerable slip effects with decreasing density are reported for gas molecules, consistent with Maxwell's slip-boundary conditions given in Equation (6.1). The literature includes conflicting findings regarding the validity of the no-slip conditions and the appropriate viscosity of liquids in nanoscale channels (see Section 2.7 in Gad-el-Hak, 1999). In a recent study by Granick (1999), the behavior of complex liquids with chain-molecule structures in nanoscales has been reported. Confined fluid behavior, solidification, melting, and rapid deformation of liquid thin films can also be studied by the molecular dynamics method.

MD is restricted to very small (nanoscale) volumes, and the maximum integration time is also limited to a few thousand mean collision times. Hence, molecular simulations should be used whenever the corresponding continuum equations are suspected of failing or are expected to fail, as in the case of fast timescale processes, thin films, or interfaces and in the presence of geometric singularities [Koplik and Banavar, 1995]. To apply MD to larger scale thermal/fluidic transport problems, Hadjiconstantinou and Patera (1997) developed coupled atomistic/continuum simulation methods and extended this work to include multifluid interfaces [Hadjiconstantinou, 1999]. Due to the computational and practical restrictions of MD, it is crucial to develop hybrid atomistic/continuum simulation methodologies for further studies of liquid transport in micron and nanoscales.

6.4 Summary and Conclusions

In this chapter molecular-based simulation methodologies for liquid and gas flows in micron and submicron scales were presented. For simulation of gas flows, the main emphasis was given to the direct simulation Monte Carlo (DSMC) method. Its algorithmic details, limitations, advantages, and disadvantages were presented. Although the DSMC is quite popular for analysis of high-speed rarefied gas flows, it is not as effective for simulation of gas microflows. It suffers from slow convergence and large statistical noise, and it requires an extensive number of simulated molecules. These disadvantages can be eliminated to some degree by using the newly developed information preservation (IP) technique. However, the IP-DSMC is still undergoing development and validation. An alternative to the DSMC method is solution of the Boltzmann transport equation, which is an integro-differential equation with seven independent variables. It is clear that the Boltzmann equation algorithms are very complicated to implement for general engineering applications, but they can be used for simple geometry cases, such as in microchannels. A final alternative for simulation of gas microflows is the lattice Boltzmann method (LBM), which has been developed extensively within the past decade. The LBM has relatively simpler algorithms that can handle both the rarefied and continuum gas flows from a kinetic theory point of view, and the ability of the LBM to capture the incompressible flow limit can make this method a great candidate for microfluidic simulations.

The molecular dynamics (MD) method was introduced for liquid flows. Because MD requires modeling of every molecule, it is computationally expensive and is usually applied to very small volumes in order to verify the onset of the continuum behavior in liquids. MD is general enough to handle the interactions of long-chain molecules with each other and the surfaces in very thin gaps. The wall-surface roughness and its molecular structure can also be included in the simulations. Thus, realistic molecule–surface interactions can be obtained using the MD method. The main drawback of MD is its prohibitively large computational cost.

The DSMC, MD, and Boltzmann equation models are numerically more expensive than solution of the Navier–Stokes equations. Considering that the microtransport applications cover a wide range of length scales from submillimeter to tens of nanometers, it is numerically more efficient to implement hybrid continuum-atomistic models, where the atomistic simulations take place only at a small section of the entire computational domain. References to developments of hybrid schemes were given for each model.

All numerical methods are inherently model based, including the constitutive laws and the boundary conditions of the Navier–Stokes equations, as well as the molecular interaction models of the MD and the DSMC methods. Although it may seem that the molecular simulation methods are more fundamental, they require assumptions and models of more fundamental levels. For example, the molecular dynamics method requires specification of the Lennart-Jones potentials and their coefficients. Physical insight about a problem is of utmost importance for any model. After all, numerical models can only deliver the physics implemented within them.

References

Alexander, F.J., Garcia, A.L., and Alder, B.J. (1998) "Cell Size Dependence of Transport Coefficients in Stochastic Particle Algorithms," *Phys. Fluids* **10**, pp. 1540–42.

Allen, M.P., and Tildesley, D.J. (1994) *Computer Simulation of Liquids*, Oxford Science Publications, New York.

Angelopoulos, A.D., Paunov, V.N., Burganos, V.N., and Payatakes, A.C. (1998) "Lattice Boltzmann Simulation of Non-Ideal Vapor–Liquid Flow in Porous Media," *Phys. Rev. E* **57** 3, pp. 3237–45.

Aoki, K. (1989) "Numerical Analysis of Rarefied Gas Flows by Finite-Difference Method," in *Rarefied Gas Dynamics: Theoretical and Computational Techniques*, E.P. Muntz, D.P. Weaver, and D.H. Campbell, eds., AIAA, New York, pp. 297–322.

Arkilic, E.B. (1997) Measurement of the Mass Flow and Tangential Momentum Accommodation Coefficient in Silicon Micro-Machined Channels, Ph.D. thesis, Massachusetts Institute of Technology.

Arkilic, E., Schmidt, M.A., and Breuer, K.S. (1997) "Gaseous Flow in Long Microchannels," *J. MEMS* **6**, pp. 2–7.

Bernsdorf, J., Durst, F., and Schafer, M. (1999) "Comparison of Cellular Automata and Finite Volume Techniques for Simulation of Incompressible Flows in Complex Geometries," *Int. J. Numer. Meth. Fluids* **29**, pp. 251–64.

Beskok, A. (1996) Simulations and Models for Gas Flows in Microgeometries, Ph.D thesis, Princeton University.

Beskok, A., Trimmer, W., and Karniadakis, G.E. (1995) "Rarefaction Compressibility and Thermal Creep Effects in Gas Microflows," in *IMECE 95, Proc. ASME Dynamic Systems and Control Division*, DSC-Vol. 57-2, pp. 877–92.

Beskok, A., Karniadakis, G.E., and Trimmer, W. (1996) "Rarefaction and Compressibility Effects in Gas Microflows," *J. Fluids Eng.* **118**, pp. 448–56.

Beskok, A., and Karniadakis, G.E. (1997) "Modeling Separation in Rarefied Gas Flows," 28th AIAA Fluid Dynamics Conference, AIAA 97-1883, June 29–July 2.

Beskok, A., and Karniadakis, G.E. (1999) "A Model for Flows in Channels, Pipes, and Ducts at Micro and Nano Scales," *Microscale Thermophys. Eng.* **3**, pp. 43–77.

Beskok, A., and Warburton, T.C. (2001) "Arbitary Lagrangian Eulerian Analysis of a Bi-Directional Micro-Pump Using Spectral Elements", *Int. J. Comput. Eng. Sci.* **2**, No. 1, pp. 43–57.

Beskok, A., and Warburton, T.C. (2001) "An Unstructured H/P Finite Element Scheme for Fluid Flow and Heat Transfer in Moving Domains," *J. Comput. Phys.* **174**, pp. 492–509.

Bhatnagar, P.L., Gross, E.P., and Krook, K. (1954) "A Model for Collision Processes in Gases," *Phys. Rev.* **94**, pp. 511–24.

Bird, G. (1976) *Molecular Gas Dynamics*, Clarendon Press, Oxford.

Bird, G. (1978) "Monte Carlo Simulation of Gas Flows," *Ann. Rev. Fluid Mech.* **10**, pp. 11–31.

Bird, G. (1994) *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*, Oxford Science Publications, New York.

Bird, G.A. (1998) "Recent Advances and Current Challenges for DSMC," *Comput. Math. Appl.* **35**, pp. 1–14.

Breuer, K.S., Piekos, E.S., and Gonzales, D.A. (1995) "DSMC Simulations of Continuum Flows," AIAA Paper 95-2088, Thermophysics Conf., June 19–22, American Institute of Aeronautics and Astronautics, San Diego.

Cai, C.P., Boyd, I.D., Fan, J., and Candler, G.V. (2000) "Direct Simulation Methods for Low Speed Microchannel Flows," *AIAA J. Thermophys. Heat Transf.* **14**, pp. 368–78.

Chapman, S., and Cowling, T.G. (1970) *The Mathematical Theory of Non-Uniform Gases*, Cambridge University Press, New York.

Chen, G., and Boyd, I.D. (1996) "Statistical Error Analysis for the Direct Simulation Monte Carlo Technique," *J. Comput. Phys.* **126**, pp. 434–48.

Chen, S., and Doolen, G.D. (1998) "Lattice Boltzmann Method for Fluid Flows," *Annu. Rev. Fluid Mech.* **30**, pp. 329–64.

Chou, S.Y., and Baganoff, D. (1997) "Kinetic Flux-Vector Splitting for the Navier–Stokes Equations," *J. Comput. Phys.* **130**, pp. 217–30.

Cummings, E.B., Griffiths, S.K., and Nilson, R.H. (1999) "Irrotationality of Uniform Electroosmosis," *Proc. of SPIE Microfluidic Devices and Systems II*, September 20–21, 1999, Santa Clara, California, Vol. 3877, C.H. Ahn, A.B. Frazier eds., pp. 248–256. SPIE, Bellingham, WA. pp. 180–89.

Cybyk, B.Z., Oran, E.S., Boris, J.P., and Anderson, J.D., Jr. (1995) "Combining the Monotonic Lagrangian Grid with a Direct Simulation Monte Carlo Model," *J. Comput. Phys.* **122**, pp. 323–34.

Deitrich, S., and Boyd, I.D. (1996) "Scalar and Parallel Optimized Implementation of the Direct Simulation Monte Carlo Method," *J. Comput. Phys.* **126**, pp. 328–42.

Drexler, K.E. (1990) *Engines of Creation: The Coming Era of Nanotechnology*, Anchor Books/Doubleday, New York.

Dutta, P., and Beskok, A. (2001) "Analytical Solution of Combined Electro-osmotic/Pressure Driven Flows in Two-Dimensional Straight Channels: Finite Debye Layer Effects" *Anal. Chem.* **73**, pp. 1979–86.

Dutta, P., Warburton, T.C., and Beskok, A. (1999) "Numerical Modeling of Electroosmotically Driven Micro Flows," Proc. ASME IMECE Meeting, *J. MEMS* **1**, pp. 467–74.

Dutta, P., Beskok, A., and Warburton, T.C. (2002) "Electroosmotic Flow Control in Complex Micro-Geometries," *Journal of MEMS*, Vol. **11**, No. 1, pp. 36–44, February 2002.

Ermakov, S.V., Jacobson, S.C., and Ramsey, J.M. (1998) "Computer Simulations of Electrokinetic Transport in Microfabricated Channel Structure," *Anal. Chem.* **70**, pp. 4494–505.

Fan, J., and Shen, C. (1999) "Statistical Simulation of Low-Speed Unidirectional Flows in Transition Regime," in *Rarefied Gas Dynamics*, Vol. 2, R. Brun et al., eds., Cepadues-Editions, Toulouse, France, pp. 245–52.

Frisch, U., Hasslacher, B., and Pomeau, Y. (1986) "Lattice-Gas Automata for the Navier–Stokes Equation," *Phys. Rev. Lett.* **56**, pp. 1505–8.

Fukui, S., and Kaneko, R. (1988) "Analysis of Ultra Thin Gas Film Lubrication Based on Linearized Boltzmann Equation: First Report, Derivation of a Generalized Lubrication Equation Including Thermal Creep Flow," *J. Tribol.* **110**, pp. 253–62.

Fukui, S., and Kaneko, R. (1990) "A database for Interpolation of Poiseuille Flow Rates for High Knudsen Number Lubrication Problems," *J. Tribol.* **112**, pp. 78–83.

Gad-el-Hak, M. (1999) "The Fluid Mechanics of Microdevices: The Freeman Scholar Lecture," *J. Fluids Eng.* **121**, pp. 5–33.

Garcia, A.L., Bell, J.B., Crutchfield, W.Y., and Alder, B.J. (1999) "Adaptive Mesh and Algorithm Refinement Using Direct Simulation Monte Carlo," *J. Comput. Phys.* **154**, pp. 134–55.

Granick, S. (1999) "Soft Matter in a Tight Spot," *Phys. Today* **52**, pp. 26–31.

Hadjiconstantinou, N.G. (1999) "Hybrid Atomistic-Continuum Formulations and the Moving Contact-Line Problem," *J. Comput. Phys.* **154**, pp. 245–65.

Hadjiconstantinou, N.G. (2000) "Analysis of Discretization in the Direct Simulation Monte Carlo," *Phys. Fluids* **12**, pp. 2634–38.

Hadjiconstantinou, N.G., and Patera, A.T. (1997) "Heterogeneous Atomistic-Continuum Representations for Dense Fluid Systems," *Int. J. Mod. Phys.* **C8**, pp. 967–76.

Harley, J.C., Huang, Y., Bau, H.H., and Zemel, J.N. (1995) "Gas Flow in Micro-Channels," *J. Fluid Mech.* **284**, pp. 257–74.

Hash, D.B., and Hassan, H.A. (1995) "A Hybrid DSMC/Navier–Stokes Solver," AIAA Paper No. 95-0410, American Institute of Aeronautics and Astronautics, Reston, VA.

He, X.Y., Shan, X.W., and Doolen, G.D. (1998) "Discrete Boltzmann Equation Model for Nonideal Gases," *Phys. Rev. E* **57**, pp. R13–R16.

Ho, C.M., and Tai, Y.C. (1998) "Micro-Electro-Mechanical Systems (MEMS) and Fluid Flows," *Annu. Rev. Fluid Mech.* **30**, pp. 579–612.

Huang, A.B., Giddens, D.P., and Bagnal, C.W. (1966) "Rarefied Gas Flow Between Parallel Plates Based on the Discrete Ordinate Method," *Phys. Fluids* **10**, pp. 498–502.

Inamuro, T., Yoshino, M., and Ogino, F. (1997) "Accuracy of Lattice Boltzmann Method for Small Knudsen Number with Finite Reynolds Number," *Phys. Fluids* **9**, pp. 3535–42.

Jamamoto, K., and Sanryo, T. (1991) "Flow of Rarefied Gas Past 2-D Body of an Arbitrary Shape at Small Mach Numbers," in *Seventeenth Rarefied Gas Dynamics Symp.*, A.E. Beylich, ed., VCH Verlag, Weinheim, p. 273.

Kannenberg, K.C., and Boyd, I.D. (1996) "Monte Carlo Computation of Rarefied Supersonic Flow into a Pitot Probe," *AIAA Journal* **34**(1), pp. 83–88, January 1996.

Kennard, E.H. (1938) *Kinetic Theory of Gases*, McGraw-Hill, New York.

Koplik, J., and Banavar, J.R. (1995a) "Continuum Deductions from Molecular Hydrodynamics," *Annu. Rev. Fluid Mech.* **27**, pp. 257–92.

Koplik, J., and Banavar, J.R. (1995b) "Corner Flow in the Sliding Plate Problem," *Phys. Fluids* **7**, pp. 3118–25.

Koplik, J., Banavar, J.R., and Willemsen, J.F. (1989) "Molecular Dynamics of Fluid Flow at Solid Surfaces," *Phys. Fluids A* **1**, pp. 781–94.

Lavallée, P., Boon, J.P., and Noullez, A. (1991) "Boundaries in Lattice Gas Flows," *Physica D* **47**, pp. 233–40.

Liu, H.F. (2001) 2D and 3D Unstructured Grid Simulation and Coupling Techniques for Micro-Geometries and Rarefied Gas Flows, Ph.D. thesis, Brown University.

Liu, J.Q., Tai, Y.C., Pong, K.C., and Ho, C.M. (1993) "Micromachined Channel/Pressure Sensor Systems for Micro Flow Studies," in *7th Int. Conf. on Solid-State Sensors and Actuators — Transducers '93*, pp. 995–98, Yokohama, Japan, June 1993.

Liu, F., Gatsonis, N.A., Beskok, A., and Karniadakis, G.E. (1999) "Simulation Models for Rarefied Flow Past Sphere in a Pipe," *Rarefied Gas Dynamics*, Vol. 1, pp. 679–86, Cepadues-Editions, Toulouse, France.

Lord, R.G. (1976) "Tangential Momentum Coefficients of Rare Gases on Polycrystalline Surfaces," in *Proc. 10th Int. Symp. on Rarefied Gas Dynamics*, pp. 531–38, New York: American Institute of Aeronautics and Astronautics, c. 1977.

Lou, T., Dahlby, D.C., and Baganoff, D. (1998) "A Numerical Study Comparing Kinetic Flux Vector Splitting for the Navier–Stokes Equations with a Particle Method," *J. Comput. Phys.* **145**, pp. 489–510.

Loyalka, S.K., and Hamoodi, S.A. (1990) "Poiseuille Flow of a Rarefied Gas in a Cylindrical Tube: Solution of Linearized Boltzmann Equation," *Phys. Fluids A* **2**, pp. 2061–65.

Luo, L.S. (1998) "Unified Theory of Lattice Boltzmann Models for Non-Ideal Gases," *Phys. Rev. Lett.* **81**, pp. 1618–21.

Lyklema, J., Rovillard, S., and Coninck, J.D. (1998) "Electrokinetics: The Properties of the Stagnant Layer Unraveled," *J. Surface. Colloid.* **14**, pp. 5659–63.

Mavriplis, C., Ahn, J.C., and Goulard, R. (1997) "Heat Transfer and Flow Fields in Short Microchannels Using Direct Simulation Monte Carlo," *J. Thermophys. Heat Transf.* **11**, pp. 489–96.

Michael, A.A., Spaid, M.A.A., and Phelan, F.R. (1997) "Lattice Boltzmann Methods for Modeling Microscale Flow in Fibrous Porous Media," *Phys. Fluids* **9**, pp. 2468–74.

Muntz, E.P. (1989) "Rarefied Gas Dynamics," *Annu. Rev. Fluid Mech.* **21**, pp. 387–417.

Myong, R.S. (1999) "Thermodynamically Consistent Hydrodynamic Computational Models for High-Knudsen-Number Gas Flows," *Phys. Fluids* **11**, pp. 2788–2802.

Nance, R.P., Hash, D., and Hassan, H.A. (1997) "Role of Boundary Conditions in Monte Carlo Simulation of MEMS Devices," *J. Thermophys. Heat Transf.* **11**, pp. 497–505.

Noble, D.R., Chen, S., Georgiadis, J.G., and Buckius, R.O. (1995) "A Consistent Hydrodynamic Boundary Condition for the Lattice Boltzmann Method," *Phys. Fluids* **7**, pp. 203–9.

Oh, C.K., Oran, E.S., and Sinkovits, R.S. (1997) "Computations of High-Speed, High Knudsen Number Microchannel flows," *J. Thermophys. Heat Transf.* **11**, pp. 497–505.

Ohwada, T., Sone, Y., and Aoki, K. (1989) "Numerical Analysis of the Poiseuille and Thermal Transpiration Flows between Two Parallel Plates on the Basis of the Boltzmann Equation for Hard Sphere Molecules," *Phys. Fluids A* **1**, pp. 2042–49.

Oran, E.S., Oh, C.K., and Cybyk, B.Z. (1998) "Direct Simulation Monte Carlo: Recent Advances and Applications," *Annu. Rev. Fluid Mech.* **30**, pp. 403–41.

Pan, L.S., Liu, G.R., and Lam, K.Y. (1999) "Determination of Slip Coefficient for Rarefied Gas Flows Using Direct Simulation Monte Carlo," *J. Micromech. Microeng.* **9**, pp. 89–96.

Pfahler, J., Harley, J., Bau, H., and Zemel, J. (1991) "Gas and Liquid Flow in Small Channels," *Proc. of ASME Winter Annu. Mtg.* (DSC) **32**, pp. 49–59.

Piekos, E.S., and Breuer, K.S. (1996) "Numerical Modeling of Micromechanical Devices Using the Direct Simulation Monte Carlo Method," *J. Fluids Eng.* **118**, pp. 464–69.

Pong, K.C., Ho, C.M., Liu, J., and Tai, Y.C. (1994) "Non-Linear Pressure Distribution in Uniform Microchannels" *Appl. Microfabrication Fluid Mech.* pp. 51–56.

Probstein, R.F. (1994) *Physiochemical Hydrodynamics: An Introduction*, 2nd ed., John Wiley & Sons, New York.

Qian, Y.H. (1993) "Simulating Thermohydrodynamics with Lattice BGK Models," *J. Sci. Comp.* **8**, pp. 231–41.

Qian, Y.H., D'Humiéres, D., and Lallemand, P. (1992) "Lattice BGK Models for Navier–Stokes Equation," *Europhys. Lett.* **17**, pp. 479–82.

Roveda, R., Goldstein, D.B., and Varghese, P.L. (1998) "Hybrid Euler/Particle Approach for Continuum/Rarefied Flows," *J. Spacecraft Rockets* **35**, pp. 258–65.

Saripov, F., and Seleznev, V. (1998) "Data on Internal Rarefied Gas Flows," *J. Phys. Chem. Ref. Data* **27**, pp. 657–706.

Schaaf, S.A., and Chambre, P.L. (1961) *Flow of Rarefied Gases*, Princeton University Press, Princeton, NJ.

Seidl, M., and Steinheil, E. (1974) "Measurement of Momentum Accommodation Coefficients on Surfaces Characterized by Auger Spectroscopy, SIMS, and LEED," in *9th Int. Symp. on Rarefied Gas Dynamics*, pp. E9.1–E9.2, Symposium Proceedings, M. Becker, and M. Fiedig eds., DFVLR-Press, Porz-Wahn.

Shan, X., and Chen, H. (1994) "Simulation of Non-Ideal Gases and Liquid Gas Phase Transitions by the Lattice Boltzmann Equation," *Phys. Rev. E* **49**, pp. 2941–48.

Sone, Y. (1989) "Analytical and Numerical Studies of Rarefied Gas Flows on the Basis of the Boltzmann Equation for Hard Sphere Molecules," *Phys. Fluids A* **1**, pp. 2042–49.

Sone, Y. (2000) "Flows Induced by Temperature Fields in a Rarefied Gas and Their Ghost Effect on the Behavior of a Gas in the Continuum Limit," *Annu. Rev. Fluid Mech.* **32**, pp. 779–811.

Spaid, M.A.A., and Phelan, F.R. Jr. (1997) "Lattice Boltzmann Methods for Modeling Microscale Flow in Fibrous Porous Media," *Phys. Fluids* **9**, pp. 2468–74.

Tehver, R., Toigo, F., Koplik, J., and Banavar, J.R. (1998) "Thermal Walls in Computer Simulations," *Phys. Rev. E* **57**, pp. R17–R20.

Tison, S.A. (1993) "Experimental Data and Theoretical Modeling of Gas Flows through Metal Capillary Leaks," *Vacuum* **44**, pp. 1171–75.

Vangenabeek, O., and Rothman, D.H. (1996) "Macroscopic Manifestations of Microscopic Flows through Porous Media Phenomenology from Simulation," *Annu. Rev. Earth Planet Sci.* **24**, pp. 63–87.

Vargo, S.E., and Muntz, E.P. (1996) "A Simple Micromechanical Compressor and Vacuum Pump for Flow Control and Other Distributed Applications," Thirty-Fourth Aerospace Sciences Meeting and Exhibit, January 15–18, Reno, NV.

Vargo, S.E., Muntz, E.P., Shiflett, G.R., and Tang, W.C. (1998) "Knudsen Compressor as a Micro- and Macroscale Vacuum Pump without Moving Parts or Fluids," *J. Vac. Sci. Technol. A Vacuum Surfaces Films* **17**, pp. 2308–13.

Veijola, T., Kuisma, H., and Lahdenperä, J. (1995) "Equivalent Circuit Model of the Squeezed Gas Film in a Silicon Accelerometer," *Sensor. Actuator. A* **48**, pp. 239–48.

Veijola, T., Kuisma, H., and Lahdenperä, J. (1998) "The Influence of Gas Surface on Gas Film Damping in a Silicon Accelerometer," *Sensor. Actuator. A* **66**, pp. 83–92.

Vergeles, M., Keblinski, P., Koplik, J., and Banavar, J.R. (1996) "Stokes Drag and Lubrication Flows: A Molecular Dynamics Study," *Phys. Rev. E* **53**, pp. 4852–64.

Vijayakumar, P., Sun, Q., and Boyd, L.D. (1999) "Vibrational-Translational Energy Exchange Models for the Direct Simulation Monte Carlo Method," *Phys. Fluids* **11**, pp. 2117–26.

Vincenti, W.G., and Kruger, C.H. (1977) *Introduction to Physical Gas Dynamics*, Robert E. Krieger Publishing, Huntington, NY.

Yang, C., Li, D., and Masliyah, J.H. (1998) "Modeling Forced Liquid Convection in Rectangular Microchannels with Electrokinetic Effects," *Int. J. Heat Mass Transf.* **41**, pp. 4229–49.

Zou, Q., and He, X. (1997) "On Pressure and Velocity Boundary Conditions for the Lattice Boltzmann BGK Model," *Phys. Fluids* **9**, pp. 1591–98.

7

Hydrodynamics of Small-Scale Internal Gaseous Flows

Nicolas G. Hadjiconstantinou
Massachusetts Institute of Technology

7.1 Introduction

7.1.1 Overview

Small-scale, atmospheric-pressure internal gaseous flows have received significant attention in recent years in connection with micro-and nanoscale science and technology [Ho and Tai, 1998; Karniadakis and Beskok, 2001]. In addition to a number of applications of practical interest, small-scale gaseous hydrodynamics is attractive to researchers because of the scientific challenges it poses. It is well known [Vincenti, and Kruger, 1965; Cercignani, 1988] that as the characteristic hydrodynamic lengthscale approaches the fluid internal lengthscale, in this case the molecular mean free path, the Navier–Stokes description fails. Extensive discussions and additional background on this subject can be found in reviews in this handbook (e.g., [Gad-el-Hak, 2002]) or elsewhere [Karniadakis and Beskok, 2001]. The objective of this chapter is to present and discuss some of the *recent* progress in modeling small scale *internal* gaseous flows of *engineering interest* where the Navier–Stokes description cannot be applied. This is an area in which until recently little was known beyond the classical shear, pressure-driven, and thermal-creep-driven duct flows, primarily because previous efforts had focused on external high-speed flows associated with flight in the upper atmosphere. Gaining fundamental understanding in this regime is important for facilitating the design of small-scale devices and also for educational purposes. For this reason, particular emphasis is given here to theoretical results on basic archetypal problems which, although perhaps simplified to some extent, can provide fundamental understanding of the flow physics.

7.1.2 Background

The failure of the Navier–Stokes description in gas flows is quantified by the Knudsen number, $Kn = \lambda/H$, where λ is the molecular mean free path and H is a characteristic hydrodynamic length scale. When the Knudsen number is small ($Kn \ll 1$), transport is dominated by intermolecular collisions and can be characterized as diffusive. Hydrodynamic gradients over length scales characterized by $Kn \ll 1$ lead to small nonequilibrium, which can be described by the Chapman–Enskog theory [Chapman and Cowling, 1970]; within this approximation, the gas response can be described by linear-gradient transport, which leads to the Navier–Stokes description[1] [Chapman and Cowling, 1970; Gad-el-Hak, 2002].

The system walls, however, introduce an inhomogeneity that leads to rather strong nonequilibrium effects. In the $Kn \ll 1$ limit, these effects remain localized to small regions of space in contact with the walls; these regions have a thickness of the order of one mean free path and are known as Knudsen layers. At the Navier–Stokes description level, the effect of the Knudsen layers for $Kn \lesssim 0.1$ manifests itself in the form of apparent hydrodynamic property slip/jump at the boundaries that can be captured by slip-flow boundary conditions; as a result, the regime $Kn \lesssim 0.1$ is known as slip flow. For $Kn \ll 1$, Knudsen layers are present irrespectively of the characteristic system lengthscale H; however, as H grows, their effect becomes less pronounced, as one would expect, to the extent that in the limit $Kn \lll 1$ their effect is for all practical purposes negligible, and the classical no-slip boundary condition becomes an excellent approximation.

When the Knudsen number is large ($Kn \to \infty$), the rate of intermolecular collisions is very small compared to the rate of molecule–wall collisions. As a result, transport at high Knudsen numbers is ballistic. Ballistic transport is typically assumed to take place for $Kn \gtrsim 10$. The regime $0.1 \lesssim Kn \lesssim 10$ is known as the transition regime and is typically the most challenging to model. In this regime, ballistic transport is important while collisions between molecules are not negligible.

Gaseous flows beyond the Navier–Stokes regime ($Kn \gtrsim 0.1$) are sometimes referred to as rarefied. The origin of this terminology can be found in the rarefied gas dynamics literature [Kogan, 1969]. These flows were first extensively studied in connection with high altitude aerodynamics in which the gas was at low density. Perhaps some what misleading is the term *noncontinuum* frequently used to refer to flows for which the Navier–Stokes description breaks down. This term is very common within the rarefied gas dynamics literature [Bird, 1994] and, now, the MEMS literature and may lead to confusion in a mechanics setting where the expression *noncontinuum* will most likely be associated with a breakdown of the continuum assumption. One may surmise that in rarefied gas dynamics the term *noncontinuum* is a result of the view that the continuum approach culminates in the Navier–Stokes equations, and that consequently when the latter fails, the continuum approach fails without necessarily implying the failure of the continuum assumption.

This chapter is dedicated to the hydrodynamics of such systems, that is, systems that cannot be modeled by the Navier–Stokes description but can still be meaningfully described by hydrodynamic fields. The view taken here is the one typically adopted within the rarefied gas dynamics community and described in [Vincenti and Kruger, 1965]: conservation laws for mass, momentum, and energy follow naturally from moments of the Boltzmann equation. Defining the molecular distribution function as the expected value of a large ensemble of systems leads to a meaningful description in terms of conservation laws (in the presence or absence of Navier–Stokes closures) for a quite wide range of conditions including very small length and time scales.

[1] In the interests of simplicity, limiting cases of this description (e.g., Stokes flow) will not be denoted separately but will be understood to apply under the appropriate conditions.

7.2 Flow Physics

This section discusses recent developments in modeling small-scale internal gaseous hydrodynamics. One of its basic assumptions is that air at atmospheric pressure can be treated as a dilute gas [Bird, 1994], the hydrodynamics of which can be described for all Knudsen numbers using the Boltzmann equation [Cercignani, 1988, Bird, 1994]. This assumption is shown to be satisfied (albeit by a narrow margin) in [Bird (1994)]. The theoretical developments discussed in this chapter have been aided by the direct simulation Monte Carlo (DSMC), a stochastic simulation method for solving the nonlinear Boltzmann equation. Comprehensive descriptions of this method can be found in this handbook [Beskok, 2002] or in the monograph by the method's inventor, Graeme Bird (1994). An augmented DSMC formulation that extends the applicability of DSMC to gases of moderate densities where molecular size effects are not negligible has also been developed [Alexander et al., 1995] and is known as the consistent Boltzmann algorithm. The majority of theoretical developments presented here use DSMC for verification purposes. In some cases, however, DSMC provides the only solution available to the problem of interest.

Unless otherwise stated, DSMC simulations will use the hard-sphere model as a matter of computational convenience. The hard-sphere model provides acceptable models of rarefied gas flows [Cercignani, 1988], and for the purposes of this discussion it provides a good compromise between simplicity and realistic modeling. The mean free path of a hard sphere gas is given by

$$\lambda = \frac{1}{\sqrt{2}\pi n\sigma^2} \tag{7.1}$$

while the first order approximations to the viscosity and thermal conductivity of the hard-sphere gas within the Chapman–Enskog theory are given by

$$\mu = \frac{5}{16\sigma^2}\sqrt{\frac{mk_bT}{\pi}} \tag{7.2}$$

and

$$\kappa = \frac{75k_b}{64\sigma^2}\sqrt{\frac{k_bT}{m\pi}} \tag{7.3}$$

respectively [Chapman and Cowling, 1970]. Here m is the molecular mass, T is the gas temperature k_b is Boltzmann's constant, n is the gas number density and σ is the hard-sphere molecular diameter. These rational approximations to the transport coefficients are typically preferred over the more accurate infinite-order approximations from which they differ by only approximately 2% [Chapman and Cowling, 1970]. One of the disadvantages of the hard-sphere model is that it predicts transport coefficients that are proportional to $T^{0.5}$, whereas real gases exhibit a slightly higher exponent of approximately $T^{0.7}$. To remedy this, collision models with variable collision cross-sections have been proposed [Bird, 1994]; one example is the variable hard-sphere (VHS) model in which the collision cross-section is a function of the relative velocity of the colliding molecules. The work presented here can be easily extended to these modified collision models.

One of the basic geometries that we will visit frequently in this chapter is a two-dimensional channel such as the one shown in Figure 7.1. The two-dimensional channel geometry has been widely studied in the context of small-scale flows due to its direct relevance to typical small-scale applications but also due to its simplicity, which enables investigations aimed at the physics of transport at small scales. Let us introduce the following notation that we will use throughout this chapter: the gas velocity field will be denoted $\vec{\mathbf{u}} = \vec{\mathbf{u}}(x, y) = (u_x(x, y), u_y(x, y), u_z(x, y))$, while $T = T(x, y)$, $P = P(x, y)$ and $\rho = \rho(x, y)$ denote the temperature, pressure, and density fields respectively.

7.2.1 Preliminaries

In this section we will briefly review some basic results for rarefied internal flows. We will discuss the velocity slip and temperature jump relations used to obtain Navier–Stokes solutions in the slip-flow

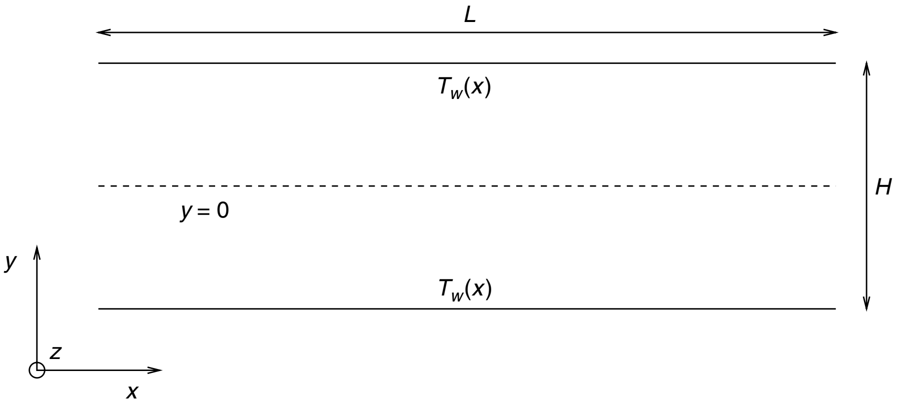


FIGURE 7.1 A two-dimensional channel and nomenclature.

regime, and we will consider solutions of the Boltzmann equation for isothermal pressure-driven flows for arbitrary Knudsen numbers.

7.2.1.1 Slip-Flow Boundary Conditions

Despite the failure of the Navier–Stokes description within the Knudsen layers, for $Kn \lesssim 0.1$ the effect of these layers on the remainder of the flow field (within the Navier–Stokes approximation) can be captured by a set of effective slip/jump boundary conditions. In particular, according to slip flow, the velocity of the gas at the wall, $u_{\text{gas}}|_{\text{wall}}$, differs from the velocity of the wall u_{w} by an amount that is proportional to the normal velocity gradient at the wall. More precisely,

$$u_{\text{gas}}|_{\text{wall}} - u_{\text{w}} = \alpha \frac{2 - \sigma_v}{\sigma_v} \lambda \frac{du}{d\eta} |_{\text{wall}} \tag{7.4}$$

where σ_v is the momentum accommodation coefficient [Beskok and Karniadakis, 1999], and η is the coordinate normal to the wall and pointing into the gas. The temperature jump at the wall is given by the following analogous expression

$$T_{\text{gas}}|_{\text{wall}} - T_{\text{w}} = \zeta \frac{2\gamma}{\gamma + 1} \frac{2 - \sigma_T}{\sigma_T} \frac{\lambda}{Pr} \frac{dT}{d\eta} |_{\text{wall}} \tag{7.5}$$

where σ_T is the energy accommodation coefficient, Pr is the gas Prandtl number and γ is the ratio of specific heats.

The coefficients α and ζ introduce corrections to the original results of Maxwell ($\alpha = \zeta = 1$) that were obtained through an approximate method [Cercignani, 1988]. These coefficients are weak functions of the interaction model [Cercignani, 1988]; for air, α and ζ are usually taken to be equal to unity, although recent theoretical results suggest that this may lead to *additional* error of the order of 15%. In particular, direct Monte Carlo simulations [Wijesinghe and Hadjiconstantinou, 2001], molecular dynamics simulations and linearized solutions of the Boltzmann equation [Ohwada et al., 1989b] show that for hard spheres and fully accommodating conditions, $\alpha \approx 1.11$ and $\zeta \approx 1.13$.

A few further comments:

1. Slip-flow theory naturally reduces to the standard no-slip boundary conditions in the limit $Kn \lll 1$. This can be easily seen by nondimensionalizing η in equations (7.4) and (7.5) using the characteristic lengthscale H.
2. The above slip-flow relations remain accurate for time-dependent flows evolving at *hydrodynamic* timescales (for $Kn \lesssim 0.1$); this suggests that the hydrodynamic evolution time scales for problems characterized by $Kn \lesssim 0.1$ are sufficiently long for the behavior of the Knudsen layer to be effectively quasi-static. This is verified by theoretical treatments of the Boltzmann equation, at least in the BGK approximation [Sone, 1964], where slip-flow relations equivalent (at least formally) to the above are obtained by assuming that the evolution time scale is long compared to the molecular

collision time $\tau_c = \lambda/\bar{c}$ where $\bar{c} = \sqrt{8k_bT/(\pi m)}$ is the mean thermal speed. As we show in section 7.2.2.2, this quasi-static behavior seems to hold beyond the slip-flow regime and up to $Kn \lesssim 0.4$ where hydrodynamic evolution timescales may be as low $O(5\tau_c)$.

3. The above slip-flow expressions assume that the wall surface has no curvature; corrections due to wall curvature are given in [Cercignani, 1988; Grad, 1969].
4. The velocity slip expression equation (7.4) does not include the thermal creep contribution in the presence of a temperature gradient along the wall. A discussion of this form of velocity slip can be found in [Fukui and Kaneko, 1988]. Thermal creep phenomena extend beyond the slip-flow regime; thermal creep flow for the hard-sphere model and the associated thermal creep coefficient have been characterized in [Ohwada et al., 1989a].

Unless otherwise stated, we will assume that both accommodation coefficients are equal to unity; this appears to be a reasonable approximation for atmospheric-pressure engineering surfaces [Bird, 1994; Cercignani, 1988; Ohwada et al., 1989a].

7.2.1.2 Isothermal Pressure-Driven Flows in Two-Dimensional Channels

Isothermal pressure-driven flow in two-dimensional ducts for large Knudsen numbers was originally studied by Knudsen (1909). This pioneering work showed the existence of a minimum in the flow rate when it is normalized by the driving pressure difference and plotted against the average pressure in the channel [Karniadakis and Beskok, 2001]. Following Knudsen's work, a theoretical description of this phenomenon remained for many years one of the ultimate challenges within the rarefied gas community. Following the development of semianalytical solutions of simple models of the Boltzmann equation [Cercignani, 1988], numerical solutions of the linearized Boltzmann equation for the more realistic hard-sphere gas for various two-dimensional geometries were finally developed [Ohwada et al., 1989a]. For two-dimensional channels (as in Figure 7.1) the gas response for arbitrary Knudsen numbers is typically expressed in kinetic terms through the following expression for the bulk velocity u_b

$$\dot{Q} = u_b H = -\frac{1}{P}\frac{dP}{dx}H^2\sqrt{\frac{RT}{2}}\,\bar{Q} \tag{7.6}$$

where $\dot{Q}$ is the flow rate per unit depth, u_b is the bulk (average over the channel width) velocity, $R = k_b/m$ is the gas constant, and $\bar{Q} = \bar{Q}(Kn)$ is a proportionality coefficient. Similarly defined $\bar{Q}$ parameters have now been tabulated for a variety of two-dimensional duct geometries [Karniadakis and Beskok, 2001].

As shown in Fig. 7.2, $\bar{Q}(Kn)$ for a two-dimensional channel in the transition regime varies slowly about its minimum value occurring at $Kn \approx 1$. Numerical solutions, such as linearized solutions of the Boltzmann equation for hard-spheres [Cercignani, 1988; Ohwada et al., 1989a; Beskok and Karniadakis, 1999] and molecular simulations [Beskok and Karniadakis, 1999], have been shown to be in good agreement with experiments [Cercignani, 1988; Beskok and Karniadakis, 1999], even when the former use simpler models such as that of Maxwellian molecules.

7.2.2 Isothermal flows

7.2.2.1 Second-Order Velocity Slip

A variety of slip-flow approaches to problems of interest suggest that slip-flow theory is remarkably robust (see also comparisons between DSMC and slip-flow solutions for a variety of problems in this chapter), in the sense that it continues to be reasonably accurate, at least in a qualitative sense, well beyond its expected limits of applicability. Robust slip-flow models will always be preferable to alternatives such as molecular simulations or numerical solutions of the Boltzmann equation because the numerical cost associated with solutions of the Navier–Stokes equations is negligible compared with the cost of these alternative methods. For this reason a variety of researchers [Cercignani, 1964; Sone and Onishi, 1978; Beskok and Karniadakis, 1999; Deissler, 1964; Aubert and Colin, 2001; Maurer et al., 2003] have attempted to develop or evaluate slip models that can be used beyond $Kn \approx 0.1$. A review of these approaches can be found in Karniadakis and Beskok (2001).

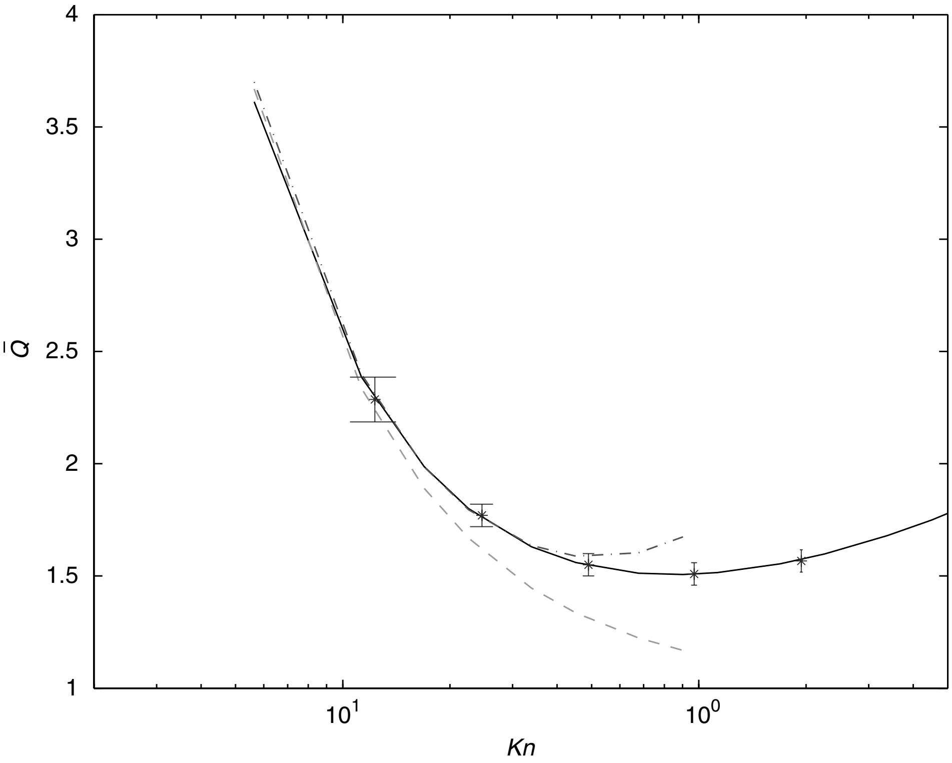


FIGURE 7.2 Nondimensional flow rate as a function of the Knudsen number for fully developed pressure-driven flow. The solid line denotes $\bar{Q}$ as determined by solution of the linearized Boltzmann equation for hard-sphere gases [Ohwada et al., 1989a], and the dash-dotted line denotes the second-order slip model discussed in section 7.2.2.2. The stars denote DSMC simulation results, and the dashed line a first-order slip model.

The general idea behind these approaches is a second-order slip law of the form

$$u|_{\text{wall}} - u_{\text{w}} = \alpha\lambda \frac{\partial u}{\partial \eta}|_{\text{wall}} - \beta\lambda^2 \frac{\partial^2 u}{\partial \eta^2}|_{\text{wall}} \tag{7.7}$$

which naturally extends the first-order slip concept; here β is the second-order slip coefficient. The first rigorous approach in this subject is the one by Cercignani (1964) who, using the BGK model of the linearized Boltzmann equation and the assumptions of zero wall curvature, steady flow, and no variation in the flow direction ($\partial u/\partial x = 0$ in Figure 7.1), showed that the contributions from the Knudsen layers to the velocity field are of order Kn^2 and thus need to be taken into account when using a second-order slip model. These findings explain why the contribution of the Knudsen layers does not need to be considered when using a first-order slip model, and also why simplistic approaches that comprise just Equation (7.7) with β chosen to fit DSMC flow profiles do not work. This can be illustrated by considering that the thickness of the Knudsen layer is approximately 1.5λ [Hadjiconstantinou, 2005] and thus, in a one-dimensional flow, at $Kn = 0.2$ the Knudsen layers from both walls penetrate 60% of the physical domain. Since within the slip-flow approximation the Navier–Stokes description only captures the solution outside the Knudsen layers, models that extract the second-order slip coefficient β by fitting the velocity profile from DSMC simulation throughout the flow domain are destined to fail. In fact, the contribution of the Knudsen layer is sufficiently large that for $Kn \gtrsim 0.3$ direct comparison between the Navier–Stokes and the true flowfield is impossible.

Sone and Onishi (1978) later obtained the same results as Cercignani. Despite being very useful, at least in a qualitative sense, this model has been neglected, perhaps because the BGK model does not lead to good agreement with experimental data.

7.2.2.2 A Second-Order Slip Model for the Hard-Sphere Gas

Recently, Hadjiconstantinou (2003a) has shown how the above theoretical results can be used to develop a slip model for the hard-sphere gas that is a more realistic model of isothermal gaseous flows[2]. In one-dimensional flows ($\partial u/\partial z = \partial u/\partial x = 0$), this new model reduces to Equation (7.7) with $\alpha = 1.11$ and $\beta = 0.61$ (in higher dimensions the second-order term is more complex [Hadjiconstantinou, 2003a]); however, the model also includes a method for quantitatively accounting for the contribution of the Knudsen layer; this is discussed below. This model has been tested in a variety of *low-speed one-dimensional flows*, both steady and transient, and has yielded results that are in excellent agreement with DSMC simulations. As shown below, it also seems to provide a reasonable explanation for the recent experimental findings of [Maurer et al., 2003], who measured the second-order slip coefficient in two-dimensional channel flow.

What sets this model apart from all other approaches is its ability to *quantitatively* account for the effect of the Knudsen layers on the flow; this, in fact, holds the key to obtaining an accurate second-order slip model. As we show below, the effect of the Knudsen layers can be accounted for such that the second-order slip model remains *quantitatively* correct at least up to $Kn \approx 0.4$ and *qualitatively* correct well beyond that.

The contribution of the Knudsen layer can be most conveniently accounted for in an average sense (i.e., when calculating averages over the domain). For the purposes of this discussion (Equations [7.8]–[7.10]), let us differentiate between the true Boltzmann equation solution for the flow field and the Navier–Stokes approximation to this solution by denoting the latter by û and recalling that û does not contain any information about the Knudsen layer close to the walls. Then, in a one-dimensional geometry, according to the slip model [Hadjiconstantinou, 2005], the average (bulk) flow velocity is given by

$$u_b = \frac{1}{H}\int_{-H/2}^{H/2} u\,dy = \frac{1}{H}\int_{-H/2}^{H/2}\left[\hat{u} + \chi\lambda^2\frac{\partial^2\hat{u}}{\partial y^2}\right]dy \tag{7.8}$$

where for a hard-sphere gas $\chi = 0.296$.

A direct consequence of the above relation is that in Poiseuille-type flows where the velocity curvature is a constant, experimental measurement of the flow rate (mean flow velocity) yields an effective second-order slip coefficient $\beta - \chi$ (see also [Hadjiconstantinou, 2003a]). In other words, while the *average value* of a Poiseuille profile subject to second-order slip of the form (7.7) is given by

$$\hat{u}_b = \frac{1}{H}\int_{-H/2}^{H/2}\hat{u}\,dy = -\frac{H^2}{2\mu}\frac{dP}{dx}\left(\frac{1}{6} + \alpha Kn + 2\beta Kn^2\right) \tag{7.9}$$

the *true* bulk flow speed (as inferred by an experiment measuring the flowrate) is given by equation (7.8) which leads to

$$u_b = \frac{1}{H}\int_{-H/2}^{H/2}\left[\hat{u} + \chi\lambda^2\frac{\partial^2\hat{u}}{\partial y^2}\right]dy = -\frac{H^2}{2\mu}\frac{dP}{dx}\left(\frac{1}{6} + \alpha Kn + 2\varepsilon Kn^2\right) \tag{7.10}$$

or

$$\bar{Q} = \frac{4}{15\sqrt{\pi}}\,\frac{1 + 6\alpha Kn + 12\varepsilon Kn^2}{Kn} \approx \frac{\sqrt{\pi}}{12}\,\frac{1 + 6\alpha Kn + 12\varepsilon Kn^2}{Kn} \tag{7.11}$$

with $\varepsilon = \beta - \chi = 0.31$. (The above two expressions for $\overline{Q}$ differ by less than 2%; the difference is due to the use of slightly different approximations for the hard-sphere gas viscosity [Chapman and Cowling, 1970; Cercignani, 1964].) As shown in Figure 7.2, the above equation captures the flow rate in isothermal pressure-driven flow very accurately up to $Kn \approx 0.4$. This is also demonstrated in section 7.2.2.4 where the pressure-driven flow-rate is used to determine the wave propagation constant in two-dimensional channels (under the long wavelength approximation).

Most importantly, the above model explains the findings of recent experiments [Maurer et al., 2003] on helium and nitrogen flow in small-scale channels; these experiments find the second-order slip coefficient

[2] A discussion of the limitations of this model is provided at the end of this section.

to be approximately 0.25 ± 0.1. Of course, since the slip coefficient was determined by measuring the *flow rate*, these experiments were in fact determining the effective second-order slip coefficient ε, which is in good agreement with the value 0.31 given above.

We now present a calculation that further illustrates the capabilities of the above second-order slip model. The results provide additional evidence that this model rigorously extends the slip-flow approach into the early transition regime. Of particular importance is that the stress field is accurately captured for arbitrary flows with **no adjustable parameters** up to $Kn \approx 0.4$, suggesting that any correction due to the presence of the Knudsen layer is small; recall that at this Knudsen number, the domain half-width is 1.25λ, which is smaller than the typical size of the Knudsen layer.

Consider the following one-dimensional test problem, which is periodic in the x and z directions (referring to Figure 7.1): both channel walls impulsively start to move parallel to their planes with velocity U at time $t = 0$; the velocity is small compared to the most probable molecular velocity. Below we show a comparison between a Navier–Stokes solution using the second-order slip model and DSMC simulations of this problem. Comparisons for the velocity profile as a function of position at two representative times, the average (bulk) velocity as function of time, and the shear stress τ_{xy} as a function of position at two representative times are shown. Figure 7.3 shows that the effect of the Knudsen layer at $Kn = 0.21$ is already visible; however, the velocity field outside the Knudsen layer, the bulk velocity as a function of time as given by Equation (7.8), and the shear stress *throughout* the physical domain are accurately captured. The comparison at $Kn = 0.42$ (Figure 7.4) shows that the slip model is still reasonably accurate, although the Knudsen layers have penetrated to the middle of the domain leading to the impression that the velocity prediction is incorrect. However, when Equation (7.8) is used to calculate the bulk flow speed, the agreement between Navier–Stokes and DSMC simulations is very good (Figure 7.4, middle). The agreement between the stress fields (Figure 7.4, bottom) is also good suggesting that any correction due to the presence of the Knudsen layer is small.

This comparison also shows that the above slip model can be used in transient problems provided the evolution time scale is long compared to the molecular collision time. Comparisons for a different one-dimensional problem that exhibits no symmetry about the channel centerline can be found in [Hadjiconstantinou, 2005]; the level of agreement exhibited is similar to the one observed here. This suggests that the excellent agreement observed, at least in one-dimensional flows, is not limited to symmetric flowfields.

Discussion of limitations: It appears that a number of the assumptions on which this model is based do not significantly limit its applicability. For example, it would be reasonable to assume that the assumption of steady flow would be satisfied by flows that appear quasi-static at some time scale. Our results above suggest that this time scale is the molecular collision time; in other words, the slip model is valid for flows that evolve at time scales that are long compared to the molecular collision time, which can be satisfied by the vast majority of practical flows of interest.

The model was also derived under the assumption of flat walls and no variations in directions other than the normal to the wall. Of course approaches based on assumptions of slow variation in the axial direction (x in Figure 7.1), such as the widely used locally-fully-developed assumption or long wavelength approximation, are expected to yield excellent approximations when used for two-dimensional problems. This is verified by comparison of solutions of such problems to DSMC simulations (see section 7.2.2.4 for example) or experiments (e.g., [Maurer et al., 2003]).

Extension of the model to the case $\partial u/\partial z \neq 0$ within the BGK approximation has been considered by Cercignani (see [Hadjiconstantinou, 2003a]). Validation of this and other solutions [Sone, 1969] (after they have been appropriately modified using the approach described by the author in [Hadjiconstantinou, 2003a]) that take wall curvature[3], three-dimensional flow fields and nonisothermal conditions into account should be undertaken. The exact conditions under which Equation (7.8) can be generalized also need to be clarified. While the contribution of the Knudsen layer can always be found by a Boltzmann equation analysis, the value of Equation (7.8) lies in the fact that it relates this contribution to the

[3] Due to wall curvature, the second-order slip coefficient for flow in cylindrical capillaries is different from flow in two-dimensional channels.

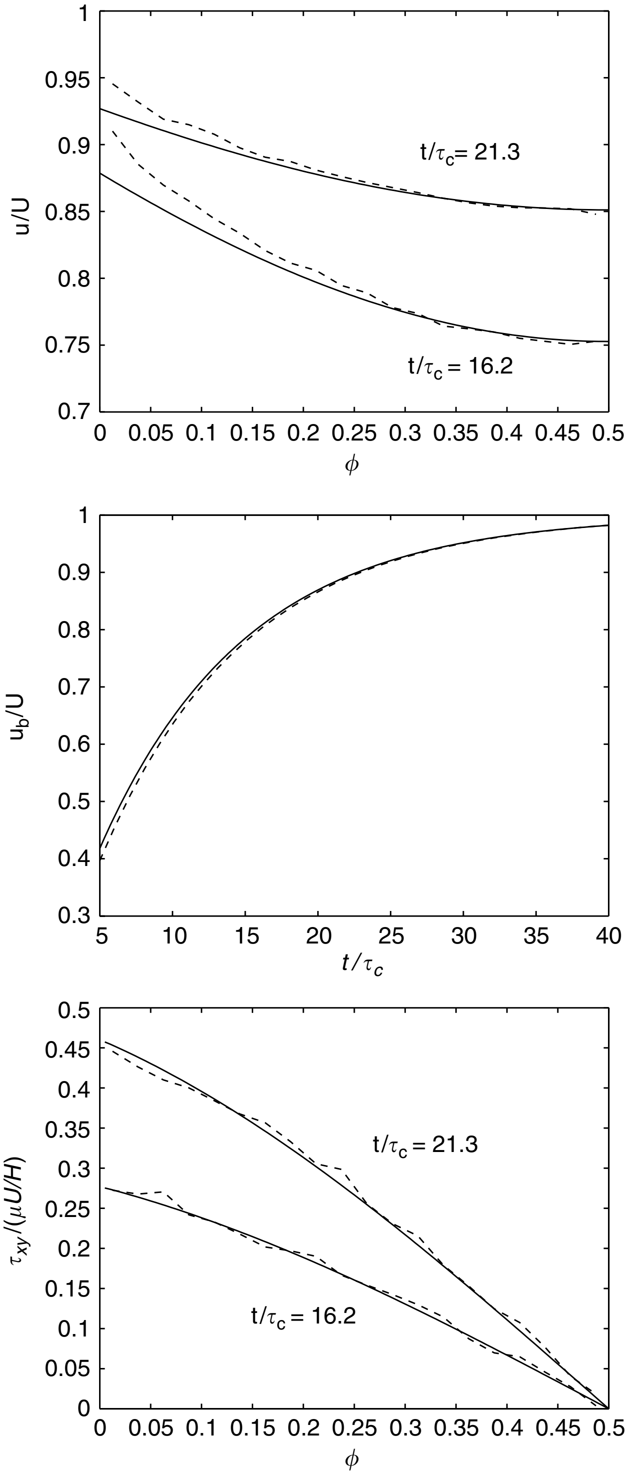


FIGURE 7.3 The impulsive start problem at $Kn = 0.21$. Comparison between the second-order slip model and DSMC simulations for the velocity field (top), the average velocity (Equation [7.8]) as a function of time (middle), and the stress field (bottom). Here, $(\phi) = (y + H/2)/H$ is a shifted nondimensional channel transverse coordinate.

Navier–Stokes solution, and thus it requires no solution of the Boltzmann equation. Finally, recall that the linearized conditions ($Ma \ll 1$) under which the second-order model is derived imply $Re \ll 1$ since $Ma \approx ReKn$ and $Kn > 0.1$. Here *Ma* is the Mach number and *Re* is the Reynolds number, based on the same characteristic lengthscale as *Kn*.

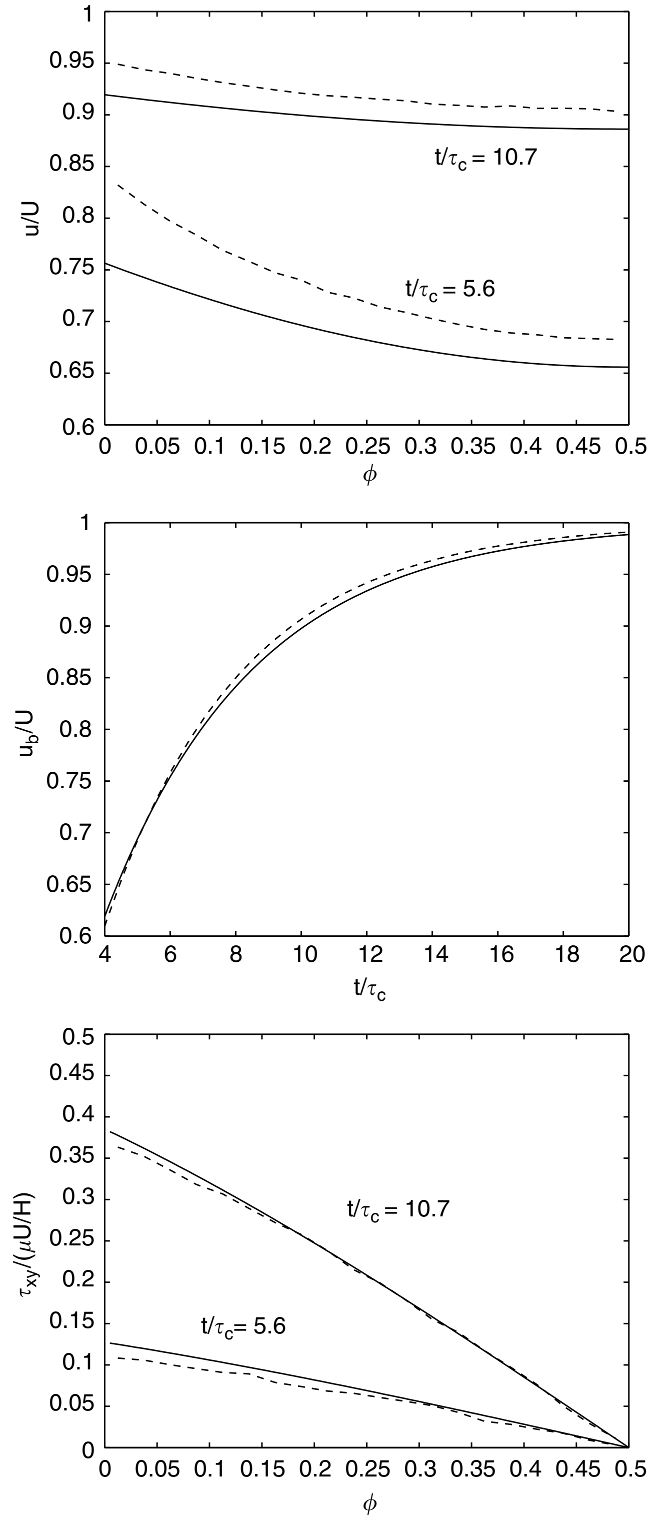


FIGURE 7.4 The impulsive start problem at $Kn = 0.42$. Comparison between the second-order slip model and DSMC simulations for the velocity field (top), the average velocity (Equation [7.8]) as a function of time (middle), and the stress field (bottom). Here, $(\phi) = (y + H/2)/H$ is a shifted nondimensional channel transverse coordinate.

7.2.2.3 Oscillatory Shear Flows

Oscillatory shear flows are very common in MEMS and have been characterized as being of "tremendous importance in MEMS devices" [Breuer, 2002]. A comprehensive study of rarefaction effects on oscillatory shear (Couette) flows was recently conducted by Park et al. (2004). Due to the linear velocity profile observed in the quasi-static regime ($\sqrt{\omega H^2/\nu} \ll 1$ where $\nu = \mu/\rho$ is the kinematic viscosity and ω is the wave angular frequency) Park et al. used an extended first-order slip-flow relation to describe the velocity field (in essence the amount of slip) for all Knudsen numbers, provided the flow was quasi-static. Note that the quasi-static assumption is not at all restrictive due to the very small size of the gap, H. This extended slip-flow relation is fitted to DSMC data and reduces to the first-order slip model Equation (7.4) for $Kn < 0.1$. Park et al. also solved the linearized Boltzmann equation [Cercignani, 1964] in the collisionless ($Kn \to \infty$) limit; they found that in this limit the solution at the wall is identical to the steady Couette flow solution in the sense that the value of the velocity and shear stress at the wall is the same in both cases.

The oscillatory Couette flow problem was used in [Hadjiconstantinou, 2005] as a validation test problem for the second-order slip model of section 7.2.2.2. Relatively high frequencies were used, such that the flow was not in the quasi-static regime. The agreement obtained was excellent up to $Kn \approx 0.4$ in complete analogy with the findings of the test problem presented in section 7.2.2.2.

7.2.2.4 Wave Propagation in Small-Scale Channels

In this section we discuss a theory of axial-plane wave propagation under the long wavelength approximation in two-dimensional channels (such as the one shown in Figure 7.1) for *arbitrary Knudsen numbers*. The theory is based on the observation that within the Navier–Stokes approximation wave propagation in small-scale channels for most frequencies of practical interest is viscous dominated. The importance of viscosity can be quantified by a narrow channel criterion $\delta = \sqrt{2\nu/\omega}/H \gg 1$. When $\delta \gg 1$ (whereby the channel is termed narrow) the viscous diffusion length based on the oscillation frequency is much larger than the channel height; viscosity is expected to be dominant and inertial effects will be negligible. This observation has two corollaries. First, because the inertial effects are negligible, the flow is governed by the steady equation of motion, that is, the flow is effectively quasi-steady [Hadjiconstantinou, 2002]. Second, since for gases the Prandtl number is of order one, the flow is also isothermal (for a discussion see [Hadjiconstantinou and Simek, 2003]). This was first realized by Lamb [Crandall, 1926], who used this approach to describe wave propagation in small-scale channels using the Navier–Stokes description. Lamb's prediction for the propagation constant using this theory is identical to Kirchhoff's more general theory [Kirchhoff, 1868] when the narrow channel limit is taken in the latter.

The author has recently [Hadjiconstantinou, 2002] used the fact that wave propagation in the narrow channel limit[4] is governed by the steady equation of motion to provide a prediction for the propagation constant for *arbitrary Knudsen numbers* without explicitly solving the Boltzmann equation. This is achieved by rewriting Equation (7.6) in the form

$$\tilde{u}_b = -\frac{1}{\mathcal{R}(Kn)}\frac{d\tilde{P}}{dx} \tag{7.12}$$

where tilde denotes the amplitude of a sinusoidally time-varying quantity. This equation *locally* describes wave propagation because, as we argued above, in the narrow channel limit the flow is isothermal and quasi static and governed by the steady-flow equation of motion. Using the long wavelength approximation, which implies a constant pressure across the channel width, allows us to integrate mass conservation, written here as a kinematic condition [Hadjiconstantinou, 2002],

$$\frac{\partial P}{\partial x} = -\left(\frac{\partial P}{\partial \rho}\right)_T \rho_0 \frac{\partial^2 \xi}{\partial x^2} \tag{7.13}$$

across the channel width. Here $(\partial P/\partial \rho)_T$ indicates that this derivative is evaluated under isothermal conditions appropriate to a narrow channel. Additionally, ρ_0 is the average density, and ξ is the fluid-particle displacement defined by

$$u_x(x, y, t) = \frac{\partial \xi(x, y, t)}{\partial t} \tag{7.14}$$

[4] The narrow channel limit needs to be suitably redefined in the transition regime where viscosity loses its meaning. However, the work in [Hadjiconstantinou, 2002; Hadjiconstantinou and Simek, 2003] shows that d as defined here remains a conservative criterion for the neglect of inertia and thermal effects.

Combining Equations (7.12) and (7.13), we obtain [Hadjiconstantinou, 2002]

$$i\omega\xi_b = \frac{\rho_0\,(\partial P/\partial\rho)_T}{\mathcal{R}(Kn)}\,\frac{\partial^2\xi_b}{\partial x^2} \tag{7.15}$$

where ξ_b is the bulk (average over the channel width) fluid-particle displacement. From the above we can obtain the propagation constant

$$(m_m + ik)^2 = \frac{i\omega\mathcal{R}(Kn)}{P_0} \tag{7.16}$$

where P_0 is the average pressure, m_m is the attenuation coefficient, and k is the wave number.

From Equation (7.6) we can identify

$$\mathcal{R}(Kn) = \frac{P_0}{H\bar{Q}\sqrt{RT_0/2}} \tag{7.17}$$

leading to

$$(m_m + ik)^2\lambda^2 = \frac{8i\sqrt{\pi}Kn}{\bar{Q}}\,\frac{\tau_c}{\tau} \tag{7.18}$$

where $\tau = 2\pi/\omega$ is the oscillation period.

This result is expected to be of very general use because the narrow channel requirement is easily satisfied in the transition regime [Hadjiconstantinou, 2002]. A more convenient expression for use in the early transition regime that does not require a lookup table (for $\bar{Q}$) can be obtained using the second-order slip model discussed in section 7.2.2.2. Using this model we obtain

$$(m_m + ik)^2\lambda^2 = \frac{96iKn^2}{1 + 6\alpha Kn + 12\varepsilon Kn^2}\,\frac{\tau_c}{\tau} \tag{7.19}$$

which as can be seen in Figure 7.5 remains reasonably accurate up to $Kn \approx 1$ (aided by the square root dependence of the propagation constant on $\mathcal{R}$). This expression for $Kn \to 0$ reduces to the well known narrow-channel result obtained using the no-slip Navier–Stokes description [Rayleigh, 1896].

Figure 7.5 shows a comparison between Equation (7.19) (Equation [7.18]), DSMC simulations, and the Navier–Stokes result. (DSMC simulations of wave propagation are discussed in [Hadjiconstantinou, 2002].) The theory is in excellent agreement with simulation results. As noted above, the second-order slip model provides an excellent approximation for $Kn \lesssim 0.5$ and a reasonable approximation up to $Kn \approx 1$. The no-slip Navier–Stokes result clearly fails as the Knudsen number increases. The theory presented here can be easily generalized to ducts of arbitrary cross-sectional shape and has been extended [Hadjiconstantinou and Simek, 2003] to include the effects of inertia and heat transfer in the slip-flow regime where closures for the shear stress and heat flux exist.

7.2.2.5 Reynolds Equation for Thin Films

The approach of section 7.2.2.4 is reminiscent of lubrication theory approaches used in describing the flow in thin films [Hamrock, 1994]. In lubrication-theory-type approaches, the small transverse system dimension allows the neglect of inertial and thermal effects; this approximation allows quasi-steady solutions to be used for predicting the flow field in the film. Application of conservation of mass leads to an equation for the pressure in the film known as the Reynolds equation. The Reynolds equation and its applications to small-scale flows is extensively covered in a different chapter of this handbook [Breuer, 2002] and other publications [Karniadakis and Beskok, 2001]. Our objective here is to briefly discuss the opportunities provided by the lubrication approximation for obtaining analytical solutions for arbitrary Knudsen numbers to various MEMS problems.

Because the Reynolds equation is essentially a height (gap) averaged description, its formulation requires only knowledge of the flow rate (average flow speed) in response to a pressure field; it can, therefore, be easily generalized to arbitrary Knudsen numbers in a fashion that is exactly analogous to the procedure used in section 7.2.2.4. This was realized by Fukui and Kaneko (1988), who formulated such a generalized Reynolds equation. Fukui and Kaneko were also able to include the flow rate due to thermal

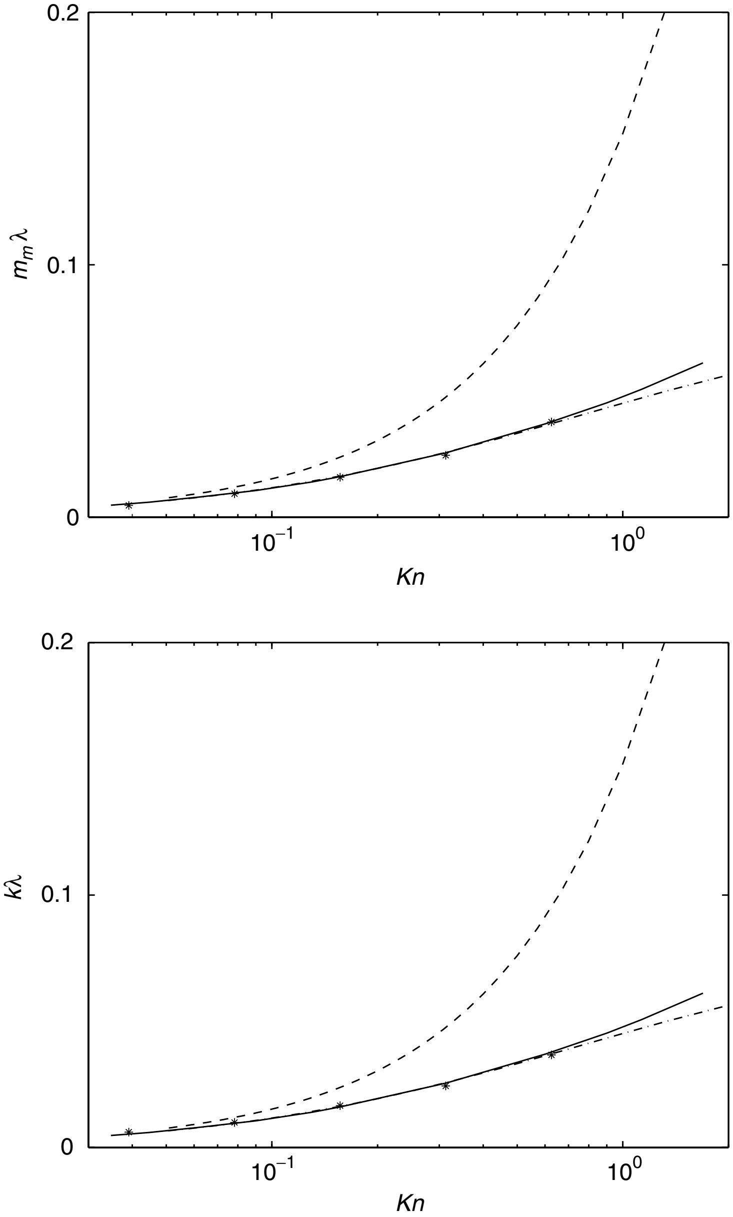


FIGURE 7.5 Comparison between the theoretical predictions of Equation (7.18) shown as a solid line and the simulation results denoted by stars as a function of the Knudsen number at a constant frequency given by $\tau/\tau_c \approx 6400$. The dash-dotted line denotes the prediction of equation (7.19). The no-slip Navier–Stokes solution (dashed lines) is also included for comparison.

creep into the Reynolds equation and thus account for the effects of an axial temperature gradient. Comparison between the formulation of Fukui and Kaneko and DSMC simulations can be found in [Alexander et al., 1994].

More recent work by Veijola and collaborators (see [Karniadakis and Beskok, 2001]) uses fits of the quantity $\overline{Q}$ to define an effective viscosity for integrating the Reynolds equation. It is hoped that the discussion of this chapter and section 7.2.2.2 in particular clarify the fact that the concept of an effective viscosity is not very robust. For $Kn \gg 0.1$ the physical mechanism of transport changes completely, and there is no reason to expect the concept of linear-gradient transport to hold. Even in the early transition regime, the concept of an effective viscosity is contradicted by a variety of findings. To be more specific, an effective viscosity can only be viewed as a particular choice of absorbing the non-Poiseuille part of the flow rate $(1 + 6\alpha Kn + 12\varepsilon Kn^2)$ in Equation (7.10) into another proportionality constant, namely the viscosity. However, section 7.2.2.2 has shown that the correct way of interpreting Equation (7.10) is that, provided correct boundary conditions are supplied, viscous behavior extends to $Kn \approx 0.4$, with the **value of viscosity remaining unchanged**. If, instead, the effective viscosity approach is adopted, the following problems arise:

- The non-Poiseuille part of the flow rate is problem-dependent (flow[5], geometry) while the viscosity is not. In other words, an effective viscosity fitted from the Poiseuille flow rate in a tube is different from the effective viscosity fitted from the Poiseuille flow rate in a channel.
- The fitted effective viscosity does not give the correct stress through the linear constitutive law.

The effective viscosity approach has another disadvantage in the context of its application to the Reynolds equation: it requires neglecting the effect of pressure on the local Knudsen number because the fits used for $\bar{Q}$ result in very complex expressions that cannot be directly integrated, unless the assumption $Kn \neq Kn(P)$ is made. This approach is thus only valid for small pressure changes. Use of equation (7.11) for $Kn \lesssim 0.5$, on the other hand, should not suffer from this disadvantage.

7.2.3 Flows Involving Heat Transfer

In this section we review flows in which heat transfer is important. We give particular emphasis to *convective* heat transfer in internal flows, which has only recently been investigated within the context of rarefied gas dynamics. We also summarize the investigation of Gallis and coworkers on thermophoretic forces on small particles in gas flows.

7.2.3.1 The Graetz Problem for Arbitrary Knudsen Numbers

Since its original solution in 1885 [Graetz, 1885], the Graetz problem has served as an archetypal convective heat transfer problem both from a process modeling viewpoint and an educational viewpoint. In the Graetz problem a fluid is flowing in a long channel whose wall temperature changes in a step fashion. The channel is assumed to be sufficiently long so that the fluid is in an isothermal and hydrodynamically fully developed state before the wall temperature changes.

The gas-phase Graetz problem subject to slip-flow boundary conditions was studied originally by Sparrow and Lin (1962); this study, however, did not include the effects of axial heat conduction, which cannot be neglected in small-scale flows. Here we review the solution by the author [Hadjiconstantinou and Simek, 2002] in which the extended Graetz problem (including axial heat conduction) is solved in the slip-flow regime, and the solution is compared to DSMC simulations in a wide range of Knudsen numbers; the DSMC solutions serve to verify the slip-flow solution but also extend the Graetz solution to the transition regime. The DSMC simulations were performed at sufficiently low speeds for the effects of viscous heat dissipation to be small; this is very important since high speeds typically used in DSMC simulations to alleviate signal-to-noise issues may introduce sufficient viscous heat dissipation effects to render the simulation results useless. (The effect of viscous dissipation on convective heat transfer for a model problem is discussed in the next section.)

In [Hadjiconstantinou and Simek, 2002] a complete solution of the Graetz problem in the slip-flow regime for all Peclet $[Pe = Re\,Pr = (\rho u_b 2H/\mu)Pr]$ numbers was presented. The solution in [Hadjiconstantinou and Simek, 2002] showed that in the presence of axial heat conduction characteristic of small scale devices $(Pe < 1)$, the Nusselt number defined by

$$Nu_T = \frac{q2H}{\kappa(T_w - T_b)} \tag{7.20}$$

[5] The dependence on the flow field comes from the second term in the right hand side of equation (7.8).

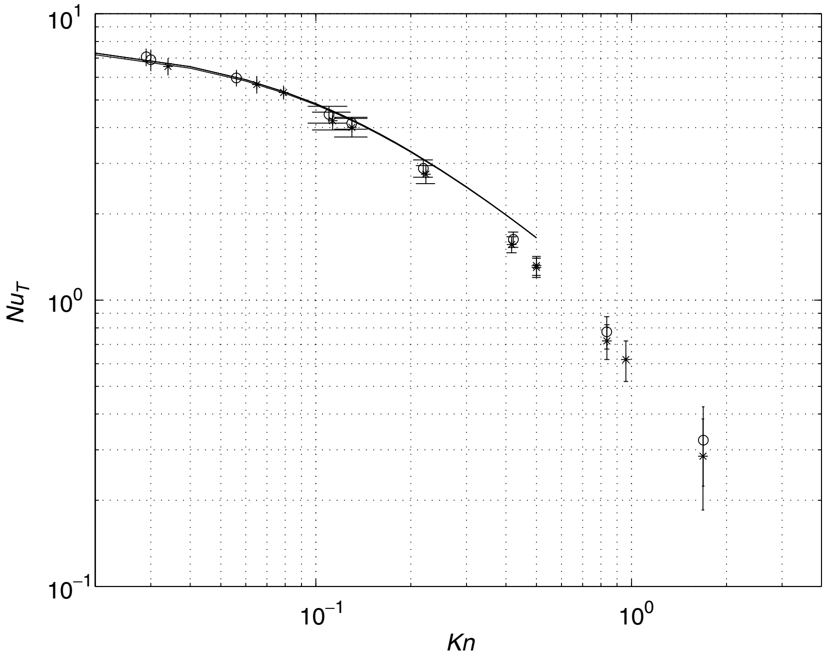


FIGURE 7.6 Variation of Nusselt number Nu_T with Knudsen nunber Kn (from [Hadjiconstantinou and Simek, 2002]). The stars denote DSMC simulation data with a positive wall temperature step, and the circles denote DSMC simulation data with a negative temperature step. The solid lines denote hard-sphere slip-flow results for $Pe = 0.01$, 0.1, and 1.0.

is fairly insensitive to the Peclet number in the small Peclet number limit but higher (by about 10%) than the corresponding Nusselt number in the absence of axial heat conduction ($Pe \rightarrow \infty$). Here q is the wall heat flux and T_b is the bulk temperature defined by

$$T_b = \frac{\int_{-H/2}^{H/2} \rho u_x T \, dy}{\int_{-H/2}^{H/2} \rho u_x \, dy} \tag{7.21}$$

This solution was complemented by low-speed DSMC simulations in both the slip-flow and transition regimes (Fig. 7.6). Comparison of the two solutions in the slip-flow regime shows that the effects of thermal creep are negligible for typical conditions and also that the velocity slip and temperature jump coefficients provide good accuracy in this regime. The DSMC solutions in the transition regime showed that for fully accommodating walls the Nusselt number decreases monotonically with increasing Knudsen number. Solutions with accommodation coefficients smaller than one exhibit the same qualitative behavior as partially accommodating slip-flow results [Hadjiconstantinou, unpublished], namely, decreasing the thermal accommodation coefficient increases the thermal resistance and decreases the Nusselt number, whereas decreasing the momentum accommodation coefficient increases the flow velocity close to the wall, which slightly increases the Nusselt number [Hadjiconstantinou and Simek, 2002]. The similarity between the Nusselt number dependence on the Knudsen number and the dependence of the skin-friction coefficient on the Knudsen number [Hadjiconstantinou and Simek, 2002] suggests that it may be possible to develop a Reynolds-type analogy between the two nondimensional numbers.

7.2.3.2 Viscous Heat Dissipation and the Effect of Slip Flow

In this section we discuss recent results [Hadjiconstantinou, 2003b] concerning the effect of viscous heat dissipation on convective heat transfer. The objective of this discussion is twofold: first, it will illustrate that the velocity slip present at the system boundaries leads to dissipation through shear work, which needs to be appropriately accounted for in convective heat transfer calculations that include the effects of viscous heat dissipation; second, it will provide an illustration of the effects of finite Brinkman number on convective heat transfer. This analysis provides a means for interpreting DSMC simulations in which, in order to alleviate signal-to-noise issues, flow velocities are artificially increased.

It can be shown [Hadjiconstantinou, 2003b] that shear work on the boundary, similarly to viscous heat dissipation, scales with the Brinkman number $Br = \mu u_b^2/\kappa\Delta T$, where ΔT is the characteristic temperature difference in the formulation. It can also be shown that shear work on the boundary can be equally important as viscous heat dissipation in the bulk of the flow as the Knudsen number increases. Although shear work at the boundary must be included in the total heat exchange with the system walls, it has no direct influence on the temperature field because it occurs at the system boundaries. The discussion below, taken from [Hadjiconstantinou, 2003b], shows how shear work at the boundary can be accounted for in convective heat transfer calculations under the assumption of (locally) fully developed conditions.

The importance of shear work at the boundary can be seen from the mechanical energy equation written in the general form valid for all Knudsen numbers

$$0 = -u_x \frac{\partial P}{\partial x} + u_x \frac{\partial \tau_{xy}}{\partial y} = -u_x \frac{\partial P}{\partial x} + \frac{\partial (u_x \tau_{xy})}{\partial_y} - \tau_{xy} \frac{\partial u_x}{\partial y} \tag{7.22}$$

written here for a fully developed flow in a two-dimensional channel. Here τ_{xy} is the xy component of the shear stress tensor. The above equation integrates to

$$[\tau_{xy} u_x]_{-\mathrm{H}/2}^{\mathrm{H}/2} = \int_{-H/2}^{H/2} \tau_{xy} \frac{\partial u_x}{\partial y}\, dy + u_b H \frac{dP}{dx} \tag{7.23}$$

and shows that the shear work at the boundary due to the slip balances the contribution of viscous dissipation and flow work ($u_x dP/dx$) inside the channel.

Thus, as shown in [Hadjiconstantinou, 2003b], if Nu is the Nusselt number based on the thermal energy exchange between the gas and the walls, the total Nusselt number, Nu_t, based on the *total* energy exchange with the walls (thermal plus shear work) under constant-wall-heat-flux conditions in slip flow is given by

$$Nu_t = Nu + \frac{(\tau_{xy} u_x)|_{H/2} 2H}{\kappa (T_{\mathrm{w}} - T_b)} = Nu - 12Br \frac{u_s}{u_b}\left(1 - \frac{u_s}{u_b}\right) \tag{7.24}$$

The Nusselt number based on the thermal energy exchange between the gas and the wall in the case of constant wall-heat-flux was found [Hadjiconstantinou, 2003b] to be given by

$$N_u = \frac{q_o 2H}{\kappa (T_{\mathrm{w}} - T_b)} = \frac{\frac{140}{17} - 2Br\left(1 - \frac{u_s}{u_b}\right)^2 \left(\frac{54}{17} - \frac{30}{17}\frac{u_s}{u_b} + \frac{12}{51}\left(\frac{u_s}{u_b}\right)^2\right)}{1 - \frac{6}{17}\frac{u_s}{u_b} + \frac{2}{51}\left(\frac{u_s}{u_b}\right)^2 + \frac{140}{17}\zeta \frac{\gamma}{\gamma + 1}\frac{Kn}{Pr}} \tag{7.25}$$

where $Br = \mu u_b^2/(\kappa(T_{\mathrm{w}} - T_b))$, q_o is the (constant) wall-heat-flux and

$$\frac{u_s}{u_b} = \frac{6\alpha Kn}{1 + 6\alpha Kn} \tag{7.26}$$

is the normalized slip velocity at the wall.

The validity of Equation (7.24) was verified [Hadjiconstantinou, 2003b] using DSMC simulations. The results of a comparison for $Kn = 0.07$ are shown in Figure 7.7. The agreement between theory and simulation is very good considering that shear work at the wall takes place within the Knudsen layer where extrapolated Navier–Stokes fields are only approximate.

7.2.3.3 Thermophoretic Force on Small Particles

Small particles in a gas through which heat flows experience a thermophoretic force in the direction of the heat flux; this force is a result of the net momentum transferred to the particle due to the asymmetric velocity

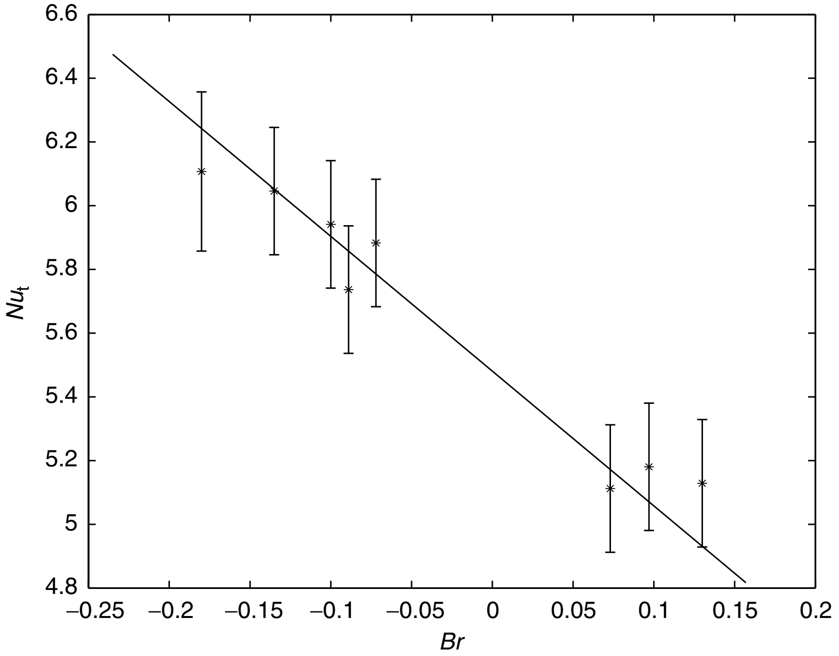


FIGURE 7.7 Variation of the fully developed Nusselt number Nu_t with Brinkman number for $Kn = 0.07$. The solid line is the prediction of equation (7.24), and the stars denote DSMC simulations.

distribution of the surrounding gas [Gallis et al., 2002] in the presence of a heat flux. This phenomenon was first described by Tyndall (1870) and has become of significant interest in connection with contamination of microfabrication processes by small solid particles. This problem appears to be particularly severe in plasma-based processes that generate small particles [Gallis et al., 2002].

Considerable progress has been made in describing this phenomenon by assuming a spherical (radius R) and infinitely conducting particle in a quiescent monoatomic gas. Provided that the particle is sufficiently small such that it has no effect on the molecular distribution function of the surrounding gas, the thermophoretic force can be calculated by integrating the momentum flux imparted by the molecules striking the particle. The particle can be considered sufficiently small when the Knudsen number based on the particle radius, $Kn_R = \lambda/R$, implies a free-molecular flow around the particle, i.e. $Kn_R \gg 1$. Based on these assumptions, Gallis et al. (2001) have also developed a general method for calculating forces on particles in DSMC simulations of arbitrary gaseous flows, provided the particle concentration is dilute. This method is briefly discussed in section 7.3.3.

In the cases where the molecular velocity distribution function is known, such as free molecular flow or the Navier–Stokes limit, the thermophoretic force can be obtained analytically. Performing the calculations in these two extremes and under the assumption that the particle surface is fully accommodating, reveals that the thermophoretic force can be expressed in the following form

$$F_{th} = \psi \pi R^2 q / \bar{c} \tag{7.27}$$

where ψ is a thermophoresis proportionality parameter that obtains the values $\psi_{FM} = 0.75$ for free-molecular flow and $\psi_{CE} = 32/(15\pi) = 0.679$ for a Chapman–Enskog distribution for a Maxwell gas. Here q is the local heat flux. Writing the thermophoretic force in the above form is, in fact, very instructive [Gallis et al., 2002]. It shows that the force is only very mildly dependent on the velocity distribution function with only a change of the order of 10% observed between $Kn \ll 1$ and $Kn \gg 1$. These conclusions extend to other collision models; for example, for a hard-sphere gas, $\psi_{CE} = 0.698$ [Gallis et al., 2002].

As a consequence of the above, the two limiting values can be used to provide bounds for the value of the thermophoretic force on fully accommodating particles close to system walls. Using the weak dependence of ψ on the distribution function, Gallis et al. (2002) provided an estimate of this quantity in the Knudsen layer,

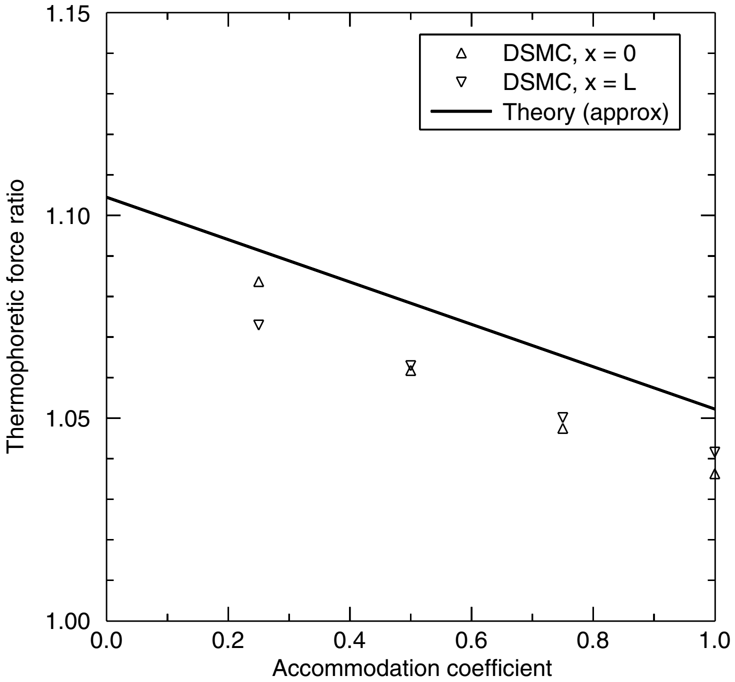


FIGURE 7.8 Comparison between the approximate theory of Gallis and coworkers shown in a straight line and one-dimensional DSMC results for ψ_{KN}/ψ_{CE}. The DSMC results represent the average value over five cells of size $\Delta x = 0.042\,\lambda$ adjacent to the wall in a $Kn = 0.0475$ calculation. (Courtesy of Dr. Gallis.)

ψ_{KN}, by assuming that the distribution function can be written as a superposition of a Chapman–Enskog (incoming and outgoing molecules) and Maxwellian distribution (outgoing molecules), with relative proportions adjusted for accommodation effects. More specifically, they consider a wall at temperature T_w with thermal accommodation coefficient σ_T. For Maxwell molecules, they find

$$\psi_{KN} = \frac{1}{2}\left[\sigma_T\psi_{CE} + (2-\sigma_T)\psi_{FM}\left(\frac{2}{1+\sqrt{T_w/T}}\right)\right] \tag{7.28}$$

which simplifies to

$$\psi_{KN} = \frac{1}{2}\,[\sigma_T\psi_{CE} + (2-\sigma_T)\psi_{FM}] \tag{7.29}$$

in the limit $T \to T_w$. In other words, the presence of a Knudsen layer has a very small effect on the thermophoresis parameter, with $\psi_{KN} = 0.5(\psi_{CE} + \psi_{FM})$ for a fully accommodating wall and $\psi_{KN} = \psi_{FM}$ in the specular reflection limit.

DSMC simulations (Figure 7.8) show [Gallis et al., 2002] that the deviation from ψ_{CE} increases with proximity to the wall, as expected, and show that ψ_{KN} serves as an upper bound to the actual thermophoresis parameter within the Knudsen layer; this is presumably because the assumed distribution function overestimates the deviation from the actual distribution.

7.3 Simulation Methods Development

In this section we briefly discuss recent developments in the simulation of dilute gaseous flows. The majority of these developments are associated with the direct simulation Monte Carlo because this is by far the most popular simulation tool for dilute gases. We also briefly discuss continuum–DSMC hybrid methods that provide computational savings by limiting the use of the molecular (DSMC) description only to the regions where it is needed. The discussion presented below also applies to hybrid methods for dense fluids; the only major difference between methods for dilute gases and dense fluids is that, in the latter, macroscopic boundary condition imposition on the molecular subdomain is significantly more challenging. A more complete discussion of hybrid methods for dense fluids can be found in [Wijesinghe and Hadjiconstantinou, 2004].

7.3.1 The Effect of Finite Discretization

DSMC has been used to capture and predict nonequilibrium gaseous hydrodynamic phenomena in all Knudsen regimes [Bird, 1994] for more than 3 decades. However, only recently has significant progress been made in its characterization as a numerical method and in understanding the numerical errors associated with it.

Recently, Wagner (1992) has shown that DSMC simulations approach solutions of the nonlinear Boltzmann equation in the limit of zero cell size and time step and infinite number of molecules. This result essentially proves consistency. Convergence results for the transport coefficients have been recently obtained by Alexander et al. (2000) for the cell size and by Hadjiconstantinou (2000) and Garcia and Wagner (2000) for the time step.

Alexander et al. (2000) used the Green–Kubo theory to evaluate the transport coefficients in DSMC when the cell size is finite but the time step is negligible. They found that because DSMC allows collisions between molecules at a distance (as long as they are within the same cell) the transport coefficients increase from the dilute-gas Chapman–Enskog values quadratically with the cell size. For example, for the viscosity Alexander et al. find for cubic cells [Alexander et al., 2000]

$$\mu = \frac{5}{16\sigma^2}\sqrt{\frac{mkT}{\pi}}\left(1 + \frac{16}{45\pi}\,\frac{\Delta x^2}{\lambda^2}\right). \tag{7.30}$$

where Δx is the cell size.

In [Hadjiconstantinou, 2000], the author considered the convergence with respect to a finite time step when the cell size is negligible. To apply the Green–Kubo formulation, the author developed a time-continuous analogue of DSMC because DSMC is discrete in time. Using this time-continuous analogue, the author was able to show that the transport coefficients deviate from the dilute-gas Chapman–Enskog values proportionally to the square of the time step. For example, for the viscosity he found

$$\mu = \frac{5}{16\sigma^2}\sqrt{\frac{mkT}{\pi}}\left(1 + \frac{16}{75\pi}\,\frac{(c_o\Delta t)^2}{\lambda^2}\right), \tag{7.31}$$

where Δt is the time step and $c_o = \sqrt{2k_b\mathrm{T}/m}$ is the most probable molecular speed. This prediction for the viscosity, and similar predictions for the thermal conductivity and diffusion coefficient were verified by DSMC simulations by Garcia and Wagner (2000). Good agreement was found between theory and simulation as illustrated in the example of Figure 7.9. The simulations show that the theoretical predictions are valid for small normalized time steps. As the time step increases, transport asymptotes to the collisionless limit prediction.

One key to obtaining the above results for the time step error is to observe that at diffusive transport time scales — which are long compared to the molecular collision time — DSMC dynamics (collisionless advection, collisions, collisionless advection, …) can be thought of as symmetric in time if one views the DSMC time step as "centered" on the middle of either the collision or the advection step. In fact, DSMC can be "symmetrized" by starting the algorithm in the middle of a collision or advection step; this would be necessary for second-order accuracy when DSMC is used for short-time explicit integrations of the Boltzmann equation [Ohwada, 1998]. To observe the above convergence rates in the transport coefficients, sampling also needs to be performed in a fashion that is consistent with the symmetry in the dynamics. Perhaps the simplest way of performing sampling that is thus symmetric is to sample *before and after* the collision part of the algorithm (e.g., see [Gallis et al., 2004]). It is noteworthy that since mass, momentum, and energy are conserved during collisions, symmetrization of sampling is expected to affect only hydrodynamic fluxes, and in fact only when those are measured as volume averages over cells; hydrodynamic fluxes measured as fluxes through surfaces during the advection part of the algorithm are naturally centered.

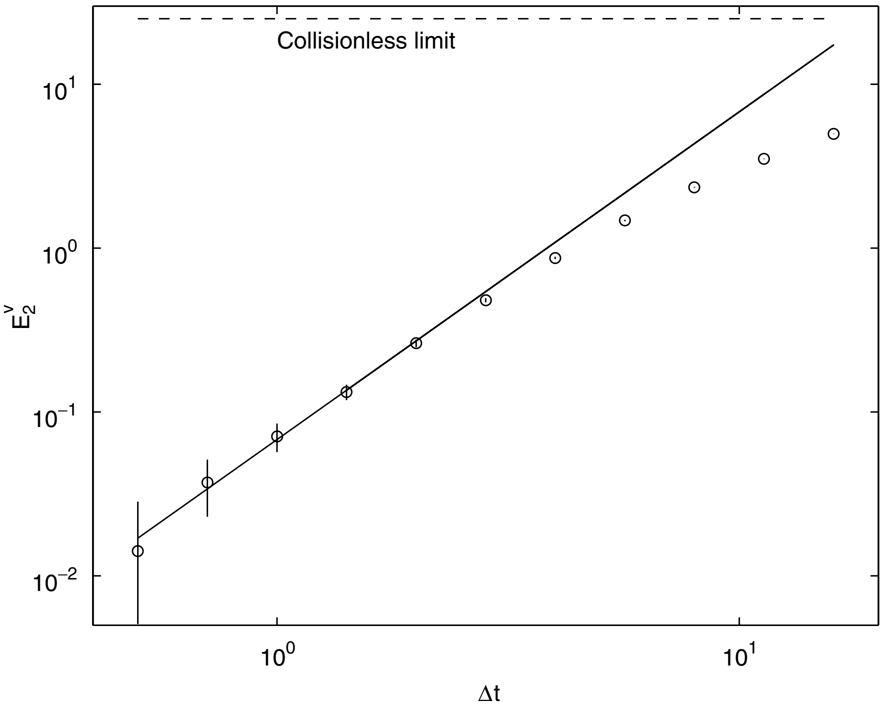


FIGURE 7.9 Error in coefficient of viscosity as a function of normalized time step $\Delta t = c_o \Delta t/\lambda$ (from [Garcia and Wagner, 2000]). Circles denote the normalized error in momentum flux (E_2^v) in the simulations of Garcia and Wagner (2000), and the solid line is the prediction of (7.31).

7.3.2 DSMC Convergence to the Chapman–Enskog Solution in the *Kn* ≪ 1 Limit

Recently Gallis et al. (2004) offered more evidence that DSMC captures the nonequilibrium distribution function corresponding to the Navier–Stokes description as predicted by the Boltzmann equation. They performed very accurate and low-noise calculations (their statistical error estimate was 0.2%) to investigate the domain of validity of the Chapman–Enskog expansion and the ability of DSMC to reproduce this distribution under the appropriate conditions. By calculating the heat flow between two parallel plates and concentrating in the middle region of the domain where wall (Knudsen layer) effects are negligible, they have shown that

1. DSMC is in excellent agreement with the infinite-approximation Chapman–Enskog expansion of the distribution function in the presence of a heat flux and for all inverse-power-law molecules investigated [Bird 1994; Gallis et al., 2004].
2. The Chapman–Enskog solution for the distribution function breaks down at $Kn_q \approx 0.01$ (Figure 7.10), where $Kn_q = q/(\rho c_o^3)$ is the Knudsen number based on the heat flux magnitude q. Note that this failure mode is different to the one associated with nonequilibrium due to the presence of walls in the system.
3. The linear relationship between the heat flux and the temperature gradient is valid independently of the magnitude of heat flux. Additionally, the coefficient of proportionality remains constant at the thermal conductivity value. This fact was proven for Maxwell molecules some years ago [Asmolov et al., 1979]. The study by Gallis et al. has verified this and demonstrated the validity of this observation for the hard-sphere gas. Note that this observation is only valid for planar geometries which are, however, quite common in MEMS.

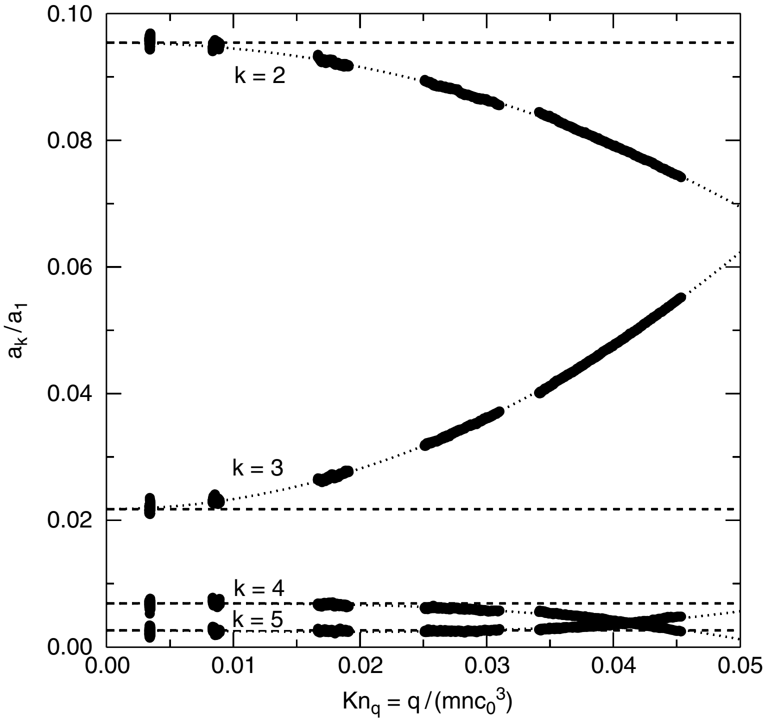


FIGURE 7.10 Comparison between theoretical and measured (DSMC) normalized Sonine polynomial coefficients (a_k/a_1, k = 2–5) [Gallis et al., 2004] as a function of the heat-flux-based Knudsen number. a_1 is proportional to the thermal conductivity and is used here as a normalization. The theoretical values are given by the dashed lines and the DSMC results by the heavy symbols. (Courtesy of Dr. Gallis.)

7.3.3 Forces on Small Spherical Particles

One of the most important challenges associated with semiconductor manufacturing is the presence of contaminants, sometimes produced during the manufacturing process, in the form of small particles. Understanding the transport of these particles is very important for their removal or for ensuring that they do not interfere with the manufacturing process. Recently, Gallis and his coworkers [Gallis et al., 2001] developed a method for calculating the force on small particles in rarefied flows simulated by DSMC. This method is based on the assumption that the particle concentration is very small and the observation that particles with sufficiently small radius such that $Kn_R = \lambda/R \gg 1$ will have a very small effect on the flow field; in this case, the effect of the flow field on the particles can be calculated from DSMC simulations that do not include the particles themselves.

Gallis and his coworkers define appropriate Green's functions that quantify the momentum $F_\delta[\tilde{\mathbf{c}}]$ and energy $Q_\delta[\tilde{\mathbf{c}}]$ transfer rates of individual molecules to the particle surface as a function of the molecule mass, momentum, and energy and degree of accommodation on the particle surface. These can then be integrated over the molecular velocity distribution function, $f(\tilde{\mathbf{c}})$, to yield the average force

$$F = \int F_\delta[\tilde{\mathbf{c}}]f(\tilde{\mathbf{c}})d\tilde{\mathbf{c}} \tag{7.32}$$

or heat flux

$$q = \int Q_\delta[\tilde{\mathbf{c}}]f(\tilde{\mathbf{c}})d\tilde{c} \tag{7.33}$$

to the particle, where $\tilde{c} = \mathbf{c} - \mathbf{u}_p$, $\mathbf{c}$ is the molecular velocity, and $\mathbf{u}_p$ is the particle speed.

For the simple case where $\sigma_v = \sigma_T = \tilde{\sigma}$, Gallis et al. [Gallis et al., 2001] find

$$F_\delta[\tilde{\mathbf{c}}] = \rho\pi R^2\tilde{\mathbf{c}}(|\tilde{\mathbf{c}}| + \tilde{\sigma}(\pi^{1/2}/3)c_p) \tag{7.34}$$

$$Q_\delta[\tilde{\mathbf{c}}] = \tilde{\sigma}\rho\pi R^2|\tilde{\mathbf{c}}|(1/2|\tilde{\mathbf{c}}|^2 - c_p^2) \tag{7.35}$$

where $c_p^2 = 2k_bT_p/m$ and T_p is the particle temperature. More complex accommodation models can also be treated; in [Gallis et al., 2001] an extended Maxwell accommodation model is presented.

In the DSMC implementation, integration of equations (7.32) and (7.33) is achieved by summing the contributions of molecules within a cell. This yields the force and heat flux to a particle as a function of position. Because the force and heat flux are a function of $\mathbf{u}_\mathrm{p}$, the former are calculated as a function of a number of values of the latter; the values of the force and heat flux at intermediate values of $\mathbf{u}_\mathrm{p}$ can be subsequently obtained by interpolation [Gallis et al., 2001].

7.3.4 Hybrid Continuum–Atomistic Methods

By limiting the molecular treatment to the regions where it is needed, a hybrid atomistic–continuum[6] method allows the simulation of complex phenomena at the microscale without the prohibitive cost of a fully molecular calculation. In this section we briefly discuss hybrid methods for multiscale hydrodynamic applications and touch upon the main challenges in developing hybrid simulations for gaseous flows. A more complete discussion including dense fluid flows as well as a more complete review of previous work can be found in Wijesinghe and Hadjiconstantinou (2004).

In Wijesinghe and Hadjiconstantinou (2004) it is shown that to a large extent the two major challenges in developing a hybrid method are the choice of a coupling method and the imposition of boundary conditions on the molecular simulation. Generally speaking, these two can be viewed as decoupled: the coupling technique can be developed on the basis of matching two compatible and equivalent (over some region of space) descriptions, while boundary condition imposition can be posed as the general problem of imposing macroscopic boundary conditions on a molecular simulation. The latter is a very challenging problem that in general has not been resolved to date completely satisfactorily for the case of dense fluids. More details on proposed approaches can be found in Wijesinghe and Hadjiconstantinou (2004). In the case of dilute gases, accurate and robust methods for imposing boundary conditions on molecular simulations exist. These typically require extending the molecular subdomain through the artifice of reservoir regions in which molecules are generated using a Chapman–Enskog distribution [Garcia and Alder, 1998] that is parametrized by the Navier–Stokes flow field being imposed. More details can be found in Wijesinghe and Hadjiconstantinou (2004).

The selection of the coupling approach between the two descriptions is the other major consideration in developing a robust hybrid method. It is becoming increasingly clear that powerful and robust hybrid methods can be developed by using already developed continuum–continuum coupling techniques (recall that the molecular and continuum description can only be coupled in regions where both are valid). Existing continuum–continuum coupling techniques have the additional advantages of being mathematically rigorous and performing optimally for the application for which they have been developed.

No general hybrid method that can be applied to all hydrodynamic problems exists. On the contrary, similarly to Navier–Stokes numerical solution methods, hybrid methods need to be tailored to the *flow physics* of the problem at hand. Perhaps the most important consideration in this respect is that of time scale decoupling originally discussed by Hadjiconstantinou (1999) explicit integration of the molecular subdomain at the molecular time step to the global solution time (or steady state) is very computationally expensive if not infeasible if the Navier–Stokes subdomain is appropriately large. This is because the molecular time step is significantly smaller (MD–dense fluids) or at best smaller (DSMC–dilute gases) than the Courant–Friedrich–Lewy (CFL) stability time step at typical discretization levels.

In Wijesinghe and Hadjiconstantinou (2004) it is shown that the above considerations are intimately linked to the flow physics; compressible flow physics have characteristic timescales that scale with the compressible CFL time step [Wesseling, 2001], which is not very different from a DSMC time step in a

[6] We use the term *continuum* here to emphasize that these approaches are not necessarily limited to the Navier–Stokes description and its breakdown.

dilute gas simulation. In this manner, explicit time integration with a finite-volume-type coupling technique is possible as a natural extension of already existing Navier–Stokes solution methods (see [Wijesinghe et al., 2003] and references therein) as long as the problem of interest is not too large. Such approaches have reached a reasonable maturity level; recent developments include techniques that extend the adaptive mesh refinement (AMR) concept to mesh and *algorithm* refinement by including the molecular description as the finest level of refinement [Garcia et al., 1999; Wijesinghe et al., 2003; 2004]. The first *fully adaptive* implementation is described in detail in [Wijesinghe et al., 2003; 2004].

On the other hand, incompressible flow physics have characteristic time scales that are much longer than the CFL time step, and thus explicit integration at the molecular time step is more prohibitive. Implicit methods are thus required that provide solutions without the need for explicit integration in time. One such implicit method for steady state problems has been proposed by the author for liquids [Hadjiconstantinou and Patera, 1997; Hadjiconstantinou, 1999] and gases [Wijesinghe and Hadjiconstantinou, 2002]; it is based on a domain decomposition approach known as the Schwarz alternating method [Lions, 1988]. A hybrid method based on this coupling approach was recently used to simulate flow through microfluidic filters [Aktas and Aluru, 2002] yielding significant computational savings.

Important prerequisites for adaptive algorithm refinement are robust criteria for Navier–Stokes or continuum assumption breakdown [Boyd, 2003] and a complete understanding of the effect of molecular fluctuations. The effect of statistical noise (resulting from molecular fluctuations) on the development of robust algorithm refinement criteria is discussed in [Wijesinghe et al., 2003; 2004]. Molecular fluctuations and the statistical noise associated with them are, of course, one of the major obstacles in obtaining DSMC solutions of low-speed flows in fully molecular or hybrid approaches. In the case of the latter, they may also influence the convergence/accuracy of various hybrid schemes. For this reason, the statistical error due to molecular fluctuations has been studied in [Hadjiconstantinou et al., 2003] and is briefly discussed in section 7.3.5.

One finding of this study that has significant bearing on the choice of coupling method for hybrid approaches is that the relative statistical error in hydrodynamic flux measurement, E_f, scales as $E_f \sim E_s/Kn$ with the relative statistical error in state property measurement E_s for low-speed gas flows. This means that in low-speed gas flows, using hydrodynamic fluxes to couple the Navier–Stokes and atomistic region (which takes place in regions where $Kn \ll 1$) is at a considerable disadvantage unless methods that are insensitive to statistical noise are developed.

7.3.5 Statistical Noise in Low-Speed Flows

In a recent paper, [Hadjiconstantinou et al., 2003] used equilibrium statistical mechanics to characterize the relative sampling error in hydrodynamic quantities in molecular simulations of flows close to equilibrium as a function of the number of samples taken. They defined the relative statistical error of some quantity Q with mean Q as $E_Q = \sigma_Q/Q$; here σ_Q is the standard deviation in the error in estimating Q. A variety of expressions for the relative statistical error for the most common hydrodynamic variables including hydrodynamic fluxes (shear stress, heat flux) were derived; for the hydrodynamic fluxes, expressions were derived when measured as volume averages and when measured as surface flux averages. The main findings of this work can be summarized as follows:

1. The two averaging methods for hydrodynamic fluxes (volume, surface) yield comparable relative statistical errors, provided that $\Delta x \approx c_o\Delta t$. Here Δt is the averaging time used in the flux method; Δx is the linear dimension, in the direction normal to the flux, of the cell in which volume averaging is performed.
2. For $Kn \ll 1$, the relative error in a particular hydrodynamic flux (e.g., shear stress) is significantly larger than the relative error in the conjugate state variable (e.g., velocity). This has significant implications in the development of hybrid methods as explained in the previous section.
3. A simple theory for incorporating the effects of correlations in volume averaging was presented. This theory is based on the theory of persistent random walks.

4. It was shown that not only the number of molecules per unit volume in an ideal gas is Poisson distributed but also that arbitrary number fluctuations of an infinite ideal gas in equilibrium are Poisson distributed.

Good agreement was found with DSMC simulations of low-speed, low Knudsen number flows where statistical noise presents the biggest challenges. This is expected because the deviation from equilibrium is small under these conditions. The results for state variables were also verified for dense fluids using molecular dynamics simulations.

7.4 Discussion

The above discussion of various phenomena involving isothermal and nonisothermal flows seems to suggest that slip flow is remarkably robust. In channel flows, slip flow seems to correctly predict *average* quantities of interest (flow rates, wave propagation constants, heat transfer coefficients) even beyond its typically acknowledged limit of applicability of $Kn \approx 0.1$ with acceptable error; moreover, in some cases it can *qualitatively* describe the behavior of such *average* quantities well into the transition regime. Methods that extend the range of applicability of the Navier–Stokes description are highly desirable. The simplicity and significant computational efficiency advantage enjoyed by the Navier–Stokes description compared to molecular approaches coupled to the effort already invested in continuum methods, make the former the approach of choice. Despite the lack of general closure models for transport in the transition regime, analytical solutions are sometimes possible through the use of the lubrication approximation and judicious use of already existing analytical results for simple flows. Rigorous high-order slip models such as the one presented in section 7.2.2.2 are proving to be valuable in this respect.

The direct simulation Monte Carlo has played and will continue to play a central role in the analysis of small-scale internal gaseous flows. The statistical sampling employed by this method and the slow convergence associated with it is, perhaps, the most serious limitation of DSMC. While the search for more efficient algorithms or sampling methods continues [Sun and Boyd, 2002], parallel efforts should be invested in developing realistic gas–surface interaction models. Unfortunately, although variable accommodation coefficient models exist [Cercignani and Lampis, 1971] and have been implemented in DSMC [Lord, 1995], experimental verification of their ability to produce physically accurate results is lacking.

Although hybrid methods provide significant savings by limiting molecular solutions only to the regions where they are needed, solution of time-evolving problems that span a large range of time scales is still not possible if the molecular domain, however small, needs to be integrated for the total time of interest. New frameworks are therefore required that allow time scale decoupling or coarse grained time evolution of molecular simulations. For steady incompressible flows, where the time scale gap is large, time-scale–decoupling hybrid methods have been proposed by the author and collaborators [Hadjiconstantinou and Patera, 1997; Hadjiconstantinou, 1999; Wijesinghe and Hadjiconstantinou, 2002].

Acknowledgments

The author wishes to thank T.R. Akylas, A. Beskok, A.L. Garcia, J. Lienhard V and A.T. Patera for helpful comments and discussions. Special thanks to M.A. Gallis for critically commenting on the manuscript. This work was supported in part by the Center for Computational Engineering and the Center for Advanced Scientific Computing, Lawrence Livermore National Laboratory, U.S. Department of Energy, W-7405-ENG-48.

References

Aktas, O., and Aluru, N.R. (2002) "A Combined Continuum/DSMC Technique for Multiscale Analysis of Microfluidic Filters," *J. Comput. Phys.* **178**, pp. 342–72.

Alexander, F.J., Garcia, A.L., and Alder, B.J. (1994) "Direct Simulation Monte Carlo for Thin-Film Bearings," *Phys. Fluids* **6**, pp. 3854–60.

Alexander, F.J., Garcia, A.L., and Alder, B.J. (1995) "A Consistent Boltzmann Algorithm," *Phys. Rev. Lett.* **74**, p. 5212–15.

Alexander, F.J., Garcia, A.L., and Alder, B.J. (1998, 2000) "Cell Size Dependence of Transport Coefficients in Stochastic Particle Algorithms," *Phys. Fluids* **10**, p. 1540; erratum, *Phys. Fluids* **12**, p. 731.

Asmolov, E.S., Makashev, N.K., and Nosik, V.I. (1979) "Heat Transfer Between Plane Parallel Plates in a Gas of Maxwellian Molecules," *Sov. Phys. Dokl.* **24**, pp. 892–94.

Aubert, C., and Colin, S. (2001) "High-Order Boundary Conditions for Gaseous Flows in Rectangular Microducts, "*Microscale Thermophys. Eng.* **5**, 41–54.

Beskok, A., and Karniadakis, G.E. (1999) "A Model for Flows in Channels and Ducts at Micro and Nano Scales," *Microscale Thermophys. Eng.* **3**, p. 43.

Beskok, A. (2002) "Molecular-Based Microfluidic Simulation Models," in *Handbook of MEMS*, 1st ed., M. Gad-el-Hak, ed., CRC press, Boca Raton.

Bird, G.A. (1994) *Molecular Gas Dynamics and the Direct Simulation of Gas Flows* Clarendon Press, Oxford.

Boyd, I.D. (2003) "Predicting Breakdown of the Continuum Equations Under Rarefied Flow Conditions," in *Proceedings of the 23rd International Rarefied Gas Dynamics Symposium* pp. 899–906.

Breuer, K.S. (2002) "Lubrication in MEMS," in *Handbook of MEMS*, 1st ed., M. Gad-el-Hak, ed., CRC Press, Boca Raton.

Cercignani, C. (1964) "Higher Order Slip According to the Linearized Boltzmann Equation," Institute of Engineering Research Report AS-64–19, University of California, Berkeley.

Cercignani, C. (1988) *The Boltzmann Equation and its Applications* Springer-Verlag, New York.

Cercignani, C., and Lampis, M. (1971) "Kinetic Models for Gas-Surface Interactions," *Transp. Theory Stat. Phys.* **1**, p. 101.

Chapman, S., and Cowling, T.G. (1970) *The Mathematical Theory of Non-Uniform Gases,* Cambridge University Press.

Crandall, I.B. (1926) *Theory of Vibrating Systems and Sound*, Van Nostrand, New York.

Deissler, R.G. (1964) "An Analysis of Second-Order Slip Flow and Temperature-Jump Boundary conditions for Rarefied Gases," *Int. J. Heat Mass Transf.* **7**, pp. 681–94.

Fukui, S., and Kaneko, R. (1988) "Analysis of Ultra Thin Gas Film Lubrication Based on Linearized Boltzmann Equation: First Report-Derivation of a Generalized Lubrication Equation Including Thermal Creep Flow," *J. Tribol.* **110**, p. 253.

Gad-el-Hak, M. (2002) "Flow Physics," in *Handbook of MEMS*, 1st ed., M. Gad-el-Hak, ed., CRC Press, Boca Raton.

Gallis, M.A., Rader, D.J., and Torczynski, J.R. (2002) "Calculations of the Near-Wall Thermophoretic Force in Rarefied Gas Flow," *Phys. Fluids* **14**, pp. 4290–301.

Gallis, M.A., Torczynski, J.R., and Rader, D.J. (2001) "An Approach for Simulating the Transport of Spherical Particles in a Rarefied Gas Flow via the Direct Simulation Monte Carlo," *Phys. Fluids* **13**, pp. 3482–92.

Gallis, M.A., Torczynski, J.R., and Rader, D.J. (2004) "Molecular Gas Dynamics Observations of Chapman-Enskog Behavior and Departures Therefrom in Nonequilibrium Gases," *Phys. Rev. E.* **69**, Art. No 042201.

Garcia, A.L., and Alder, B.J. (1998) "Generation of the Chapman Enskog Distribution," *J. Comput. Phys.* **140**, pp. 66.

Garcia, A.L., Bell, J.B., Crutchfield, W.Y., and Alder, B.J. (1999) "Adaptive Mesh and Algorithm Refinement Using Direct Simulation Monte Carlo," *J. Comput. Phys.* **54**, p. 134.

Garcia, A.L., and Wagner, W. (2000) "Time Step Truncation Error in Direct Simulation Monte Carlo," *Phys. Fluids* **12**, p. 2621.

Grad, H. (1969) "Singular and Nonuniform Limits of Solutions of the Boltzmann equation," in *Transport Theory*, vol. **1**, SIAM-AMS Proceedings.

Graetz, L. (1885) "On the Thermal Conduction of Liquids," *Ann. Phys. Chem.*, **25**, pp. 337–57.

Hadjiconstantinou, N.G., and Patera, A.T. (1997) "Heterogeneous Atomistic-Continuum Representations for Dense Fluid Systems," *Int. J. Mod. Phys. C* **8**, pp. 967–76.

Hadjiconstantinou, N.G. (1999) "Hybrid Atomistic-Continuum Formulations and the Moving Contact-Line Problem," *J. Comput. Phys.* **154**, pp. 245–65.

Hadjiconstantinou, N.G. (2000) "Analysis of Discretization in the Direct Simulation Monte Carlo," *Phys. Fluids* **12**, pp. 2634–38.

Hadjiconstantinou, N.G. (2002) "Sound Wave Propagation in Transition-Regime Micro- and Nanochannels," *Phys. Fluids* **14**, p. 802.

Hadjiconstantinou, N.G., and Simek, O. (2002) "Constant-Wall-Temperature Nusselt Number in Micro and Nano Channels," *J. Heat Transf.* **124**, pp. 356–64.

Hadjiconstantinou, N.G. (2003a) "Comment on Cercignani's Second Order Slip Coefficient," *Phys. Fluids* **15**, pp. 2352–54.

Hadjiconstantinou, N.G. (2003b) "Dissipation in Small Scale Gaseous Flows," *J. Heat Transf.* **125**, pp. 944–47.

Hadjiconstantinou, N.G. (2005) "Validation of a Second-Order Slip Model for Dilute Gas Flows," *Microscale Thermophysical Engineering* **9**, pp. 137–53.

Hadjiconstantinou, N.G., Garcia, A.L., Bazant, M.Z., and He, G. (2003) "Statistical Error in Particle Simulations of Hydrodynamic Phenomena," *J. Comput. Phys.* **187**, pp. 274–97.

Hadjiconstantinou, N.G., and Simek, O. (2003) "Sound Propagation at Small Scales under Continuum and Non-Continuum Transport," *J. Fluid Mech.* **488**, pp. 399–408.

Hadjiconstantinou, N.G. unpublished.

Hamrock, B.J. (1994) *Fundamentals of Fluid Film Lubrication*, McGraw-Hill.

Ho, C.M., and Tai, Y.C. (1998) "Micro-Electro-Mechanical Systems (MEMS) and Fluid Flows," *Ann. Rev. Fluid Mech.* **30**, p. 579.

Karniadakis, G.E., and Beskok, A. (2001) *Microflows: Fundamentals and Simulation*, Springer.

Kirchhoff, G. (1868) "Ueber den Einflufs der Warmeleitung in einem Gase auf die Schallbewegung," *Ann. Phys. Chem.* **134**, pp. 177–93.

Knudsen, M. (1909) "Die Gesetze der molecular Stromung und dieinneren Reibungstromung der Gase durch Rohren," *Ann. Phys.* **28**, p. 75.

Kogan, M.N. (1969) *Rarefied Gas Dynamics*, Plenum Press, New York.

Lions, P.L. (1988). "On the Schwarz Alternating Method," I. In *First International Symposium on Domain Decomposition Methods for Partial Differential Equations*, SIAM, Philadelphia, pp. 1–42.

Lord, R.G. (1995) "Some Further Extensions of the Cercignani-Lampis Gas-Surface Interaction Model," *Phys. Fluids* **7**, pp. 1159–61.

Maurer, J., Tabeling, P., Joseph, P., and Willaime, H. (2003) "Second-Order Slip Laws in Microchannels for Helium and Nitrogen," *Phys. Fluids* **15**, pp. 2613–21.

Ohwada, T. (1998) "Higher Order Approximation Methods for the Boltzmann Equation," *J. Comput. Phys.* **139**, p. 1.

Ohwada, T., Sone, Y., and Aoki, K. (1989a) "Numerical Analysis of the Poiseuille and Thermal Transpiration Flows between Parallel Plates on the Basis of the Boltzmann Equation for Hard-Sphere Molecules," *Phys. Fluids* **1**, p. 2042.

Ohwada, T., Sone, Y., and Aoki, K. (1989b) "Numerical Analysis of the Shear and Thermal Creep Flows of a Rarefied Gas over a Plane Wall on the Basis of the Linearized Boltzmann Equation for Hard-Sphere Molecules," *Phys. Fluids* **1**, pp. 1588–99.

Park, J.H., Bahukudumbi, P., and Beskok, A. (2004) "Rarefaction Effects on Shear Driven Oscillatory Gas Flows: A Direct Simulation Monte Carlo Study in the Entire Knudsen regime," *Phys. Fluids* **16**, pp. 317–30.

Rayleigh, J.W.S. (1896) *The Theory of Sound*, vol. 2, Macmillan.

Sone, Y. (1964) "Kinetic Theory Analysis of Linearized Rayleigh Problem," *J. Phys. Soc. Jap.* **19**, pp. 1463–73.

Sone, Y. (1969) "Asymptotic Theory of Flow of Rarefied Gas over a Smooth Boundary I," *Proceedings of the Sixth International Symposium on Rarefied Gas Dynamics* vol. 1, pp. 243–53, Academic Press.

Sone, Y., Ohwada T., and Aoki, K. (1989) "Temperature Jump and Knudsen Layer in a Rarefied Gas over a Plane Wall: Numerical Analysis of the Linearized Boltzmann Equation for Hard-Sphere Molecules," *Phys. Fluids* **1**, pp. 363–70.

Sone, Y., and Onishi, Y. (1978) "Kinetic Theory of Evaporation and Condensation-Hydrodynamic Equation and Slip Boundary Condition," *J. Phys. Soc. Jap.* **44**, pp. 1981–94.

Sparrow, E.M., and Lin, S.H. (1962) "Laminar Heat Transfer in Tubes Under Slip Flow Conditions," *J. Heat Transf.* **84**, p. 363.

Sun, Q.H., and Boyd, I.D. (2002) "A Direct Simulation Method for subsonic, Microscale Gas Flows," *J. Comput. Phys.* **179**, pp. 400–25.

Tyndall, J. (1870) "On Dust and Disease," *Proceedings of the Royal Institution of Great Britain* **6**, pp. 1–14.

Vincenti, W.G., and Kruger, C.H. (1965) *Introduction to Physical Gas Dynamics*, Krieger, Florida.

Wagner, W. (1992) "A Convergence Proof for Bird's Direct Simulation Monte Carlo Method for the Boltzmann Equation," *J. Stat. Phys.* **66**, p. 1011.

Wesseling, P. (2001) *Principles of Computational Fluid Dynamics*, Springer.

Wijesinghe, S., and Hadjiconstantinou, N.E. (2001) "Velocity Slip and Temperature Jump in Dilute Hard Sphere Gases at Finite Knudsen Numbers," *Proceedings of the First MIT Conference on Computational Fluid and Solid Mechanics*, vol. 2, pp. 1019–21.

Wijesinghe, S., and Hadjiconstantinou, N.G. (2002) "A Hybrid Continuum-Atomistic Scheme for Viscous Incompressible Flow," in *Proceedings of the 23th International Symposium on Rarefied Gas Dynamics*, July, Whistler, British Columbia, pp. 907–14.

Wijesinghe, H.S., and Hadjiconstantinou, N.G. (2004) "A Discussion of Hybrid Atomistic-Continuum Methods for Multiscale Hydrodynamics," *International Journal of Multiscale Computational Engineering* **2**, pp. 189–202.

Wijesinghe, H.S., Hornung, R., Garcia, A.L., and Hadjiconstantinou, N.G. (2003) "3-Dimensional Hybrid Continuum-Atomistic Simulations for Multiscale Hydrodynamics," *Proceedings of the 2003 International Mechanical Engineering Congress and Exposition*, vol. 2, paper NANOT-41251.

Wijesinghe, H.S., Hornung, R., Garcia, A.L., and Hadjiconstantinou, N.G. (2004) "Three-Dimensional Hybrid Continuum-Atomistic Simulations for Multiscale Hydrodynamics," *J. Fluids Eng.* **126**, pp. 768–77.

8

Burnett Simulations of Flows in Microdevices

Ramesh K. Agarwal
Washington University in St. Louis

Keon-Young Yun
Samhongsa Co., Ltd.

8.1 Introduction

Microelectromechanical systems (MEMS) are currently attracting a great deal of interest because of their vast potential in industrial and medical applications. As a result, considerable effort is being devoted to the design and fabrication of MEMS. MEMS refers to devices that have characteristic length between 1 µm and 1 mm, that combine electrical and mechanical components, and that are fabricated using integrated-circuit batch-processing technologies. A few examples of MEMS are microsensors, microactuators, micromotors, microvalves, micropumps, and microducts. Fluid flows in microdevices, such as microvalves, micropumps and microducts, are significantly different from those in macroscopic devices, due to the microdevices' small characteristic sizes. Hence, understanding the physics of the flows in the microdevices is very important in their development and design.

Various regimes of fluid flows can be broadly classified into the continuum, continuum–transition, transition, and free molecular regimes as shown in Table 8.1. For a large class of flows, Navier–Stokes equations based on the continuum approximation are adequate to model the fluid behavior. Continuum approximation implies that the mean free path of the molecules λ in a gas is much smaller than the characteristic length L of interest (say, the body dimension); that is, the Knudsen number Kn $(= \lambda/L)$ is very small $(\ll 1)$. However, for a variety of flows, this assumption is not valid; the Knudsen number is

TABLE 8.1 Flow Regimes and Fluid Models

Knudsen Number	Fluid Model
$Kn \to 0$ (continuum, no molecular diffusion)	Euler equations
$Kn \leq 10^{-3}$ (continuum with molecular diffusion)	Navier–Stokes equations with no-slip-boundary conditions
$10^{-3} \leq Kn \leq 10^{-1}$ (continuum–transition)	Navier–Stokes equations with slip-boundary conditions
$10^{-1} \leq Kn \leq 10$ (transition)	Burnett equations with slip-boundary conditions Moment equations Direct Simulation Monte Carlo (DSMC) Boltzmann equation
$Kn > 10$ (free molecular flow)	Collisionless Boltzmann equation DSMC

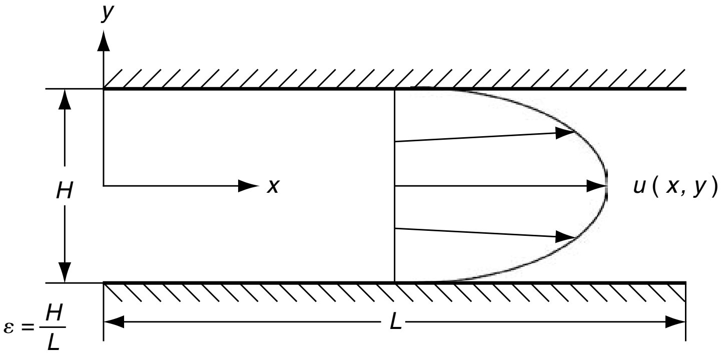


FIGURE 8.1 Flow in a microchannel. Relevant flow parameters: Mach number, Reynolds number, and Knudsen number are $M = \bar{u}/c$, $Re = \bar{\rho}\bar{u}H/\mu$, and $Kn = (\pi\gamma/2)^{0.5} M/Re$, respectively. " – " denotes the average outlet conditions.

of $O(1)$. In these flows, the gas is neither completely in the continuum regime nor in the rarefied (free molecular flow) regime. Therefore, such flows have been categorized as continuum–transition or transitional flows. Examples of such flows include the hypersonic flows about space vehicles in low earth orbit [Ivanov and Gimelshein, 1998] or flows in microchannels of MEMS [Gad-el-Hak, 1999].

In high-altitude hypersonic flows, low density gives rise to high Knudsen number effects, while in microscale flows, which usually occur at atmospheric conditions, small length scales create regions of high Knudsen numbers. In the case of high-altitude hypersonic flows, the shock layer thickness at the nose of a space vehicle (shuttle) is much thicker than that predicted from the Navier–Stokes equations. In a long microchannel, the pressure gradient is observed to be nonconstant and the experimentally measured mass flow rate is higher than that predicted from the conventional continuum flow model [Arkilic et al., 1997; Harley et al., 1995; Liu et al., 1993; Pong et al., 1994]. In such a microscale flow, the mean free path of the molecules can be of the same order of magnitude as the characteristic length of the microchannel: $Kn \sim O(1)$. For a microchannel defined by ratio $\varepsilon = H/L$, where H and L are width and length of the channel respectively as shown in Figure 8.1, Arkilic et al. (1997) have characterized various flow regimes depending upon the Reynolds number Re and Mach number M of the flow as shown in Table 8.2. Tables 8.1 and 8.2 together now can be used to select an appropriate fluid model for simulation of the flow field in a microchannel. Both low-density and microscale effects can be local in a flow so that the entire flow is in both the continuum and transition regimes.

As shown in Table 8.1, Navier–Stokes equations are not adequate to model the flows in the continuum–transition regime; the Boltzmann equation describes the flow in all the regimes — continuum, continuum–transition, and free molecular. The techniques available for solving the Boltzmann equation can

TABLE 8.2 Flow Regimes in a Microchannel for Different Knudsen Numbers

	Re		
M	$O(\varepsilon)$	$O(1)$	$O(1/\varepsilon)$
$O(\varepsilon)$	$Kn = O(1)$; creeping microflow	$Kn = O(\varepsilon)$; moderate microflow	$Kn = O(\varepsilon^2)$; low *M* Fanno flow
O(1)	$Kn = O(1/\varepsilon)$; transonic free molecular flow	$Kn = O(1)$; transonic microflow	$Kn = O(\varepsilon)$; transonic Fanno flow
$O(1/\varepsilon)$	$Kn = O(1/\varepsilon^2)$; hypersonic free molecular flow	$Kn = O(1/\varepsilon)$; hypersonic free molecular flow	$Kn = O(1)$; hypersonic Fanno (transitional) flow

Reprinted with permission from Arkilic, E.B. et al. (1997) "Gaseous Flow in Long Microchannel," *J. MEMS* **6**, 167–78.

be classified as particulate methods and moment methods. The direct simulation Monte Carlo (DSMC) method falls in the category of particulate methods [Bird, 1994]. Moment methods derive the higher order fluid dynamics approximations beyond Navier–Stokes equations to account for departures from thermal equilibrium. The higher order fluid dynamic models are known as the extended hydrodynamic equations (EHE) or generalized hydrodynamic equations (GHE). However, both classes of methods have significant limitations — either in describing the physics or in the computational resources needed for accurate simulation — for modeling flows in the continuum–transition regime. Currently, the DSMC method can be considered the most accurate and widely used technique for computation of low-density flows. However, in the continuum–transition regime, where the densities are not low enough, the DSMC method requires a large number of particles for accurate simulation making the technique prohibitively expensive both in terms of computational time and memory requirements. For example, Koppenwallner (1987) has shown that the space shuttle's nose-up pitching moment was predicted inaccurately by the DSMC method in the continuum–transition regime due to the inadequate number of particles used in modeling. The nose-up pitching moment could be corrected by deflecting the body flap to 15 degrees — twice the amount predicted by DSMC. A similar situation may occur in microscale flows due to the relatively high density and low velocity requiring enormous computational power and resulting in large statistical scatter in the DSMC simulations [Nance et al., 1998].

The majority of the computations with the DSMC method, especially in three dimensions for $Kn = O(1)$, are beyond the currently available computing power. As an alternative, higher order extended or generalized hydrodynamic equations have been proposed that have the potential to perform reasonably well in both the continuum and continuum–transition regimes. The extended hydrodynamic equations have been derived from the Boltzmann equation using either one of the following approaches. In one approach, higher order constitutive relations (beyond Navier–Stokes) for stress and heat transfer terms are obtained using the Chapman–Enskog expansion of the Boltzmann equation with the Knudsen number as a parameter. In the Chapman–Enskog expansion, the first term represents the Maxwellian (equilibrium) distribution function f_0. The first moment of the Boltzmann equation with the collision invariant vector, with f_0 as the approximation for the distribution function, results in the Euler equations. The first two terms in the Chapman–Enskog expansion — $(f_0 + Kn\, f_1)$ — give a distribution function corresponding to the Navier–Stokes equations representing a first-order departure from thermal equilibrium. The first three terms $(f_0 + Kn\, f_1 + Kn^2\, f_2)$ in the expansion give a distribution function, which results in the so-called Burnett equations representing a second-order departure from the equilibrium.

Burnett equations have been a subject of considerable investigation in recent years and are the main subject of this chapter. Higher order approximations beyond Burnett equations, the so-called super-Burnett equations, etc., can be derived by continuing the Chapman–Enskog expansion to higher orders. Presently, however, the complexity of the highly nonlinear Burnett stress and heat transfer terms itself is enormously challenging both computationally and in terms of understanding the physics, so the consideration of super-Burnett equations and beyond is meaningless.

Burnett stress and heat transfer terms contain higher than second-order derivatives. Therefore, an additional boundary condition is necessary for the solution to the Burnett equations to be uniquely determined; different solutions can result based on the choice of boundary values [Lee, 1994]. Furthermore, it has also been shown that the conventional Burnett equations can violate the second law of thermodynamics at high Knudsen numbers [Comeaux et al., 1995]. Because the focus of this chapter is on Burnett equations, they are described in detail in Section 8.2.

In another approach, the extended hydrodynamic equations are derived using the moment method, which employs the equations of transfer instead of dealing with the distribution function. In the moment method, the distribution function *f* is expanded in moments of physical variables (density, velocity, pressure, temperature, etc.) and the evolution equations for moments are derived from the Boltzmann equation. In principle, this approach should result in a set of macroscopic equations consistent with the second law of thermodynamics, but many of the methods (for example, Grad's 13-moment method [Grad, 1949]) result in the entropy equation violating the Gibb's relation [Holway, 1964; Weiss, 1996]. This problem was addressed in the recent work of Levermore (1996) and of Levermore and Morokoff (1998) by the so-called Gaussian closure. The Gaussian closure is based on a more elegant choice of a finite-dimensional linear subspace and yields a hyperbolic system of moment equations. Because the hyperbolic equations are easier to solve numerically, Groth et al. (1995) have developed some computational models based on this closure. However, the Gaussian closure is of limited practical interest, as the primary system with ten variables admits no heat flux. Other moment systems (for example, the 35-moment system of Brown [1996]) do not yield numerical solutions above Mach numbers of approximately two. Furthermore, the application of the 13- or 35-moment systems to three-dimensional problems remains computationally prohibitive at present.

Because of the physical and numerical difficulties associated with the Burnett equations and moment equations, Myong (1999) has suggested yet another set of generalized hydrodynamic equations based on the work of Eu (1992). Eu's equations are based on a nonequilibrium canonical distribution function and a cumulant expansion of the collision integral in Boltzmann equation. These equations can be considered as the most thermodynamically consistent macroscopic equations, as the second law of thermodynamics is satisfied to every order of approximation. It also turns out that they recover the correct behavior in both the continuum and free molecular limits. Myong (1999) has developed a computational model based on Eu's evolution equations within the framework of 13 moments. This model so far has been applied to some one-dimensional problems, but the full potential of this set of equations for calculating two- and three-dimensional flows in the continuum–transition regime remains to be determined and will require several years of intensive computational effort. Furthermore, the solution of Eu's equations for a three-dimensional problem will remain computationally prohibitive in the near future.

Because of the limitations of the DSMC method, a hybrid approach has been suggested by many investigators [Oran et al., 1998; Roveda et al., 1998]. The hybrid method couples a Euler or Navier–Stokes solver with DSMC. The hybrid codes have been developed for problems that contain disconnected nonequilibrium regions embedded in a continuum flow [Roveda et al., 1998]. However, the development of a hybrid code is not simple, as two issues need to be resolved before implementation: (1) when to switch between the two methods, and (2) how to pass information from one method to the other [Boyd et al., 1995]. Furthermore, a conceptual inconsistency remains, as the hybrid method must recover both the continuum and free molecular limits.

Several modifications to the original Burnett equations that have been proposed in the literature are discussed in Section 8.2. Sections 8.3 and 8.4 describe the governing equations and the wall-boundary conditions, respectively. Section 8.5 deals with the linearized stability analysis of one-dimensional Burnett equations. Section 8.6 briefly describes the numerical scheme and other computational aspects of the three-dimensional Burnett solver. In Section 8.7, computational results are presented for one- and two-dimensional problems. They include computations for hypersonic shock structures, blunt body flows, subsonic flow past an airfoil, and subsonic and supersonic flow in a microchannel. Although the focus of this chapter is on flows in microdevices, the hypersonic flow computations for blunt body flows, etc., are presented here because traditionally the Burnett equations have been applied to compute this class of flows over the past decade, and computational results from Navier–Stokes and DSMC simulations can be used for the purpose of comparison. These solutions are instructive in providing some assessment of the accuracy and applicability of Burnett equations for computing flows in the continuum–transition regime.

8.2 History of Burnett Equations

Table 8.3 briefly traces the history of Burnett equations. In 1935, Burnett (1935) developed constitutive relationships for the stress and heat transfer terms by applying the Chapman–Enskog expansion to the Boltzmann equation for second-order departures from collisional equilibrium. These equations are referred to as the original Burnett equations. In 1939, Chapman and Cowling (1970) replaced the material derivatives in the original Burnett equations by spatial derivatives obtained from inviscid Euler equations. This alternative form of the original Burnett equations is referred to as the conventional Burnett equations. Expressing the material derivatives in terms of the spatial derivatives was considered acceptable as the Navier–Stokes and Burnett equations were considered to be first- and second-order corrections to the Euler equations. The use of Euler equations to express the material derivatives retained the second-order accuracy of the Burnett equations. For reasons unknown, the conventional Burnett equations and not the original Burnett equations became the set of higher order constitutive relations studied during the past six decades.

Fiscko and Chapman (1988) and Zhong (1991) have employed the conventional Burnett equations to extend the numerical methods for continuum flow into the continuum–transition regime by incorporating the additional linear and nonlinear stress and heat transfer terms in the standard Navier–Stokes solvers. In one of the earliest attempts to numerically solve the conventional Burnett equations, Fiscko and Chapman (1988) solved the hypersonic shock structure problem by relaxing an initial solution to steady state. They obtained solutions for a variety of Mach numbers and concluded that the conventional Burnett equations do indeed describe the normal shock structure better than the Navier–Stokes equations at high Mach numbers. However, they experienced stability problems when the computational grids were made progressively finer.

TABLE 8.3 Brief History of Burnett Set of Equations

Equations	Ref.	Comments
Burnett equations	Burnett (1935)	Derived from Boltzmann equation by considering the first three terms of the Chapman–Enskog expansion; appearance of material derivatives, D()/Dt, in the second-order (Burnett) flux vectors.
Conventional Burnett equations	Chapman and Cowling (1970)	Euler equations were used to express the material derivatives in terms of the spatial derivatives.
Conventional Burnett equations	Fiscko and Chapman (1988)	Encountered problem of small wavelength instability as the grids were refined.
Augmented Burnett equations	Zhong (1991)	Linearized third-order terms were added to stabilize the Burnett equations; not entirely successful for computing blunt body wakes and flat plate boundary layers.
Conventional Burnett equations	Welder et al. (1993)	Due to the nonlinear terms in the Burnett equations, linear stability analysis alone is not sufficient to explain the instability at high Knudsen numbers.
Conventional Burnett equations	Comeaux et al. (1995)	Burnett equations can violate the second law of thermodynamics at high Knudsen numbers.
BGK–Burnett equations	Balakrishnan and Agarwal (1996)	Nonlinear collision integral in the Boltzmann equation was simplified by representing it with the Bhatnagar–Gross–Krook (BGK) model; material derivatives expressed in terms of the spatial derivatives using Navier–Stokes equations; linear stability analysis shows unconditional stability for all Knudsen numbers; when Euler equations are used to express the material derivatives, they guarantee unconditional stability for monatomic gases; entropy consistent (satisfy the Boltzmann's H-theorem) for a wide range of Knudsen numbers.

In a subsequent attempt, Zhong (1991) showed that in order to maintain second-order accuracy the conventional Burnett equations could be stabilized by adding linear third-order terms from the super-Burnett equations to the stress and heat transfer terms in the Burnett equations. This set of equations was termed the augmented Burnett equations. The coefficients (weights) of these linear third-order terms were determined by carrying out a linearized stability analysis of the augmented Burnett equations. The augmented Burnett equations did not present any stability problems when they were used to compute the hypersonic shock structure and hypersonic blunt body flows. However, attempts at computing the flow fields for blunt body wakes and flat-plate boundary layers with the augmented Burnett equations have not been entirely successful. Furthermore, the ad hoc addition of the linear super-Burnett terms and their necessity raises the question of whether the approximation used to create the conventional Burnett equations from the original Burnett equations introduces the small wavelength instabilities. Welder et al. (1993) noted that linear stability analysis alone is not sufficient to explain the instability of Burnett equations with increasing Knudsen numbers as this analysis does not take into account many nonlinear terms, products of first- and higher-order derivatives, that are present in the conventional Burnett equations. Comeaux et al. (1995) have recently surmised that this instability may also be attributed to the fact that the conventional Burnett equations can violate the second law of thermodynamics at high Knudsen numbers.

In order to overcome the difficulties associated with the conventional Burnett equations, Balakrishnan and Agarwal (1996, 1997) have recently derived a new set of Burnett equations designated as "BGK–Burnett" equations, which are entropy consistent and satisfy the Boltzmann H-theorem. The highly nonlinear nature of the collision integral in the Boltzmann equation presents the biggest hurdle in devising a higher order distribution function. This problem can be circumvented by representing the collision integral in the Bhatnagar–Gross–Krook (BGK) form [Bhatnagar et al., 1954]. This approximation assumes that any slight departure from the equilibrium distribution will eventually settle down to the equilibrium distribution exponentially. This approximation also assumes that the gas is dilute, hence the collision processes are predominantly binary in nature. Because only binary collisions are considered, the time taken for the nonequilibrium distribution to settle down to the equilibrium level is equal to the reciprocal of the collision frequency. With the BGK approximation to the collision integral, the exact closed-form analytical expression for the distribution function to any order can be obtained.

Balakrishnan and Agarwal (1997) have derived the BGK–Burnett equations by considering the first three terms in the Chapman–Enskog expansion. In this derivation, Euler equations were used to approximate the material derivatives in the first-order distribution function. Moments of the first-order distribution function with the collision invariant vector yield the Navier–Stokes equations. In order to keep in step with the iterative refinement technique, it was conjectured that the Navier–Stokes equations could be used to approximate the material derivatives in the second-order distribution function. It has been shown that this formulation ensures a positive entropy change. The BGK–Burnett equations are obtained by taking moments of this second-order distribution function with the collision invariant vector. This set of equations contains all the stress and heat transfer terms reported by Fiscko and Chapman (1988) and has additional terms that are similar to the super-Burnett terms. Linearized stability analysis has shown that these additional terms make the BGK–Burnett equations unconditionally stable for monatomic as well as polyatomic gases. In order to check if the entropy production is positive throughout the flow field, the Boltzmann H-theorem was applied to the second-order distribution function. It was shown that H is a monotonically decreasing function thereby ensuring that the equations do not violate the second law of thermodynamics [Balakrishnan et al., 1997]. Thus the BGK–Burnett equations overcome the problems associated with the conventional Burnett equations — namely, violation of the second law of thermodynamics and instability at high Knudsen numbers.

8.3 Governing Equations

In this section, the augmented Burnett and BGK–Burnett equations are presented. For original and conventional Burnett equations, refer to the papers by Burnett (1935) and Chapman and Cowling (1970), respectively. Here, we present only the two-dimensional augmented and BGK–Burnett equations for the sake of brevity; three-dimensional augmented Burnett equations and BGK–Burnett equations are given in Yun et al. (1998a and 1998b).

The governing equations for two-dimensional, unsteady, compressible, viscous flow can be written in Cartesian coordinates as:

$$\frac{\partial \mathbf{Q}}{\partial t} + \frac{\partial \mathbf{E}}{\partial x} + \frac{\partial \mathbf{F}}{\partial y} = 0 \tag{8.1}$$

where

$$\mathbf{Q} = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ e_t \end{bmatrix} \tag{8.2}$$

In Equation (8.1), **E** and **F** are the flux vectors of the flow variables **Q** in the *x* and *y* directions respectively. These flux vectors can be written as:

$$\begin{aligned} \mathbf{E} &= \mathbf{E}_I + \mathbf{E}_V \\ \mathbf{F} &= \mathbf{F}_I + \mathbf{F}_V \end{aligned} \tag{8.3}$$

where $\mathbf{E}_I$ and $\mathbf{F}_I$ are the inviscid-flux terms and E_V and F_V are the viscous-flux terms given as follows:

$$\mathbf{E}_I = \begin{bmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ (e_t + p)u \end{bmatrix}, \quad \mathbf{E}_V = \begin{bmatrix} 0 \\ \sigma_{11} \\ \sigma_{12} \\ \sigma_{11}u + \sigma_{12}v + q_1 \end{bmatrix} \tag{8.4}$$

$$\mathbf{F}_I = \begin{bmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ (e_t + p)v \end{bmatrix}, \quad \mathbf{F}_V = \begin{bmatrix} 0 \\ \sigma_{21} \\ \sigma_{22} \\ \sigma_{21}u + \sigma_{22}v + q_2 \end{bmatrix} \tag{8.5}$$

The constitutive equations for a gas flow near thermodynamic equilibrium can be derived as approximate solutions of the Boltzmann equation using the Chapman–Enskog expansion. This method yields the general constitutive relations for the stress tensor σ_{ij} and the heat-flux vector q_i as follows:

$$\begin{aligned} \sigma_{ij} &= \sigma_{ij}^{(0)} + \sigma_{ij}^{(1)} + \sigma_{ij}^{(2)} + \sigma_{ij}^{(3)} + \cdots + \sigma_{ij}^{(n)} + O(Kn^{n+1}) \\ q_i &= q_i^{(0)} + q_i^{(1)} + q_i^{(2)} + q_i^{(3)} + \cdots + q_i^{(n)} + O(Kn^{n+1}) \end{aligned} \tag{8.6}$$

where n represents the order of accuracy with respect to Kn. Kn is defined as:

$$Kn = \lambda/L \tag{8.7}$$

where L is the macroscopic characteristic length, and the mean free path λ is given by:

$$\lambda = \frac{16\mu}{5\rho\sqrt{2\pi RT}} \tag{8.8}$$

In the case of $Kn \approx 0$, only the first terms in Equation (8.6) are important. The zeroth-order approximation ($n = 0$) results in the Euler equations with:

$$\sigma_{ij}^{(0)} = 0 \quad \text{and} \quad q_i^{(0)} = 0 \tag{8.9}$$

When $Kn < 0.1$, the first two terms in Equation (8.6) become important for the accurate representation of the stress and heat transfer properties of the gas flow. This first-order approximation represents the Navier–Stokes equations. The stress tensor and the heat-flux terms ($n = 1$) are given as:

$$\begin{aligned}
\sigma_{11}^{(1)} &= -\mu(\delta_1 u_x + \delta_2 v_y) \\
\sigma_{12}^{(1)} &= \sigma_{21}^{(1)} = -\mu(u_y + v_x) \\
\sigma_{22}^{(1)} &= -\mu(\delta_1 v_y + \delta_2 u_x) \\
q_1^{(1)} &= -\kappa T_x \\
q_2^{(1)} &= -\kappa T_y
\end{aligned} \tag{8.10}$$

where $(\ )_x = \partial/\partial x$ and $(\ )_y = \partial/\partial y$. The coefficients (δ_1, δ_2) are $(1.333, -0.666)$ and $(1.6, -0.4)$ for the augmented Burnett equations and the BGK–Burnett equations (for $\gamma = 1.4$), respectively.

As Kn becomes larger (>0.1), additional higher order terms in Equation (8.6) are required. The second-order approximation yields the Burnett equations that retain the first three terms in Equation (8.6). The expression for stress and heat-flux terms ($n = 2$) are obtained as [Yun et al., 1998b]:

$$\begin{aligned}
\sigma_{11}^{(2)} = \frac{\mu^2}{p}\Bigg(&\alpha_1 u_x^2 + \alpha_2 u_x v_y + \alpha_3 v_y^2 + \alpha_4 u_x v_y + \alpha_5 u_y^2 + \alpha_6 v_x^2 + \alpha_7 RT_{xx} + \alpha_8 RT_{yy} + \alpha_9 \frac{RT}{\rho}\rho_{xx} \\
&+ \alpha_{10}\frac{RT}{\rho}\rho_{yy} + \alpha_{11}\frac{RT}{\rho^2}\rho_x^2 + \alpha_{12}\frac{R}{\rho}T_x\rho_x + \alpha_{13}\frac{R}{T}T_x^2 + \alpha_{14}\frac{RT}{\rho^2}\rho_y^2 \\
&+ \alpha_{15}\frac{R}{\rho}T_y\rho_y + \alpha_{16}\frac{R}{T}T_y^2\Bigg)
\end{aligned} \tag{8.11}$$

$$\begin{aligned}
\sigma_{22}^{(2)} = \frac{\mu^2}{p}\Bigg(&\alpha_1 v_y^2 + \alpha_2 u_x v_y + \alpha_3 u_x^2 + \alpha_4 u_y v_x + \alpha_5 v_x^2 + \alpha_6 u_y^2 + \alpha_7 RT_{yy} + \alpha_8 RT_{xx} + \alpha_9 \frac{RT}{\rho}\rho_{yy} \\
&+ \alpha_{10}\frac{RT}{\rho}\rho_{xx} + \alpha_{11}\frac{RT}{\rho^2}\rho_y^2 + \alpha_{12}\frac{R}{\rho}T_y\rho_y + \alpha_{13}\frac{R}{T}T_y^2 + \alpha_{14}\frac{RT}{\rho^2}\rho_x^2 + \alpha_{15}\frac{R}{\rho}T_x\rho_x \\
&+ \alpha_{16}\frac{R}{T}T_x^2\Bigg)
\end{aligned} \tag{8.12}$$

$$\begin{aligned}
\sigma_{12}^{(2)} = \sigma_{21}^{(2)} = \frac{\mu^2}{p}\Bigg(&\beta_1 u_x u_y + \beta_2 u_y v_y + \beta_2 u_x v_x + \beta_1 v_x v_y + \beta_3 RT_{xy} + \beta_4 \frac{RT}{\rho}\rho_{xy} \\
&+ \beta_5 \frac{R}{T}T_x T_y + \beta_6 \frac{RT}{\rho^2}\rho_x\rho_y + \beta_7 \frac{R}{\rho}\rho_x T_y + \beta_7 \frac{R}{\rho}T_x\rho_y\Bigg)
\end{aligned} \tag{8.13}$$

$$\begin{aligned}
q_1^{(2)} = \frac{\mu^2}{\rho}\Bigg(&\gamma_1 \frac{1}{T}T_x u_x + \gamma_2 \frac{1}{T}T_x v_y + \gamma_3 u_{xx} + \gamma_4 u_{yy} + \gamma_5 v_{xy} + \gamma_6 \frac{1}{T}T_y v_x \\
&+ \gamma_7 \frac{1}{T}T_y u_y + \gamma_8 \frac{1}{\rho}\rho_x u_x + \gamma_9 \frac{1}{\rho}\rho_x v_y + \gamma_{10}\frac{1}{\rho}\rho_y u_y + \gamma_{11}\frac{1}{\rho}\rho_y v_x\Bigg)
\end{aligned} \tag{8.14}$$

$$\begin{aligned}
q_2^{(2)} = \frac{\mu^2}{\rho}\Bigg(&\gamma_1 \frac{1}{T}T_y u_y + \gamma_2 \frac{1}{T}T_y v_x + \gamma_3 v_{yy} + \gamma_4 v_{xx} + \gamma_5 u_{xy} + \gamma_6 \frac{1}{T}T_x u_y \\
&+ \gamma_7 \frac{1}{T}T_x v_x + \gamma_8 \frac{1}{\rho}\rho_y v_y + \gamma_9 \frac{1}{\rho}\rho_y u_x + \gamma_{10}\frac{1}{\rho}\rho_x v_x + \gamma_{11}\frac{1}{\rho}\rho_x u_y\Bigg)
\end{aligned} \tag{8.15}$$

Both the augmented Burnett and BGK–Burnett equations have the same forms of the stress tensor and heat-flux terms in the second-order approximation; however, the two sets of equations have different values of the coefficients. The coefficients for the augmented Burnett equations (for a hard sphere gas) are $\alpha_1 = 1.199$, $\alpha_2 = 0.153$, $\alpha_3 = -0.600$, $\alpha_4 = -0.115$, $\alpha_5 = 1.295$, $\alpha_6 = -0.733$, $\alpha_7 = 0.260$, $\alpha_8 = -0.130$, $\alpha_9 = -1.352$, $\alpha_{10} = 0.676$, $\alpha_{11} = 1.352$, $\alpha_{12} = -0.898$, $\alpha_{13} = 0.600$, $\alpha_{14} = -0.676$, $\alpha_{15} = 0.449$, $\alpha_{16} = -0.300$, $\beta_1 = -0.115$, $\beta_2 = 1.913$, $\beta_3 = 0.390$, $\beta_4 = -2.028$, $\beta_5 = 0.900$, $\beta_6 = 2.028$, $\beta_7 = -0.676$, $\gamma_1 = 10.830$, $\gamma_2 = 0.407$, $\gamma_3 = -2.269$, $\gamma_4 = 1.209$, $\gamma_5 = -3.478$, $\gamma_6 = -0.611$, $\gamma_7 = 11.033$, $\gamma_8 = -2.060$, $\gamma_9 = 1.030$, $\gamma_{10} = -1.545$, and $\gamma_{11} = -1.545$.

The coefficients for the BGK–Burnett equations (for $\gamma = 1.4$) are $\alpha_1 = -2.24$, $\alpha_2 = -0.48$, $\alpha_3 = 0.56$, $\alpha_4 = -1.20$, $\alpha_5 = 0.0$, $\alpha_6 = 0.0$, $\alpha_7 = -19.6$, $\alpha_8 = -5.6$, $\alpha_9 = -1.6$, $\alpha_{10} = 0.4$, $\alpha_{11} = 1.6$, $\alpha_{12} = -19.6$, $\alpha_{13} = -18.0$, $\alpha_{14} = -0.4$, $\alpha_{15} = -5.6$, $\alpha_{16} = -6.9$, $\beta_1 = -1.4$, $\beta_2 = -1.4$, $\beta_3 = 0.0$, $\beta_4 = -2.0$, $\beta_5 = 2.0$, $\beta_6 = 2.0$, $\beta_7 = 0.0$, $\gamma_1 = -25.241$, $\gamma_2 = -0.2$, $\gamma_3 = -1.071$, $\gamma_4 = -2.0$, $\gamma_5 = -2.8$, $\gamma_6 = -7.5$, $\gamma_7 = -11.0$, $\gamma_8 = -1.271$, $\gamma_9 = 1.0$, $\gamma_{10} = -3.0$, and $\gamma_{11} = -3.0$.

The third-order approximation ($n = 3$) represents the super-Burnett equations; however, not all of the third-order terms of the super-Burnett equations are used in the augmented Burnett and the BGK–Burnett equations. In the augmented Burnett equations, the third-order terms are added on an ad hoc basis to obtain stable numerical solutions while maintaining second-order accuracy of the solutions. The third-order terms in the augmented Burnett equations are given as [Yun et al., 1998b]:

$$\sigma_{11}^{(a)} = \frac{\mu^3}{p^2} RT(\alpha_{17}u_{xxx} + \alpha_{17}u_{xyy} + \alpha_{18}v_{xxy} + \alpha_{18}v_{yyy}) \tag{8.16}$$

$$\sigma_{22}^{(a)} = \frac{\mu^3}{p^2} RT(\alpha_{17}v_{yyy} + \alpha_{17}v_{xxy} + \alpha_{18}u_{xyy} + \alpha_{18}u_{xxx}) \tag{8.17}$$

$$\sigma_{12}^{(a)} = \sigma_{21}^{(a)} = \frac{\mu^3}{p^2} RT(\beta_8 u_{xxy} + \beta_8 u_{yyy} + \beta_8 v_{xyy} + \beta_8 v_{xxx}) \tag{8.18}$$

$$q_1^{(a)} = \frac{\mu^3}{p\rho} R\left(\gamma_{12}T_{xxx} + \gamma_{12}T_{xyy} + \gamma_{13}\frac{\mathrm{T}}{\rho}\rho_{xxx} + \gamma_{13}\frac{T}{\rho}\rho_{xyy}\right) \tag{8.19}$$

$$q_2^{(a)} = \frac{\mu^3}{p\rho} R\left(\gamma_{12}T_{yyy} + \gamma_{12}T_{xxy} + \gamma_{13}\frac{T}{\rho}\rho_{yyy} + \gamma_{13}\frac{T}{\rho}\rho_{xxy}\right) \tag{8.20}$$

The superscript (a) denotes the augmented Burnett terms. The coefficients in stress and heat-flux terms are $\alpha_{17} = 0.2222$, $\alpha_{18} = -0.1111$, $\beta_8 = 0.1667$, $\gamma_{12} = 0.6875$, and $\gamma_{13} = -0.625$.

The BGK–Burnett equations have more additional third-order terms than the augmented Burnett equations. These are not added on an ad hoc basis but are derived from the second-order Chapman–Enskog expansion of the BGK–Boltzmann equation. The third-order terms in the BGK–Burnett equations are obtained as [Yun et al., 1998b]:

$$\begin{aligned}
\sigma_{11}^{(B)} = {} & \frac{\mu^3}{p^2} RT(\theta_1 u_{xxx} + \theta_2 u_{xyy} + \theta_3 v_{xxy} + \theta_4 v_{yyy}) \\
& - \frac{\mu^3}{p^2}\frac{RT}{\rho}(\theta_1\rho_x u_{xx} + \theta_5\rho_x v_{xy} + \theta_6\rho_x u_{yy} + \theta_7\rho_y v_{xx} + \theta_8\rho_y u_{xy} + \theta_4\rho_y v_{yy}) \\
& + \frac{\mu^3}{p^2}(\theta_9 u_x^3 + 3\theta_{10}u_x^2 v_y + \theta_{11}u_x v_y^2 - \theta_4 u_x u_y^2 - 2\theta_4 u_x u_y v_x \\
& \qquad - \theta_4 u_x v_x^2 + \theta_{10}v_y^3 - \theta_{12}v_y u_y^2 - 2\theta_{12}u_y v_x v_y - \theta_{12}v_x^2 v_y) \\
& + \frac{\mu^3}{p^2} R(\theta_{13}u_x T_{xx} + \theta_{13}u_x T_{yy} + \theta_{14}v_y T_{xx} + \theta_{14}v_y T_{yy})
\end{aligned} \tag{8.21}$$

$$
\begin{aligned}
\sigma_{22}^{(B)} = {} & \frac{\mu^3}{p^2} RT(\theta_1 v_{yyy} + \theta_2 v_{xxy} + \theta_3 u_{xyy} + \theta_4 u_{xxx}) \\
& - \frac{\mu^3}{p^2}\frac{RT}{\rho}(\theta_1 \rho_y v_{yy} + \theta_5 \rho_y u_{xy} + \theta_6 \rho_y v_{xx} + \theta_7 \rho_x u_{yy} + \theta_8 \rho_x v_{xy} + \theta_4 \rho_x u_{xx}) \\
& + \frac{\mu^3}{p^2}(\theta_9 v_y^3 + 3\theta_{10} v_y^2 u_x + \theta_{11} v_y u_x^2 - \theta_4 v_y v_x^2 - 2\theta_4 v_y v_x u_y \\
& \qquad - \theta_4 v_y u_y^2 + \theta_{10} u_x^3 - \theta_{12} u_x v_x^2 - 2\theta_{12} v_x u_y u_x - \theta_{12} u_y^2 u_x) \\
& + \frac{\mu^3}{p^2} R(\theta_{13} v_y T_{yy} + \theta_{13} v_y T_{xx} + \theta_{14} u_x T_{yy} + \theta_{14} u_x T_{xx})
\end{aligned}
\tag{8.22}
$$

$$
\begin{aligned}
\sigma_{12}^{(B)} = {} & \frac{\mu^3}{p^2} RT(\theta_{15} u_{xxy} + u_{yyy} + \theta_{15} v_{xyy} + v_{xxx}) \\
& - \frac{\mu^3}{p^2}\frac{RT}{\rho}(\theta_6 \rho_{\mathrm{y}} u_{xx} + \theta_{16} \rho_y v_{xy} + \rho_y u_{yy} + \rho_x v_{xx} + \theta_{16} \rho_x u_{xy} + \theta_6 \rho_x v_{yy}) \\
& - \frac{\mu^3}{p^2}(u_y + v_x)(\theta_4 u_x^2 + 2\theta_{12} u_x v_y + 2\theta_7 u_y v_x + \theta_7 u_y^2 + \theta_7 v_x^2 + \theta_4 v_y^2) \\
& + \frac{\mu^3}{p^2} R(\theta_{17} u_y T_{xx} + \theta_{17} u_y T_{yy} + \theta_{17} v_x T_{xx} + \theta_{17} v_x T_{yy})
\end{aligned}
\tag{8.23}
$$

$$
\begin{aligned}
q_1^{(B)} = {} & \frac{\mu^3}{p\rho} R\left(\theta_{18} T_{xxx} + \theta_{18} T_{xyy} - \theta_{18}\frac{1}{\rho}\rho_x T_{xx} - \theta_{18}\frac{1}{\rho}\rho_x \mathrm{T}_{yy}\right) \\
& + \frac{\mu^3}{p\rho}(\theta_{19} u_x u_{xx} + \theta_{20} u_x v_{xy} + \theta_6 u_x u_{yy} + \theta_{21} v_y u_{xx} + \theta_{22} v_y v_{xy} + \theta_7 v_y u_{yy} \\
& \qquad + \theta_{23} u_y v_{xx} + \theta_{24} u_y u_{xy} + \theta_6 u_y v_{yy} + \theta_{23} v_x v_{xx} + \theta_{24} v_x u_{xy} + \theta_6 v_x v_{yy}) \\
& - \frac{\mu^3}{p\rho}\left(\frac{1}{\rho}\rho_x + \frac{1}{T}T_x\right)(\theta_{13} u_x^2 + 2\theta_{14} u_x v_y + 2\theta_{17} u_y v_x + \theta_{17} u_y^2 + \theta_{17} v_x^2 + \theta_{13} v_y^2) \\
& + \frac{\mu^3}{p\rho}\frac{R}{T}(\theta_{18} T_x T_{xx} + \theta_{18} T_x T_{yy})
\end{aligned}
\tag{8.24}
$$

$$
\begin{aligned}
q_2^{(B)} = {} & \frac{\mu^3}{p\rho} R\left(\theta_{18} T_{yyy} + \theta_{18} T_{xxy} - \theta_{18}\frac{1}{\rho}\rho_y T_{yy} - \theta_{18}\frac{1}{\rho}\rho_y T_{xx}\right) \\
& + \frac{\mu^3}{p\rho}(\theta_{19} v_y v_{yy} + \theta_{20} v_y u_{xy} + \theta_6 v_y v_{xx} + \theta_{21} u_x v_{yy} + \theta_{22} u_x u_{xy} + \theta_7 u_x v_{xx} \\
& \qquad + \theta_{23} v_x u_{yy} + \theta_{24} v_x v_{xy} + \theta_6 v_x u_{xx} + \theta_{23} u_y u_{yy} + \theta_{24} u_y v_{xy} + \theta_6 u_y u_{xx}) \\
& - \frac{\mu^3}{p\rho}\left(\frac{1}{\rho}\rho_y + \frac{1}{T}T_y\right)(\theta_{13} u_x^2 + 2\theta_{14} u_x v_y + 2\theta_{17} u_y v_x + \theta_{17} u_y^2 + \theta_{17} v_x^2 + \theta_{13} v_y^2) \\
& + \frac{\mu^3}{p\rho}\frac{R}{T}(\theta_{18} T_y T_{xx} + \theta_{18} T_y T_{yy})
\end{aligned}
\tag{8.25}
$$

The superscript (B) denotes third-order stress and heat-flux terms in the BGK–Burnett equations. The θ_i are given as follows for $\gamma = 1.4$: $\theta_1 = 2.56$, $\theta_2 = 1.36$, $\theta_3 = 0.56$, $\theta_4 = -0.64$, $\theta_5 = 0.96$, $\theta_6 = 1.6$, $\theta_7 = -0.4$, $\theta_8 = -0.24$, $\theta_9 = 1.024$, $\theta_{10} = -0.256$, $\theta_{11} = 1.152$, $\theta_{12} = 0.16$, $\theta_{13} = 2.24$, $\theta_{14} = -0.56$, $\theta_{15} = 3.6$, $\theta_{16} = 0.6$, $\theta_{17} = 1.4$, $\theta_{18} = 4.9$, $\theta_{19} = 7.04$, $\theta_{20} = -0.16$, $\theta_{21} = -1.76$, $\theta_{22} = 4.24$, $\theta_{23} = 3.8$, and $\theta_{24} = 3.4$.

Finally, governing Equation (8.1) is nondimensionalized by a reference length and freestream variables and is written in a curvilinear coordinate system (ξ, η) by employing a coordinate transformation:

$$\tau = t, \quad \xi = \xi(x, y), \quad \eta = \eta(x, y) \tag{8.26}$$

8.4 Wall-Boundary Conditions

The no-slip-/no-temperature-jump boundary conditions are employed at the wall when solving the continuum Navier–Stokes equations for $Kn < 0.001$. In the continuum–transition regimes, the non-slip-boundary conditions are no longer correct. First-order slip/temperature-jump boundary conditions should be applied to both the Navier–Stokes equations and Burnett equations in the range $0.001 < Kn < 0.1$. The transition regime spans the range $0.01 < Kn < 10$; the second-order slip/temperature-jump conditions should be used in this regime with the Navier–Stokes as well as the Burnett equations. The Navier–Stokes equations are first-order accurate in Kn, while the Burnett equations are second-order accurate in Kn. Both first- and second-order Maxwell–Smoluchowski slip/temperature-jump boundary conditions are generally employed on the body surface when solving the Burnett equations.

The first-order Maxwell–Smoluchowski slip-boundary conditions in Cartesian coordinates are [Smoluchowski, 1898]:

$$U_s = \frac{2-\bar{\sigma}}{\bar{\sigma}} \frac{2\mu}{\rho} \sqrt{\frac{\pi}{8RT}} \left(\frac{\partial U}{\partial y}\right)_s + \frac{3}{4} \frac{\mu}{\rho T} \left(\frac{\partial T}{\partial x}\right)_s \tag{8.27}$$

and

$$T_s - T_w = \frac{2-\bar{\alpha}}{\bar{\alpha}} \frac{2\gamma}{\gamma+1} \frac{2\mu}{\rho} \sqrt{\frac{\pi}{8RT}} \frac{1}{Pr} \left(\frac{\partial T}{\partial y}\right)_s \tag{8.28}$$

The subscript s denotes the flow variables on the solid surface of the body. First-order Maxwell–Smoluchowski slip-boundary conditions can be derived by considering the momentum and energy-flux balance on the wall surface. The reflection coefficient $\bar{\sigma}$ and the accommodation coefficient $\bar{\alpha}$ are assumed to be equal to unity (for complete accommodation) in the calculations presented in this chapter.

Beskok's slip-boundary condition [Beskok et al., 1996] is the second-order extension of the Maxwell's slip-velocity-boundary condition excluding the thermal creep terms, given as:

$$U_s = \frac{2-\bar{\sigma}}{\bar{\sigma}} \left[\frac{Kn}{1-bKn} \left(\frac{\partial U}{\partial y}\right)_s\right] \tag{8.29}$$

where b is the slip coefficient determined analytically in the slip flow regime and empirically in transitional and free molecular regimes.

Langmuir's slip-boundary condition has also been employed in the literature [Myong, 1999]. Langmuir's slip-boundary condition is based on the theory of adsorption phenomena at the solid wall. Gas molecules do not in general rebound elastically but condense on the surface, being held by the field of force of the surface atoms. These molecules may subsequently evaporate from the surface resulting in some time lag. Slip is the direct result of this time lag. The slip velocity at the wall is given as:

$$U_s = \frac{1}{1+\beta p} \tag{8.30}$$

where β is the adsorption coefficient determined empirically or by theoretical prediction.

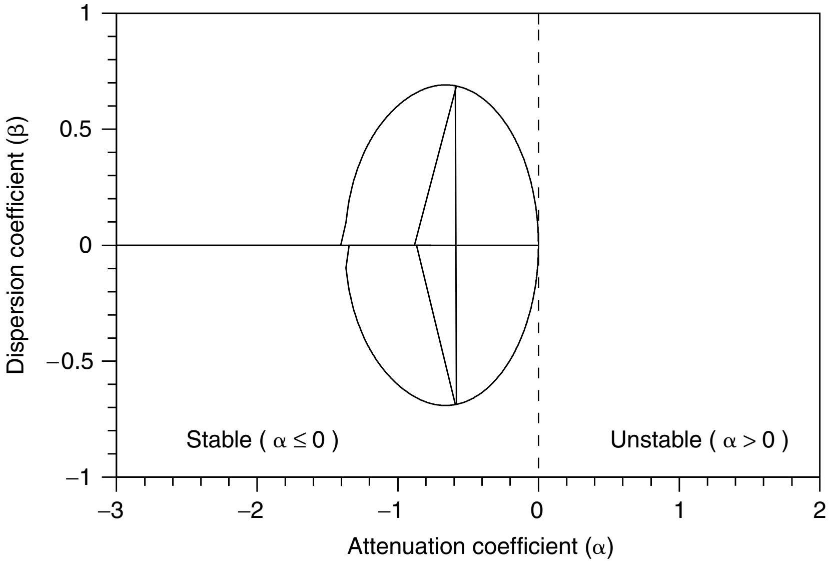


FIGURE 8.2 Characteristic trajectories of the one-dimensional Navier–Stokes equations.

In this chapter, these slip boundary conditions are applied and compared to determine their influence on the solution.

8.5 Linearized Stability Analysis of Burnett Equations

Bobylev (1982) showed that the conventional Burnett equations are not stable to small wavelength disturbances; hence, the solutions to conventional Burnett equations tend to diverge when the mesh size is made progressively finer. Balakrishnan and Agarwal (1999) performed the linearized stability of one-dimensional original Burnett equations, conventional Burnett equations, augmented Burnett equations, and the BGK–Burnett equations. They considered the response of a uniform gas subjected to small one-dimensional periodic perturbations ρ', u', and T' for density, velocity, and temperature respectively. Burnett equations were linearized by neglecting products and powers of small perturbations, and a linearized set of equations for small perturbation variables $V' = [\rho', u', T']^T$ was obtained. They assumed that the solution is of the form:

$$V' = \bar{V}e^{i\omega x}e^{\phi t} \tag{8.31}$$

where $\phi = \alpha + i\beta$, and α and β denote the attenuation and dispersion coefficients respectively. For stability, $\alpha \leqslant 0$ as the Knudsen number increases. Substitution of Equation (8.31) in the equations for small perturbation quantities V' results in a characteristic equation, $|F(\phi, \omega)| = 0$. The trajectory of the roots of this characteristic equation is plotted in a complex plane on which the real axis denotes the attenuation coefficient and the imaginary axis denotes the dispersion coefficient. For stability, the roots must lie to the left of the imaginary axis as the Knudsen number increases. Figures 8.2 to 8.5 show the trajectory of the three roots of the characteristic equations as the Knudsen number increases. The plots show that the Navier–Stokes equations, the augmented Burnett equations, and the BGK–Burnett equations (with $\gamma = 1.667$) are stable, but the conventional Burnett equations are unstable. Euler equations are employed to approximate the material derivatives in all three types of Burnett equations. The BGK–Burnett equations, however, become unstable for $\gamma = 1.4$. On the other hand, if the material derivatives are approximated using the Navier–Stokes equations, then the conventional, augmented, and BGK–Burnett equations are all stable to small wavelength disturbances.

Based on these observations, we have employed the Navier–Stokes equations to approximate the material derivatives in the conventional, augmented, and BGK–Burnett equations presented in Section 8.3. For the detailed analysis behind Figures 8.2 to 8.5, see Balakrishnan and Agarwal (1999). The linearized stability

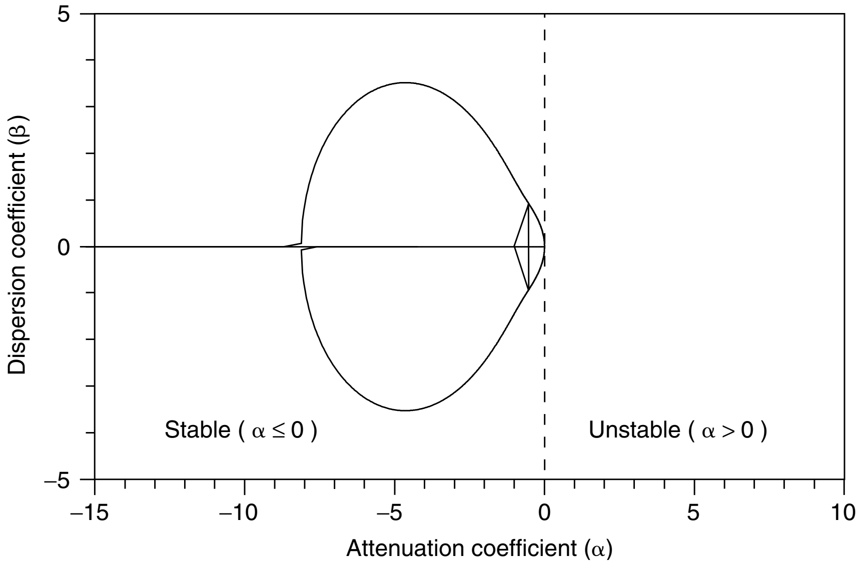


FIGURE 8.3 Characteristic trajectories of the one-dimensional augmented Burnett equations ($\gamma = 1.667$); Euler equations are used to express the material derivatives $D(\ )/Dt$ in terms of spatial derivatives.

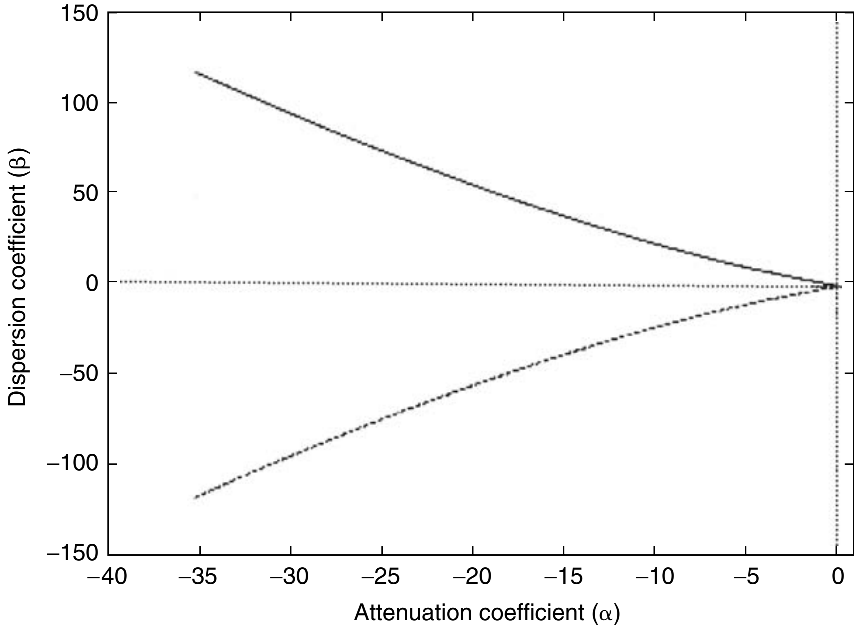


FIGURE 8.4 Characteristic trajectories of the one-dimensional BGK–Burnett equations ($\gamma = 1.667$); Euler equations are used to express the material derivatives $D(\ )/Dt$ in terms of spatial derivatives.

analysis of conventional, augmented, and super-Burnett equations has also been performed in three dimensions with similar conclusions [Yun and Agarwal, 2000].

8.6 Numerical Method

An explicit finite-difference scheme is employed to solve the governing equations of Section 8.3. The Steger–Warming flux-vector splitting method [Steger and Warming, 1981] is applied to the inviscid-flux terms. The second-order, central-differencing scheme is applied to discretize the stress tensor and heat-flux terms. Converged solutions were obtained with a reduction in residuals of six orders of magnitude.

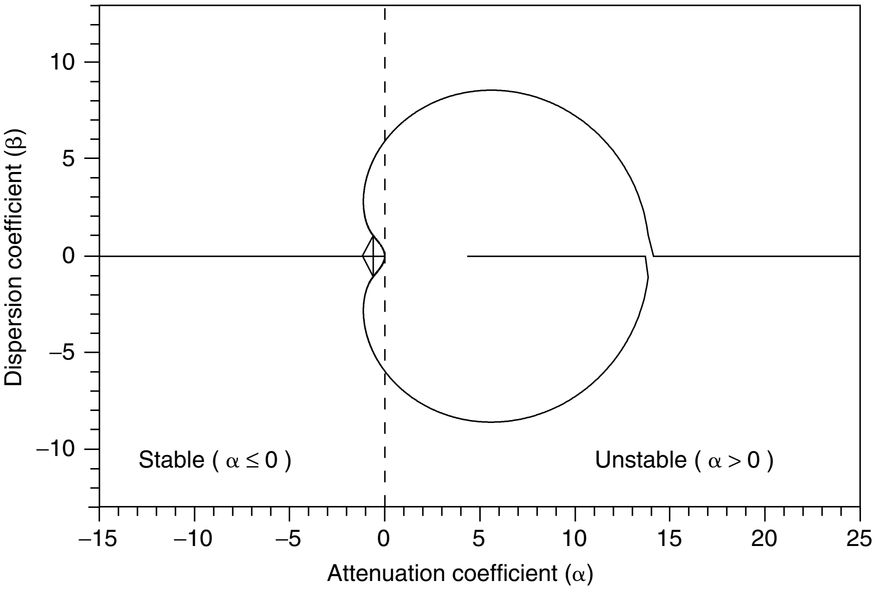


FIGURE 8.5 Characteristic trajectories of the one-dimensional conventional Burnett equations; Euler equations are used to express the material derivatives $D(\ )/Dt$ in terms of spatial derivatives.

All the calculations were performed on a sequence of successively refined grids to assure grid independence of the solutions.

8.7 Numerical Simulations

Numerical simulations have been performed for both the hypersonic flows and microscale flows in the continuum–transition regime. Hypersonic flow calculations include one-dimensional shock structure, two-dimensional and axisymmetric blunt bodies, and a space shuttle re-entry condition. Microscale flows include the subsonic flow and supersonic flow in a microchannel.

8.7.1 Application to Hypersonic Shock Structure

The hypersonic shock for argon was computed using the BGK–Burnett equations. The upstream flow conditions were specified and the downstream conditions were determined from the Rankine–Hugoniot relations. For purposes of comparison, the same flow conditions as in Fiscko and Chapman (1988) were used in the computations. The parameters used were

$$T_\infty = 300\,\text{K},\quad P_\infty = 1.01323 \times 10^5\,\text{N/m}^2,\quad \gamma_{\text{argon}} = 1.667,\quad \mu_{\text{argon}} = 22.7 \times 10^{-6}\,\text{kg/sec} \cdot \text{m}$$

The Navier–Stokes solution was taken as the initial value. This initial Navier–Stokes spatial distribution of variables was imposed on a mesh that encloses the shock. The length of the control volume enclosing the shock was chosen to be 1000 × λ_∞ where the mean free path based on the freestream parameters is given by the expression $\lambda_\infty = 16\mu/(5\rho_\infty\sqrt{2\pi RT_\infty})$. This is the mean free path that would exist in the unshocked region if the gas were composed of hard elastic spheres and had the same viscosity, density, and temperature as the gas being considered. The solution was marched in time until the observed deviations were smaller than a preset convergence criterion.

A set of computational experiments was carried out to compare the BGK–Burnett solutions with the Burnett solutions of Fiscko and Chapman (1988). Tests were conducted at Mach 20 and Mach 35. In order to test for instabilities to small wavelength disturbances, the grid points were increased from 101 to 501 points. Figures 8.6 and 8.7 show variations of specific entropy across the shock wave. The BGK–Burnett equations

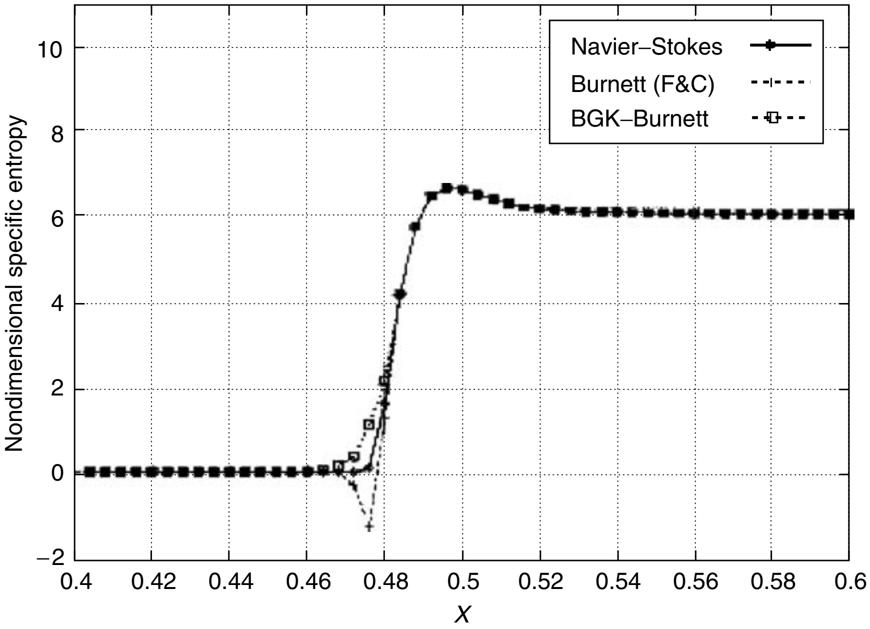


FIGURE 8.6 Specific entropy variation across a Mach 20 normal shock in a monatomic gas (argon), $\Delta x/\lambda_\infty = 4.0$ and $\gamma = 1.667$; F & C ≡ Fiscko and Chapman (1988).

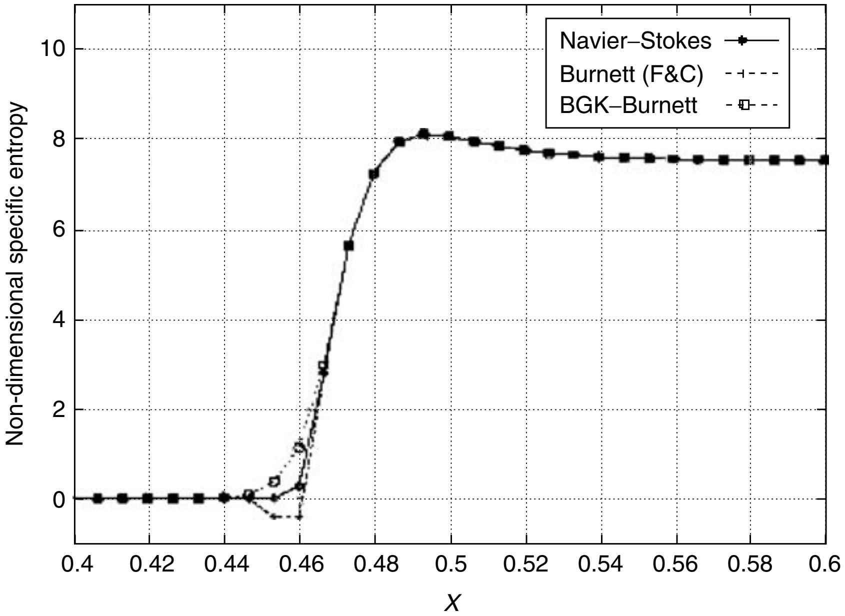


FIGURE 8.7 Specific entropy variation across a Mach 35 normal shock in a monatomic gas (argon), $\Delta x/\lambda_\infty = 4.0$ and $\gamma = 1.667$; F & C ≡ Fiscko and Chapman (1988).

show a positive entropy change throughout the flow field, while the conventional Burnett equations give rise to a negative entropy spike just ahead of the shock as the number of grid points is increased. This spike increases in magnitude until the conventional Burnett equations break down completely. The BGK–Burnett equations did not exhibit any instabilities for the range of grid points considered. Figure 8.8 shows the variation of reciprocal density thickness with Mach number. BGK–Burnett calculations compare well to those of Woods and simplified Woods equations [Reese et al., 1995] and the experimental data of Alsmeyer (1976). Extensive calculations for one-dimensional hypersonic shock structure using various higher order kinetic formulations are given in Balakrishnan (1999).

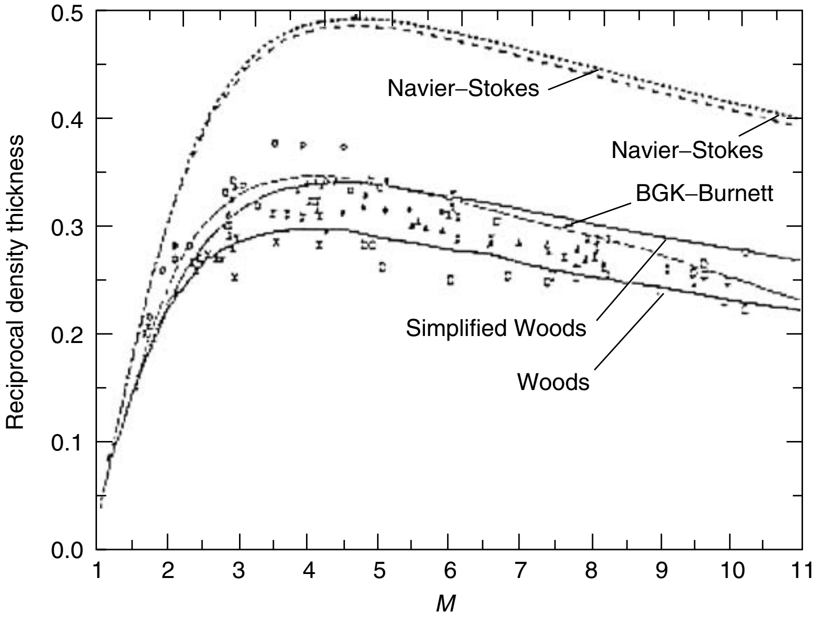


FIGURE 8.8 Plot showing the variation of reciprocal density thickness with Mach number, obtained with the Navier–Stokes, Woods and Simplified Woods [Reese et al., 1995], and BGK–Burnett equations for a monatomic gas (argon). Experimental data were obtained from Alsmeyer (1976). (Reprinted with permission from Alsmeyer, H. (1976) "Density Profiles in Argon and Nitrogen Shock Waves Measured by the Absorption of an Electron Beam," *J. Fluid Mech.* **74**, pp. 497–513.)

8.7.2 Application to Two-Dimensional Hypersonic Blunt Body Flow

The two-dimensional augmented Burnett code was employed to compute the hypersonic flow over a cylindrical leading edge with a nose radius of 0.02 m in the continuum–transition regime. The grid system (50 × 82 mesh) used in the computations is shown in Figure 8.9. The results were compared with those of Zhong (1991).

The flow conditions for this case are as follows:

$$M_\infty = 10, \quad Kn_\infty = 0.1, \quad Re_\infty = 167.9,$$

$$P_\infty = 2.3881\,\text{N/m}^2, \quad T_\infty = 208.4\,\text{K}, \quad T_w = 1000.0\,\text{K}$$

The viscosity is calculated by Sutherland's law, $\mu = c_1 T^{1.5}/(T + c_2)$. The coefficients c_1 and c_2 for air are 1.458×10^{-6} kg/(sec m K$^{1/2}$) and 110.4 K, respectively. Other constants used in this computation for air are $\gamma = 1.4$, $Pr = 0.72$ and $R = 287.04\,\text{m}^2/(\text{sec}^2\,\text{K})$.

The comparisons of density, velocity, and temperature distributions along the stagnation streamline are shown in Figures 8.10, 8.11, and 8.12 respectively. The results agree well with those of Zhong (1991) for both the Navier–Stokes and the augmented Burnett computations. Because the flow is in the continuum–transition regime ($Kn = 0.1$), the Navier–Stokes equations become inaccurate, and the differences between the Navier–Stokes and the augmented Burnett solutions are obvious. In particular, the difference between the Navier–Stokes and Burnett solutions for the temperature distribution is significant across the shock. Temperature and Mach number contours of the Navier–Stokes solutions and the augmented Burnett solutions are compared in Figures 8.13 and 8.14 respectively. The shock structure of the augmented Burnett solutions agrees well with that of Zhong (1991). The shock layer of the augmented Burnett solutions is thicker, and the shock location starts upstream of that of the Navier–Stokes solutions. However, because the local Knudsen number decreases and the flow tends toward equilibrium as it approaches the wall surface, the differences between the Navier–Stokes and augmented Burnett solutions become negligible near the

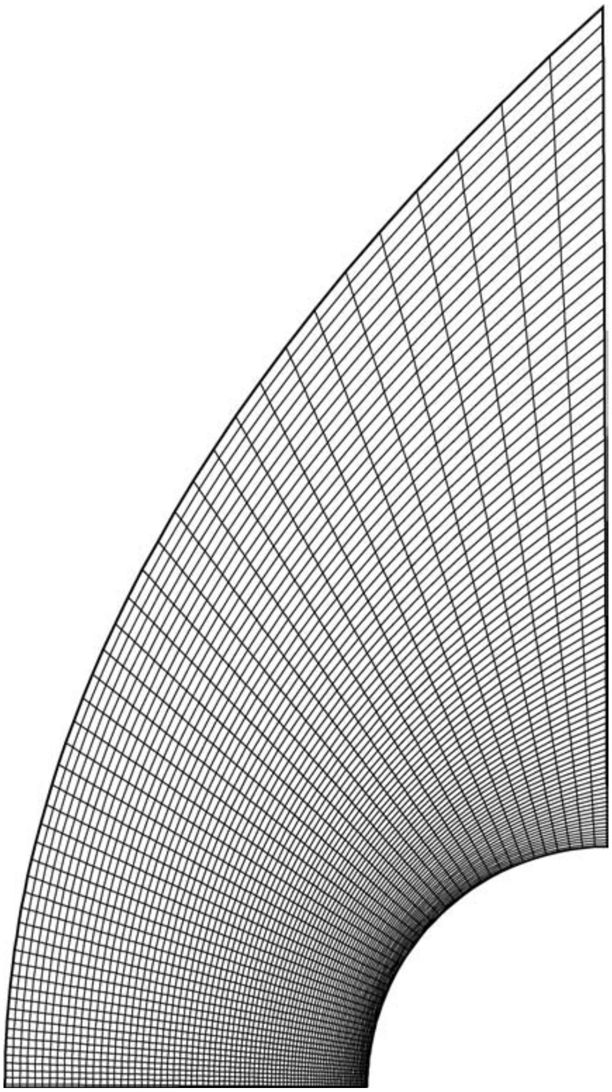


FIGURE 8.9 Two-dimensional computational grid (50 × 82 mesh) around a blunt body, $r_n = 0.02\,\text{m}$.

wall, especially near the stagnation point. Thus, the Maxwell–Smoluchowski slip boundary conditions can be applied for both the Navier–Stokes and the augmented Burnett calculations for the hypersonic blunt body.

8.7.3 Application to Axisymmetric Hypersonic Blunt Body Flow

The results of the axisymmetric augmented Burnett computations are compared with the DSMC results obtained by Vogenitz and Takara (1971) for the axisymmetric hemispherical nose. The computed results are also compared with Zhong and Furumoto's (1995) axisymmetric augmented Burnett solutions. The flow conditions for this case are

$$M_\infty = 10, \quad Kn_\infty = 0.1$$

$$\frac{T_w}{T_\infty} = 1.0, \quad \frac{T_w}{T_0} = 0.029$$

T_0 is the stagnation temperature. The gas is assumed to be a monatomic gas with a hard-sphere model. The viscosity coefficient is calculated by the power law $\mu = \mu_r\,(T/T_r)^{0.5}$. The reference viscosity μ_r and the reference temperature T_r used in this case are $2.2695 \times 10^{-5}\,\text{kg/(sec m)}$ and 300 K, respectively. Other constants used in this computation are $\gamma = 1.67$ and $Pr = 0.67$.

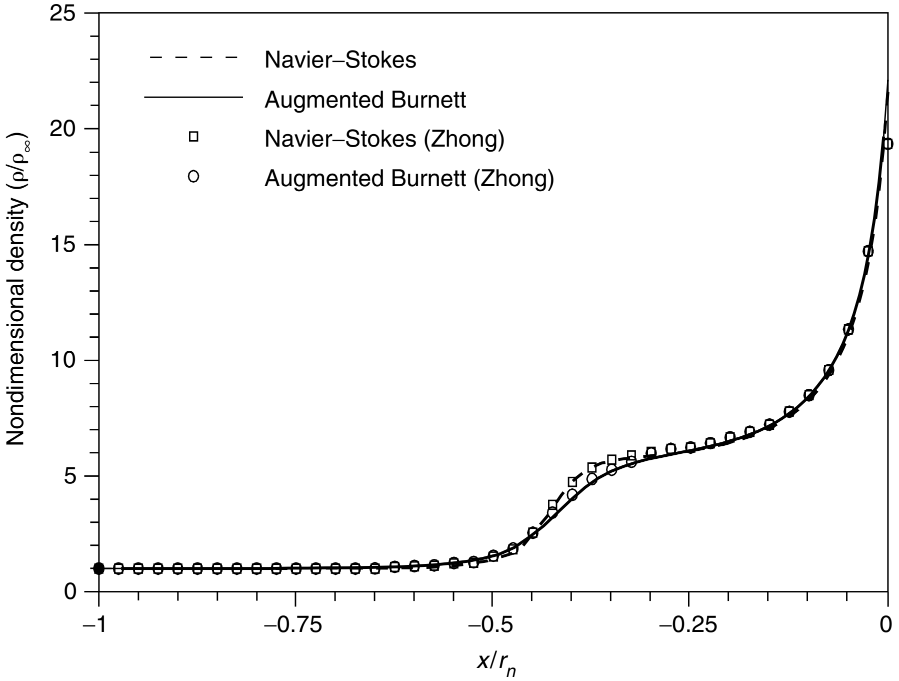


FIGURE 8.10 Density distributions along stagnation streamline for blunt body flow: air, $M_\infty = 10$, and $Kn_\infty = 0.1$.

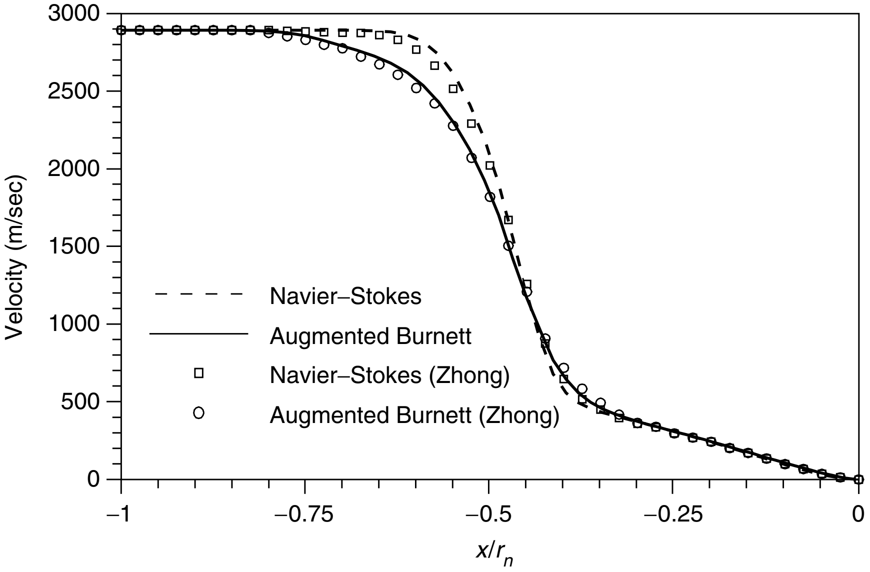


FIGURE 8.11 Velocity distributions along stagnation streamline for blunt body flow: air, $M_\infty = 10$, and $Kn_\infty = 0.1$.

The comparisons of density and temperature distributions along the stagnation streamline among the current axisymmetric augmented Burnett solutions, the axisymmetric augmented Burnett solutions of Zhong and Furumoto, and the DSMC results are shown in Figures 8.15 and 8.16, respectively. The corresponding Navier–Stokes solutions are also compared in these figures. The axisymmetric augmented Burnett solutions agree well with Zhong and Furumoto's axisymmetric augmented Burnett solutions in both density and temperature. The density distributions for both the Navier–Stokes and augmented

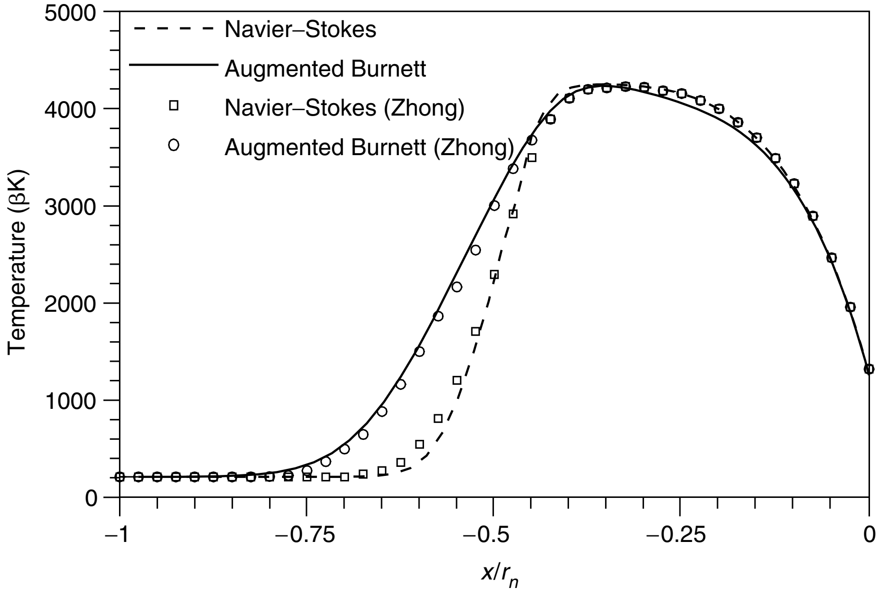


FIGURE 8.12 Temperature distributions along stagnation streamline for blunt body flow: air, $M_\infty = 10$, and $Kn_\infty = 0.1$.

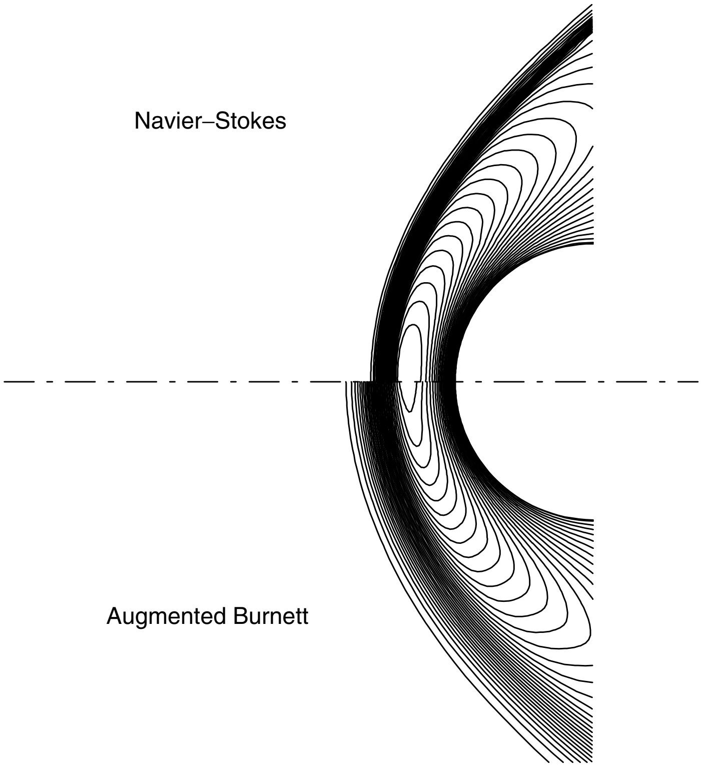


FIGURE 8.13 Comparison of temperature contours for blunt body flow: air, $M_\infty = 10$, and $Kn_\infty = 0.1$.

Burnett equations show little differences from the DSMC results. The temperature distributions, however, show that the DSMC method predicts a thicker shock than the augmented Burnett equations. The maximum temperature of the DSMC results is slightly higher than those of the augmented Burnett solutions. However, the augmented Burnett solutions show much closer agreement with the DSMC results than the Navier–Stokes solutions.

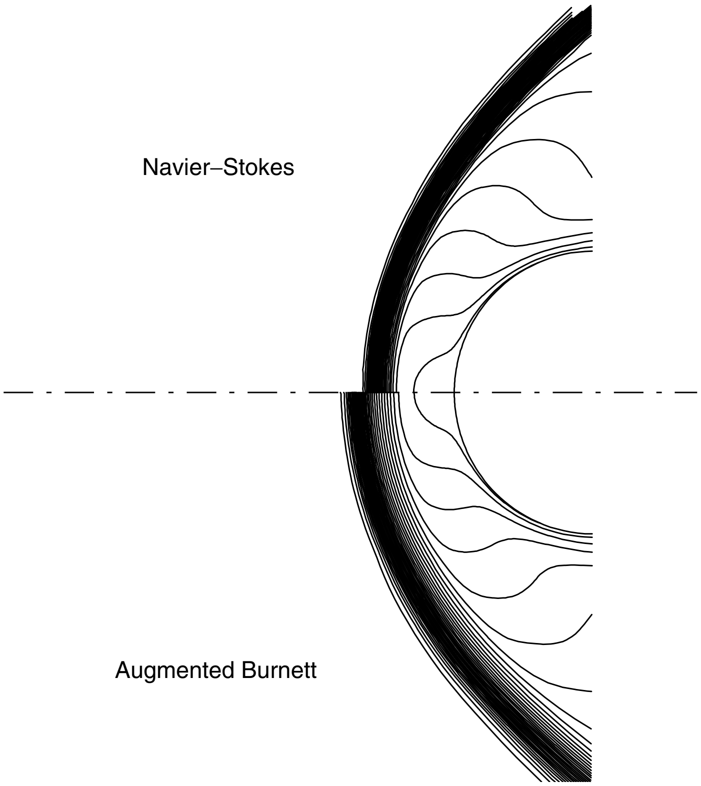


FIGURE 8.14 Comparison of Mach number contours for blunt body flow: air, $M_\infty = 10$, and $Kn_\infty = 0.1$.

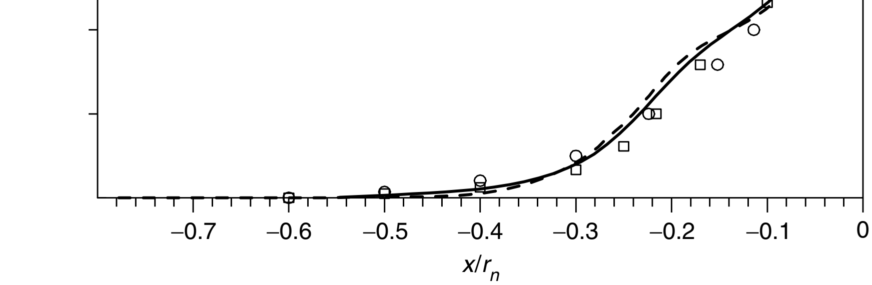


FIGURE 8.15 Density distributions along stagnation streamline for a hemispherical nose: hard-sphere gas, $M_\infty = 10$ and $Kn_\infty = 0.1$.

Overall, the axisymmetric augmented Burnett solutions presented here agree well with Zhong and Furumoto's (1995) axisymmetric augmented Burnett solutions and describe the shock structure closer to the DSMC results than the Navier–Stokes solutions.

As another application to the hypersonic blunt body, the augmented Burnett equations are applied to compute the hypersonic flow field at re-entry condition encountered by the nose region of the space shuttle.

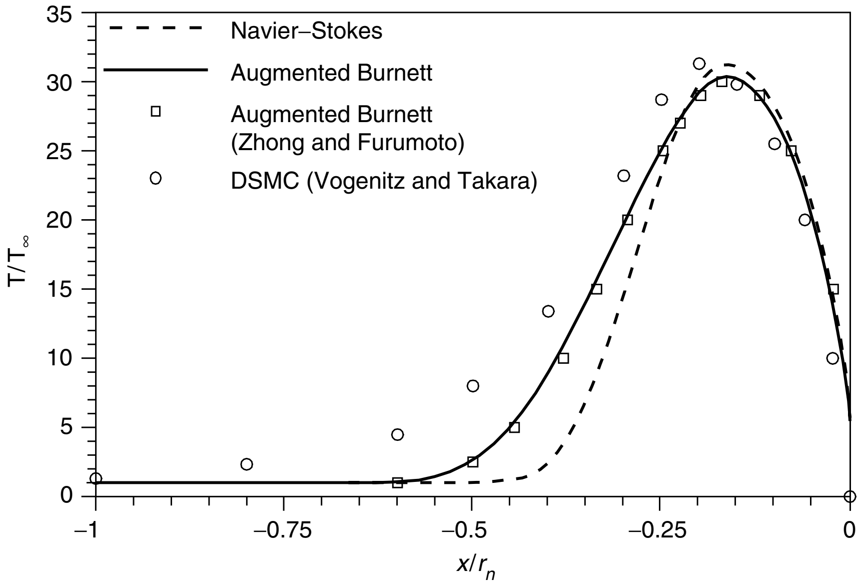


FIGURE 8.16 Temperature distribution along stagnation streamline for a hemispherical nose: hard-sphere gas, $M_\infty = 10$ and $Kn_\infty = 0.1$.

The computations are compared with the DSMC results of Moss and Bird (1984). The DSMC method accounts for the translational, rotational, vibrational, and chemical nonequilibrium effects.

An *equivalent axisymmetric body* concept [Moss and Bird, 1984] is applied to model the windward centerline of the space shuttle at a given angle of attack. A hyperboloid with nose radius of 1.362 m and asymptotic half angle of 42.5° is employed as the equivalent axisymmetric body to simulate the nose of the shuttle. Figure 8.17 shows the side view of the grid (61 × 100 mesh) around the hyperboloid. The freestream conditions at an altitude of 104.93 km as given in Moss and Bird (1984) are

$$M_\infty = 25.3, \quad Kn_\infty = 0.227, \quad Re_\infty = 163.8,$$

$$\rho_\infty = 2.475 \times 10^{-7}\,\text{kg/m}^3, \quad T_\infty = 223\,\text{K}, \quad T_w = 560\,\text{K}$$

The viscosity is calculated by the power law. The reference viscosity μ_r and the reference temperature T_r are taken as 1.47×10^{-5} kg/(sec m) and 223 K, respectively.

Figures 8.18 and 8.19 show comparisons of the density and temperature distributions along the stagnation streamline between the Navier–Stokes solutions, the augmented Burnett solutions, and the DSMC results. The differences between the augmented Burnett solutions and the DSMC results are significant in both density and temperature distributions. In Figure 8.18, the density distribution of the DSMC results is lower and smoother than that of the augmented Burnett solutions. In Figure 8.19, the DSMC method predicts about 30% thicker shock layer and 9% lower maximum temperature than the augmented Burnett equations. The DSMC results can be considered to be more accurate than the augmented Burnett solutions as the DSMC method accounts for all the effects of translational, rotational, vibrational, and chemical nonequilibrium, while the augmented Burnett equations do not. However, the augmented Burnett solutions agree much better with the DSMC results than the Navier–Stokes computations. The difference between the Navier–Stokes solutions and the augmented Burnett solutions in temperature distributions is very significant. The shock layer of the augmented Burnett solutions is almost two times thicker than the Navier–Stokes solutions. The augmented Burnett solutions predict about 11% less maximum temperature than the Navier–Stokes solutions.

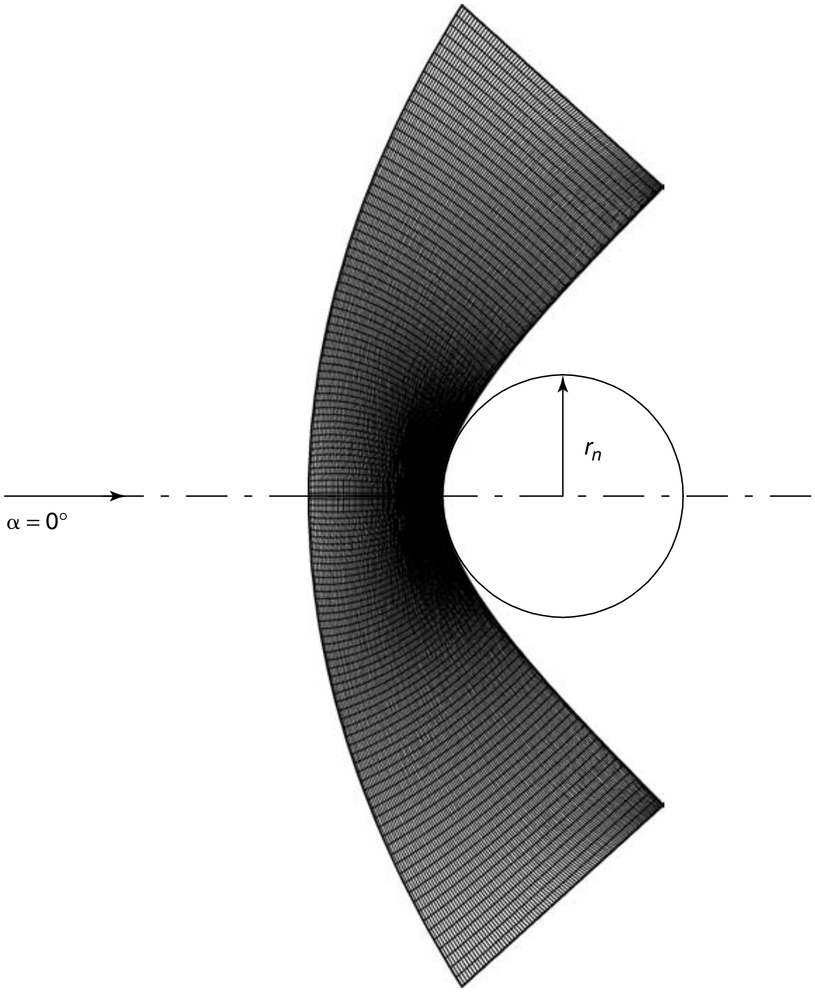


FIGURE 8.17 Side view of the grid (61 × 100 mesh) around a hyperboloid nose of radius $r_n = 1.362$ m.

8.7.4 Application to NACA 0012 Airfoil

The Navier–Stokes equations are applied to compute the rarefied subsonic flow over a NACA 0012 airfoil with chord length of 0.04 m. The grid system in the physical domain is shown in Figure 8.20. The flow conditions are

$$M_\infty = 0.8, \quad Re_\infty = 73, \quad \rho_\infty = 1.116 \times 10^{-4}\,\text{kg/m}^3, \quad T_\infty = 257\,\text{K}, \quad \text{and} \quad Kn_\infty = 0.014$$

Various constants used in the calculation for air are $\gamma = 1.4$, $Pr = 0.72$, and $R = 287.04\,\text{m}^2/(\text{sec}^2\,\text{K})$.

Figure 8.21 shows the density contours of the Navier–Stokes solution with the first-order Maxwell–Smoluchowski slip-boundary conditions. These contours using the continuum approach agree well with those of Sun et al. (2000) using the information preservation (IP) particle method. At these Mach and Knudsen numbers, the contours from the DSMC calculations are not smooth due to the statistical scatter. The comparison of pressure distribution along the surface between our Navier–Stokes solution with a slip-boundary condition and the DSMC calculation [Sun et al., 2000] is shown in Figure 8.22; the agreement between the solutions is good. Figure 8.23 compares the surface slip velocity from the DSMC, IP, and Navier–Stokes methods as calculated by Sun et al. and by our Navier–Stokes calculation. The slip velocity

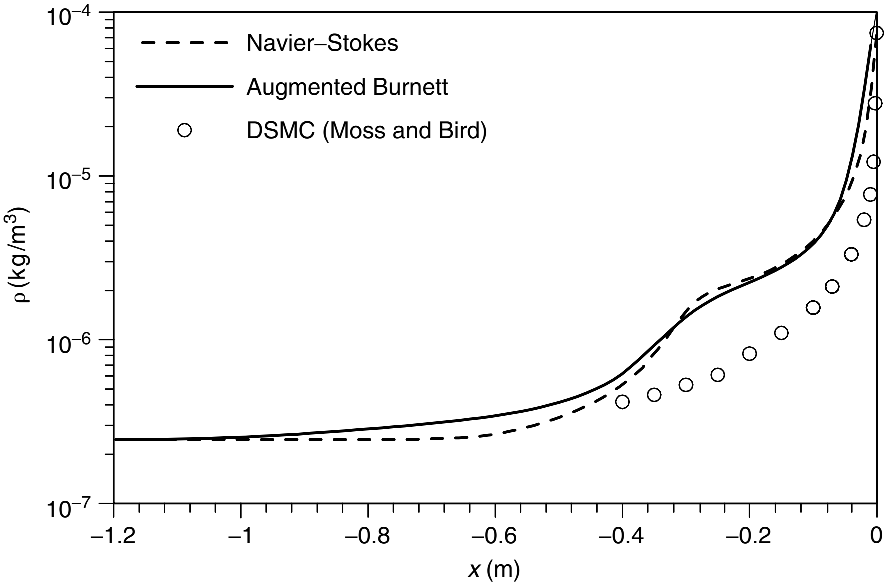


FIGURE 8.18 Density distributions along stagnation streamline for a hyperboloid nose: air, $M_\infty = 25.3$ and $Kn_\infty = 0.227$.

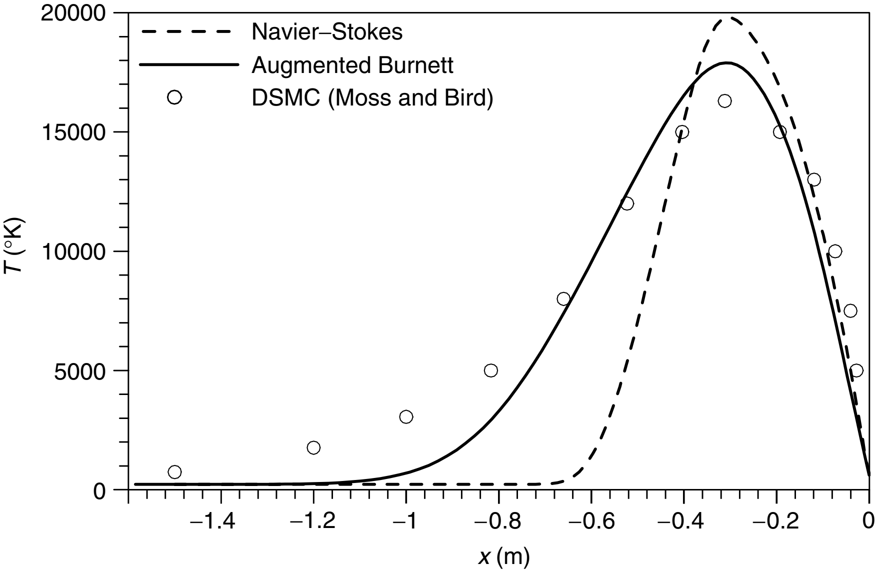


FIGURE 8.19 Temperature distributions along stagnation streamline for a hyperboloid nose: air, $M_\infty = 25.3$ and $Kn_\infty = 0.227$.

distribution from our Navier–Stokes calculation shows good agreement with that obtained from the DSMC and IP methods except near the trailing edge. However, our Navier–Stokes results disagree considerably with those reported in Sun et al. (2000). This calculation again demonstrates that Navier–Stokes equations with slip-boundary conditions can provide accurate flow simulation $0.001 < Kn < 0.1$.

8.7.5 Subsonic Flow in a Microchannel

The augmented Burnett equations are employed for computation of subsonic flow in a microchannel with a ratio of channel length to height of 20 ($L/H = 20$). For the wall boundary conditions, Beskok's and

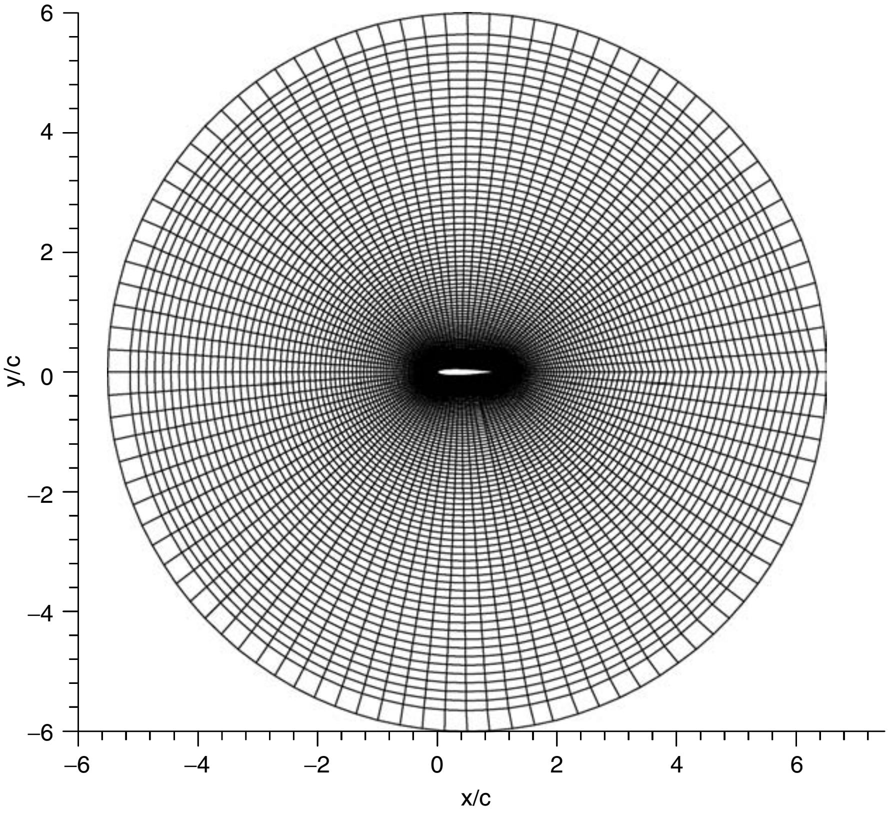


FIGURE 8.20 Grid (101 × 91 mesh) around a NACA 0012 airfoil, $c = 0.04\,\text{m}$.

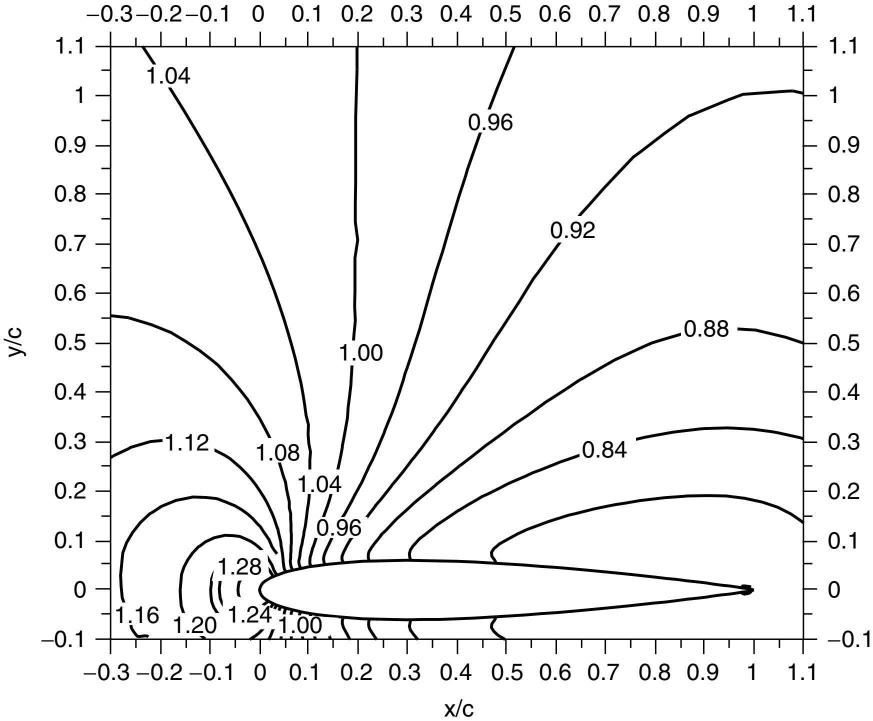


FIGURE 8.21 Density contours for NACA 0012 airfoil: air, $M_\infty = 0.8$ and $Kn_\infty = 0.014$; Navier–Stokes solution with first-order slip boundary condition.

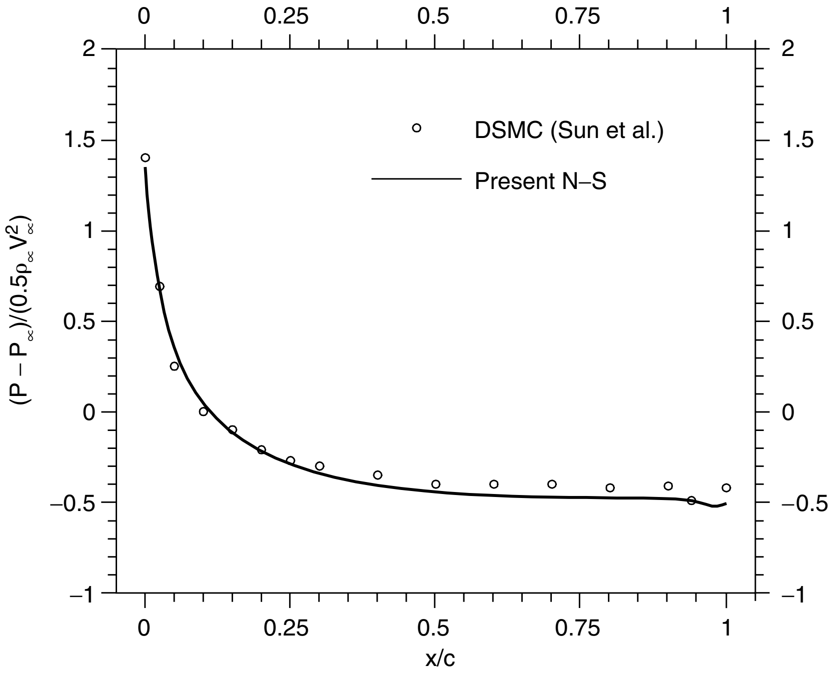


FIGURE 8.22 Pressure distributions along NACA 0012 airfoil surface: air, $M_\infty = 0.8$ and $Kn_\infty = 0.014$.

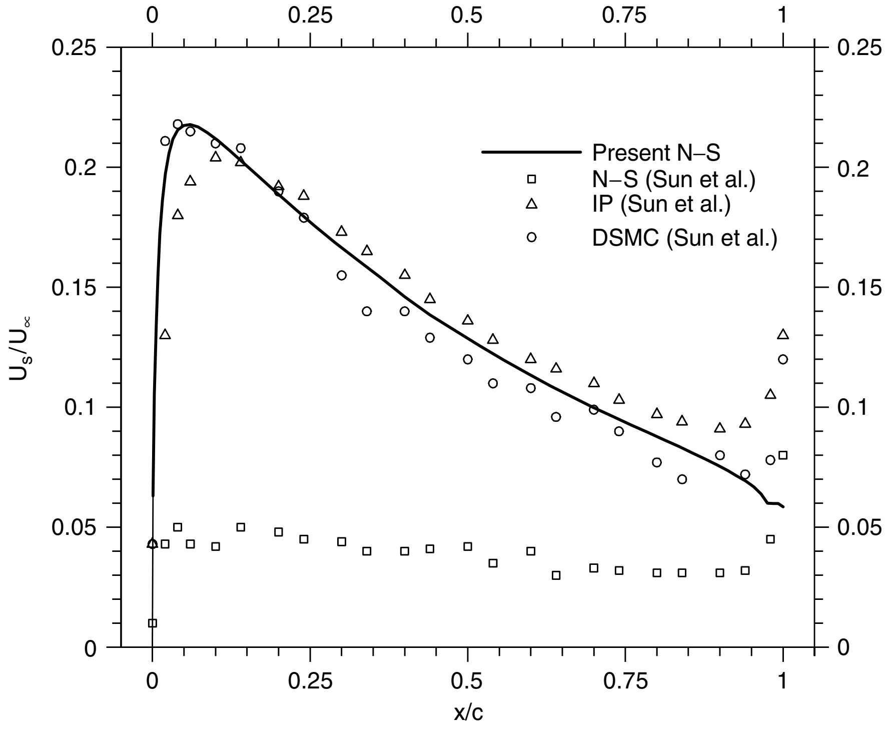


FIGURE 8.23 Slip velocity distributions along NACA 0012 airfoil surface: air, $M_\infty = 0.8$ and $Kn_\infty = 0.014$.

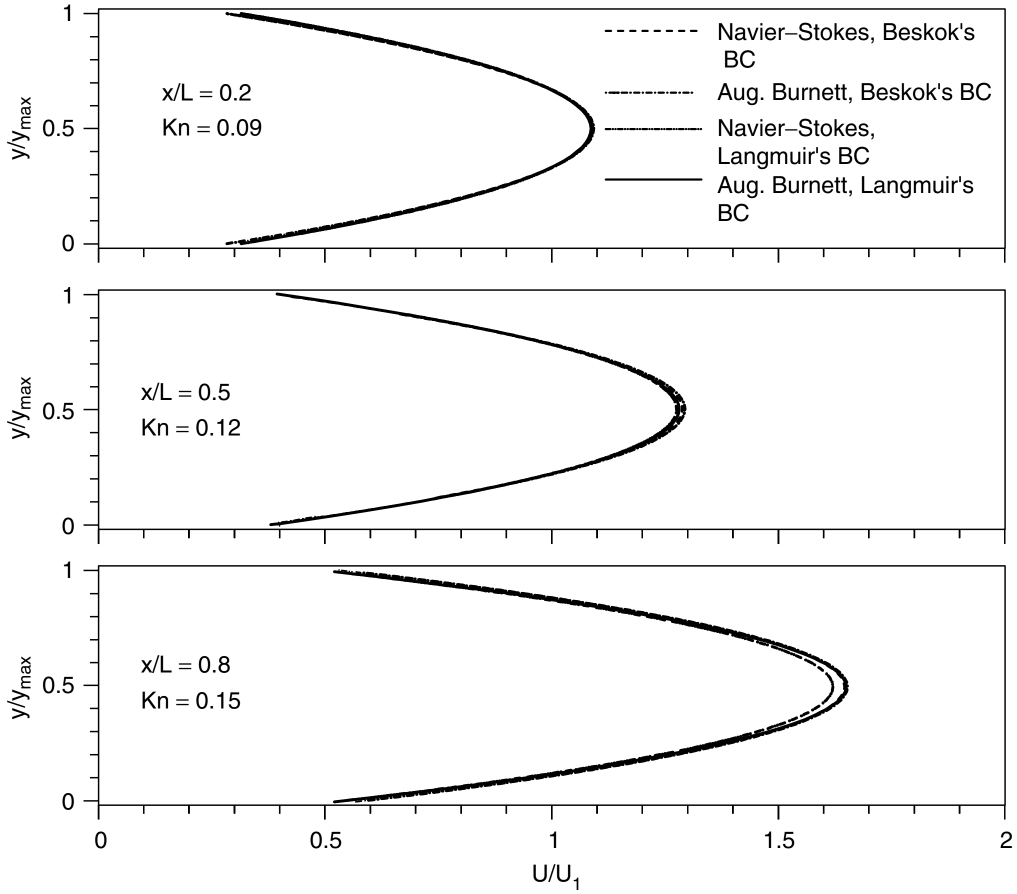


FIGURE 8.24 Comparison of velocity profiles at various streamwise locations: $Kn_{in} = 0.088$, $Kn_{out} = 0.2$ and $P_{in}/P_{out} = 2.28$.

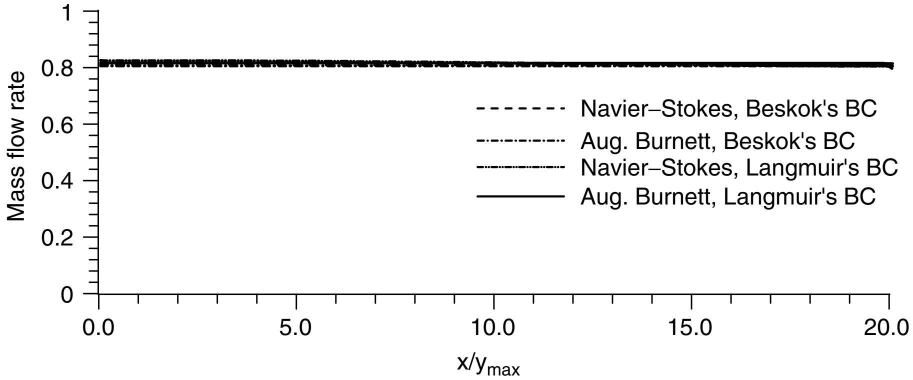


FIGURE 8.25 Comparison of mass flow rates along the microchannel: $Kn_{in} = 0.088$, $Kn_{out} = 0.2$ and $P_{in}/P_{out} = 2.28$.

Langmuir's boundary conditions are employed and compared. The augmented Burnett solutions are also compared with the Navier–Stokes solutions. Flow conditions at the entrance and exit of the channel are $Kn_{in} = 0.088$, $Kn_{out} = 0.2$ and $P_{in}/P_{out} = 2.28$.

Figure 8.24 compares the velocity profiles at various streamwise locations. Both Navier–Stokes and augmented Burnett equations using either Beskok's or Langmuir's boundary conditions show almost identical velocity profiles. These velocity profiles agree well with the velocity profiles from the micro-flow calculation by Beskok and Karniadakis (1999). Nondimensional mass flow rates along the microchannel are shown in Figure 8.25. All the mass flow rates from both equations and both slip-boundary conditions are about 0.76 and almost constant along the channel, as should be the case. This mass flow rate is 13% higher than that

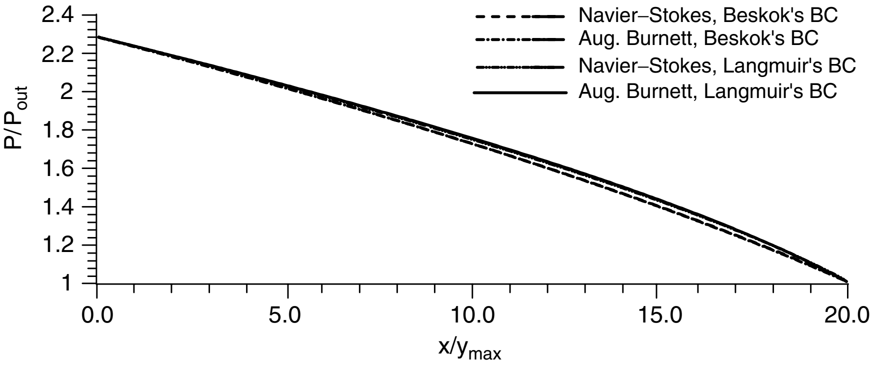


FIGURE 8.26 Comparison of pressure distribution along the centerline: $Kn_{\text{in}} = 0.088$, $Kn_{\text{out}} = 0.2$ and $P_{\text{in}}/P_{\text{out}} = 2.28$.

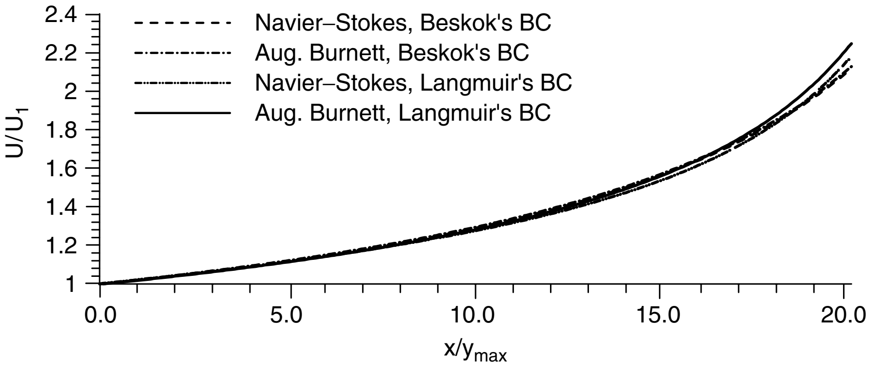


FIGURE 8.27 Comparison of streamwise velocity distributions along the centerline: $Kn_{\text{in}} = 0.088$, $Kn_{\text{out}} = 0.2$ and $P_{\text{in}}/P_{\text{out}} = 2.28$.

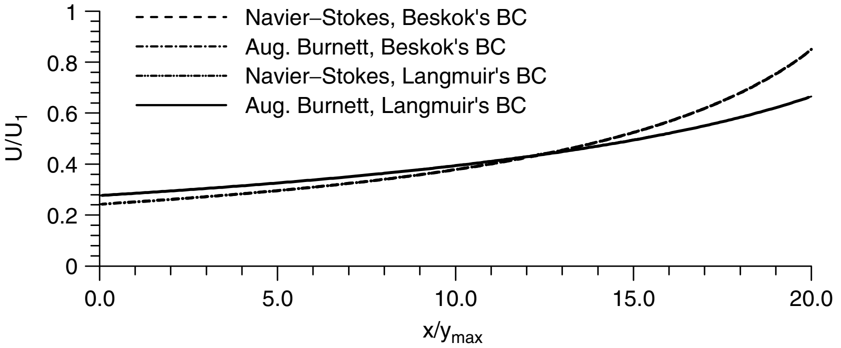


FIGURE 8.28 Comparison of slip velocity distributions along the wall: $Kn_{\text{in}} = 0.088$, $Kn_{\text{out}} = 0.2$ and $P_{\text{in}}/P_{\text{out}} = 2.28$.

predicted with a no-slip-boundary condition, which is 0.667. Figure 8.26 shows comparison of pressure distribution along the centerline. Both the Navier–Stokes and the augmented Burnett equations show a nonconstant pressure gradient along the channel. The solutions using Beskok's slip-boundary condition show less change in pressure gradient than those from the Langmuir's boundary condition. Figure 8.27 compares the streamwise velocity distributions along the centerline. The streamwise velocity distributions are almost identical except near the exit. Figure 8.28 compares the slip velocity distributions along the wall. The slip velocity profiles obtained from both Navier–Stokes and augmented Burnett equations are identical when the same wall-boundary conditions are employed. However, the Beskok's slip-boundary condition and Langmuir's slip-boundary condition show a large difference. The Beskok's slip-boundary

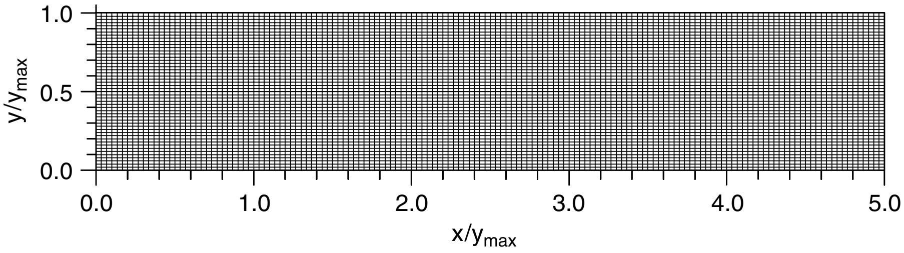


FIGURE 8.29 Microchannel geometry and grid.

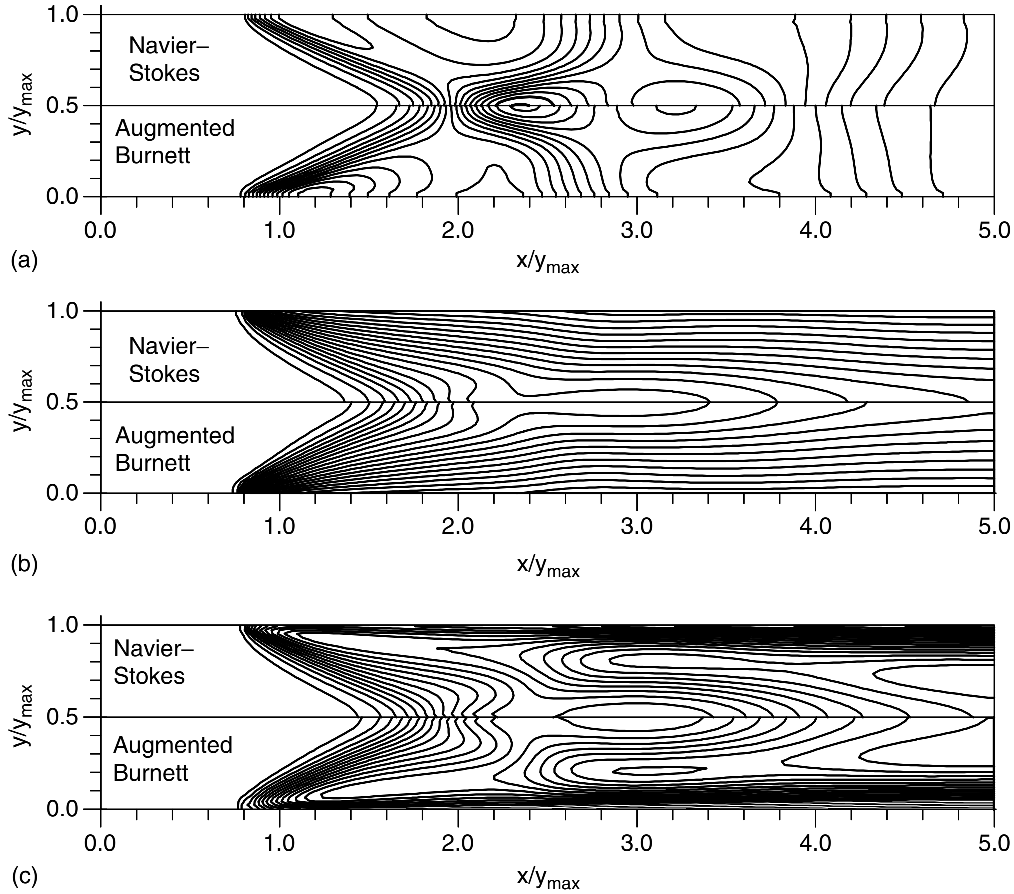


FIGURE 8.30 Comparisons of contours between Navier–Stokes and augmented Burnett equations: helium, $M_\infty = 5$ and $Kn_\infty = 0.7$.

condition predicts lower slip velocity near the entrance and higher slip velocity near the exit. As the figures show, there is very little difference between the Navier–Stokes solutions and the augmented Burnett solutions at the entrance, but as the local Knudsen number increases toward the exit of the channel, the difference between the Navier–Stokes solutions and the augmented Burnett solutions increases as expected.

8.7.6 Supersonic Flow in a Microchannel

The Navier–Stokes equations and the augmented Burnett equations are applied to compute the supersonic flow in a microchannel. The geometry and grid of the microchannel are shown in Figure 8.29. As the flow enters the channel, the tangential velocity component to the wall is retained, while the velocity component normal to the wall is neglected at wall boundaries in the region $0 \leq x \leq 1\,\mu$m. The first-order

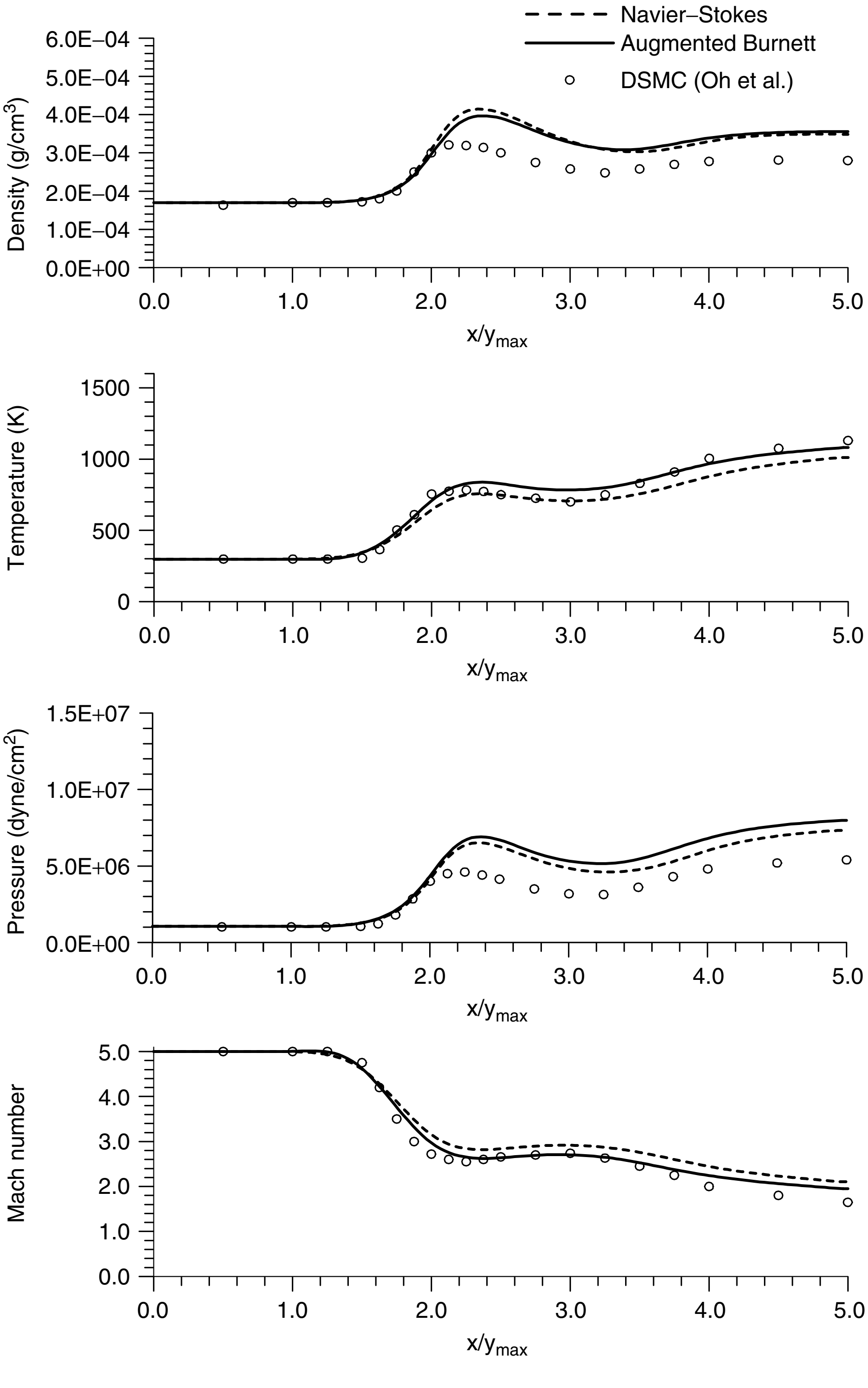


FIGURE 8.31 Comparisons of density, temperature, pressure, and Mach number profiles along the centerline of the channel: helium, $M_\infty = 5$ and $Kn_\infty = 0.7$.

Maxwell–Smoluchowski slip-boundary conditions are employed at the rest of the wall boundaries. The channel height and length are 2.4 and 12 μm, respectively. The flow conditions at the entrance for the helium flow are

$$M_\infty = 5.0, \quad P_\infty = 1.01 \times 10^6\,\text{dyne/cm}^2,$$

$$Kn_\infty = 0.07, \quad T_\infty = 298\,\text{K}$$

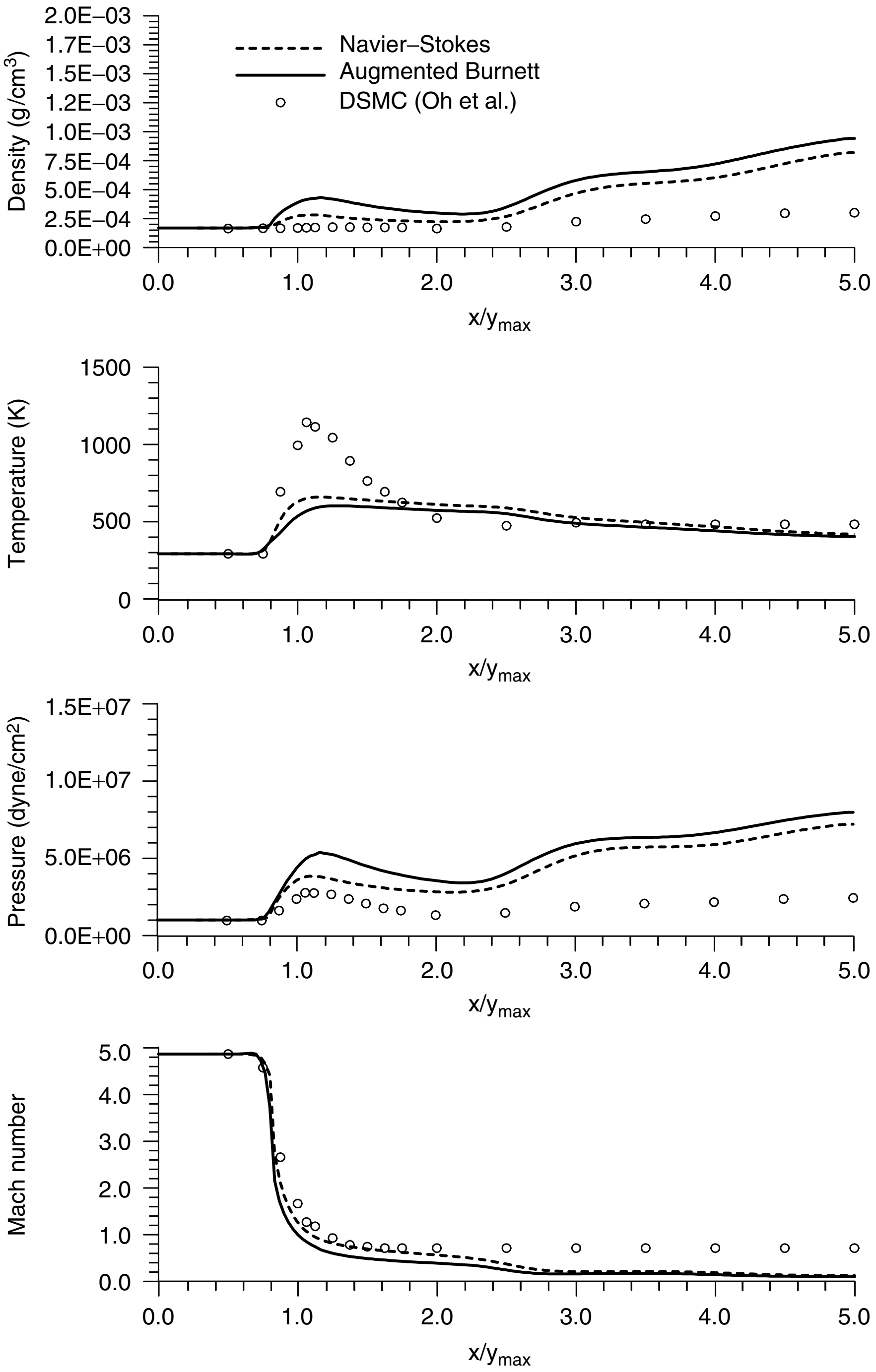


FIGURE 8.32 Comparisons of density, temperature, pressure, and Mach number profiles along the wall of the channel: helium, $M_\infty = 5$ and $Kn_\infty = 0.7$.

Figure 8.30 compares the pressure, Mach number, and temperature contours obtained from the Navier–Stokes and augmented Burnett equations. Solutions from the Navier–Stokes and augmented Burnett equations do not show significant differences. These flow property contours also agree well with the DSMC solutions obtained by Oh et al. (1997). Figure 8.31 compares the density, temperature, pressure, and Mach number profiles along the centerline of the channel using the Navier–Stokes, augmented Burnett, and DSMC formulations [Oh et al., 1997]. The profiles generally agree well with the DSMC results. The temperature and Mach number profiles especially show very close agreement with the DSMC

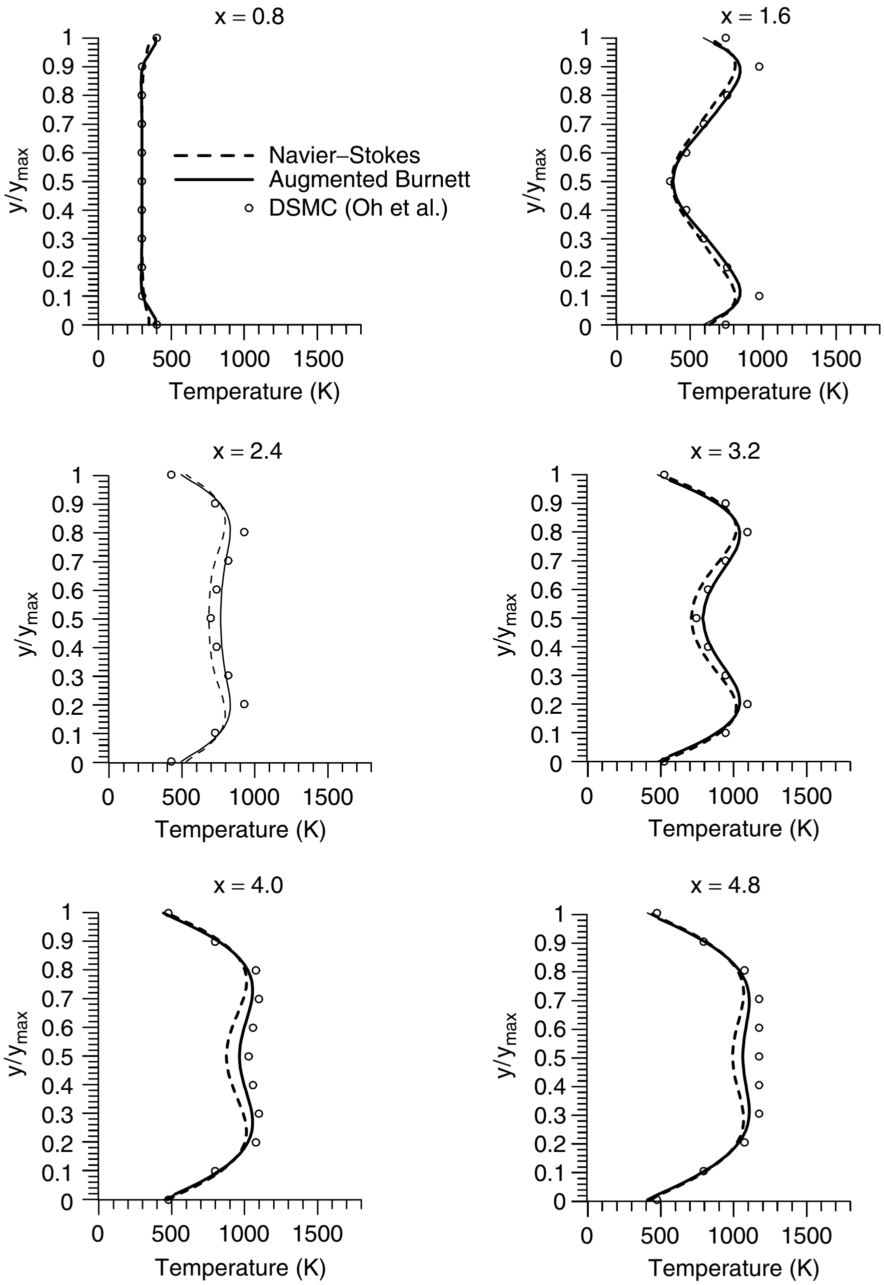


FIGURE 8.33 Comparisons of temperature profiles across the channel at various streamwise locations: helium, $M_\infty = 5$ and $Kn_\infty = 0.7$.

results. The augmented Burnett solutions are closer to the DSMC solutions than the Navier–Stokes solutions in the temperature and Mach number profiles. Figure 8.32 compares the density, temperature, pressure, and Mach number profiles along the channel wall using the Navier–Stokes, augmented Burnett, and DSMC formulations. Both the Navier–Stokes and augmented Burnett solutions show some difference with the DSMC solutions. Figures 8.33 and 8.34 compare the temperature and velocity profiles across the channel at various streamwise locations using the Navier–Stokes, augmented Burnett, and DSMC formulations respectively. The profiles obtained from the augmented Burnett solutions are closer to the DSMC results than the Navier–Stokes solutions.

Unfortunately, experimental data are not available to assess the accuracy of the Navier–Stokes, Burnett, and DSMC models for computing the microchannel flows. A substantial amount of both experimental

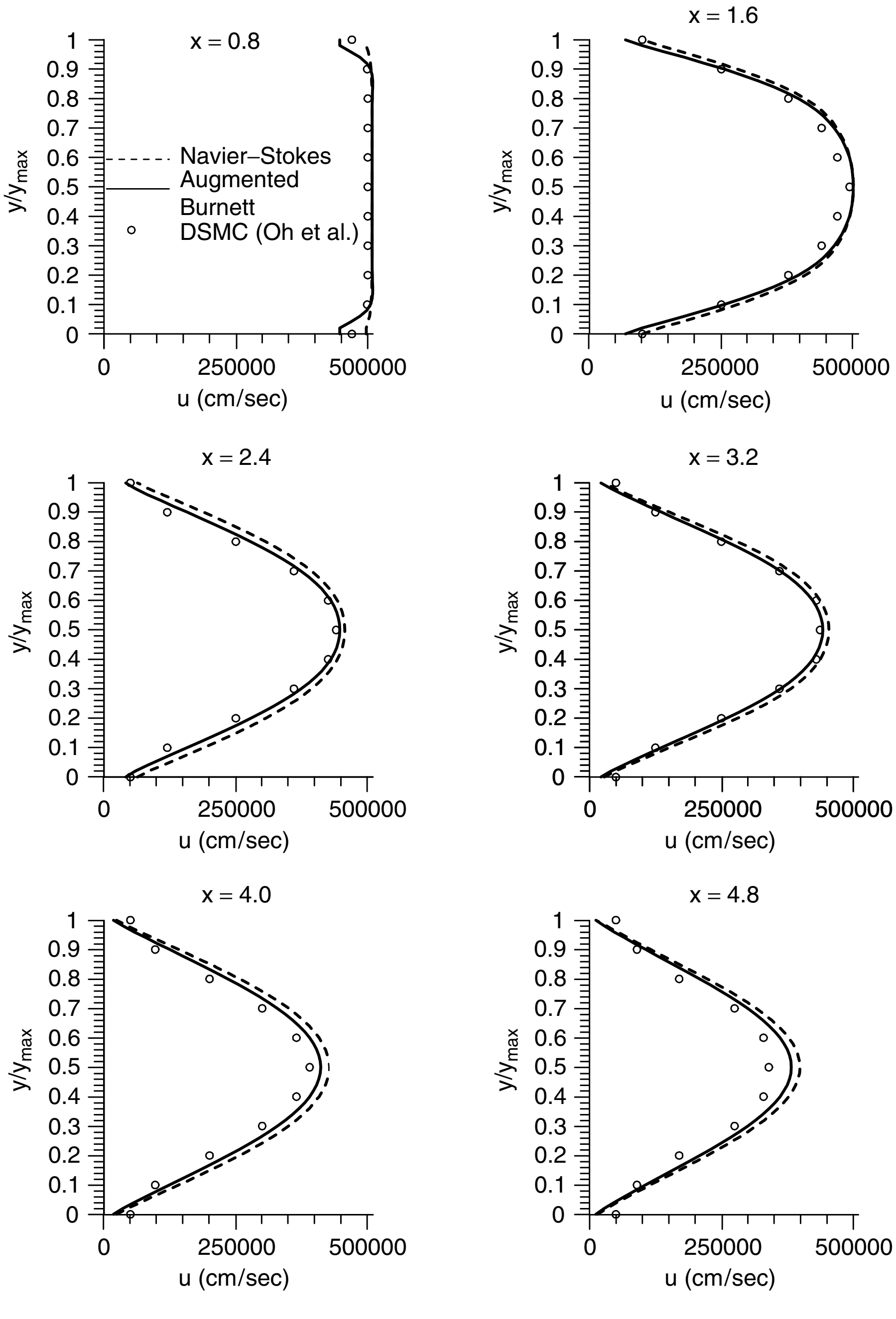


FIGURE 8.34 Comparisons of velocity profiles across the channel at various streamwise locations: helium, $M_\infty = 5$ and $Kn_\infty = 0.7$.

and computational simulation work is needed to determine the applicability and accuracy of various fluid models for computing high Knudsen number flow in microchannels.

8.8 Conclusions

For computing flows in the continuum–transition regime, higher order fluid dynamics models beyond Navier–Stokes are needed. These models are known as extended, or generalized, hydrodynamics models in the literature. Some of these models are the Burnett equations; 13-moment Grad's equations; Gaussian closure or Levermore moment equations; and Eu's equations. An alternative to generalized hydrodynamic models is the hybrid approach, which combines a Euler or Navier–Stokes solver with the DSMC method. Most of the generalized hydrodynamic models proposed to date suffer from the following drawbacks: they do not capture the physics properly or they are too computationally intensive, or both. In this chapter, the history of a set of extended hydrodynamics equations based on the Chapman–Enskog expansion of Boltzmann equations to $O(Kn^2)$ known as the Burnett equations has been reviewed. The various sets known in the literature as conventional, augmented, and BGK–Burnett equations have been considered and critically examined. Computations for hypersonic flows and microscale flows show that both the augmented and BGK–Burnett equations can be effectively applied to accurately compute flows in the continuum–transition regime. However, a great deal of additional work is needed, both experimentally and computationally, to assess the range of applicability and accuracy of Navier–Stokes, Burnett, and DSMC approximations for simulating the flows in transition regime.

References

Alsmeyer, H. (1976) "Density Profiles in Argon and Nitrogen Shock Waves Measured by the Absorption of an Electron Beam," *J. Fluid Mech.* **74**, pp. 497–513.

Arkilic, E.B., Schmidt, M.A., and Breuer, K.S. (1997) "Gaseous Flow in Long Microchannel," *J. MEMS* **6**, pp. 167–78.

Balakrishnan, R. (1999) Entropy Consistent Formulation and Numerical Simulation of the BGK–Burnett Equations for Hypersonic Flows in the Continuum–Transition Regime, Ph.D. thesis, Wichita State University.

Balakrishnan, R., and Agarwal, R.K. (1996) "Entropy Consistent Formulation and Numerical Simulation of the BGK–Burnett Equations for Hypersonic Flows in the Continuum–Transition Regime," in *Proc. of the 15th Int. Conf. on Numerical Methods in Fluid Dynamics*, Lecture Note in Physics, Springer–Verlag, New York, pp. 480–85.

Balakrishnan, R., and Agarwal, R.K. (1997) "Numerical Simulation of the BGK–Burnett for Hypersonic Flows," *J. Thermophys. Heat Transf.* **11**, pp. 391–99.

Balakrishnan, R., and Agarwal, R.K. (1999) "A Comparative Study of Several Higher-Order Kinetic Formulations Beyond Navier–Stokes for Computing the Shock Structure," AIAA Paper 99-0224, American Institute of Aeronautics and Astronomics, Reno, NV.

Balakrishnan, R., Agarwal, R.K., and Yun, K.Y. (1997) "Higher-Order Distribution Functions, BGK–Burnett Equations, and Boltzmann's H-Theorem," AIAA Paper 97-2552, American Institute of Aeronautics and Astronomics, Atlanta, GA.

Beskok, A., and Karniadakis, G. (1999) "A Model for Flows in Channels, Pipes, and Ducts at Micro and Nano Scales," *Microscale Thermophys. Eng.* **8**, pp. 43–77.

Beskok, A., Karniadakis, G., and Trimmer, W. (1996) "Rarefaction and Compressibility Effects in Gas Microflows," *J. Fluid Eng.* **118**, pp. 448–56.

Bhatnagar, P.L., Gross, E.P., and Krook, M. (1954) "A Model for Collision Process in Gas," *Phys. Rev.* **94**, pp. 511–25.

Bird, G.A. (1994) *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*, Oxford Science Publications, New York.

Bobylev, A.V. (1982) "The Chapman–Enskog and Grad Methods for Solving the Boltzmann Equation," *Sov. Phys. Doklady* **27**, pp. 29–31.

Boyd, I., Chen, G., and Candler, G. (1995) "Predicting Failure of the Continuum Fluid Equations in Transitional Hypersonic Flows," *Phys. Fluids* **7**, pp. 210–19.

Brown, S. (1996) Approximate Riemann Solvers for Moment Models of Dilute Gases, Ph.D. thesis, University of Michigan.

Burnett, D. (1935) "The Distribution of Velocities and Mean Motion in a Slight Non-Uniform Gas," *Proc. London Math. Soc.* **39**, pp. 385–430.

Chapman, S., and Cowling, T.G. (1970) *The Mathematical Theory of Non-Uniform Gases,* Cambridge University Press, New York.

Comeaux, K.A., Chapman, D.R., and MacCormack, R.W. (1995) "An Analysis of the Burnett Equations Based in the Second Law of Thermodynamics," AIAA Paper 95-0415, American Institute of Aeronautics and Astronautics, Reno, NV.

Eu, B.C. (1992) *Kinetic Theory and Irreversible Thermodynamics*, John Wiley & Sons, New York.

Fiscko, K.A., and Chapman, D.R. (1988) "Comparison of Burnett, Super-Burnett and Monte Carlo Solutions for Hypersonic Shock Structure," in *Proc. 16th Int. Symp. on Rarefied Gas Dynamics*, Pasadena, CA, July 1988, American Institute of Physics, pp. 374–95.

Gad-el-Hak, M. (1999) "The Fluid Mechanics of Microdevices," *J. Fluids Eng.* **121**, pp. 5–33.

Grad, H. (1949) "On the Kinetic Theory of Rarefied Gases," *Comm. Pure Appl. Math.* **2**, pp. 325–31.

Groth, C.P.T., Roe, P.L., Gombosi, T.I., and Brown, S.L. (1995) "On the Nonstationary Wave Structure of 35-Moment Closure for Rarefied Gas Dynamics," AIAA Paper 95-2312, American Institute of Aeronautics and Astronautics, San Diego, CA.

Harley, J.C., Huang, Y., Bau, H.H., and Zemel, J.N. (1995) "Gas Flow in Microchannels," *J. Fluid Mech.* **284**, pp. 257–74.

Holway, L.H. (1964) "Existence of Kinetic Theory Solutions to the Shock Structure Problem," *Phys. Fluids* **7**, pp. 911–13.

Ivanov, M.S., and Gimelshein, S.F. (1998) "Computational Hypersonic Rarefied Flows," *Annu. Rev. Fluid Mech.* **30**, pp. 469–505.

Koppenwallner, G. (1987) "Low Reynolds Number Influence on the Aerodynamic Performance of Hypersonic Lifting Vehicles," *Aerodyn. Hypersonic Lifting Vehicles* (AGARD, CP-428) **11**, pp. 1–14.

Lee, C.J. (1994) "Unique Determination of Solutions to the Burnett Equations," *AIAA J.* **32**, pp. 985–90.

Levermore, C.D. (1996) "Moment Closure Hierarchies for Kinetic Theory," *J. Stat. Phys.* **83**, pp. 1021–65.

Levermore, C.D., and Morokoff, W.J. (1998) "The Gaussian Moment Closure for Gas Dynamics," *SIAM J. Appl. Math.* **59**, pp. 72–96.

Liu, J.Q., Tai, Y.C., Pong, K.C., and Ho, C.M. (1993) "Micromachined Channel/Pressure Sensor Systems for Micro Flow Studies," in *Proc. of the 7th Int. Conf. on Solid-State Sensors and Actuators—Transducers '93*, Yokohama, Japan, 7–10 June, pp. 995–98.

Moss, J.N., and Bird, G.A. (1984) "Direct Simulation of Transitional Flow for Hypersonic Reentry Conditions," AIAA Paper 84-0223, American Institute of Aeronautics and Astronautics, Reno, NV.

Myong, R. (1999) "A New Hydrodynamic Approach to Computational Hypersonic Rarefied Gas Dynamics," AIAA Paper 99-3578, American Institute of Aeronautics and Astronautics, Norfolk, VA.

Nance, R.P., Hash, D.B., and Hassan, H.A. (1998) "Role of Boundary Conditions in Monte Carlo Simulation of Microelectromechanical Systems," *J. Spacecraft Rockets* **12**, pp. 447–49.

Oh, C.K., Oran, E.S., and Sinkovits, R.S. (1997) "Computations of High-Speed, High Knudsen Number Microchannel Flows," *J. Thermophys. Heat Transf.* **11**, pp. 497–505.

Oran, E.S., Oh, C.K., and Cybyk, B.Z. (1998) "Direct Simulation Monte Carlo: Recent Advances and Application," *Annu. Rev. Fluid Mech.* **30**, pp. 403–41.

Pong, K.C., Ho, C.M., Liu, J.Q., and Tai, Y.C. (1994) "Non-Linear Pressure Distribution in Uniform Microchannels," *ASME FED* **197**, pp. 51–56.

Reese, J.M., Woods, L.C., Thivet, F.J.P., and Candel, S.M. (1995) "A Second-Order Description of Shock Structure," *J. Comput. Phys.* **117**, pp. 240–50.

Roveda, R., Goldstein, D.B., and Varghese, P.L. (1998) "Hybrid Euler/Particle Approach for Continuum/Rarefied Flows," *J. Spacecraft Rockets* **35**, pp. 258–65.

Smoluchowski, von M. (1898) "Veder Warmeleitung in Verdumteu Gasen," *Ann. Phys. Chem.* **64**, pp. 101–30.

Steger, J.L., and Warming, R.F. (1981) "Flux Vector Splitting of the Inviscid Gas Dynamics Equations with Application to Finite-Difference Methods," *J. Comput. Phys.* **40**, pp. 263–93.

Sun, Q., Boyd, I.D., and Fan, J. (2000) "Development of Particle Methods for Computing MEMS Gas Flows," *J. MEMS* **2**, pp. 563–69.

Vogenitz, F.W., and Takara, G.Y. (1971) "Monte Carlo Study of Blunt Body Hypersonic Viscous Shock Layers," *Rarefied Gas Dynamics* **2**, pp. 911–18.

Weiss, W. (1996) "Comments on Existence of Kinetic Theory Solutions to the Shock Structure Problem," *Phys. Fluids* **8**, pp. 1689–90.

Welder, W.T., Chapman, D.R., and MacCormack, R.W. (1993) "Evaluation of Various Forms of the Burnett Equations," AIAA Paper 93-3094, American Institute of Aeronautics and Astronautics, Orlando, FL.

Yun, K.Y., and Agarwal, R.K. (2000) "Numerical Simulation of 3-D Augmented Burnett Equations for Hypersonic Flow in Continuum–Transition Regime," AIAA Paper 2000-0339, American Institute of Aeronautics and Astronautics, Reno, NV.

Yun, K.Y., Agarwal, R.K., and Balakrishnan, R. (1998a) Three-Dimensional Augmented and BGK–Burnett Equations, unpublished report, Wichita State University.

Yun, K.Y., Agarwal, R.K., and Balakrishnan, R. (1998b) "Augmented Burnett and Bhatnagar–Gross–Krook–Burnett for Hypersonic Flow," *J. Thermophys. Heat Transf.* **12**, pp. 328–35.

Zhong, X. (1991) Development and Computation of Continuum Higher Order Constitutive Relations for High-Altitude Hypersonic Flow, Ph.D. thesis, Stanford University.

Zhong, X., and Furumoto, G. (1995) "Augmented Burnett Equation Solutions over Axisymmetric Blunt Bodies in Hypersonic Flow," *J. Spacecraft Rockets* **32**, pp. 588–95.

9

Lattice Boltzmann Simulations of Slip Flow in Microchannels

Ramesh K. Agarwal
Washington University in St. Louis

9.1 Introduction

Historically originating from the seminal work of Frisch, Hasslacher, and Pomeau (1986) on lattice gas automata (LGA), the lattice Boltzmann method (LBM) has recently developed into an alternative and very promising numerical scheme for simulating fluid flows [Chen and Doolen, 1998]. The lattice Boltzmann algorithms are simple, fast, and very suitable for parallel computing. It is also easy to incorporate complex boundary conditions with the LBM. The algorithms have been successfully applied to compute flows modeled by the incompressible Navier–Stokes equations including reactive and multiphase flows.

Unlike the conventional numerical methods, which directly discretize the continuum equations of fluid dynamics on a finite-difference, finite-volume, or finite-element mesh, the LBM derives its basis from the kinetic theory that models the microscopic behavior of gases. The fundamental idea behind LBM is to construct the simplified kinetic models that capture the essential physics of microscopic behavior so that the macroscopic flow properties (calculated from the microscopic quantities) obey the desired continuum equations of fluid dynamics. Thus LBM is based on the particle dynamics governed by a simplified model of the Boltzmann equation; the simplification is usually to the nonlinear collision integral. In 1992, a major simplification of the original LBM was achieved by Chen et al. (1992) and Qian et al. (1992) by employing a single relaxation time approximation due to Bhatnagar, Gross, and Krook (BGK) to the collision operator in the lattice Boltzmann equation. In this lattice BGK (LBGK) model, one solves the evolution equations of the distribution functions of fictitious fluid particles colliding and moving synchronously on a symmetric lattice. The symmetric lattice space is a result of the discretization of the particle velocity space and the condition for synchronous motions. That is, the discretizations of time and particle phase space are coherently coupled. This makes the evolution of the lattice Boltzmann equation very simple consisting of only two steps, collision and advection. Furthermore, the advection operator in phase space (velocity space) is linear in contrast to the nonlinear convection terms in the macroscopic continuum equations of fluid dynamics. Thus, this simple linear advection operator in LBM combined with the simplified BGK collision operator results in the recovery of nonlinear macroscopic convection. Qian et al. (1992) and others using multiple scale expansion have shown that the local equilibrium particle distribution function obtained from the BGK-Boltzmann equation can recover the Navier–Stokes equations, and the incompressible Navier–Stokes equations can be obtained in the nearly incompressible limit of the LBGK method.

Thus, three essential ingredients in the development of a lattice Boltzmann method for a single physics or multiphysics fluid-flow problem must be completely specified: (1) a discrete lattice on which the fluid particles reside, (2) a set of discrete velocities $\mathbf{e}_i$ to represent particle advection from one node of the lattice to its nearest neighbor, and (3) a set of rules for the redistribution of particles on a node to mimic collision processes in the fluid. These rules are provided by the distribution functions f_i of the particles; the evolution of distribution functions in time (for a discrete time step Δt) is obtained by solving the LBGK equation. The LBGK equation for f_i requires the knowledge of the equilibrium distribution function $f_i^{(0)}$. The discrete velocities $\mathbf{e}_i$ are determined so that the macroscopic density and momentum satisfy the constraints $\rho = \sum_i f_i$ and $\rho V = \sum_i f_i \mathbf{e}_i$ respectively, where V is the macroscopic-averaged fluid velocity. Therefore, the determination of appropriate equilibrium particle distribution function for a given fluid flow problem is essential for solving the problem by LBM.

For multiphysics problems, sometimes it is not a straightforward process to determine the distribution function. Magneto-hydrodynamics is one such area where it has been difficult to develop the LBM in a systematic and straightforward manner because of the difficulty in determining the distribution functions that correspond to the MHD continuum flow equations and magnetic field equations. Nevertheless there have been many attempts to solve the MHD equations by LBM. The earliest dates to 1988 [Chen et al., 1988] using lattice gas automata (LGA). A summary of the previous work in lattice Boltzmann MHD is given in Dellar (2002) and will not be repeated here. In this paper, we employ the LBGK method to compute the slip flow in a pressure-driven microchannel flow without and with magnetic field. To allow for the rarefaction effects (variation in Knudsen number along the length of the channel), we follow the approach developed by Lim et al. (2002). For MHD flow in a microchannel, this paper extends the previous work of Agarwal (2001) on lattice Boltzmann MHD for continuum flows; it combines the lattice kinetic scheme of Dellar (2002) with the approach of Lim et al. (2002). We provide the LBGK formulation for MHD slip flow; the formulation of Navier–Stokes slip flow is a subset of MHD formulation simply obtained by equating the magnetic field to zero.

9.2 3-D Compressible Viscous MHD Equations

Three-dimensional viscous MHD equations in tensor notation can be written as:

$$\frac{\partial \rho}{\partial t} + \frac{\partial \rho u_\alpha}{\partial x_\alpha} = 0 \tag{9.1}$$

$$\frac{\partial(\rho u_\alpha)}{\partial t} + \frac{\partial(\rho u_\alpha u_\beta)}{\partial x_\beta} = -\frac{\partial}{\partial x_\alpha}\left(P + \frac{B^2}{8\pi}\right) + \frac{\partial}{\partial x_\beta} \left[\rho \nu\left(\frac{\partial u_\beta}{\partial x_\alpha} + \frac{\partial u_\alpha}{\partial x_\beta} + \frac{2}{3}\delta_{\alpha\beta}\frac{\partial u_m}{\partial x_m}\right) + \frac{B_\alpha B_\beta}{4\pi}\right] \tag{9.2}$$

$$\frac{\partial B_\alpha}{\partial t} + \frac{\partial}{\partial x_\beta}\left(u_\beta B_\alpha - u_\alpha B_\beta\right) = \eta \frac{\partial^2 B_\alpha}{\partial x_\beta \partial x_\beta} \tag{9.3}$$

$$\frac{\partial B_\alpha}{\partial x_\alpha} = 0 \tag{9.4}$$

In Equations (1)–(4), ν and η are the kinematic viscosity and magnetic resistivity of the fluid and are assumed constant. Note that $\eta = \frac{1}{\sigma\mu_f}$, where σ is the electrical conductivity and μ_f is the magnetic permeability. Equation (9.4) is the solenoidal condition on the magnetic field. Sections 9.3 and 9.4 develop the LBGK method for solving Equations (9.1)–(9.4). In the absence of magnetic field, Equations (9.3) and (9.4) become identically zero, and Equation (9.2) changes to the Navier–Stokes equation for a compressible viscous fluid.

9.3 LBGK Equation and Equilibrium Particle Distribution Function f_i^0 for MHD Flow Equations

In two dimensions, as shown in Figure 9.1, we consider a square lattice with unit spacing on which each node has eight nearest neighbors connected by eight links. Particles can reside only on the nodes and move to their nearest neighbors along the links in unit time. There are two types of moving particles: the particles that move along the axis with velocity$|\mathbf{e}_i| = 1$, $i = 1, 2, 3, 4$ and the particles that move along the diagonals with velocity$|\mathbf{e}_i| = \sqrt{2}$, $i = 5, 6, 7, 8$. Also, there are rest particles with speed zero at each node. The occupation of these three types of particles is described by the single particle distribution function f_i where the subscript i indicates the velocity direction. The distribution function f_i is the probability of finding a particle i at node $\mathbf{x}$ at time t with velocity $\mathbf{e}_i$. We assume that the particle distribution function f_i evolves according to the LBGK equation:

$$\frac{\partial f_i}{\partial t} + \mathbf{e}_i \cdot \nabla f_i = -\frac{1}{\tau}\left(f_i - f_i^{(0)}\right), \tag{9.5}$$

where $f_i^{(0)}$ is the equilibrium particle distribution function for MHD equations and τ is the single relaxation time that controls the rate of approach to equilibrium.

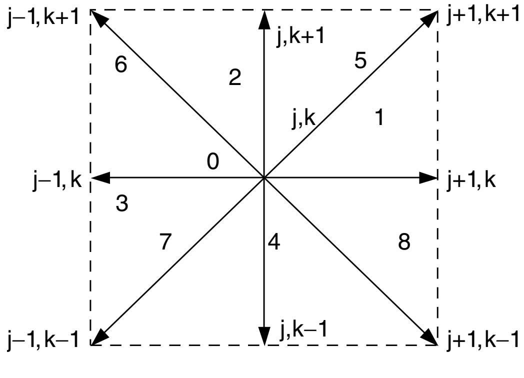


FIGURE 9.1 Nine-velocity square lattice.

For determination of $f_i^{(0)}$ in Equation (9.5) for MHD flow Equations (9.1) and (9.2), Dellar (2002) accounts for the influence of Lorentz force in the hydrodynamic equilibrium particle distribution function (derived from the Maxwellian) by changing the second moment of the equilibrium distribution function to include the divergence of Maxwell stress by adding appropriate magnetic field terms to the hydrodynamic lattice Boltzmann distribution function. In our notation, the equilibrium particle distribution function of [Dellar, 2002] in two dimensions for $\boldsymbol{B} = (B_x, B_y, 0)$ can be written as

$$f_i^{(0)} = \rho\omega_i\left[1 + 3(\mathbf{e}_i \cdot \mathrm{V}) + \frac{9}{2}(\mathbf{e}_i \cdot \mathrm{V})^2 - \frac{3}{2}|\mathrm{V}|^2\right] + \frac{9}{2}\omega_i\left[\frac{1}{2}|\mathbf{e}_i|^2|\boldsymbol{B}|^2 - (\mathbf{e}_i \cdot \boldsymbol{B})^2\right], \tag{9.6}$$

where $\omega_0 = 4/9$, $\omega_1 = \omega_2 = \omega_3 = \omega_4 = 1/9$, and $\omega_5 = \omega_6 = \omega_7 = \omega_8 = 1/36$. In the absence of magnetic field, Equation (9.6) reduces to standard hydrodynamic LBGK equation for simulating Navier–Stokes flows. The relaxation time τ in Equation (9.5) is related to the kinematic viscosity ν by $\tau = 3\nu$.

9.4 LBGK Equation and Equilibrium Particle Distribution Function $\mathbf{g}_i^{(0)}$ for Magnetic Induction Equation

It is not possible to construct a kinetic formulation for the magnetic induction Equation (9.3) using a scalar distribution function. Croisille et al. (1995) introduced a vector-valued distribution function for the magnetic field $\boldsymbol{B}$ in their development of a kinetic scheme for the MHD equations. However, their formulation involves an integral equation and does not explicitly describe the equilibrium distribution function. Dellar (2002) has now developed a vector-valued particle distribution function $\mathbf{g}_i$ that obeys the vector LBGK equation:

$$\frac{\partial \mathbf{g}_i}{\partial t} + \zeta_i \cdot \nabla \mathbf{g}_i = -\frac{1}{\tau_m}\left(\mathbf{g}_i - \mathbf{g}_i^{(0)}\right), \tag{9.7}$$

where τ_m is the relaxation time related to the magnetic resistivity η by $\tau_m = 2\eta$. Dellar (2002) has shown that the suitable and simplest choice for $\mathbf{g}_i^{(0)}$ is

$$g_{i\beta}^{(0)} = W_i\left[B_\beta + 2\zeta_{i\alpha}\left(V_\alpha B_\beta - B_\alpha V_\beta\right)\right], \tag{9.8}$$

where $W_0 = 1/3$ and $W_i = 1/6$, $i = 1, 2, 3, 4$ for the symmetric square lattice shown in Figure 9.1. The magnetic field is given by $\boldsymbol{B} = \Sigma_i\, \mathbf{g}_i$. Dellar (2002) also has shown that this formulation makes it possible to maintain $\nabla \cdot \boldsymbol{B} = 0$ to machine round-off error as required by Equation (9.4). A Chapman–Enskog procedure can be applied to determine the macroscopic behavior of the model represented by Equations (9.6) and (9.8). It can be shown that the macroscopic behavior matches Equations (9.1)–(9.3) as shown by Dellar (2002).

9.5 Solution of Coupled LBGK Equations for Particle Distribution Functions

LBGK Equations (9.5) and (9.7) are solved by writing the equations in fully discretized form. Again following Dellar (2002), these equations can be discretized for the evolution of particle distribution functions at time step Δt as:

$$\bar{f}_i(\mathbf{x} + \mathbf{e}_i\Delta t, t + \Delta t) = \bar{f}_i(\mathbf{x}, t) - \frac{\Delta t}{\tau + \Delta t/2}\left\{\bar{f}_i(\mathbf{x}, t) - f^{(0)}(\mathbf{x}, t)\right\}, \text{ and} \tag{9.9}$$

$$\bar{\mathbf{g}}_i\,(\mathrm{x} + \mathbf{e}_i\Delta t, t + \Delta t) = \bar{\mathbf{g}}_i(\mathrm{x}, t) - \frac{\Delta t}{\tau_m + \Delta t/2}\left\{\bar{\mathbf{g}}_i(\mathrm{x}, t) - \mathbf{g}^{(0)}(\mathrm{x}, t)\right\}, \text{ where} \tag{9.10}$$

$$\bar{f}_i\,(\mathbf{x}, t) = f_i\,(\mathbf{x}, t) + \frac{\Delta t}{2\tau}\left\{f_i\,(\mathbf{x}, t) - f_i^{(0)}(\mathbf{x}, t)\right\}, \text{ and} \tag{9.11}$$

$$\bar{\mathbf{g}}_i(\mathrm{x}, t) = \mathbf{g}_i\,(\mathrm{x}, t) + \frac{\Delta t}{2\tau_m}\left\{\mathbf{g}_i\,(\mathrm{x}, t) - \mathbf{g}_i^{(0)}(\mathrm{x}, t)\right\} \tag{9.12}$$

From Equations (9.9)–(9.12), f_i and $\mathbf{g}_i$ can be obtained explicitly at each time step Δt. For steady state computations, solution can be marched in time until a specified convergence criterion is met.

9.6 Pressure-Driven Slip Flow in a Microchannel without and with Magnetic Field

9.6.1 Analytical and Numerical Solutions

We consider the pressure-driven MHD slip flow in a long constant area microchannel as shown in Figure 9.2 subjected to a constant magnetic field B_0 in y-direction and a constant electric field E_0 in the z-direction. Let the bar "¯" over a flow quantity denote the average value at the exit of the channel so that, $\bar{p}, \bar{\rho}$ and $\bar{u}$ are average pressure, density, and velocity respectively at the exit. Then all the relevant non-dimensional parameters — Mach number M, Knudsen number Kn, Reynolds number Re, and Hartmann number Ha — obtained in terms of the exit variables are also shown in Figure 9.2. The pressure at the inlet and outlet of the channel is different, and Knudsen number becomes an important parameter to account for rarefaction. Now we define the non-dimensional variables as $\tilde{p} = (p/\bar{p})$, $\tilde{\rho} = (\rho/\bar{\rho})$, $\tilde{u} = (u/\bar{u})$, $\tilde{x} = (x/L)$, and $\tilde{y} = (y/H)$. The first- and second-order boundary conditions for the slip velocity at the wall can be written as follows.

First-order Maxwell–Smoluchowski (1879) boundary condition:

$$\tilde{u}\big|_{\text{wall}} = \alpha \mathrm{K} \left.\frac{\partial \tilde{u}}{\partial \tilde{y}}\right|_{\text{wall}} \tag{9.13}$$

Second-order Beskok (1999) boundary condition:

$$\tilde{u}\big|_{\text{wall}} = \alpha\left[\frac{\mathrm{K}}{1 - \mathrm{bK}}\left(\frac{\partial \tilde{u}}{\partial \tilde{y}}\right)_{\text{wall}}\right] \tag{9.14}$$

In Equations (9.13) and (9.14), K is the local Knudsen number and b is the slip coefficient.

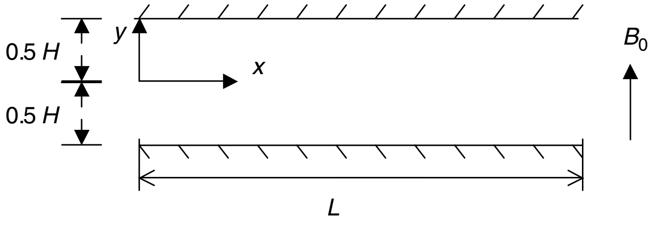


FIGURE 9.2 MHD slip flow in a microchannel. E_0 is a constant electric field in the z-direction, B_0 is the constant magnetic field in y-direction. $\varepsilon = \frac{H}{L}$, $\mathrm{Ha} = \frac{B_0^2 H^2 \sigma}{\bar{\rho}\,\upsilon}$, $\mathrm{Kn} = \sqrt{\frac{\pi\gamma}{2}} \cdot \frac{\mathrm{M}}{\mathrm{Re}}$, $\mathrm{Re} = \frac{\bar{u}H}{\nu}$, $\mathrm{M} = \frac{\bar{u}}{c}$, "¯" denotes the average outlet condition.

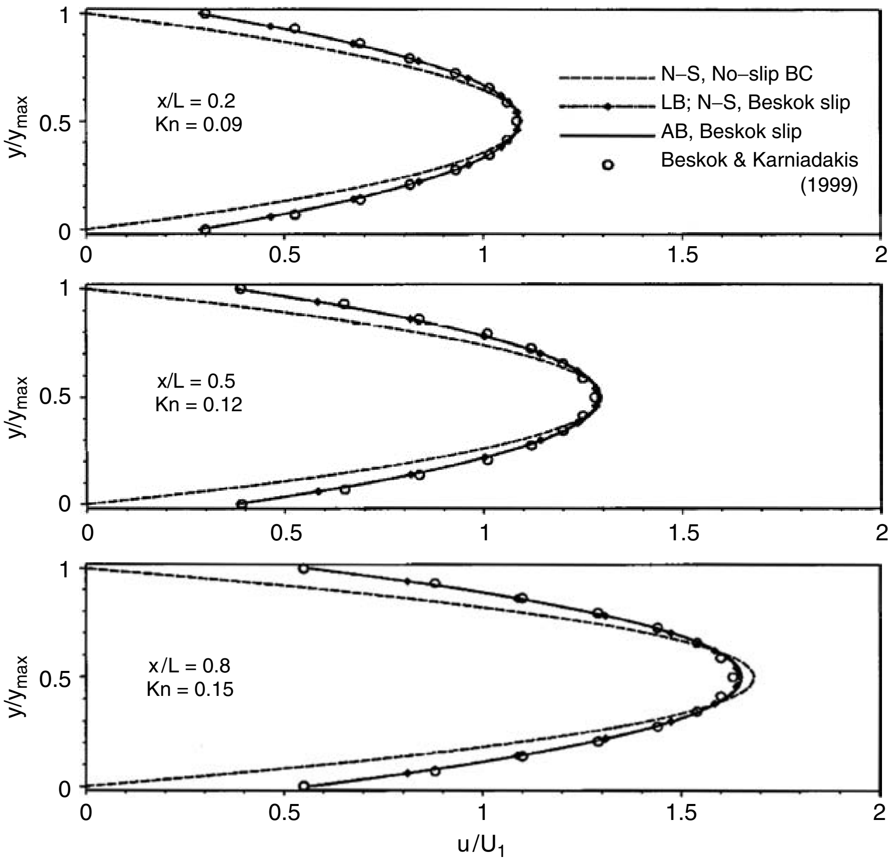


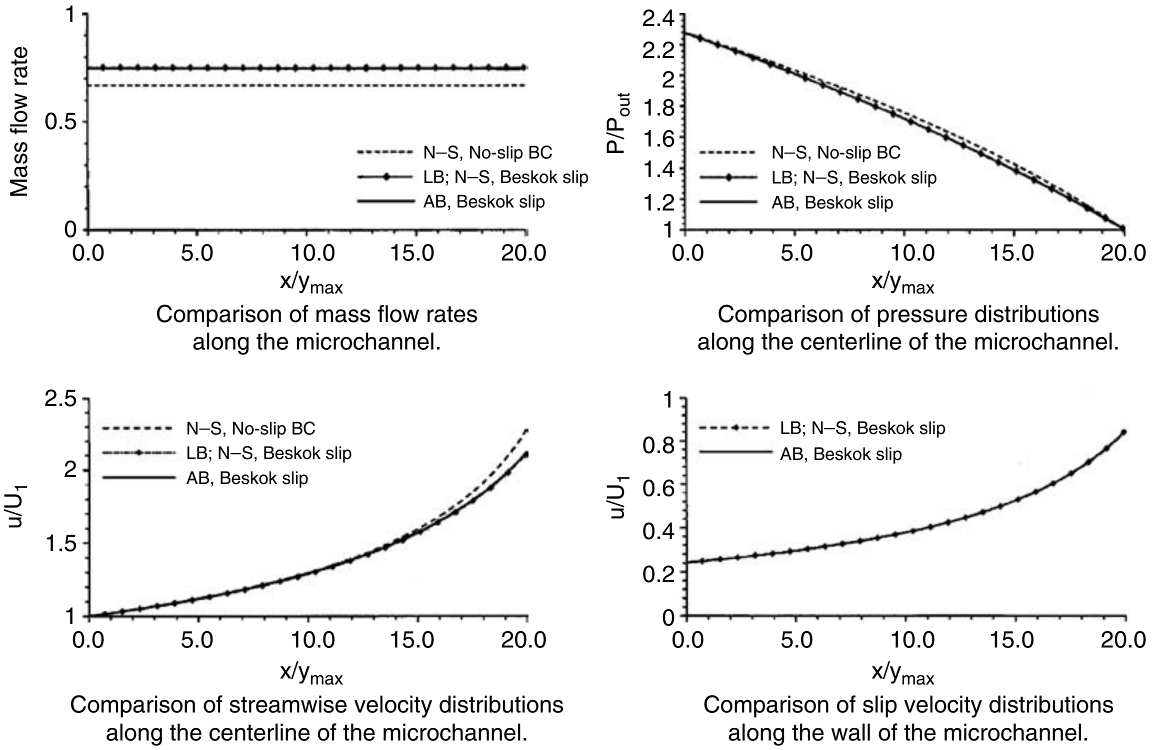
FIGURE 9.3 Comparison of velocity profiles at various streamwise locations of the microchannel: $\mathrm{Kn_{in}} = 0.088$, $\mathrm{Kn_{out}} = 0.2$, and $\mathrm{P_{in}/P_{out}} = 2.28$; N–S ≡ Navier–Stokes, LB ≡ Lattice-Boltzmann, AB ≡ Augmented Burnett.

In Equations (9.13) and (9.14), α is the accommodation coefficient. $\alpha = 1$ for a noncatalytic wall. Agarwal (2005) has recently obtained an exact analytical solution for velocity along the channel as follows:

$$\tilde{u} = \left[\sqrt{\frac{\pi}{2\gamma}}\,\frac{\varepsilon}{\mathrm{MKnHa}^2}\,\frac{\partial\tilde{p}}{\partial\tilde{x}} + \frac{\mathrm{E}_0}{B_0\tilde{u}}\right] \left[\left\{\frac{2\sinh\left(\frac{\mathrm{Ha}}{2}\right) + 2\alpha\,\mathrm{KHacosh}\left(\frac{\mathrm{Ha}}{2}\right)}{(1+\alpha^2\mathrm{K}^2\mathrm{Ha}^2)\sinh(\mathrm{Ha}) + 2\alpha\mathrm{KHacosh}(\mathrm{Ha})}\right\}\cosh\,(\mathrm{Ha}\tilde{y}) - 1\right] \tag{9.15}$$

The solution given by Equation (9.15) is valid for both first-order Maxwell–Smoluchowki (1879) and second-order Beskok (1999) slip boundary conditions. In Equation (9.15), K is the local Knudsen number, $= (\mathrm{Kn}/\tilde{p})$ and Kn is the Knudsen number at the outlet. The solution (9.15) as written has been obtained with first-order boundary condition (9.13); however, if K is replaced by $[\mathrm{K}/(1 - \mathrm{bK})]$, it becomes valid for second-order boundary condition (9.14). In the absence of magnetic field, the solution given by Equation (9.15) is the same as given in Beskok and Karniadakis (1999).

We now compute the analytical solution give by equation (9.15) for M (Mach number at outlet) = 0.1, $\mathrm{Kn_{out}}$ (Knudsen number at outlet) = 0.3, $\mathrm{Kn_{in}}$ (Knudsen number at inlet) = 0.088, and $\mathrm{P_{in}/P_{out}}$ (inlet pressure/outlet pressure) = 2.28. Figure 9.3 shows the velocity profiles at three streamwise locations $\tilde{x} = 0.2$, 0.5, and 0.8 computed using the LBGK method and their comparison with profiles computed using the Navier–Stokes (N–S) equations with and without slip, augmented Burnett (AB) equations with slip, and



Comparison of mass flow rates along the microchannel.

Comparison of pressure distributions along the centerline of the microchannel.

Comparison of streamwise velocity distributions along the centerline of the microchannel.

Comparison of slip velocity distributions along the wall of the microchannel.

FIGURE 9.4 Comparison of mass flow rates, pressure distributions, and velocity distributions with no-slip and slip boundary conditions. $\mathrm{Kn_{in}} = 0.088$, $\mathrm{Kn_{out}} = 0.3$, $\mathrm{P_{in}/P_{out}} = 2.28$, $\varepsilon = H/L = 0.05$, $\alpha = 1$; N–S ≡ Navier–Stokes, LB ≡ Lattice–Boltzmann, AB ≡ Augmented Burnett.

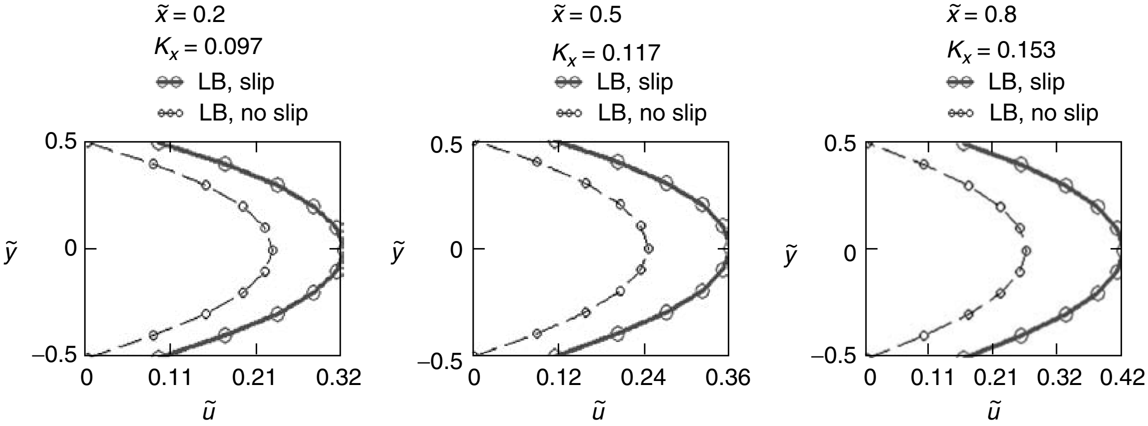


FIGURE 9.5 **(See color insert following** page 10-34.) Velocity profiles for MHD flow in a microchannel. $\mathrm{Kn_{in}} = 0.088$, $\mathrm{Kn_{out}} = 0.3$, $\mathrm{P_{in}/P_{out}} = 2.28$, $\varepsilon = H/L = 0.05$, $\alpha = 1$, $\mathrm{M} = 0.1$, $\mathrm{Ha} = 0.054$, $\mathrm{E_0} = 0$.

the analytical solution given by Beskok and Karniadakis (1999) and Equation (9.15). All the solutions are in excellent agreement with each other except the Navier–Stokes solution with no-slip boundary condition, which is expected because all other methods employ the slip boundary condition. Figure 9.4 shows that the mass flow is conserved in the computations. Also the LBGK solution agrees quite well with the Navier–Stokes (N–S) and augmented Burnett (AB) solutions with slip boundary conditions for streamwise velocity distribution along the centerline of the channel, pressure distribution along the centerline

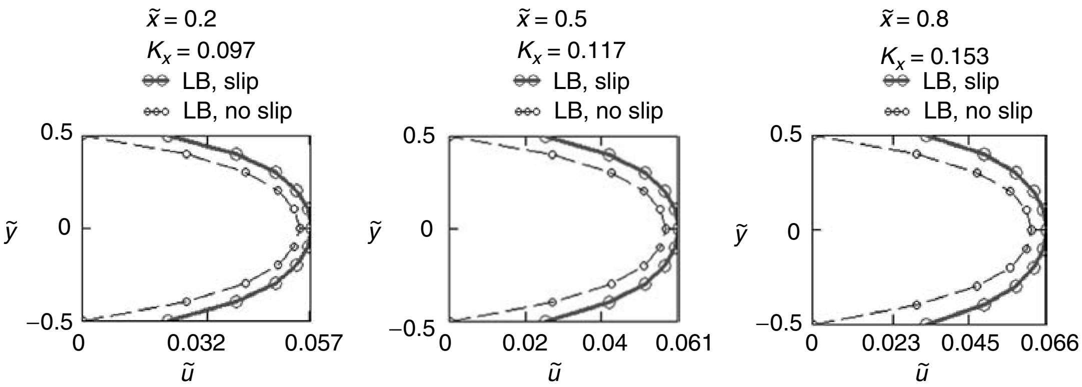


FIGURE 9.6 **(See color insert following page 10-34.)** Velocity profiles for MHD flow in a microchannel. $\mathrm{Kn_{in}} = 0.088$, $\mathrm{Kn_{out}} = 0.3$, $\mathrm{P_{in}/P_{out}} = 2.28$, $\varepsilon = H/L = 0.05$, $\alpha = 1$, $\mathrm{M} = 0.1$, $\mathrm{Ha} = 5.4$, $\mathrm{E_0} = 0$.

of the channel, and slip velocity distribution on the channel wall. The same computations are performed in the presence of magnetic field for various Hartmann number Ha. Figures 9.5–9.7 show the velocity profiles at three streamwise locations $\tilde{x} = 0.2$, 0.5, and 0.8 for Ha = 0.054, 5.4, and 54.0 respectively computed by the LBGK method and the analytical solution given by Equation (9.15). Again the agreement between the two sets of solutions is excellent. These solutions have been obtained using both the no slip ($\tilde{u} = 0$ at the wall) and second-order slip boundary condition.

The next section briefly describes the procedure employed in obtaining the solutions using the LBGK method.

9.6.2 LBGK Solution Procedure

For computing the LBGK solution, a uniform lattice with equally spaced points is created by generating a 1001 × 51grid with square cells. The following steps are followed in obtaining the LBGK solution.

1. The characteristic velocity and length scale are chosen to be $\overline{u} = \mathrm{M}c$ (where c is the speed of sound $= \sqrt{\gamma RT}$, the flow is assumed to be isothermal) and H respectively.
2. Reynolds number at the exit is calculated from $\mathrm{Re} = \sqrt{\frac{\pi\gamma}{2}}\frac{\mathrm{M}}{\mathrm{Kn_{out}}}$ from which the kinematic viscosity is calculated as $\nu = \frac{\overline{u}H}{\mathrm{Re}}$.
3. The relaxation time is then determined as $\tau = 3\nu$.
4. The electrical conductivity is calculated from the expression $\sigma = \frac{1}{\rho\nu}\left(\frac{B_o H}{\mathrm{Ha}}\right)^2$ for a given Hartmann number Ha and magnetic field B_0.
5. The relaxation time τ_m is then determined by $\tau_m = 2\eta$.
6. After determining all the relevant parameters as described in steps (1)–(5), flow field is initialized by assuming a distribution of density, velocity and magnetic field.
7. The initial values of distribution functions (as equilibrium distribution functions $f_i^{(0)}$ and $\mathbf{g}_i^{(0)}$ at $t = 0$) are then determined on the lattice from Equations (9.6) and (9.8) respectively.
8. The updating of the particle distribution functions f_i and $\mathbf{g}_i$ at subsequent time steps is done as described in Equations (9.9)–(9.12).
9. Step 8 is repeated until the convergence of both the distribution functions is obtained.
10. The macroscopic variables are then calculated from the converged distribution functions as

$$\rho = \sum_i f_i, \rho V = \sum_i f_i \mathbf{e}_i, \quad \text{and} \quad B = \sum_i \mathbf{g}_i. \tag{9.16}$$

In Equation (9.16), i represents summation over all lattice points. The treatment of boundary conditions in LBGK method is similar to that described in Lim et al. (2002).

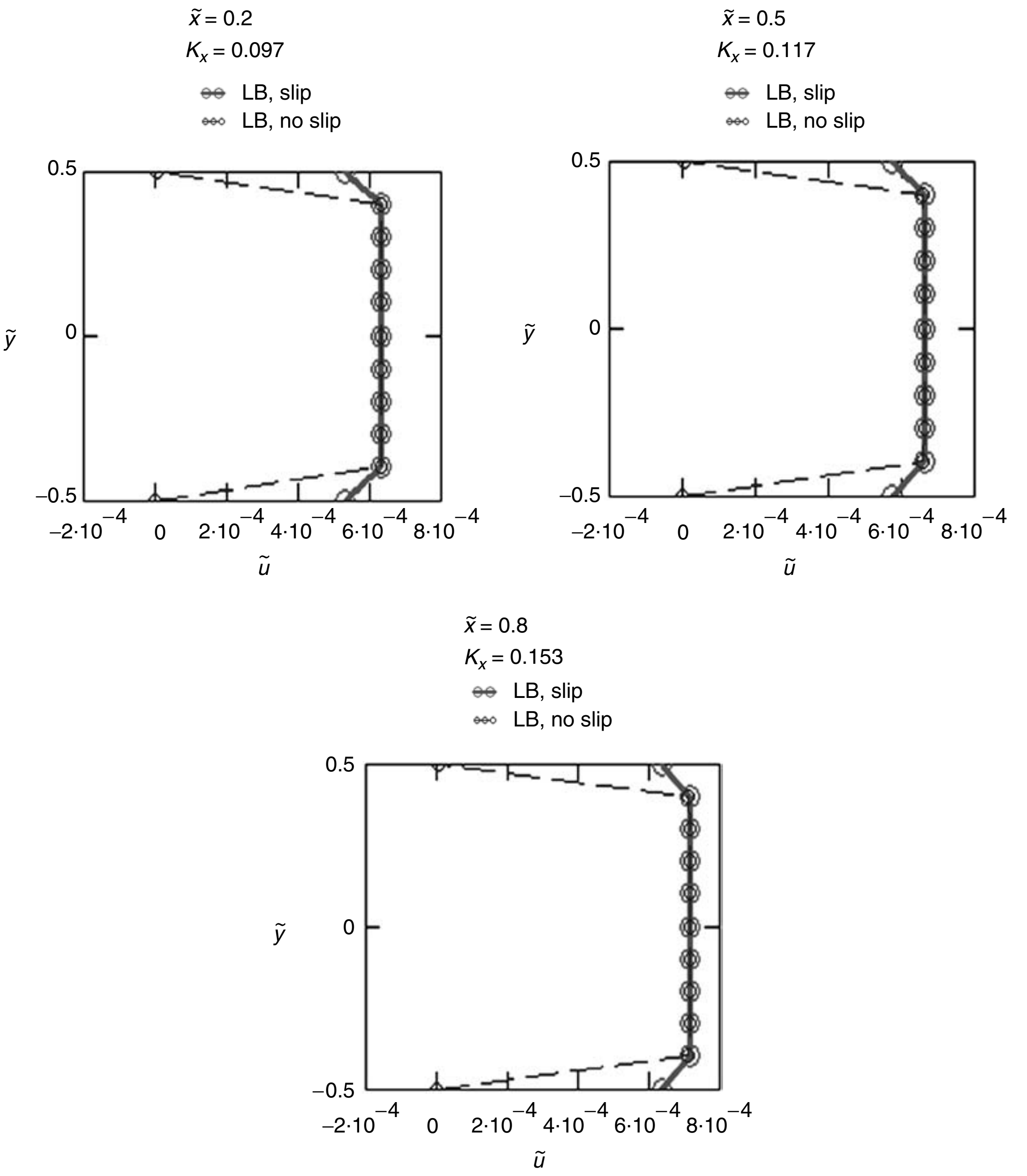


FIGURE 9.7 (**See color insert following page 10-34.**) Velocity profiles for MHD flow in a microchannel. $\mathrm{Kn_{in}} = 0.088$, $\mathrm{Kn_{out}} = 0.3$, $\mathrm{P_{in}/P_{out}} = 2.28$, $\varepsilon = H/L = 0.05$, $\alpha = 1$, $\mathrm{M} = 0.1$, $\mathrm{Ha} = 54$, $\mathrm{E_0} = 0$.

9.7 Conclusions

A Lattice-BGK (LBGK) formulation has been developed for incompressible Navier–Stokes and viscous MHD flows. The method has been successfully applied to compute the slip flow in a microchannel without and with magnetic field. The results show the strong potential of the LBGK method for achieving high efficiency as well as accuracy on a lattice comparable to a finite-difference grid.

Acknowledgments

The author wants to gratefully acknowledge the contributions of his former students Dr. Keon-Young Yun and Jose Vadillo, who have performed some of the computations reported in this chapter.

References

Agarwal, R.K. (2001) "Lattice Boltzmann Simulation of Magnetohydrodynamic Flows," in *Proc. Int. Conf. on Computational Fluid Dynamics* (ICCFD1), Satofuka, N., Editor, Springer-Verlag, Berlin, p. 511.

Agarwal, R.K. (2005) "Lattice Boltzmann Simulations of Magnetohydrodynamic Slip Flow in Microchannels," AIAA Paper 2005-0163, 43rd AIAA Aerospace Sciences Meeting & Exhibit, Reno, NV, 10–13 January 2005.

Beskok, A., and Karniadakis, G.E. (1999) "A Model for Flows in Channels, Pipes and Ducts at Micro and Nanoscales," *J. Microscale Thermophys. Eng*. **3**, p. 43.

Chen, S., and Doolen, G.D. (1998) "Lattice Boltzmann Method for Fluid Flows," *Annu. Rev. Fluid Mech.* **30**, p. 329.

Chen, H., Chen, S., and Matthaeus, W.H. (1992) "Recovery of Navier-Stokes Equations Using a Lattice-Gas Boltzmann Method, *Phys. Rev. A*, **45**, p. R 5339.

Chen, H., Matthaeus, W.H., and Klein, L.W. (1988) "An Analysis Theory and Formulation of a Local Magnetohydrodynamic Lattice Gas Model, *Phys. Fluids* **31**, p. 1439.

Croisille, J.-P., Khanfir, R., and Chanteur, G. (1995) "Numerical Simulation of MHD Equations by a Kinetic-Type Method*, J. Sci. Comput*. **10**, p. 81.

Deller, P.J. (2002) "Lattice Kinetic Schemes for Magnetohydrodynamics," *J. Comp. Phys*. **179**, p. 95.

Frisch, U., Hasslacher, B., and Pomeau,Y. (1986) "Lattice-Gas Automata for the Navier-Stokes Equations," *Phys. Rev. Lett*. **56**, p. 1505.

Lim, C.Y., Shu, C., Niu, X.D., and Chew, Y.T. (2002) "Application of Lattice Boltzmann Method to Simulate Microchannel Flows," *Phys. Fluids* **14**, p. 2299.

Maxwell, J.C. (1879) "On Stresses in Rarefied Gases Arising from Inequalities of Temperature, *Phil. Trans. Royal Soc., London* **170**, p. 231.

Qian, Y.H., D'Humieres, D., and Lallemand, P. (1992) "Lattice BGK Models for Navier–Stokes Equations," *Europhys. Lett*. **17**, p. 479.

10

Liquid Flows in Microchannels

Kendra V. Sharp
Pennsylvania State University

Ronald J. Adrian
Arizona State University

Juan G. Santiago
Stanford University

Joshua I. Molho
Caliper Life Sciences Incorporated

10.1 Introduction

Nominally, microchannels can be defined as channels whose dimensions are less than 1 millimeter and greater than 1 micron. Above 1 millimeter the flow exhibits behavior that is the same as most macroscopic flows. Currently, microchannels have characteristic dimensions anywhere from the submicron scale to hundreds of microns. Microchannels can be fabricated in many materials — glass, polymers, silicon, metals — using various processes including surface micromachining, bulk micromachining, molding, embossing, and conventional machining with microcutters. These methods and the characteristics of the resulting flow channels are discussed elsewhere in this handbook.

Microchannels offer advantages due to their high surface-to-volume ratio and their small volumes. The large surface-to-volume ratio leads to high rate of heat and mass transfer, making microdevices excellent tools for compact heat exchangers. For example, the device in Figure 10.1 is a cross-flow heat exchanger constructed from a stack of 50 14 mm × 14 mm foils, each containing 34 200 µm wide by 100 µm deep

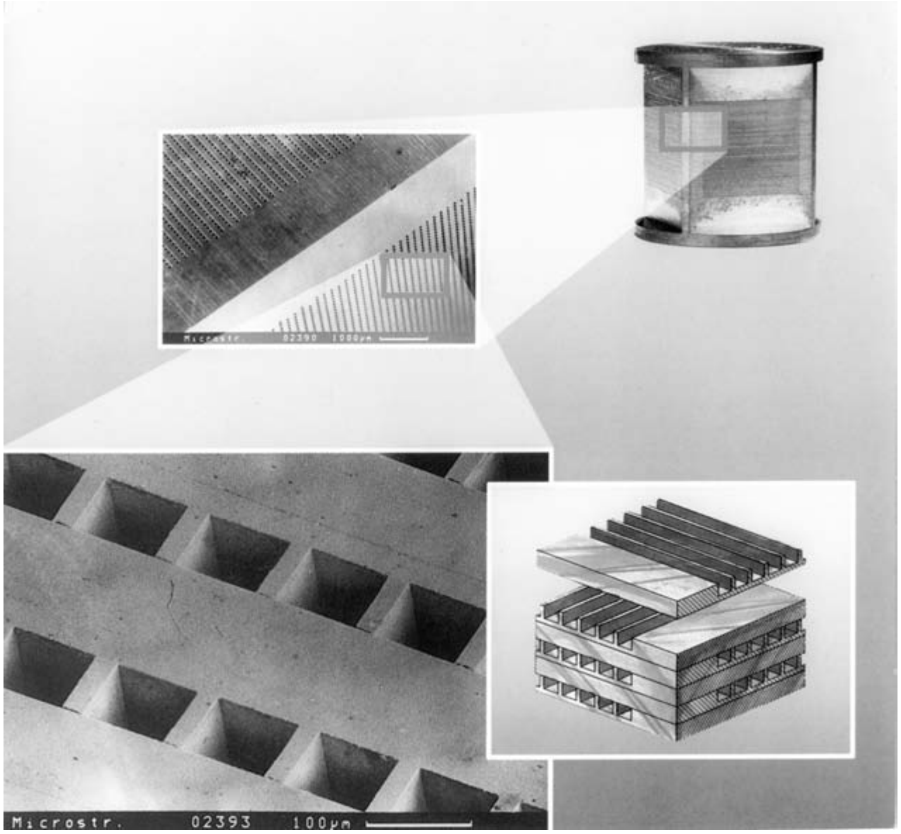


FIGURE 10.1 Micro heat exchanger constructed from rectangular channels machined in metal. (Reprinted with permission from K. Schubert and D. Cacuci, Forschungszentrum, Karlsruhe.)

channels machined into the 200 µm thick stainless steel foils by the process of direct, high-precision mechanical micromachining [Brandner et al., 2000; Schaller et al., 1999]. The direction of flow in adjacent foils is alternated 90°, and the foils are attached by means of diffusion bonding to create a stack of cross-flow heat exchangers capable of transferring 10 kW at a temperature difference of 80 K using water flowing at 750 kg/hr. The impressively high rate of heat transfer is accomplished mainly by the large surface area covered by the interior of the microchannel: approximately 3,600 mm^2 packed into a 14 mm cube.

A second example of the application of microchannels is in the area of MEMS devices for biological and chemical analysis. The primary advantage of microscale devices in these applications are the good match with the scale of biological structures and the potential for placing multiple functions for chemical analysis on a small area; that is, the concept of a chemistry laboratory on a chip.

Microchannels are used to transport biological materials such as (in order of size) proteins, DNA, cells, and embryos or to transport chemical samples and analytes. Typical of such devices is the i-STAT blood sample analysis cartridge shown in Figure 10.2. The sample is taken onboard the chip through a port and moved through the microchannels by pressure to various sites where it is mixed with analyte and moved to a different site where the output is read. Flows in biological devices and chemical analysis microdevices are usually much slower than those in heat transfer and chemical reactor microdevices.

10.1.1 Unique Aspects of Liquids in Microchannels

Flows in microscale devices differ from their macroscopic counterparts for two reasons: the small scale makes molecular effects such as wall slip more important, and it amplifies the magnitudes of certain ordinary continuum effects to extreme levels. Consider, for example, strain rate and shear rate, which scale in proportion to the velocity scale U_s and inverse proportion to the length scale L_s. Thus, 100 mm/sec flow in a 10 µm channel experiences a shear rate of the order of 10^4 sec^{-1}. Acceleration scales as U_s^2/L_s and is

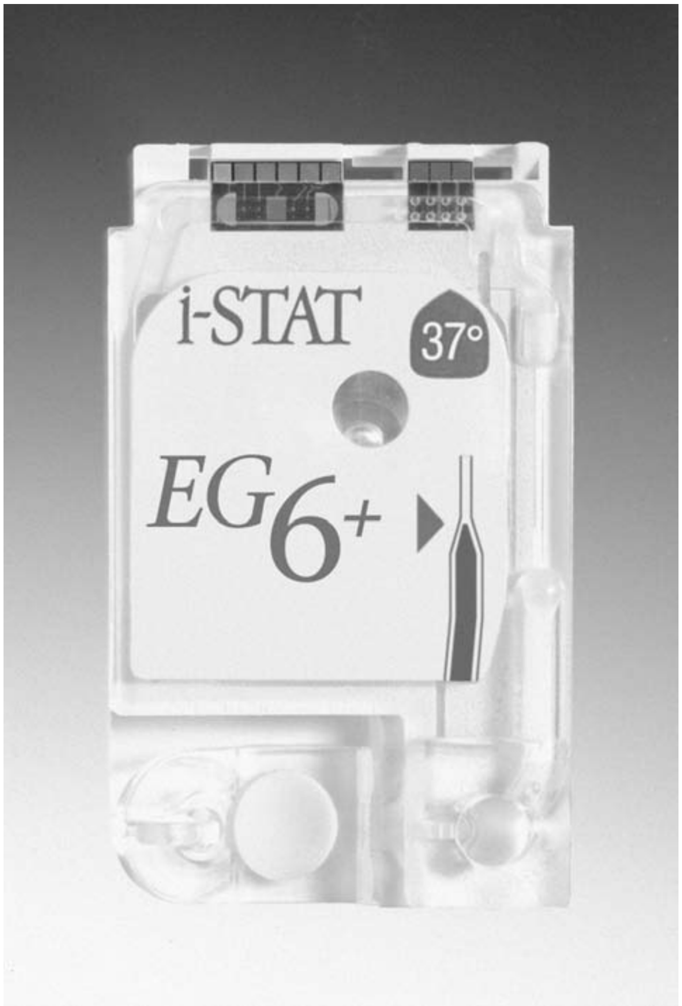


FIGURE 10.2 (**See color insert following page 10-34.**) Blood sample cartridge using microfluidic channels. (Reprinted with permission from i-Stat, East Windsor, NJ, 2000.)

similarly enhanced. The effect is even more dramatic if one tries to maintain the same volume flux while scaling down. The flux scales as $Q \sim U_s L_s^2$, so at constant flux $U_s \sim L_s^{-2}$ and both shear and acceleration go as L_s^{-3}. Fluids that are Newtonian at ordinary rates of shear and extension can become non-Newtonian at very high rates. The pressure gradient becomes especially large in small cross section channels. For fixed volume flux, the pressure gradient increases as L_s^{-4}.

Electrokinetic effects occur at the interface between liquids and solids such as glass due to chemical interaction. The result is an electrically charged double layer that induces a charge distribution in a very thin layer of fluid close to the wall. Application of an electric field to this layer creates a body force capable of moving the fluid as if it were slipping over the wall. The electroosmotic effect and the electrophoretic effect (charges around particles) will be discussed in detail in a later section. Neither occurs in gases.

The effects of molecular structure are quite different in gases and liquids. If the Knudsen number (defined as $\text{Kn} = \lambda/L_s$, where λ is the mean free path in a gas and L_s is the characteristic channel dimension) is greater than 10^{-3} [Janson et al., 1999, Gad-el-Hak, 1999], nonequilibrium effects may start to occur. Modified slip boundary conditions can be used in continuum models for Knudsen numbers between 10^{-1} and 10^{-3} [Gad-el-Hak, 1999]. As the Knudsen number continues to increase, continuum assumptions and fluid theory are no longer applicable. Analysis of such flow requires consideration of different physical phenomena (see the chapters on Analytical and Computational Models for Microscale Flows in this book, Gad-el-Hak, 1999, Janson et al., 1999, Arkilic et al. 1997, and Harley et al., 1995).

Because the density of liquids is about 1000 times the density of gases, the spacing between molecules in liquids is approximately 10 times less than the spacing in gases. Liquid molecules do not have a mean free path, but following Bridgman (1923), the lattice spacing δ may be used as a similar measure. The lattice spacing δ is defined as [Probstein, 1994]

$$\delta \sim \left(\frac{\overline{V}_1}{N_A}\right)^{1/3}, \tag{10.1}$$

where $\overline{V}_1$ is the molar volume and N_A is Avogadro's number. For water, this spacing is 0.3 nm. In a 1 μm gap and a 50 μm diameter channel, the equivalent Knudsen numbers are 3×10^{-4} and 6×10^{-6} respectively, well within the range of obeying continuum flow. In gases, effects such as slip at the wall occur when the mean free path length of the molecules is more than about one-tenth the flow dimension (i.e., flow dimensions of order less than 650 nm in air at STP). (Note that the mean free path length of a gas is longer than the mean spacing between its molecules; see the chapter Flow-Physics by Gad-el-Hak in this book for a detailed discussion.) In liquids this condition will not occur unless the channels are smaller than approximately 3 nm, and continuum hydrodynamics may provide a useful description at scales even smaller than this because the forces of interaction between molecules in liquids are long range. For example, Stokes' classical result for drag on a sphere is routinely applied to particles whose diameters are well below 100 nm. Thus, liquid flow in micro devices should be described adequately by continuum hydrodynamics well below dimensions of one micron.

Molecular effects in liquids are difficult to predict because the transport theory is less well developed than the kinetic theory of gases. For this reason, studies of liquid microflows in which molecular effects may play a role are much more convincing if done experimentally.

Liquids are generally considered incompressible. Consequently, the density of a liquid in microchannel flow remains very nearly constant as a function of distance along the channel, despite the very large pressure gradients that characterize microscale flow. This behavior greatly simplifies the analysis of liquid flows relative to gas flows, wherein the large pressure drop in a channel leads to large expansion and large heat capacity.

The large heat capacity of liquids relative to gases implies that the effects of internal heating due to viscous dissipation are much less significant in liquid flows. The pressure drop in microchannel flow can be very large, and since all of the work of the pressure difference against the mean flow ultimately goes into viscous dissipation, effects due to internal heating by viscous dissipation may be significant. However they will be substantially lower in liquids than in gases, and they can often be ignored allowing one to treat the liquid as a constant density, constant property fluid.

The dynamic viscosity μ of a liquid is larger than that of a gas by a factor of about 100 (c.f., Table 10.1). This implies much higher resistance to flow through the channels. The kinematic viscosity of a liquid is typically much less than the kinematic viscosity of a gas, owing to the much higher density of liquids (c.f. Table 10.1) qualitatively to the thermal conductivity and the thermal diffusivity.

Liquids in contact with solids or gases have surface tension in the interface. At the microscale, the surface tension force becomes one of the most important forces, far exceeding body forces such as gravity and electrostatic fields.

TABLE 10.1 Dynamic and Kinematic Viscosities of Typical Liquids Compared to Air at 1 Atmosphere

Fluid	Dynamic Viscosity μ [gm/cm-s]	Kinematic Viscosity ν [cm^2/s]	Thermal Conductivity k [J/K s cm]	Thermal Diffusivity κ [cm^2/s]
Water @15°C	0.0114	0.0114	0.0059	0.00140
Ethyl Alcohol @ 15°C	0.0134	0.0170	0.00183	0.00099
Glycerin @15°C	23.3	18.50	0.0029	0.00098
Air @15°C	0.000178	0.145	0.000253	0.202

Bubbles can occur in liquids for good or ill. Unwanted bubbles can block channels or substantially alter the flow. But bubbles can also be used to apply pressure and to perform pumping by heating and cooling the gases inside the bubble.

Particulates and droplets suspended in liquids have densities that match those of liquids more closely. Settling is much less rapid in liquids, and suspensions have the ability to follow the accelerations of the flow. This effect can also keep suspended impurities in suspension for much longer, thereby increasing the probability that an impurity will introduce unwanted behavior.

Liquids can interact with solids to form an electric double layer at the interface. This is the basis for the phenomena of electroosmosis and electrophoresis, both of which can be used to move fluid and particles in channels. These topics will be discussed in detail in a later section. Liquids can be non-Newtonian especially at the high shear rates encountered in microchannels.

10.1.2 Continuum Hydrodynamics of Pressure-Driven Flow in Channels

The general continuum description of the flow of an incompressible, Newtonian fluid flow with variable properties and no body forces other than gravity (i.e., no electrical forces) consists of the incompressible continuity equation

$$\frac{\partial u_j}{\partial x_j} = 0, \tag{10.2}$$

and the momentum equation

$$\rho\left(\frac{\partial u_i}{\partial t} + u_j\frac{\partial u_i}{\partial x_j}\right) = \frac{\partial \tau_{ij}}{\partial x_j} + \rho b_i, \tag{10.3}$$

where the fluid stress is given by Stokes' law of viscosity

$$\tau_{ij} = -p\delta_{ij} + \mu\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)\cdot \tag{10.4}$$

Here u_i is the *ith* component of the velocity vector $\mathbf{u}(\mathbf{x}, t)$; ρ is the mass density [kg/m^3]; b_i is the body force per unit mass m/s^2 (often $b_i = g_i$, the gravitational acceleration), and τ_{ij} is the stress tensor N/m^2. The corresponding enthalpy equation is

$$\rho c_p\left(\frac{\partial T}{\partial t} + u_j\frac{\partial T}{\partial x_j}\right) = -\frac{\partial q_j}{\partial x_j} + \Phi, \tag{10.5}$$

where T is the temperature, and q is the heat flux J/s m^2 given by Fourier's law of heat conduction by molecular diffusion k,

$$q_i = -k\frac{\partial T}{\partial x_i}\cdot \tag{10.6}$$

The rate of conversion of mechanical energy into heat due to internal viscous heating, is

$$\Phi = \mu\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)\frac{\partial u_i}{\partial x_j}\cdot \tag{10.7}$$

Consider a long parallel duct or channel with the x-direction along the axis of the channel and the coordinates y and z in the plane perpendicular to the axis of the channel (Figure 10.3). The entering flow undergoes a transient response in which the velocity and temperature profiles change in the streamwise direction. This process continues until the flow properties become independent of the streamwise position. In this state of *fully developed velocity profile*, the velocity field is unidirectional, $u(x) = [u(y, z), 0, 0]$, and there is no acceleration of the fluid. Thus, for fully developed flow with gravitational body force g the equations become very simply

$$\rho\frac{\partial u}{\partial t} = -\frac{dp}{dx} + \rho g_x + \frac{\partial}{\partial y}\left(\mu\frac{\partial u}{\partial y}\right) + \frac{\partial}{\partial z}\left(\mu\frac{\partial u}{\partial z}\right) \tag{10.8}$$

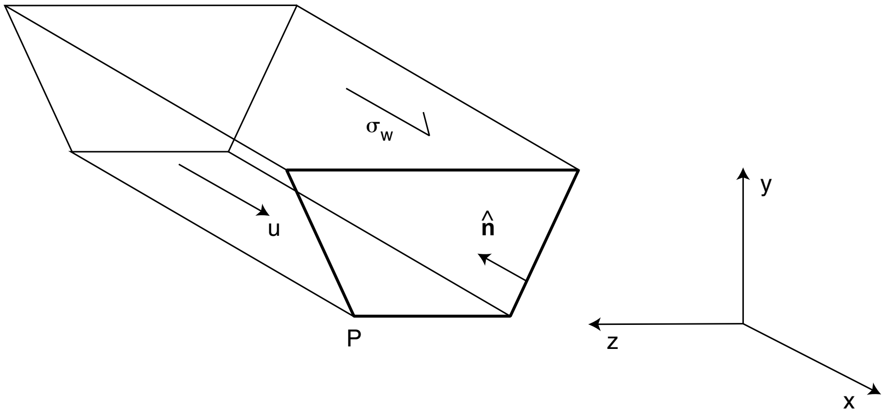


FIGURE 10.3 Flow in a duct of arbitrary cross-section A. P is the perimeter and τ_w is the wall shear stress.

$$\rho c_p \frac{\partial T}{\partial t} = \frac{\partial}{\partial y}\left(k\frac{\partial T}{\partial y}\right) + \frac{\partial}{\partial z}\left(k\frac{\partial T}{\partial z}\right) + \Phi. \tag{10.9}$$

Lastly, if the flow is steady and the temperature and properties are constant, then the equation for streamwise velocity profiles becomes a simple Poisson equation

$$\frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = \frac{1}{\mu}\frac{d}{dx}(p - \rho g_x x). \tag{10.10}$$

In the absence of electrokinetic effects and for shear rates less than about $10^{12}\,\mathrm{s}^{-1}$ the appropriate boundary condition is the no-slip condition

$$u = 0 \quad \textit{on the boundary P}. \tag{10.11}$$

10.1.3 Hydraulic Diameter

Control volume analysis of fully developed flow leads naturally to the concept of the *hydraulic diameter*. Figure 10.3 shows flow in a duct of arbitrary cross-section. Since the flow is fully developed and unidirectional (assuming a straight duct), the acceleration is zero and control volume analysis of the momentum reduces to a simple force balance in the streamwise direction,

$$-\frac{dp}{dx}A = \overline{\tau}_w P \tag{10.12}$$

wherein

$$\overline{\tau}_w = \frac{1}{P}\oint_P \tau_w \, dl \tag{10.13}$$

is the wall shear stress averaged around the perimeter, and the local wall shear stress is given by

$$\tau_w = \mu \left.\frac{\partial u}{\partial n}\right|_{n=0}. \tag{10.14}$$

Equation (10.12) displays the relevance of the ratio of the area A to the perimeter P. In practice, the hydraulic diameter is defined to be

$$D_h = \frac{4A}{P} \tag{10.15}$$

so that, when the cross-section is a circle, D_h equals its diameter. The hydraulic diameter provides a convenient way to characterize a duct with a single length scale and a basis for comparison between ducts of different shapes. A common approximation is to also estimate the flow resistance in a duct or channel as the resistance of a round duct whose diameter is equal to the hydraulic diameter. This approximation is useful but subject to errors of order 10–20%. Since solution of Poisson's equation to obtain the exact wall shear stress is accomplished readily by numerical means, the approximation is not necessary.

10.1.4 Flow in Round Capillaries

Flow in a round tube is the archetype for all duct and channel flows. While microfabrication characteristically yields channels of noncircular cross-section, the round cross-section is a useful and familiar point of reference, and microcapillaries are not uncommon. Extensive macroscale research on pipe flows dates back to Hagen's (1839), Poiseuille's (1841), and Reynolds' (1883) original studies in the 19th century. Independently, both Hagen (1839) and Poiseuille (1841) observed the relation between pressure head and velocity and its inverse proportionality to the fourth power of tube diameter.

In a round capillary of radius $a = D/2$ and radial coordinate r, it is well known that the velocity profile across a diameter is parabolic

$$u = u_{\max}\left(1 - \frac{r^2}{a^2}\right) \tag{10.16}$$

where the maximum velocity is given by

$$u_{\max} = \frac{a^2}{4\mu}\left(-\frac{dp}{dx}\right). \tag{10.17}$$

The volume flow rate Q is given by:

$$Q = \overline{U}A \tag{10.18}$$

where the average velocity $\overline{U}$ defined by

$$\overline{U} = \frac{1}{\pi a^2}\int_0^a u(r)2\pi r\, dr \tag{10.19}$$

is numerically equal to

$$\overline{U} = \frac{1}{2}u_{\max}. \tag{10.20}$$

Using these relations it is easily shown that the pressure drop in a length L, $\Delta p = (-dp/dx)L$, is given by

$$\Delta p = \frac{8\mu LQ}{\pi a^4}. \tag{10.21}$$

The Darcy friction factor f is defined so that

$$\Delta p = f\frac{L}{D}\rho\frac{\overline{U}^2}{2} \tag{10.22}$$

(The Fanning friction factor is one-fourth of the Darcy friction factor). The Reynolds number is defined in terms of a characteristic length scale[1] L_s by

$$\text{Re} = \frac{\rho\overline{U}L_s}{\mu}. \tag{10.23}$$

For a round pipe, the characteristic length scale is the diameter of the pipe D. The friction factor for laminar flow in a round capillary is given by

$$f = \frac{64}{\text{Re}}. \tag{10.24}$$

[1] In the remainder of this chapter, the characteristic length scale used in calculating Re is to be inferred from context, e.g., generally D_h for a rectangular channel and D for a circular tube.

The Poiseuille number is sometimes used to describe flow resistance in ducts of arbitrary cross-section. It is defined by

$$\begin{aligned} P_O &= f\,\mathrm{Re}/4 \\ &= -\frac{1}{\mu}\frac{dp}{dx}\frac{D_h^2}{2\overline{U}}, \end{aligned} \tag{10.25}$$

where L_s used in the calculation of Re is D_h. The Poiseuille number has a value of 16 for a round capillary.

The inverse relationship between friction factor and Reynolds number has been well documented on the macroscale. It means that the pressure drop is linearly proportional to the flow rate, Q. In the laminar region there is no dependence on surface roughness.

The pressure drops due to pressure-driven flow in microchannels are quite large. For example, water (nominally, $\mu = 10^{-3}$ kg/m-s) flowing at $Q = 0.01$ cc/sec in a $D = 100$ micron diameter, $L = 10$ mm long tube creates a pressure drop of $\Delta p = 40.7\,\mathrm{kN/m^2}$. Under these conditions the mean velocity is $1.27\,\mathrm{msec^{-1}}$, and the Reynolds number is Re = 127. If the tube diameter is reduced to 10 microns keeping all other factors constant, the mean velocity is $127\,\mathrm{msec^{-1}}$, the Reynolds number is Re = 1270, and the pressure drop increases to $407\,\mathrm{MN/m^2}$, or 4070 atmospheres.

As the Reynolds number increases above 2000 in a circular duct, the flow begins to transition to turbulence. At this point, the friction factor increases dramatically, and the flow resistance ultimately becomes proportional to Q^2 rather than Q.

10.1.5 Entrance Length Development

Before the flow reaches the state of a fully developed velocity profile, it must transition from the profile of the velocity at the entrance to the microduct, whatever that is, to the fully developed limit. This transition occurs in the *entrance length* (L_e) of the duct. In this region the flow looks like a boundary layer that grows as it progresses downstream. Ultimately, the viscously retarded layers meet in the center of the duct at the end of the entrance length.

The pressure drop from the beginning of the duct to a location x is given by

$$p_0 - p(x) = \left(f\frac{x}{D_h} + K(x)\right)\frac{\rho\overline{U}^2}{2} \tag{10.26}$$

wherein *K(x)* is the pressure-drop parameter given in Figure 10.4 for a circular duct and for parallel plates [White, 1991]. The flow development is largely completed by *x/D* = 0.065 Re.

10.1.6 Transition to Turbulent Flow

In 1883, Reynolds found a critical value of velocity, u_{crit}, above which the form of the flow resistance changes. The corresponding dimensionless parameter is the critical Reynolds number, $\mathrm{Re}_{\mathrm{crit}}$, below which disturbances in the flow are not maintained. Such disturbances may be caused by inlet conditions like a sharp edge or unsteadiness in the flow source. Depending on the Reynolds number, disturbances may also be introduced by natural transition to turbulent flow.

Reynolds found $\mathrm{Re}_{\mathrm{crit}}$ to be approximately 2000, and this value has been generally accepted. Once the flow is fully turbulent, the empirical relationship often used to correlate friction factor and Reynolds number for smooth pipes and initially proposed by Blasius is

$$f \approx 0.3164\,\mathrm{Re}^{-0.25}. \tag{10.27}$$

For rough pipes, the friction factor departs from the Blasius relation in the turbulent region. This departure occurs at different values of Re depending on the magnitude of the surface roughness. The Moody chart summarizes the traditional friction factor curves and is readily available in any basic fluids textbook [e.g., White, p. 318, 1994].

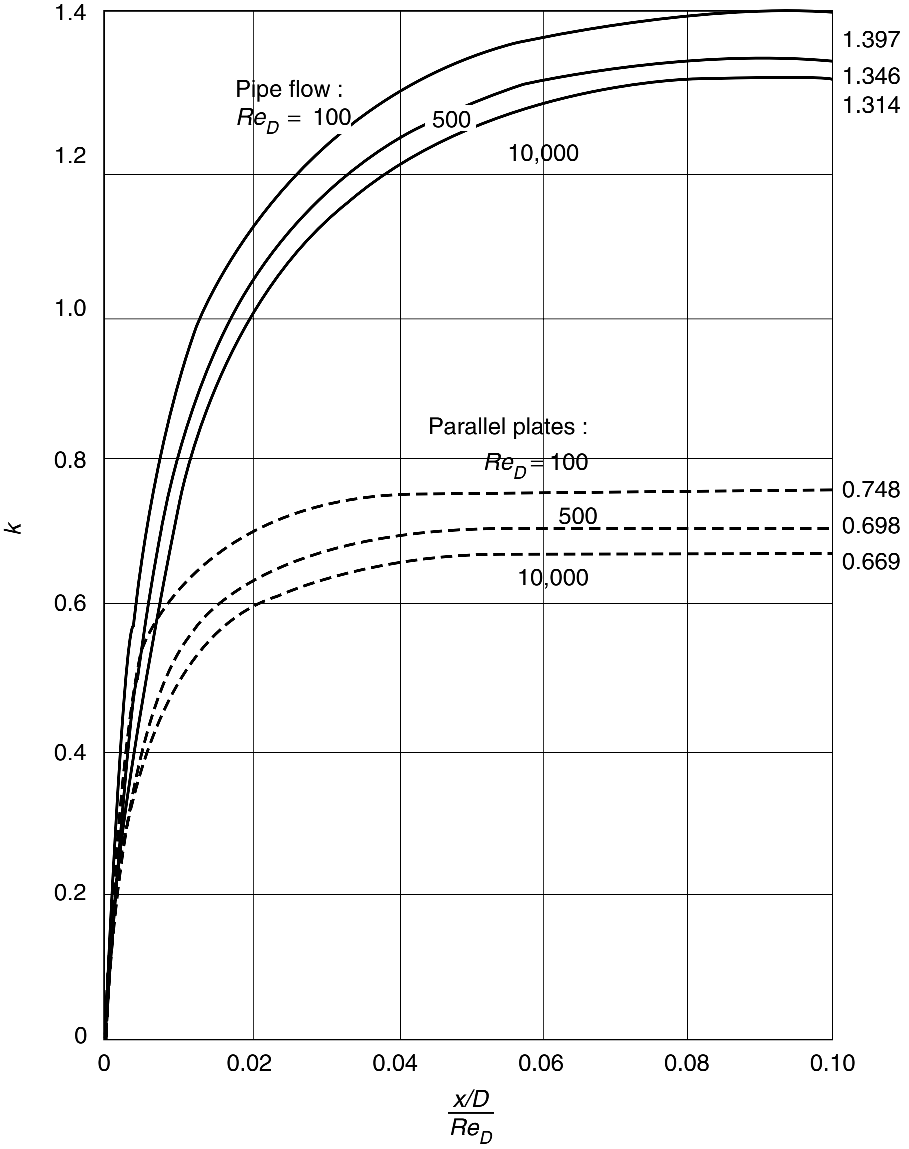


FIGURE 10.4 Entrance length parameter *K* for laminar flow in the inlet of a duct. (Reprinted with permission from White, F.M. [1991] *Viscous Flow*, 2nd ed., p. 292, McGraw-Hill, New York.).

10.1.7 Noncircular Channels

Microfluidic channels are generally formed by micromachining open channels on a planar substrate and closing the channels by covering the substrate with a thin plate, such as a microscope slide or cover slip. The method of attaching the plate to the substrate must be very strong for pressure-driven flows because the pressure gradients, and hence the maximum pressures, can be very large. As an example, microchannels are often cast into the surface of blocks of polydimethylsiloxane (PDMS), a transparent flexible polymer commercially known as Sylgard. Closed channels are formed using a cover glass slip bonded to the PDMS surface by oxidizing both the surface of the PDMS replica and the glass by oxygen plasma treatment (70 W, 85 mTorr for 20 sec). When the two oxidized surfaces are brought into contact they bond covalently, creating a seal that can withstand up to 5 bars. Since the surfaces of the glass and the PDMS are each hydrophilic, filling the channels with aqueous liquids is relatively easy.

Most microfluidic channels have noncircular cross-sections whose shape is associated with the method of fabrication. Isotropic etching in glass or silicon produces cross-sections that are anywhere from semicircular to rectangular with rounded corners in the bottom. Anisotropic etching in Si creates shapes defined by the crystallographic planes. A common case is Si with its <100> plane (Miller index) coincident with the planar surface to be machined. The <111> planes are inclined at 54.74° so that anisotropic etching creates either trapezoidal cross-sections with slanted sidewalls or triangular cross-sections. Laser machining of polymers creates roughly semicircular channels, while molding PDMS creates rectangular channels with slightly rounded corners. The various types of microfabrication and their characteristics are discussed in great detail throughout this handbook.

Fully developed flow in noncircular ducts is found by solving the Poisson Equation (10.10). Frequently, analytical solutions can also be found, but the numerical approach is so reliable that there is little need for exact solutions. Developing flow in the entrance region is more difficult, but here again numerical approaches are relatively straightforward. Table 10.2 summarizes the flow resistance for various laminar flows. One sees that the effect of the shape of the channel is relatively weak.

As mentioned earlier, a common approximation made in analyzing flow in ducts of noncircular cross-section is to use the results for circular ducts but replacing the hydraulic diameter of the noncircular duct with that of the round duct. For example, this can be done to estimate the flow resistance of fully developed flow and the resistance in the entrance region.

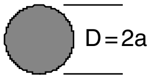
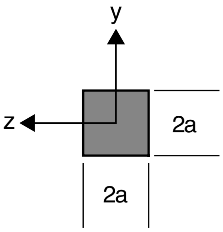
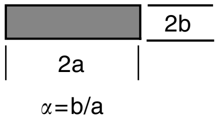
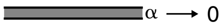
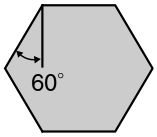
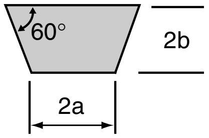
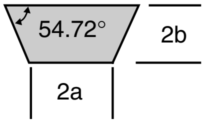
10.1.8 Experimental Studies of Flow-Through Microchannels

Despite the fundamental simplicity of laminar flow in straight ducts, experimental studies of microscale flow have often failed to reveal the expected relationship between friction factor and Reynolds number. The frictional resistance of the flow has been reported, under *certain* conditions, to be *consistent* with predictions based on conventional macroscale Hagen-Poiseuille theory [Celata et al., 2002; Flockhart and Dhariwal, 1998; Jiang et al., 1995; Judy et al., 2002; Li et al., 2003; Liu and Garimella, 2004; Phares and Smedley, 2004; Sharp and Adrian, 2004; Wilding et al., 1994; Wu and Little, 1983], *increased* as compared to conventional macroscale predictions [Brutin and Tadrist, 2004; Celata et al., 2002; Cui et al., 2004; Hsieh et al.; 2004; Li et al., 2003; Mala and Li, 1999; Papautsky et al., 1999a, 1999b; Peng et al., 1994; Pfund et al., 2000; Phares and Smedley, 2004; Qu et al., 2000; Ren et al., 2001; Wu and Little, 1983] and *decreased* as compared to conventional macroscale predictions [Choi et al., 1991; Peng et al., 1994; Pfahler et al., 1990a, 1990b, 1991; Yu et al., 1995].

A brief summary is presented herein; for detailed historical summaries of the experiments that have been conducted to investigate the behavior of fluid flow in microchannels, see the recent reviews of microchannel fluid flows in both tabular [Sobhan and Garimella, 2001] and text format [Koo and Kleinstreuer, 2003; Obot, 2002]. Flow resistance experiments in microscale channels or tubes have been conducted over a large range of Reynolds numbers, geometries, and experimental conditions, and in the subsequent discussion of results, they will be grouped according to the results of friction factor measurements (follows macroscale predictions, higher than predictions, and lower than predictions).

The first experimental investigations of flow through microchannels in the early 1980s were motivated by the interest in high-performance heat sinking. The large surface-to-volume ratios of microchannels make them excellent candidates for efficient heat transfer devices. Tuckerman and Pease (1981) studied flow through an array of microchannels with approximately rectangular cross-sections (height range 50–56 μm, width range 287–320 μm). Although their study focused primarily on heat transfer characteristics, they "confirmed that the flow rate obeyed Poiseuille's equation." Shortly thereafter, a study of microchannels for use in small Joule-Thomson refrigerators was performed [Wu and Little, 1983]. Significant roughnesses were present in some of these etched silicon or glass channels, but friction factors measured in the smoothest channel showed reasonable agreement with theoretical macroscale predictions. A number of other experiments also have shown general agreement with the macroscale theoretical predictions for friction factor in the flow of a Newtonian fluid in at least certain parameter ranges in circular microtubes [Celata et al., 2002; Jiang et al., 1995; Judy et al., 2002; Li et al., 2003; Phares and Smedley, 2004; Sharp and Adrian, 2004], rectangular microchannels [Judy et al., 2002; Liu and Garimella, 2004], and channels with other cross-sectional shapes including the trapezoidal cross-section commonly

TABLE 10.2 Resistance to Flow in Fully Developed Flow-Through Straight Microchannels of Various Cross-Sectional Geometry

Cross Section	fRe	$u_{\max}/u_B$
	64	2.000
	56.92	2.0962
	$96[1 - 1.3553\alpha + 1.9467\alpha^2 - 1.7012\alpha^3 + 0.9564\alpha^4 - 0.2537\alpha^5]$	—
	96	1.5000
	60	—
	2b/2a	
	4.000 55.66	2.181
	2.000 55.22	2.162
	1.000 56.60	2.119
	0.500 62.77	1.969
	0.250 72.20	1.766
	1.000 56.15	2.137

Data from Shah, R.K., and London, A.L. (1978) *Laminar Flow Forced Convection in Ducts,* Adv. in Heat Transfer series, Supp. 1, Academic Press, New York.

encountered in microfluidic applications due to anisotropic etching in the fabrication process [Flockhart and Dhariwal, 1998; Wilding et al., 1994].

In circular fused silica microchannels with diameters from approximately 50 to 250 μm and Reynolds numbers less than 1800, the results of more than 1500 measurements of pressure drop versus flow rate confirm agreement between macroscale Poiseuille theory and microscale measurements of the friction factor to within −1% systematic and ±2.5% rms random error [Sharp and Adrian, 2004]. Similar agreement was also obtained using a 20% solution of glycerol and 1-propanol. Good agreement between conventional Poiseuille theory and experimental results has been reported for other microscale flows including: smooth circular microtubes with diameters of 80 to 200 microns [Li et al., 2003]; circular

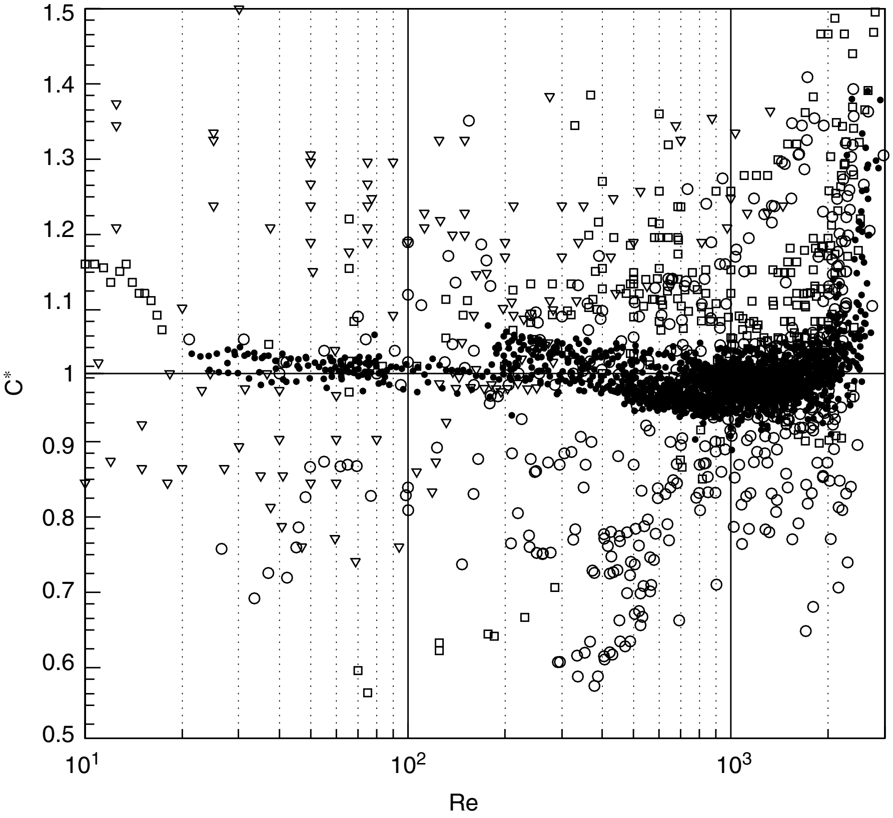


FIGURE 10.5 Comparison of C* vs. Reynolds number in the literature. Symbols indicate geometry of channel and the following data are shown: (·) circular microtubes Sharp and Adrian (2004); (○) circular microtubes Yu et al. (1995), Choi et al. (1991), Judy et al. (2002), Mala and Li (1999); (▽) trapezoidal microchannels: Pfahler et al. (1991), Flockhart and Dhariwal (1998), Wilding et al. (1994), Qu et al. (2000); (□) rectangular microchannels: Pfahler et al. (1991), Pfahler et al. (1990b), Papautsky et al. (1999a), Pfund et al. (2000), Celata et al. (2002), Liu and Garimella (2004), Hsieh et al. (2004).

microtubes with diameters of 130 microns and Reynolds numbers less than 600 [Celata et al., 2002]; circular and square microtubes with diameter or hydraulic diameter of 15 to 150 microns and Reynolds numbers of 8–2300 [Judy et al., 2002]; smooth circular microtubes with diameters of 119 and 152 microns [Phares and Smedley, 2004]; and rectangular channels with hydraulic diameters from 244–974 microns [Liu and Garimella, 2004].

An increase in the frictional resistance of liquid flows in microchannels over theoretical predictions based on conventional macroscale theory has been reported in some studies [Brutin and Tadrist, 2004; Celata et al., 2002; Cui et al., 2004; Hsieh et al., 2004; Li et al., 2003; Mala and Li, 1999; Papautsky et al., 1999a, 1999b; Peng et al., 1994; Pfund et al., 2000; Phares and Smedley, 2004; Qu et al., 2000; Ren et al., 2001; Wu and Little, 1983], including increases of as much as 38% [Qu et al., 2002], 37% [Li et al., 2003], and 27% [Brutin and Tadrist, 2003] over conventional Poiseuille theoretical predictions. Another group of studies found the flow resistance to be less than theoretical macroscale predictions for certain conditions [Choi et al., 1991; Peng et al., 1994; Pfahler et al., 1990a, 1990b, 1991; and Yu et al., 1995].

To aid in comparing the results of these studies, a normalized friction factor C^* is defined as

$$C^* = \frac{(f\,\mathrm{Re})_{\mathrm{experimental}}}{(f\,\mathrm{Re})_{\mathrm{theoretical}}}. \tag{10.28}$$

The wide variability of results is illustrated in Figure 10.5. There is also wide variability in experimental conditions, microchannel geometries, and methodology. The inconsistencies demonstrate the need for both detailed velocity measurements and careful study of potential microscale effects such as surface roughness or electrical effects in order to conclusively understand the flow behavior in microscale channels.

10.1.9 Proposed Explanations for Measured Behavior

Thus far the explanations offered in the literature for anomalous behavior of friction factor and flow resistance in microchannels include surface/roughness effects and electrical charge, variations in viscosity, "early" transition to turbulence, entrance effects, inaccuracies in measuring channel dimensions, microrotational effects of individual fluid molecules, and geometry effects.

The increase in frictional resistance has often been reasonably linked to surface roughness. For example, in some studies [Li et al., 2003; Phares and Smedley, 2004; Wu and Little, 1983], experimental results on frictional resistance agreed well with Hagen-Poiseuille theory for smooth channels, but significant deviations were reported for flows through similar microchannels or tubes with increased surface roughnesses.

In macroscale theory, the surface roughness does not affect the flow resistance relationships in the laminar region [White, 1994]. Flow resistance results in microscale geometries have shown both a strong increase due to roughness [Li et al., 2003; Phares and Smedley, 2004; Wu and Little, 1983] and no effect due to roughness [Choi et al., 1991].

In terms of viscosity effects, a roughness viscosity model (RVM) has been proposed [Mala and Li, 1999, based on work by Merkle et al., 1974]. Assuming that surface roughness increases the momentum transfer near the wall, the roughness viscosity μ_r as a function of r is proposed to be higher near the wall and proportional to the Reynolds number [Mala and Li, 1999]. Implementing this roughness-viscosity model for water flowing through trapezoidal channels, reasonable agreement with model prediction and experimental results was found in most cases, but the model did not accurately depict the increased slope in the relationship between pressure drop and Reynolds number observed in the same experiments for $\mathrm{Re} \geqslant 500$ [Qu et al., 2000]. Direct measurement of viscosity in very thin layers, or thin films, was performed by Israelachvili (1986). The viscosity of water was found to retain its bulk viscosity value to within 10% even in a film as thin as 5 nm. Concentrated and dilute NaCl/KCl solutions were also tested to assess the impact of double-layer forces on the value of viscosity near a surface. The viscosity of these dilute NaCl/KCl solutions remained only minimally affected until the last molecular layer near the wall. Based on these measurements, the viscosity of fluid in the wall region is not expected to vary significantly from the bulk value even in the presence of possible charging effects, somewhat contrary to the proposed explanations given by Mala and Li (1999) and Qu et al. (2000).

Other changes in viscosity are suggested to occur for liquids under extremely high pressure. A decrease in experimental flow rate for isopropanol and carbon tetrachloride as compared to conventional Hagen-Poiseuille theory has been reported and attributed to viscosity changes in isopropanol and carbon tetrachloride at pressures greater than 10 MPa, but in similar experiments any effects of high pressure on the viscosity of water could not be conclusively established [Cui et al., 2004]. It is also possible that the very high shear rates in these microchannels cause normally Newtonian fluids to behave in a non-Newtonian fashion. The shear rates in Sharp and Adrian (2004) were as high as $7.2 \times 10^5\,\mathrm{sec}^{-1}$. Measuring the rheology of fluids at very high shear rates is challenging. Using a flat plate rheometer, Novotny and Eckert (1974) determined that the relationship between shear stress and shear rate is still linear for water at a shear rate of $10{,}000\,\mathrm{sec}^{-1}$, but the possibility that anomalous effects are caused by non-Newtonian behavior above shear rates of $10^4\,\mathrm{sec}^{-1}$ has not been adequately explored.

Electroviscous effects have been cited as a possible cause for increases in frictional resistance [Brutin and Tadrist, 2003; Ren et al., 2001]. Interestingly, while Brutin and Tadrist (2003) ruled out effects of surface roughness and attributed increases in resistance to ionic effects, Phares and Smedley (2004) ruled out electroviscous effects and indicate that surface roughness effects are a more likely explanation for departures from conventional Poiseuille theory.

Some dependence of flow resistance on channel geometry has been observed [Papautsky et al., 1999b; Peng et al., 1994; Pfahler et al., 1991; Qu et al., 2000]. One of the challenges of measuring the dependence of flow resistance on aspect ratio or thinness of one dimension is the accurate characterization of the flow channels. Both Papautsky et al. (1999b) and Pfahler et al. (1991) acknowledge the difficulty of ensuring accurate size measurement and the corresponding difficulty of conclusively establishing a geometrical effect in their experiments. In at least one case, the departure from conventional macroscale theoretical predictions is found to depend on both Reynolds number and hydraulic diameter of the trapezoidal channels [Qu et al., 2000].

The critical Reynolds numbers for transition to turbulence in microchannel flows have been reported or modeled for certain flow conditions as below [Hsieh et al., 2004; Mala and Li, 1999; Morini, 2004; Peng et al., 1994; Pfund et al., 2000; Wu and Little, 1983] or consistent [Celata et al., 2002; Liu and Garimella, 2004; Sharp and Adrian, 2004] with nominal values for macroscale conduit flows, such as near 2000 for circular pipe flow [Darbyshire and Mullin, 1995]. The ranges of critical Reynolds numbers cited in the literature include values both dramatically lower than the nominal macroscale values, such as 240 [Hsieh et al., 2004], 200–700 [Peng et al., 1994], 400 [Wu and Little, 1983], 300–900 [Mala and Li, 1999], and values slightly lower than nominal macroscale values [Pfund et al., 2000]. Conclusive causes of observed early transition to turbulence have not been established, but the reported trend in the most recent experiments, Hsieh et al.'s (2004) work not withstanding, has been that transition is occurring at critical Reynolds numbers consistent with those in macroscale experiments. In earlier experiments and in Hsieh et al.'s (2004) experiments, the early-transition observations were based primarily on data obtained from bulk flow measurements. More recently the transitional Reynolds number range was established using flow visualization [Liu and Garimella, 2004; Hsieh et al., 2004] and by quantifying the magnitude of spatial and temporal velocity variations measured using micro-particle image velocimetry [Sharp and Adrian, 2004]. Reports of irregular tracer motion are used to justify the conclusion that transition is occurring at Reynolds numbers lower than Re $\sim$ 470 in Hsieh et al. (2004); in contrast, the spatial and temporal velocity variations indicated an onset of transition for Reynolds numbers of 1800–2000 in circular microtubes [Sharp and Adrian, 2004], and flow visualization indicated an onset of transition at Reynolds numbers of approximately 1800–2200 in rectangular channels [Liu and Garimella, 2003] consistent with nominal critical Reynolds numbers of 2000–3000 in macroscale rectangular conduits, where the critical Reynolds number in a rectangular channel can depend on aspect ratio [Hanks and Ruo, 1966].

Certainly, the inclusion or exclusion of entrance effects can affect the magnitude of the measured friction factor and is generally considered in careful experimental studies. Regardless of the geometry, to accurately measure the dimensions of these microchannels is extremely difficult, particularly when one of the dimensions is on the order of a couple of microns. The pressure drop in a round capillary is inversely proportional to D^4 (Equation [10.21]), so an inaccuracy of 5% in measuring D can bias resistance results by 20%, enough to explain the majority of the early discrepancies between the conventional macroscopic resistance predictions and the observed values in Figure 10.5.

The validity of the no-slip assumption for liquids in contact with a solid surface has also been brought into question [Choi et al., 2003; Tretheway and Meinhart, 2002; 2004; Zhu and Granick, 2001], particularly in the case of coated microchannels. Documented slip lengths are at most 1 mm [Tretheway and Meinhart, 2002] and as low as tens of nm [Choi et al., 2003] and could be an additional source of error in flow resistance experiments.

The details of other models incorporating micropolar fluid theory, cross-sectional geometry, roughness, entrance, and viscous dissipation effects may be found in the literature [Koo and Kleinstreuer, 2003; Morini, 2004; Papautsky, et al. 1999a].

10.1.10 Measurements of Velocity in Microchannels

Along with the growth of research in microdevices, rapid development of experimental techniques for investigating flows in such devices is also underway including modification of experimental techniques commonly applied at the macroscale and development of new techniques. Measurements of velocities in microchannels have been obtained using bulk flow, pointwise, and field measurements. Each technique has certain advantages that may make it more suitable to providing a specific type of flow field information. A brief summary is included herein. For more detailed discussions of diagnostic techniques, including the most recent advances in acquisition and processing techniques, Nguyen and Wereley (2002, chap. 4), and Devasenathipathy et al. (2003).

The majority of flow resistance data in microscale geometries to date has been obtained through the use of bulk flow measurements. Typical methods used to measure bulk flow rate include an in-line flowmeter or the timed collection of fluid at the outlet and pressure taps located at the inlet and outlet or simply at the inlet if the pressure at the outlet is known. Bulk flow measurements require neither optical access to the microchannel nor seeding, and there are no restrictions on the geometrical parameters of the channel. However, given the disagreement in results regarding microscale effects on flow resistance in particular, bulk flow measurements lack sufficient detail to discern potential mechanisms causing deviation from macroscale theory. Detailed measurements of flow velocity are also useful for optimizing the design of complex microdevices for mixing, separation, reaction, and thermal control. For examples of these devices, consult the section on applications of MEMS in this handbook.

The first micro-Particle Image Velocimetry (micro-PIV) measurements were made in a Hele-Shaw cell [Santiago et al., 1998]. These velocity field measurements were resolved to 6.9 μm in the lateral directions and 1.5 μm in the depth direction and demonstrated the applicability of the well-established PIV technique for microflows. Micro-PIV measurements in a rectangular glass microchannel with 200 nm fluorescent tracer particles ($\mathrm{Re} < 1$) have been described in Meinhart et al. (1999). With improved acquisition and analysis, the lateral resolution was 13.6 μm in the streamwise direction significantly better 0.9 μm in the cross-stream direction, the direction of highest velocity variation. The first demonstration of micro-PIV within a circular capillary was performed in a 236 μm diameter channel with Reynolds number $\ll 1$ [Koutsiaris et al., 1999]. The seeding particles were 10 μm glass spheres, and the resolution of the measurements was 26.2 μm in the cross-stream direction. The measured velocity profiles agreed well with the predicted laminar parabolic profiles. More recently, micro-PIV has been used to study the velocity profiles and turbulence statistics of water flows in circular channels with $D \sim 100$–250 μm and Reynolds numbers up to 3000 using 2 μm fluorescent particles [Sharp and Adrian, 2004].

Alternate visual methods applied to microchannel velocity measurements have been demonstrated by numerous researchers [Brody et al., 1996; Maynes and Webb, 2002; Ovryn, 1999; Paul et al., 1998b; Taylor and Yeung, 1993]. Molecular tagging velocimetry (MTV) was adapted to the microscale, and velocity profiles were obtained in circular tubes with $D = 705$ μm, and for $\mathrm{Re} = 600$–5000 [Maynes and Webb, 2002]. The spatial resolution of these measurements was approximately 10 μm. The measured velocity profiles were consistent with macroscale laminar predictions for $\mathrm{Re} \leqslant 2000$ and show indications of transition at a Reynolds number of approximately 2100. Relevant development issues for microscale MTV are similar to those for PIV, namely optical access and index of refraction compensation, particularly for curved surfaces, and optimized detection of the tracking particles (PIV) or beams (MTV).

Particle tracking, streak quantification, or dye visualization can be implemented given optical access and the ability to illuminate the flow [Brody et al., 1996; Devasenathipathy et al., 2002: Taylor and Yeung, 1993]. Care must be exercised in the extraction of quantitative data, particularly if there is a large depth of field of the imaging device or optical complications due to complex microchannel geometries or if the particles or molecules are not accurately following the flow due to charge, size, or density effects.

Novel three-dimensional measurement techniques for microchannel flows are currently in development [Hitt and Lowe, 1999; Ovryn, 1999]. Building upon a technique already developed for the study of microscale structures, Hitt and Lowe (1999) used confocal imaging to build a three-dimensional map of the *separation surface* following a bifurcation, where the separation surface describes the interfacial boundary between two components from different branches of the bifurcation. Using two laser-scanning confocal microscopes, a series of thin (4.5 or 7.1 μm) horizontal slices were acquired and reconstruction software was used to combine these slices into a three-dimensional map. Again, optical access and effects are primary issues in the implementation of this method, and it is not suitable for unsteady flows. Ovryn

(1999) sought to resolve and interpret the scattering pattern of a particle to determine its three-dimensional position, and has applied this technique to laminar flow.

X-ray imaging techniques do not require optical access in the channel, though a contrast medium detectable by X-rays must be used as the working fluid. Lanzillotto et al. (1996) obtained flow displacement information from microradiograph images of emulsion flow through a 640 μm diameter tube and iododecane flow through a silicon V-groove chip.

The level of complexity increases when electrokinetic flows are considered. A few of the earliest visual measurements of electrokinetic flows are described in Paul et al. (1998b); Cummings (1999); and Taylor and Yeung (1993). Paul et al. (1998b) seeded the flow with an uncaged fluorescent dye. Once the dye was uncaged by an initial ultraviolet (UV) laser pulse, the flow was illuminated by succeeding pulses of blue light for Charge Coupled Device (CCD) image acquisition, causing the excitation of only the uncaged dye molecules. This technique was applied to both pressure-driven and electrokinetic flows in circular capillaries with diameters of the order 100 μm. Since the dye transport represents represented both convection and diffusion, requisite care is necessary to separate the effects [Paul et al., 1998b]. This method can be used also to acquire quantitative information regarding diffusion effects. More recently, particle tracking techniques have been adapted to electrokinetic flows [Devasenathipathy et al., 2002].

Pointwise techniques were used to acquire early velocity measurements in microfluidic systems [Chen et al., 1997; Tieu et al., 1995; Yazdanfar et al., 1997]. Optical doppler tomography combines elements of Doppler velocimetry with optical coherence tomography in an effort to develop a system that can quantify the flow in biological tissues [Chen et al., 1997]. Chen et al. (1997) applied the technique to a 580 μm diameter conduit seeded with 1.7 μm particles. An approximate parabolic profile was measured in the first test, and in the second test, it was shown that fluid particle velocities could be measured even with the conduit submerged in a highly scattering medium, as would be the case for particles in biological tissues. A similar measurement technique has been used for in vivo measurements [Yazdanfar et al., 1997]. An adaptation of laser doppler anemometry (LDA) techniques to microscale flows was demonstrated by Tieu et al. (1995), and pointwise data were obtained in a 175 μm channel.

10.1.11 Non-Linear Channels

For practical MEMS applications, it is often useful to consider mixing or separation of components in microchannels. Numerous designs have been proposed, including T- and H-shaped channels, zigzag-shaped channels, 2-D and 3-D serpentine channels, and multilaminators.

For example, Weigl and Yager (1999) have designed a T-sensor for implementation of assays in microchannels, as shown in Figure 10.6. A reference stream, a detection stream, and a sample stream have been introduced through multiple T-junctions into a common channel. The design relied upon the differential diffusion of different sized molecules to separate components in the sample stream. Differential diffusion rates are also fundamental to the design of the H-filter, used to separate components [Schulte et al., 2000]. Application of a slightly different T-channel design has been demonstrated for measurement of diffusion coefficients of a species in a complex fluid [Galambos and Forster, 1998].

A layering approach has been implemented by Branebjerg et al. (1996), splitting the streams and relayering to increase interfacial area, thus promoting mixing. Adding complexity to the flow field also has potential to increase the amount of mixing between streams, as demonstrated by Branebjerg et al.'s (1995) zigzag channel and the serpentine channels introduced by Liu et al. (2000). The 3-D serpentine channel in Liu et al. (2000) was designed to introduce chaotic advection into the system and further enhance mixing over a 2-D serpentine channel. A schematic of the 3-D serpentine channel is shown in Figure 10.7.

For further information on the use of nonlinear channels in microdevices, consult the applications of MEMS section of this handbook.

10.1.12 Capacitive Effects

While liquids are incompressible, the systems through which they flow may expand or contract in response to pressure in the liquid. This behavior can be described by analogy to flow in electrical circuits.

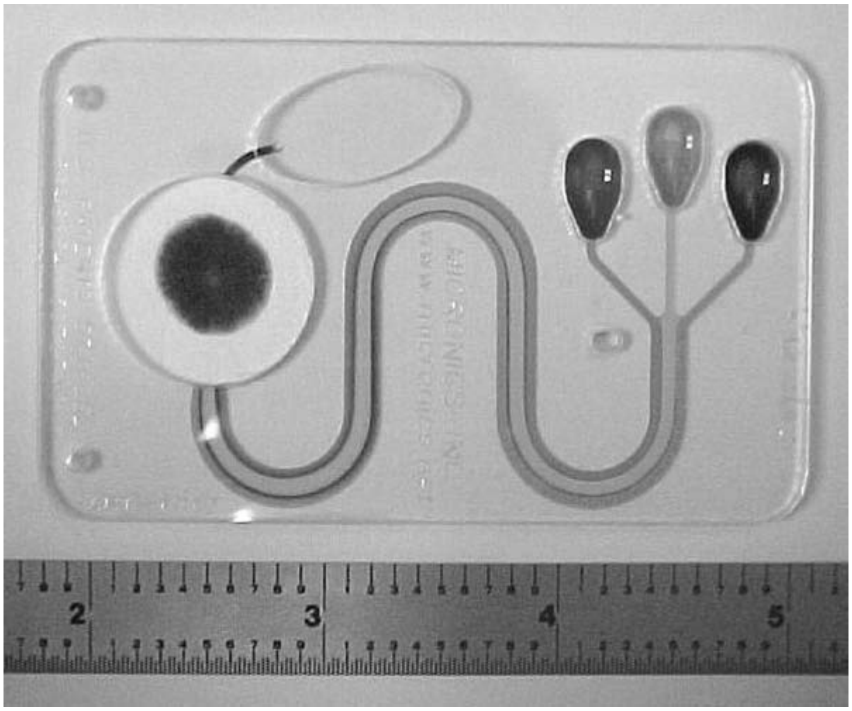


FIGURE 10.6 T-Sensor, self-calibrating microchemical reactor and sensor. This design allows for self-calibration through the simultaneous flow of a reference solution on the opposite flank of the indicator stream from the sample to be analyzed. (Reprinted with permission from Micronics, Inc., Redmond, WA, 2000.)

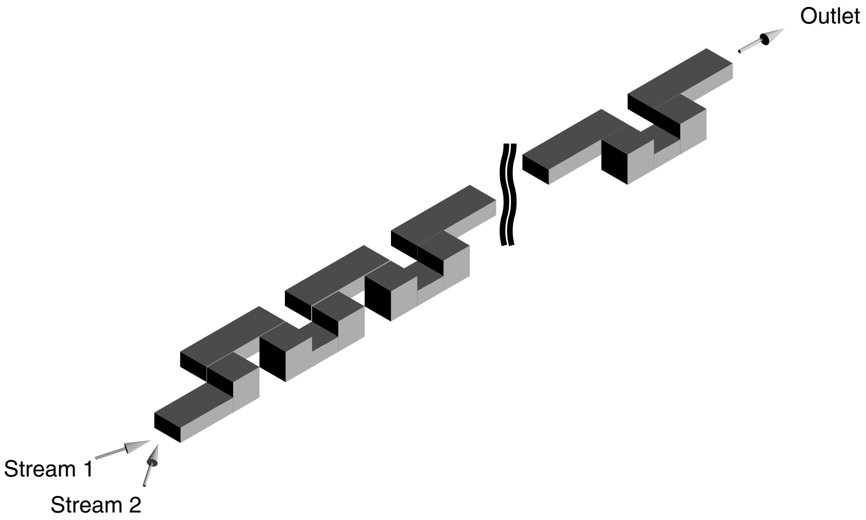


FIGURE 10.7 Three-dimensional serpentine channel. (Reprinted with permission from Liu et al. [2000], personal communication.)

In this analogy, fluid pressure corresponds to electrical voltage $p \sim V$; the volume flow rate corresponds to electrical current $Q \sim I$; and the flow resistance through a fluid element corresponds to an electrical resistor $R_{\text{flow}} \sim R_{\text{elec.}}$ Thus for capillary flow, $\Delta p = R_{\text{flow}}Q$, where $R_{\text{flow}} = 8\,\mu L/\pi a^4$ (c.f., Equation [10.21]), whereas in the electrical analogy, $\Delta V = R_{\text{elec}}I$. If a fluid element is able to change its volume (expansion of plastic tubing, flexing in pressure transducer diaphragm, etc.), fluid continuity implies that:

$$\Delta Q = C_{\text{flow}}\frac{dp}{dt} \tag{10.29}$$

where C_{flow} is the capacitance of the fluid element. The corresponding electrical law is $I = C_{\text{elec}}\ dV/dt$, where C_{elec} is the electrical capacitance.

It is well known in the context of electrical circuits that a resistor and capacitor in combination cause transients whose time constant τ is proportional to $R_{\text{elec}}C_{\text{elec}}$. In a microfluidic circuit, any capacitive element in combination with a flow resistance leads to analogous transients whose time constant is proportional to $R_{\text{flow}}C_{\text{flow}}$. Since R_{flow} can be very large in microchannels, the time constant can be surprisingly large, that is 10^3 seconds. Consequently, capacitive effects can cause significant and inconveniently long transients.

10.1.13 Applications of Particle/Cell Manipulation in Microfluidics

A number of research efforts are underway to develop particle separation, sorting, and detection capabilities in microfluidic networks with a particular emphasis on biological applications [Berger et al., 2001; Blankenstein and Larsen, 1998; Cho and Kim, 2003; Chou et al., 2000; Glückstad, 2004; Lee et al., 2001, 2003; Mirowski et al., 2004]. Cell or particle separation and sorting techniques have been proposed using concepts from electrokinetics [Cho and Kim, 2003; Fu et al., 2004], optical methods [Glückstad, 2004], magnetics [Mirowski et al., 2004; Berger et al., 2001], and hydrodynamic-based manipulation [Blankenstein and Larsen, 1998; Chou et al., 2000; Lee et al., 2001, 2003].

10.1.14 Recommended Review Papers on Microfluidics

For further information on research and development trends in microfluidics, the reader is referred to two review papers, Stone et al. (2004) and Ho and Tai (1998).

10.2 Electrokinetics Background

The first demonstration of electrokinetic phenomena is attributed to F.F. Reuss, who demonstrated electroosmotic flow of water through a clay column in a paper published in the Proceedings of the Imperial Society of Naturalists of Moscow in 1809 [Probstein, 1994]. In the latter part of the 20th century, the main applications of electrokinetic phenomena have been fairly wide-ranging from the dewatering of soils and waste sludges using electric fields [Hiemenz and Rajagopalan, 1997] to the study of the stability of colloidal suspensions for household paint and to devices that use electrophoretic mass transfer of colloidal suspensions to produce images on a planar substrate [Kitahara and Watanabe, 1984]. A community that has paid particular attention to the study of mass and momentum transport using electrokinetic effects is the developers of capillary electrophoresis (CE) devices [Khaledi, 1998; Landers, 1994; Manz et al., 1994]. CE devices are used to separate biological and chemical species by their electrophoretic mobility, which is roughly proportional to their mass-to-charge ratio. CE devices that employ a sieving matrix separate macromolecules based on size (e.g., DNA separations or surfactant-coated, denatured protein separations). These traditional CE systems incorporate on-line detection schemes such as ultraviolet radiation scatter/absorption and laser-induced fluorescence [Baker, 1995; Landers, 1994].

Electrokinetics is the general term describing phenomena that involve the interaction between solid surfaces, ionic solutions, and macroscopic electric fields. Two important classes of electrokinetics are electrophoresis and electroosmosis where the motions of particles and electrolyte liquids, respectively, occur when an external electric field is applied to the system. Electrophoresis is the induced drift motion of colloidal particles or molecules suspended in liquids that results from the application of an electric field. Electroosmosis describes the motion of electrolyte liquids with respect to a fixed wall that results when an electric field is applied parallel to the surface. An example of electroosmosis is the liquid pumping that occurs in a microcapillary when an electric field is applied along the axis of the capillary [Hunter, 1981; Levich, 1962; Probstein, 1994]. Two other phenomena also classified under electrokinetics are flows with a finite streaming potential and sedimentation potential. These phenomena are counter-examples of electroosmosis and electrophoresis respectively. Streaming potential is the spontaneous generation of an electric potential from a pressure-driven flow in a charged microchannel [Hunter, 1981; Scales et al., 1992]. Sedimentation potential is the generation of an electric potential that results from the sedimentation (e.g., due to gravity) of a charged particle [Russel et al., 1999]. All of the phenomena classified under the term electrokinetics are manifestations of the electrostatic component of the Lorentz force (on ions and surface charges) and Newton's second law of motion. These interactions between charged particles and electric fields often involve electric double layers formed at liquid/solid interfaces, and an introduction to this phenomenon is presented below. Electrokinetic flows are in general a subclass of electrohydrodynamic flows [Melcher, 1981; Saville, 1997], which describe the general coupling between electric fields and fluid flow. Electrokinetic systems are distinguishable in that they involve liquid electrolyte solutions and the presence of electrical double layers (i.e., involve electrophoresis and electroosmosis).

10.2.1 Electrical Double Layers

Most solid surfaces acquire a surface electric charge when brought into contact with an electrolyte (liquid). Mechanisms for the spontaneous charging of surface layers include the differential adsorption of ions from an electrolyte onto solid surfaces (e.g., by ionic surfactants), the differential solution of ions from the surface to the electrolyte, and the deprotonation/ionization of surface groups [Hunter, 1981]. The most common of these in microfluidic electrokinetic systems is the deprotonation of surface groups on the surface of materials like silica, glass, acrylic, and polyester. In the case of glass and silica, the deprotonation of surface silanol groups (SiOH) determines the generated surface charge density. The magnitude of the net surface charge density at the liquid/solid interface is a function of the local pH. The equilibrium reaction associated with this deprotonation can be represented as

$$\mathrm{SiOH} \Leftrightarrow \mathrm{SiO}^- + \mathrm{H}^+ \tag{10.30}$$

Models describing this reaction have been proposed for several types of glass and silica [Hayes et al., 1993; Huang et al., 1993; Scales et al., 1992]. In practice, the full deprotonation of the glass surface, and therefore the maximum electroosmotic flow mobility, is achieved for pH values greater than about 9.

In response to the surface charge generated at a liquid–solid interface, nearby ions of opposite charge in the electrolyte are attracted by the electric field produced by the surface charge, and ions of like charge are repelled. The spontaneously formed surface charge therefore forms a region near the surface called an electrical double layer (EDL) that supports a net excess of mobile ions with a charge opposite to that of the wall. Figure 10.8 shows a schematic of the EDL for a negatively charged wall (e.g., as in the case of a glass surface). The region of excess charge formed by the counterions shielding the wall's electric field can be used to impart a force on the bulk fluid through ion drag.

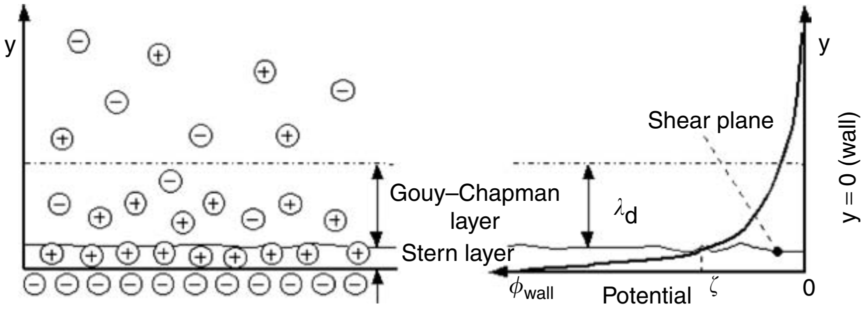


FIGURE 10.8 Schematic of the electrical double layer (EDL): (a) Distribution of co- and counterions near a charged wall. The Stern and Gouy–Chapman layers are shown with the Gouy–Chapman thickness roughly approximated as the Debye length of the solution. (b) A plot of the negative potential distribution near a glass wall indicating the zeta potential, the wall potential, and the location of the shear plane.

As shown in Figure 10.8, counterions reside in two regions divided into the Stern and Gouy–Chapman diffuse layers [Adamson and Gast, 1997; Hunter, 1981]. The Stern layer counterions are adsorbed onto the wall, while the ions of the Gouy–Chapman diffuse layer are free to diffuse into the bulk fluid and therefore available to impart work on the fluid. The plane separating the Stern and Gouy–Chapman layers is called the shear plane. The bulk liquid far from the wall is assumed to have net neutral charge. Also in Figure 10.8 is a sketch of the potential associated with the EDL. The magnitude of the potential is a maximum at the wall and drops rapidly through the Stern layer. The potential at the shear plane, which is also the boundary of the liquid flow problem, is called the "zeta potential" ζ Because of the difficulties associated with predicting the properties of the EDL from first principles [Hunter, 1981], the zeta potential is typically viewed as an empirical parameter determined using electroosmotic or streaming potential flow measurements.

A simple treatment of the physics of the diffuse portion of the EDL is presented here; it assumes a liquid with constant properties (i.e., constant viscosity and electrical permittivity). A more detailed model of the diffuse portion of the electrical double layer should include non-continuum effects such as finite-ion size effects and gradients in the dielectric strength and viscosity of the fluid [Hunter, 1981]. The width of the diffuse portion of the EDL is determined by the opposing forces of electrostatic attraction and thermal diffusion. This balance between electromigration and diffusive fluxes, together with the Nernst–Einstein equation relating ion diffusivity and mobility [Hiemenz and Rajagopalan, 1997], can be used to show that the concentration profile is described by a Boltzmann distribution. For an EDL on a flat plate, the Boltzmann distribution of ions of species i, c_i, is

$$c_i(y) = c_{\infty,i} \exp\left(-\frac{ze\phi(y)}{kT}\right), \tag{10.31}$$

where $c_{\infty,\,i}$ is the molar concentration of ion i in the bulk, z is the valance number of the ion, ϕ is the local potential, T is temperature, e is the charge of an electron, and k is Boltzmann's constant. The coordinate y is perpendicular to the wall and the origin is at the shear plane of the EDL. The net charge density in the EDL, ρ_E, is related to the molar concentrations of N species using the relation

$$\rho_E = F\sum_{i=1}^{N} z_i c_i, \tag{10.32}$$

where F is Faraday's constant. The net charge density can also be related to the local potential in the diffuse EDL by the Poisson equation

$$\nabla^2\phi = -\frac{\rho_E}{\varepsilon} \tag{10.33}$$

where ε is the permittivity of the liquid. Substituting Equations (10.31) and (10.32) into Equation (10.33), we find that

$$\frac{d^2\phi}{dy^2} = \frac{-F}{\varepsilon}\sum_{i=1}^{N} z_i c_{\infty,i} \exp\left(-\frac{ze\phi(y)}{kT}\right) \tag{10.34}$$

For the simple case of a symmetric electrolyte with (two) monovalent ions, this relation becomes

$$\frac{d^2\phi}{dy^2} = \frac{2Fz_i c_\infty}{\varepsilon}\sinh\left(\frac{ze\phi(y)}{kT}\right) \tag{10.35}$$

where c_∞ is the molar concentration of each of the two ion species in the bulk. This relation is the nonlinear Poisson–Boltzmann equation. A closed form, analytical solution of this equation for the EDL on a flat wall is given by Adamson and Gast (1997) and Hunter (1981).

A well-known approximation to the Poisson-Boltzmann solution known as the Debye–Hückel limit is the case where the potential energy of ions in the EDL is small compared to their thermal energy so that the argument of the hyperbolic sine function in Equation (10.35) is small. Applying this approximation, Equation (10.35) becomes

$$\frac{d^2\phi}{dy^2} = \frac{\phi(y)}{\lambda_D^2} \tag{10.36}$$

where λ_D is the Debye length of the electrolyte defined as

$$\lambda_D \equiv \left(\frac{\varepsilon kT}{2z^2 F^2 c_\infty}\right)^{\frac{1}{2}} \tag{10.37}$$

for a symmetric monovalent electrolyte. The Debye length describes the characteristic thickness of the EDL, which varies inversely with the square root of ion molar concentration. At typical biochemical, singly ionized buffer concentrations of 10 mM, the thickness of the EDL is therefore on the order of a few nanometers [Hiemenz and Rajagopalan, 1997]. In analyzing electrokinetic flow in microchannels, the Debye length should be compared to the characteristic dimension of the microchannel in order to classify the pertinent flow regime. Overbeek (1952) points out that the Debye–Hückel approximation of the potential of the EDL holds remarkably well for values of the ratio $ze\phi/(kT)$ up to approximately 2. This value is equivalent to a zeta potential of about 50 mV, which is within the typical range of microfluidic applications.

Models of the physics of the EDLs can be used to extrapolate zeta potential of particles and microchannels across a significant range of buffer concentration, fluid viscosity, electrical permittivity of electrolytes, and field strengths given only a few measurements. One of the most difficult zeta potential extrapolations to make is across different values of pH because pH changes the equilibrium reactions associated with the charge at the liquid–solid interface.

A full formulation of the coupled system of equations describing electroosmotic and electrokinetic flow includes the convective diffusion equations for each of the charged species in the system, the Poisson equation for both the applied electric field and the potential of the EDL, and the equations of fluid motion. A few solutions to this transport problem relevant to microfluidic systems are presented below.

10.2.2 EOF with Finite EDL

Electroosmotic flow (EOF) results when an electric field is applied through a liquid-filled microchannel having an EDL at the channel surfaces, as described above. This applied electric field introduces an electrostatic Lorentz body force

$$\rho \mathbf{b} = \rho_E \mathbf{E} \tag{10.38}$$

into the equation of motion for the fluid, Equation (10.3). Within the EDL, the electric field exerts a net force on the liquid causing the liquid near the walls to move. Alternately, one can describe the effect as simply the ion drag on the liquid associated with the electrophoresis of the ions in the EDL. The fluid in the EDL exerts a viscous force on the rest of the (net zero charge) liquid in the bulk of the channel. For EDLs much smaller than the channel dimension D, the fluid velocity reaches steady state in a short time t that is on the order of D^2/ν, where ν is kinematic viscosity of the fluid. The resulting bulk electroosmotic flow is depicted schematically in Figure 10.9.

The equation of motion for steady low-Reynolds-number flow in the microchannel is given by Equation (10.39).

$$\nabla p = \mu \nabla^2 \mathbf{u} + \rho_E \mathbf{E} \tag{10.39}$$

Substituting Equation (10.33) for the charge density, results in

$$\nabla^2\left(\mathbf{u} - \frac{\varepsilon \mathbf{E}}{\mu}\phi\right) = \frac{\nabla p}{\mu}. \tag{10.40}$$

In Equation (10.40), the electric field, $\mathbf{E}$, can be brought into the Laplace operator because $\nabla \cdot \mathbf{E} = \nabla \times \mathbf{E} = 0$. Equation (10.40) is linear so that the velocities caused by the pressure gradient and the electric field can be considered separately and then superposed as follows:

$$\nabla^2\left(\mathbf{u}_{\text{EOF}} - \frac{\varepsilon \mathbf{E}}{\mu}\phi\right) = 0, \tag{10.41}$$

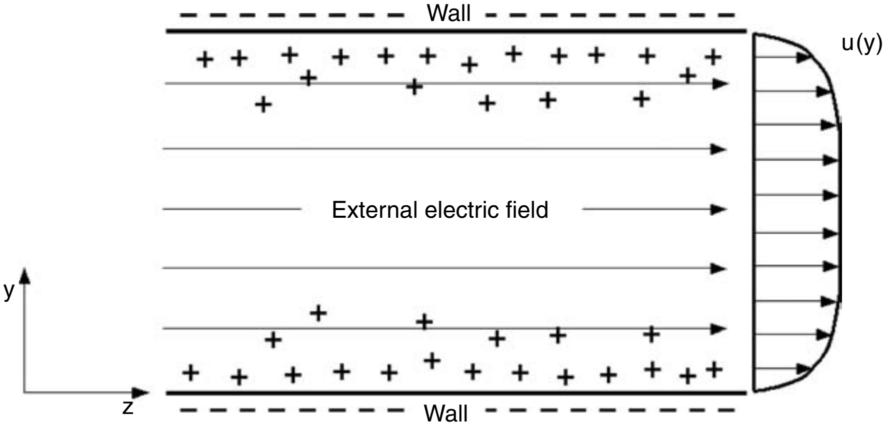


FIGURE 10.9 Schematic of an electroosmotic flow channel with a finite EDL. The charges drawn in the figure indicate net charge. The boundary layers on either wall have a thickness on the order the Debye length of the solution. For non-overlapping EDLs, the region near the center of the channel is net neutral.

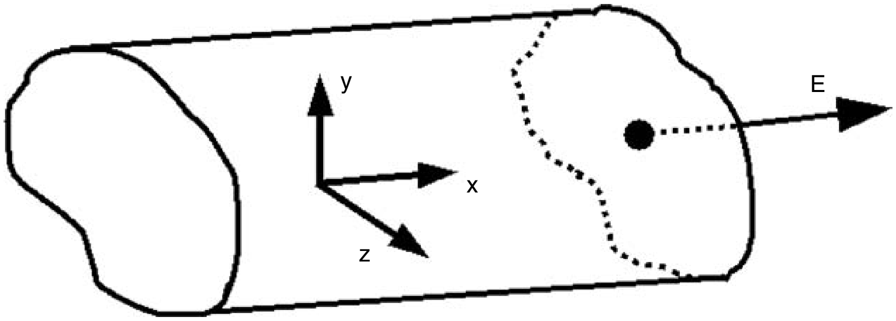


FIGURE 10.10 Section of a long, straight channel having an arbitrary cross-section.

$$\nabla^2 \mathbf{u}_{\text{pressure}} = \frac{\nabla p}{\mu}. \tag{10.42}$$

Together with Equation (10.2), these are the general equations for electroosmotic flow in a microchannel. Evaluation of the pressure-driven flow component of velocity in a microchannel can leverage analytical solutions available for channels of various cross-sections [White, 1991]. The pressure gradient can be applied externally or may arise internally because of variations in the zeta potential at the channel walls [Anderson and Idol, 1985; Herr et al., 2000].

Now consider electroosmosis in a long straight microchannel with a finite width electrical double layer and an arbitrary cross-section that remains constant along the flow direction (x-axis), as shown in Figure 10.10. The applied electric field is assumed to be uniform and along the x-axis of the microchannel. For the case where the potential at the wall is uniform, the solution to Equation (10.41) is

$$u_{\text{EOF}} - \frac{\varepsilon E \zeta}{\mu}\phi = \frac{-\varepsilon E \zeta}{\mu}, \tag{10.43}$$

with the zeta-potential ζ, being the value of ϕ at the top of the double layer. In Equation (10.43) u_{EOF} and E are the unidirectional velocity and unidirectional applied electric field, respectively. The general expression for the electroosmotic velocity, implicit in the potential is then

$$u_{\text{EOF}}\,(y, z) = \frac{-\varepsilon E \zeta}{\mu}\left(1 - \frac{\phi(y,z)}{\zeta}\right). \tag{10.44}$$

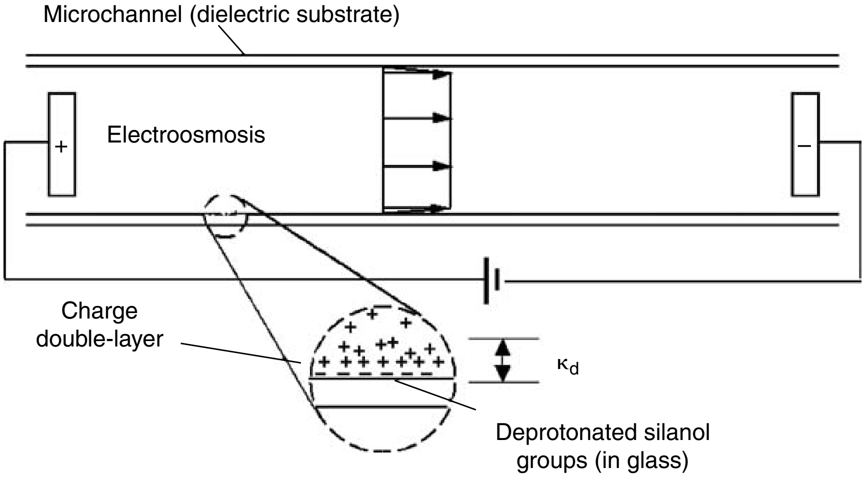

Microchannel (dielectric substrate)
Electroosmosis
+
–
Charge double-layer
κ_d
Deprotonated silanol groups (in glass)

FIGURE 10.11 Schematic of electroosmotic flow in a glass microchannel with a thin EDL. A zero pressure gradient plug flow is shown. The electrodes on the ends of the channel indicate the polarity of the electric field.

To compute values for the velocity given in Equation (10.44), an expression for the potential $\phi(y, z)$ is required. In general $\phi(y, z)$ can be computed numerically from Equation (10.34), but analytical solutions exist for several geometries. Using the Boltzmann equation for a symmetric analyte and the Debye–Hückel approximation discussed in the previous section, Rice and Whitehead (1965) give the solution for electroosmosis in a long cylindrical capillary.

$$u_{\mathrm{EOF}}(r) = \frac{-\varepsilon E \zeta}{\mu}\left(1 - \frac{I_o(r/\lambda_D}{I_o(r/a)}\right). \tag{10.45}$$

In Equation (10.45), I_o is the zero-order modified Bessel function of the first kind; r is the radial direction; and a is the radius of the cylindrical capillary. This solution can be superposed with the solution of Equation (10.42) for a constant pressure gradient. The resulting composite solution is

$$u(r) = \frac{-\varepsilon E \zeta}{\mu}\left(1 - \frac{I_o(r/\lambda_D)}{I_o(r/a)}\right) - \frac{dp}{dx}\frac{a^2}{4\mu}\left(1 - \frac{r^2}{a^2}\right). \tag{10.46}$$

Burgeen and Nakache (1964) give a general solution for electroosmotic flow between two long, parallel plates, for a finite EDL thickness (but with nonoverlapping EDLs). For other more complex geometries and many unsteady problems, numerical solutions for the electroosmotic flow are required [Arulanandam and Li, 2000; Bianchi et al., 2000; Dutta et al., 2002; Myung-Suk and Kwak, 2003; Patankar and Hu, 1998; Yao, 2003a].

However, when the Debye length is finite but much smaller than other dimensions (e.g., the width of the microchannel) the disparate length scales can make numerical solutions difficult [Bianchi et al., 2000, Patankar and Hu, 1998]. In many cases, EOF in complex geometries can be determined numerically using a thin double layer assumption described in the next section.

10.2.3 Thin EDL Electroosmotic Flow

This section presents a brief analysis of electroosmotic flow in microchannels with thin EDLs. Figure 10.11 shows a schematic of an electroosmotic flow in a microchannel with zero pressure gradient. As shown in the figure, the Debye length of typical electrolytes used in microfabricated electrokinetic systems is much smaller than the hydraulic diameter of the channels. Typical Debye-length-to-channel diameter ratios are less than 10^{-4}. For low Reynolds number electroosmotic flow in a cylindrical channel in the presence of a constant axial pressure gradient and a Debye length much smaller than the capillary radius, the solution of the velocity field is simply

$$u(r) = -\frac{\varepsilon\zeta E}{\mu} - \frac{dp}{dx}\frac{(a^2 - r^2)}{4\mu}. \tag{10.47}$$

This equation can be derived by evaluating Equation (10.46) in the limit of a thin EDL (i.e., a small value of λ_D/a).

The zeta potential typically determines flow velocities and flow rates in common thin EDL systems. As mentioned above, this quantity can often be interpreted as an empirically measured mobility parameter that determines the local velocity of the flow at the top of the electrical double layer. The zeta potential can be approximately related to the local surface charge density on the wall and the bulk fluid properties by applying continuum field and flow theory. Theoretically, the zeta potential is defined as the value of the electrostatic potential at the plane that separates double layer counterions that are mobile from those that are fixed. For the case of zero applied pressure gradients, Equation (10.47) reduces to the well-known Helmholtz-Smoluchowski relation for electroosmotic flow: $u = \varepsilon\zeta E/\mu$ [Probstein, 1994]. Other thin EDL solutions include that of Ghosal (2002) for slowly varying zeta potential and cross-sectional area channels, and Oddy and Santiago (2004) for a rectangular channel with four different wall zeta potentials and an applied AC electric field.

10.2.4 Electrophoresis

Many electrokinetic microfluidic systems leverage the combination of electroosmotic and electrophoresis to achieve biological separations and to transport charged particles (e.g., biological assay microbeads) and ions. Because of this, we present here a short introduction to electrophoresis. Electrophoresis is the induced drift motion of colloidal particles or molecules suspended in polar solutions that results from the application of an electric field. Two important regimes of electrophoresis depicted in Figure 10.12 are for the electromigration of species that that are either large or small compared to the Debye length of the ionic solution in which they are suspended.

Electrophoresis of ionic molecules and macromolecules can be described as a simple balance between the electrostatic force on the molecule and the viscous drag associated with its resulting motion. As a result, the electrophoretic mobility (velocity-to-electric field ratio) of molecules is a function of the molecule's size/molecular weight and directly proportional to their valence number or

$$u = \frac{qE}{3\pi\mu d}(d \ll \lambda_d) \tag{10.48}$$

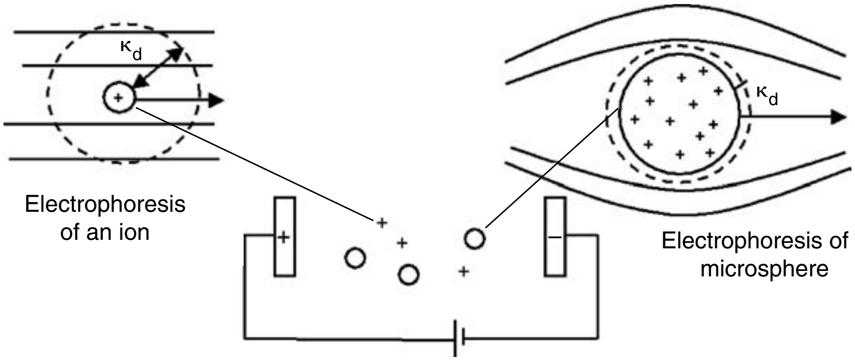


FIGURE 10.12 Two limiting limits of electrophoresis in an electrolyte. Shown are electrophoretic particles in the electric field generated between two electrodes. On the left is the detail of a charged ion with a characteristic dimension much smaller than the Debye length of the electrolyte. On the right is a charged microsphere with a diameter much larger than the Debye length.

where q is the total molecule charge and d is the particle's Stokes diameter (the diameter of a sphere of equal drag). In comparison, the electrophoresis of relatively large solid particles such as 100–10,000 nm diameter polystyrene spheres, clay particles, and single-celled organisms is a function of the electrostatic forces on the surface charge, the electrostatic forces on their charge double layers, and the viscous drag associated with both the motion of the body as well as the motion of the ionic cloud. For a wide range of cases where the particle-diameter-to-Debye-length ratio is large so that locally the ionic cloud near the particle surface can be approximated by the EDL relations for a flat plate, the velocity of an electrophoretic particle reduces simply to

$$u = \frac{\varepsilon \zeta E}{\mu}\left(d \gg \lambda_d\right) \tag{10.49}$$

where the dimension d in the inequality condition is a characteristic dimension of the particle (e.g., its Stokes diameter). This equation was shown by Smoluchowski (1903) to be independent of particle shape. This is the Helmholtz-Smoluchowski equation introduced earlier (with a change of sign).

The two expressions above describing the electrophoresis of particles can be expressed in terms of a mobility ν_{eph} equal to $q/(3\pi\mu d)$ and $\varepsilon\zeta/\mu$ for characteristic particle dimensions much smaller and much larger than the Debye length, respectively. Note also that for the simple case of a fluid with uniform properties, the solution of the drift velocity of electrophoretic particles with respect to the bulk liquid are similar (i.e., parallel and directly proportional) to lines of electric field.

Several solutions of the particle velocity and velocity field in the region of an electrophoretic particle with a finite EDL exist [Hunter, 1981; Russel et al., 1999]. A well-known solution is that of Henry (1948) for the flow around an electrophoretic sphere in the Debye–Hückel limit. The $d \gg \lambda_d$ limit of Henry's solution results in Equation (10.49).

10.2.5 Similarity between Electric and Velocity Fields for Electroosmosis and Electrophoresis

The previous sections have described the solution for electroosmotic velocity field in straight, uniform cross-section channels. In general, solving for the electroosmotic velocity field in more complex geometries requires a solution of the electric field and charge density in the microchannel, together with a solution to the Navier–Stokes equations. A simplification of this flow problem first proposed by Overbeek (1952) suggests that the electroosmotic velocity is everywhere parallel to the electric field for simple electroosmotic flows at low Reynolds numbers. This concept is also discussed by Cummings et al. (2000) and Santiago (2001). Santiago (2001) describes a set of sufficient conditions for which there exist a velocity field solution that is similar to the electric field:

- Uniform zeta potential
- Electric double-layers thin compared to channel dimension
- Electrically insulating channel walls
- Low Reynolds number
- Low product of Reynolds and Strouhal numbers
- Parallel flow at inlets and outlets
- Uniform electrolyte properties (including temperature)

When these conditions are met, the electroosmotic streamlines exactly correspond to the electric field lines. The approximation is applicable to systems with a microchannel length scale less than 100 µm, a Debye length less than 10 nm, a velocity scale less than 1 mm/sec, and a characteristic forcing function time scale greater than 10 msec [Santiago, 2001]. An important part of this similarity proof is to show the applicability of the Helmholtz–Smoulochowski equation in describing the local velocity field at the slip surface that bounds the internal flow of the microchannel that excludes the EDL. The Helmholtz–Smoulochowski equation can be shown to hold for most microfluidic systems where the motion of the EDL is dominated by the Lorentz and viscous forces. In such systems, we can consider the velocity field of the fluid outside of the EDL as a three-dimensional, unsteady flow of a viscous fluid of zero net charge that is bounded by the following slip velocity condition:

$$u_{slip} = \frac{-\varepsilon\zeta}{\mu} E_{slip} \tag{10.50}$$

where the subscript *slip* indicates a quantity evaluated at the slip surface at the top of the EDL (in practice, a few Debye lengths from the wall). The velocity along this slip surface is, for thin EDLs, similar to the electric field. This equation and the condition of similarity also hold for inlets and outlets of the flow domain that have zero imposed pressure-gradients.

The complete velocity field of the flow bounded by the slip surface (and inlets and outlets) can be shown to be similar to the electric field [Santiago, 2001]. We nondimensionalize the Navier–Stokes equations by a characteristic velocity and length scale U_s and L_s, respectively. The pressure p is nondimensionalized by the viscous pressure $\mu U_s/L_s$. The Reynolds and Strouhal numbers are $\mathrm{Re} = \rho L_s U_s/\mu$ and $\mathrm{St} = L_s/\tau U_s$, respectively, where τ is the characteristic time scale of a forcing function. The equation of motion is

$$\mathrm{ReSt}\,\frac{\partial \mathbf{u}'}{\partial t'} + \mathrm{Re}(\mathbf{u}' \cdot \nabla \mathbf{u}') = -\nabla p' + \nabla^2 \mathbf{u}' \tag{10.51}$$

Note that the right-most term in Equation (10.51) can be expanded using a well-known vector identity

$$\nabla^2 \mathbf{u}' = \nabla(\nabla \cdot \mathbf{u}') - \nabla \times \nabla \times \mathbf{u}'. \tag{10.52}$$

We can now propose a solution to Equation (10.52) that is proportional to the electric field and of the form

$$\mathbf{u}' = \frac{c_o}{U_s}\mathbf{E} \tag{10.53}$$

where c_o is a proportionality constant, and $\mathbf{E}$ is the electric field driving the fluid. Since we have assumed that the EDL is thin, the electric field at the slip surface can be approximated by the electric field at the wall. The electric field bounded by the slip surface satisfies Faraday's and Gauss' laws,

$$\nabla \cdot \mathbf{E} = \nabla \times \mathbf{E} = 0 \tag{10.54}$$

Substituting Equation (10.53) and Equation (10.54) into Equation (10.51) yields

$$\mathrm{ReSt}\,\frac{\partial \mathbf{u}'}{\partial t'} + \mathrm{Re}(\mathbf{u}' \cdot \nabla \mathbf{u}') = -\nabla p' \tag{10.55}$$

This is the condition that must hold for Equation (10.53) to be a solution to Equation (10.51). One limiting case where this holds is for very high Reynolds number flows where inertial and pressure forces are much larger than viscous forces. Such flows are found in, for example, high speed aerodynamics regimes and are not applicable to microfluidics. Another limiting case applicable here is when Re and ReSt are both small, so that the condition for Equation (10.53) to hold becomes

$$\nabla p' = 0. \tag{10.56}$$

Therefore we see that for small Re and ReSt and the pressure gradient at the inlets and outlets equal to zero, Equation (10.53) is a valid solution to the flow bounded by the slip surface, inlets, and outlets (note that these arguments do not show the uniqueness of this solution). We can now consider the boundary conditions required to determine the value of the proportionality constant c_o. Setting Equation (10.50) equal to Equation (10.53) we see that $c_o = \varepsilon\zeta/\eta$. So that, if the simple flow conditions are met, then the velocity everywhere in the fluid bounded by the slip surface is given by Equation (10.57).

$$\mathbf{u}(x, y, z, t) = -\frac{\varepsilon\zeta}{\mu}\,\mathbf{E}(x, y, z, t) \tag{10.57}$$

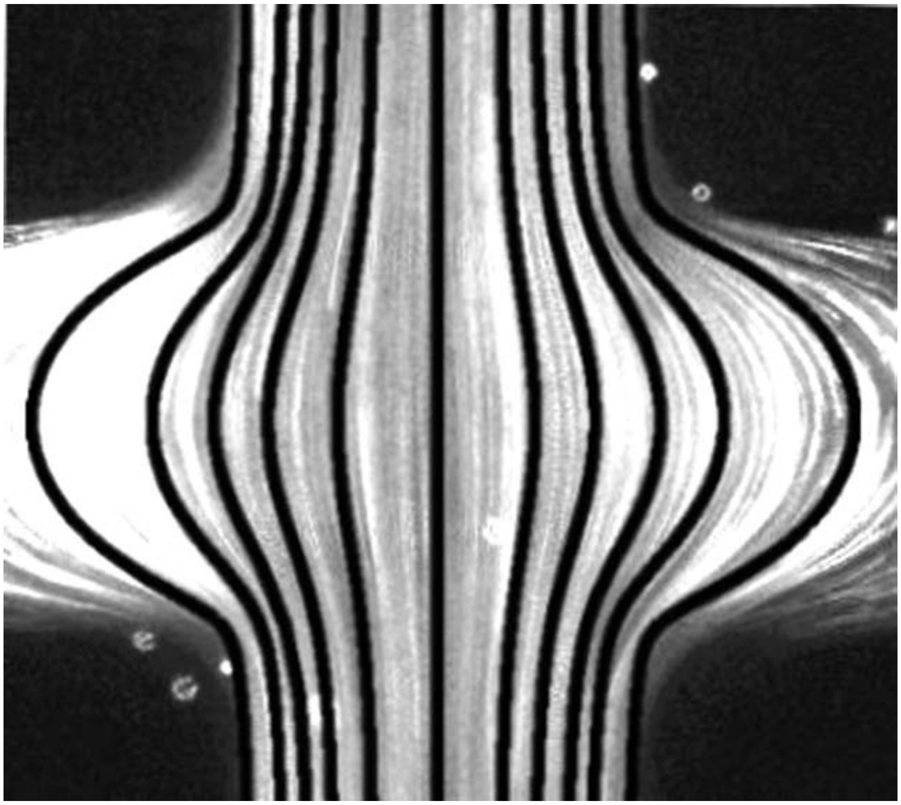


FIGURE 10.13 Comparison between experimentally determined electrokinetic particle pathlines at a microchannel intersection and predicted electric field lines. The light streaks show the path lines of 0.5 μm diameter particles advecting through an intersection of two microchannels. The electrophoretic drift velocities and electroosmotic flow velocities of the particles are approximately equal. The channels have a trapezoidal cross-section having a hydraulic diameter of 18 μm (130 μm wide at the top, 60 μm wide at the base, and 50 μm deep). The superposed heavy black lines correspond to a prediction of electric field lines in the same geometry. The predicted electric field lines very closely approximate the experimentally determined pathlines of the flow. (Reprinted with permission from Devasenathipathy, S., and Santiago, J.G. [2000] unpublished results, Stanford University.)

Equation (10.57) is the Helmholtz–Smoluchowski equation shown to be a valid solution to the quasi-steady velocity field in electroosmotic flow with ζ the value of the zeta potential at the slip surface. This result greatly simplifies the modeling of simple electroosmotic flows since simple Laplace equation solvers can be used to solve for the electric potential and then using Equation (10.57) for the velocity field. This approach has been applied to the optimization of microchannel geometries and verified experimentally [Bharadwaj et al., 2002; Devasenathipathy et al., 2002; Mohammadi et al., 2003; Molho et al., 2001; Santiago, 2001]. An increasing number of researchers have recently applied this result in analyzing electrokinetic microflows [Bharadwaj et al., 2002; Cummings and Singh, 2003; Devasenathipathy et al., 2002; Dutta et al., 2002; Fiechtner and Cummings, 2003; Griffiths and Nilson, 2001; MacInnes et al., 2003; Santiago, 2001]. Figure 10.13 shows the superposition of particle pathlines/streamlines and predicted electric field lines [Santiago, 2001] in a steady flow that meets the simple electroosmotic flow conditions summarized above. As shown in the figure, the electroosmotic flow field streamlines are very well approximated by electric field lines.

For the simple electroosmotic flow conditions analyzed here, the electrophoretic drift velocities (with respect to the bulk fluid) are also similar to the electric field, as mentioned above. Therefore, the time-averaged, total (local drift plus local liquid) velocity field of electrophoretic particles can be shown to be

$$\mathbf{u}_{\text{particle}} = \left(\nu_{\text{eph}} - \frac{\varepsilon\zeta}{\mu}\right)\mathbf{E}\cdot \qquad (10.58)$$

Here, we use the electrophoretic mobility ν_{eph} that was defined earlier, and $\varepsilon\zeta/\mu$ is the electroosmotic flow mobility of the microchannel walls. These two flow field components have been measured by Devasenathipathy et al. (2002) in two- and three-dimensional electrokinetic flows.

10.2.6 Electrokinetic Microchips

The advent of microfabrication and microelectromechanical systems (MEMS) technology has seen an application of electrokinetics as a method for pumping fluids on microchips [Auroux et al., 2002; Bruin, 2000; Jacobson et al., 1994; Manz et al., 1994; Reyes et al., 2002; Stone et al., 2004]. On-chip electroosmotic pumping is easily incorporated into electrophoretic and chromatographic separations, and laboratories on a chip offer distinct advantages over the traditional, freestanding capillary systems. Advantages include reduced reagent use, tight control of geometry, the ability to network and control multiple channels on chip, the possibility of massively parallel analytical process on a single chip, the use of chip substrate as a heat sink (for high field separations), and the many advantages that follow the realization of a portable device [Khaledi, 1998; Stone et al. 2004]. Electrokinetic effects significantly extend the current design space of microsystems technology by offering unique methods of sample handling, mixing, separation, and detection of biological species including cells, microparticles, and molecules.

This section presents typical characteristics of an electrokinetic channel network fabricated using microlithographic techniques (see description of fabrication in the next section). Figure 10.14 shows a top view schematic of a typical microchannel fluidic chip used for capillary electrophoresis [Bruin, 2000; Manz et al., 1994; Stone et al., 2004]. In this simple example, the channels are etched on a dielectric substrate and bonded to a clear plate of the same material (e.g., coverslip). The circles in the schematic represent liquid reservoirs that connect with the channels through holes drilled through the coverslip. The parameters V_1 through V_4 are time-dependent voltages applied at each reservoir well. A typical voltage switching system may apply voltages with on/off ramp profiles of approximately 10,000 V/s or less so that the flow can often be approximated as quasi-steady.

The four-well system shown in Figure 10.14 can be used to perform an electrophoretic separation by injecting a sample from well #3 to well #2 by applying a potential difference between these wells. During this injection phase, the sample is confined, or pinched, to a small region within the separation channel by flowing solution from well #1 to #2 and from well #4 to well #2. The amount of desirable pinching is generally a tradeoff between separation efficiency and sensitivity. Ermakov et al. (2000), Alarie et al. (2000), and Bharadwaj et al. (2002) all present optimizations of the electrokinetic sample injection process. Next, the injection phase potential is deactivated and a potential is applied between well 1 and well #4 to dispense the injection plug into the separation channel and begin the electrophoretic separation. The potential between wells #1 and #2 is referred to as the separation potential. During the separation phase, potentials are applied at wells #2 and #3, which "retract," or "pull back," the solution-filled streams on either side of the separation channel. As with the pinching described above, the amount of "pull back" is a trade-off between separation efficiency and sensitivity. As discussed by Bharadwaj et al.

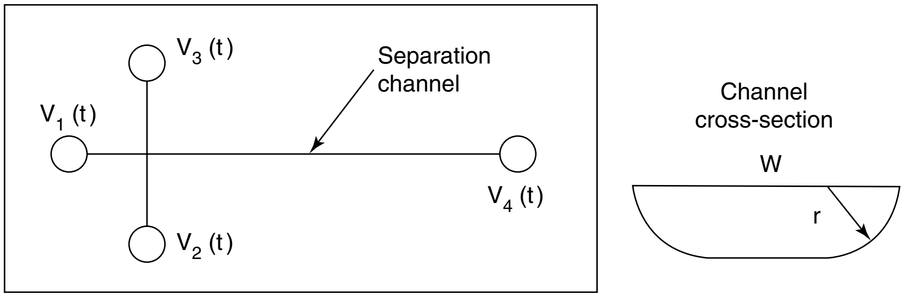


FIGURE 10.14 Schematic of a typical electrokinetic microchannel chip. V_1 through V_5 represent time-dependent voltages applied to each microchannel. The channel cross-section shown is for the (common case) of an isotropically etched glass substrate with a mask line width of $(w - 2r)$.

(2002), additional injection steps (such as a reversal of flow from well #2 to #1)for a short period prior to injection and pull back) can minimize the dispersion of sample during injection.

Figure 10.15 shows a schematic of a system that was used to perform and image an electrophoretic separation in a microfluidic chip. The microchip depicted schematically in Figure 10.15 is commercially available from Micralyne, Inc., Alberta, Canada. The width and depth of the channels are 50 μm and 20 μm respectively. The separation channel is 80 mm from the intersection to the waste well (well #4 in Figure 10.14). A high voltage switching system allows for rapid switching between the injection and separation voltages and a computer, epifluorescent microscope, and CCD camera are used to image the electrophoretic separation. The system depicted in Figure 10.15 is used to design and characterize electrokinetic injections; in a typical electrophoresis application, the CCD camera would be replaced with a point detector (e.g., a photo-multiplier tube) near well #4.

Figure 10.16 shows an injection and separation sequence of 200 μM solutions of fluorescein and Bodipy dyes (Molecular Probes, Inc., Eugene, Oregon). Images 10.16a through 10.16d are each 20 msec exposures separated by 250 msec. In Figure 16a, the sample is injected applying 0.5 kV and ground to well #3 and well #2, respectively. The sample volume at the intersection is pinched by flowing buffer from well #1 and well #4. Once a steady flow condition is achieved, the voltages are switched to inject a small sample plug into the separation channel. During this separation phase, the voltages applied at well #1 and well #4 are 2.4 kV and ground respectively. The sample remaining in the injection channel is retracted from the intersection by applying 1.4 kV to both well #2 and well #3. During the separation, the electric field strength in the separation channel is about 200 V/cm. The electrokinetic injection introduces an approximately 400 pL volume of the homogeneous sample mixture into the separation channel, as seen in Figure 10.16b. The Bodipy dye is neutral, and therefore its species velocity is identical to that of the electroosmotic flow velocity. The relatively high electroosmotic flow velocity in the capillary carries both the neutral Bodipy and negatively charged fluorescein toward well #4. The fluorescein's negative electrophoretic mobility moves it against the electroosmotic bulk flow, and therefore it travels more slowly than the Bodipy dye. This difference in electrophoretic mobilities results in a separation of the two dyes into distinct analyte bands, as seen in Figures 10.16c and 10.16d. The zeta potential of the microchannel walls for the system used in this experiment was estimated at −50 mV from the velocity of the neutral Bodipy dye [Bharadwaj and Santiago, 2002]. The inherent trade-offs between initial sample plug length, electric field,

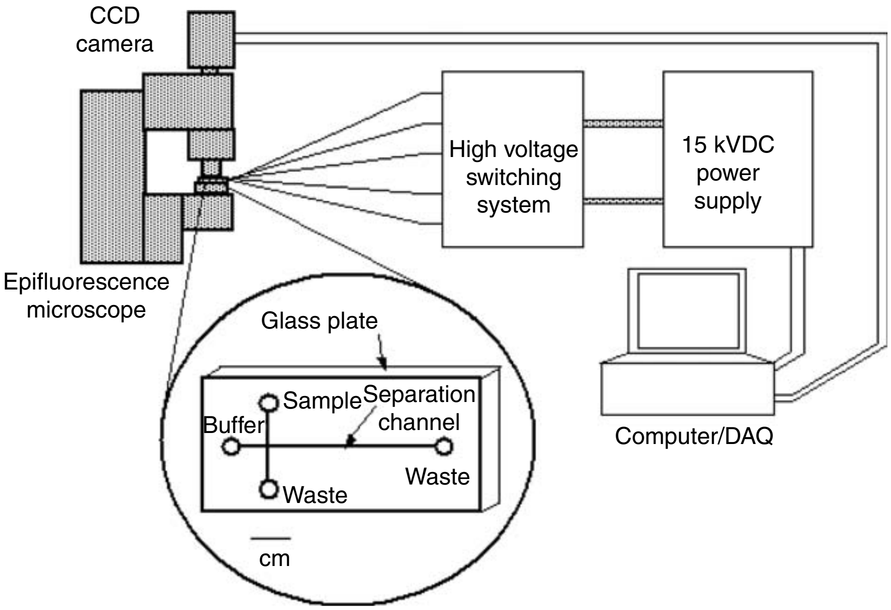


FIGURE 10.15 Schematic of microfabricated capillary electrophoresis system, flow imaging system, high voltage control box, and data acquisition computer.

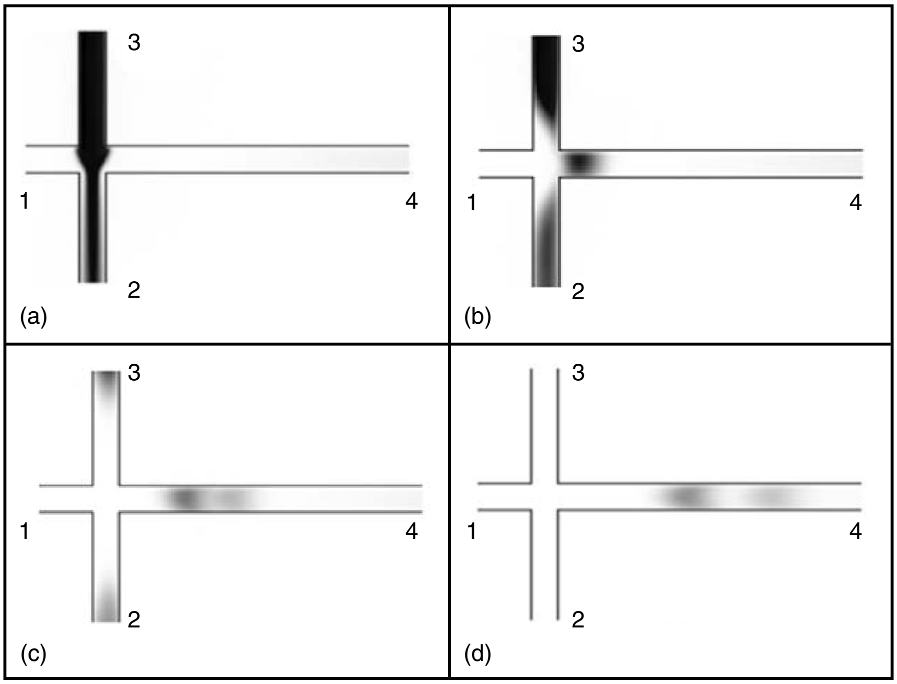


FIGURE 10.16 Separation sequence of Bodipy and fluorescein in a microfabricated capillary electrophoresis system. The channels shown are 50 µm wide and 20 µm deep. The fluoresceine images are 20 msec exposures and consecutive images are separated by 250 msec. A background image has been subtracted from each of the images, and the channel walls were drawn in for clarity. (Reprinted with permission from Bharadwaj, R., and Santiago, J.G. [2000] unpublished results, Stanford University.)

channel geometry, separation channel length, and detector characteristics are discussed in detail by Bharadwaj et al. (2002). Kirby and Hasselbrink (2004) present a review of electrokinetic flow theory and methods of quantifying zeta potentials in microfluidic systems. Ghosal (2004) presents a review of band-broadening effects in microfluidic electrophoresis.

10.2.7 Engineering Considerations: Flow Rate and Pressure of Simple Electroosmotic Flows

As we have seen, the velocity field of simple electrokinetic flow systems with thin EDLs is approximately independent of the location in the microchannel and is therefore a "plug flow" profile for any cross-section of the channel. The volume flow rate of such a flow is well approximated by the product of the electroosmotic flow velocity and the cross-sectional area of the inner capillary:

$$\mathrm{Q} = -\frac{\varepsilon\zeta EA}{\mu}. \tag{10.59}$$

For the typical case of electrokinetic systems with a bulk ion concentration in excess of about 100 µM and characteristic dimension greater than about 10 µm, the vast majority of the current carried within the microchannel is the electromigration current of the bulk liquid. For such typical flows, we can rewrite the fluid flow rate in terms of the net conductivity of the solution, σ,

$$\mathrm{Q} = -\frac{\varepsilon\zeta I}{\mu\sigma}, \tag{10.60}$$

where I is the current consumed, and we have made the reasonable assumption that the electromigration component of the current flux dominates. The flow rate of a microchannel is therefore a function of the current carried by the channel and otherwise independent of geometry.

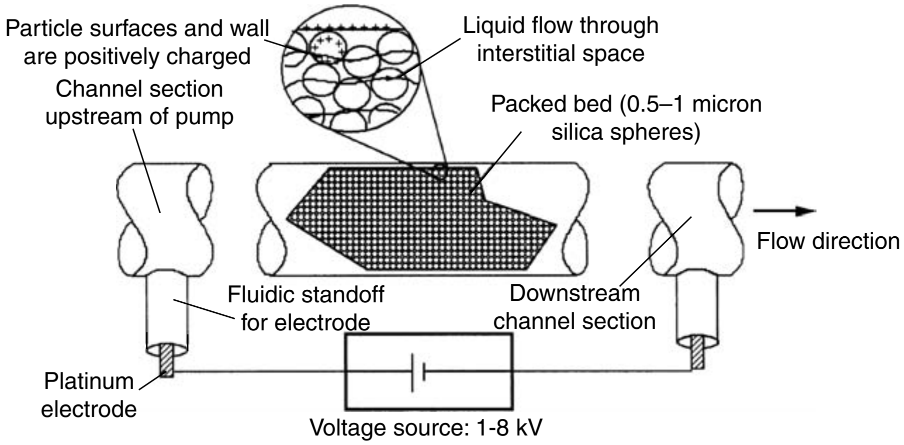

Particle surfaces and wall are positively charged
Liquid flow through interstitial space
Channel section upstream of pump
Packed bed (0.5–1 micron silica spheres)
Flow direction
Fluidic standoff for electrode
Downstream channel section
Platinum electrode
Voltage source: 1-8 kV

FIGURE 10.17 Schematic of electrokinetic pump fabricated using a glass microchannel packed with silica spheres. The interstitial spaces of the packed bed structure create a network of submicron microchannels that can be used to generate pressures in excess of 5000 psi.

Another interesting case is that of an electrokinetic capillary with an imposed axial pressure gradient. For this case, we can use Equation (10.47) to show the magnitude of the pressure that an electrokinetic microchannel can achieve. To this end, we solve Equation (10.47) for the maximum pressure generated by a capillary with a sealed end and an applied voltage ΔV, noting that the electric field and the pressure gradient can be expressed as $\Delta V/L$ and $\Delta p/L$ respectively. Such a microchannel will produce zero net flow but will provide a significant pressure gradient in the direction of the electric field (in the case of a negatively charged wall). Imposing a zero net flow condition $Q = \int_A \mathbf{u} \cdot dA = 0$ the solution for pressure generated in a thin EDL microchannel is then

$$\Delta p = -\frac{8\varepsilon\zeta\Delta V}{a^2} \tag{10.61}$$

which shows that the generated pressure will be directly proportional to voltage and inversely proportional to the square of the capillary radius. Equation (10.61) dictates that decreasing the characteristic radius of the microchannel will result in higher pressure generation. The following section discusses a class of devices designed to generate both significant pressures and flow rate using electroosmosis.

10.2.8 Electroosmotic Pumps

Electroosmotic pumps are devices that generate both significant pressure and flow rate using electroosmosis through pores or channels. A review of the history and technological development of such electroosmotic pumps is presented by Yao and Santiago (2003a). The first electroosmotic pump structure (generating significant pressure) was demonstrated by Theeuwes in 1975. Other notable contributions include that of Gan et al. (2000), who demonstrated pumping of several electrolyte chemistries; and Paul et al. (1998a) and Zeng et al. (2000), who demonstrated of order 10 atm and higher. Yao et al. (2003b) presented experimentally validated, full Poisson–Boltzmann models for porous electroosmotic pumps. They demonstrated a pumping structure less than 2 cm^3 in volume that generates 33 ml/min and 1.3 atm at 100 V.

Figure 10.17 shows a schematic of a packed-particle bed electroosmotic pump of the type discussed by Paul et al. (1998a) and Zeng et al. (2000). This structure achieves a network of submicron diameter microchannels by packing 0.5–1 micron spheres in fused silica capillaries, using the interstitial spaces in these packed beds as flow passages. Platinum electrodes on either end of the structure provide applied potentials on the order of 100 to 10,000 V. A general review of micropumps that includes sections on electroosmotic pumps is given by Laser and Santiago (2004).

10.2.9 Electrical Analogy and Microfluidic Networks

There is a strong analogy between electroosmotic and electrophoretic transport and resistive electrical networks of microchannels with long axial-to-radial dimension ratios. As described above, the electroosmotic flow rate is directly proportional to the current. This analogy holds provided that the previously described conditions for electric/velocity field similarity also hold. Therefore, Kirkoff's current and voltage laws can be used to predict flow rates in a network of electroosmotic channels given voltage at endpoint nodes of the system. In this one-dimensional analogy, all of the current, and hence all of the flow, entering a node must also leave that node. The resistance of each segment of the network can be determined by knowing the cross-sectional area, the conductivity of the liquid buffer, and the length of the segment. Once the resistances and applied voltages are known, the current and electroosmotic flow rate in every part of the network can be determined using Equation (10.60).

10.2.10 Electrokinetic Systems with Heterogenous Electrolytes

The previous sections have dealt with systems with uniform properties such as ion-concentrations (including pH), conductivity, and permittivity. However, many practical electrokinetic systems involve heterogeneous electrolyte systems. A general transport model for heterogenous electrolyte systems (and indeed for general electrohydrodynamics) should include formulations for the conservation of species, Gauss' law, and the Navier–Stokes equations describing fluid motion [Castellanos, 1998; Melcher, 1981; Saville, 1997]. The solutions to these equations can in general be a complex nonlinear coupling of these equations. Such a situation arises in a wide variety of electrokinetic flow systems. This section presents a few examples of recent and ongoing work in these complex electrokinetic flows.

10.2.10.1 Field Amplified Sample Stacking (FASS)

Sensitivity to low analyte concentrations is a major challenge in the development of robust bioanalytical devices. Field amplified sample stacking (FASS) is one robust way to carry out on-chip sample preconcentration. In FASS, the sample is prepared in an electrolyte solution of lower concentration than the background electrolyte (BGE). The low-conductivity sample is introduced into a separation channel otherwise filled with the BGE. In these systems, the electromigration current is approximately nondivergent so that $\nabla \cdot (\sigma \bar{E}) = 0$, where σ is ionic conductivity. Upon application of a potential gradient along the axis of the separation channel, the sample region is therefore a region of low conductivity (high electric field) in series with the BGE region(s) of high conductivity (low electric field). Sample ions migrate from the high-field–high-drift-velocity of the sample region to the low-field–low-drift-velocity region and accumulate, or stack, at the interface between the low and high conductivity regions.

The seminal work in the analysis of unsteady ion distributions during electrophoresis is that of Mikkers et al. (1979), who used the Kohlrausch regulating function (KRF) [Beckers and Bocek, 2000; Kohlrausch, 1897] to study concentration distributions in electrophoresis. There have been several review papers on FASS, including discussions of on-chip FASS devices, by Quirino et al. (1999), Osborn et al. (2000), and Chien (2003). FASS has been applied by Burgi and Chien (1991), Yang and Chien (2001), and Lichtenberg et al. (2001) to microchip-based electrokinetic systems. These three studies demonstrated maximum signal enhancements of 100-fold over nonstacked assays. More recently, Jung et al. (2003) demonstrated a device that avoids electrokinetic instabilities associated with conductivity gradients and achieves a 1,100-fold increase in signal using on-chip FASS. Recent modeling efforts include the work of Sounart and Baygents (2001), who developed a multicomponent model for electroosmotically driven separation processes. They performed two-dimensional numerical simulations and demonstrated that nonuniform electroosmosis in these systems causes regions of recirculating flow in the frame of the moving analyte plug. These recirculating flows can drastically reduce the efficiency of sample stacking. Bharadwaj and Santiago (2004) present an experimental and theoretical investigation of FASS dynamics. Their model analyzes dispersion dynamics using a hybrid analysis method that combines area-averaged, convective-diffusion equations with regular perturbation methods to provide a simplified equation set for FASS. They also present model validation data in the form of full-field epifluorescence images quantifying the spatial and temporal dynamics of concentration fields in FASS.

The dispersion dynamics of nonuniform electroosmotic flow FASS systems results in concentration enhancements that are a strong function of parameters such as electric field, electroosmotic mobility, diffusivity, and the background electrolyte-to-sample conductivity ratio γ. At low γ and low electroosmotic mobility, electrophoretic fluxes dominate transport and concentration enhancement increases with γ. At γ and significant electroosmotic mobilities, increases in γ increase dispersion fluxes and lower sample concentration rates. The optimization of this process is discussed in detail by Bharadwaj and Santiago (2004).

10.2.10.2 Isotachophoresis

Isothachopheresis [Everaerts et al., 1976] uses a heterogenous buffer to achieve both concentration and separation of charged ions or macromolecules. Isotachophoresis (ITP) occurs when a sample plug containing anions (or cations) is sandwiched between a trailing buffer and a leading buffer such that all the sample anions (cations) are slower than the anion (cation) in the leading buffer and faster than all the anion (cation) in the trailing buffer. When an electric field is applied in this situation, all the sample anions (cations) will rapidly form distinct zones that are arranged by electrophoretic mobility. In the case where each sample ion carries the bulk of the current in its respective zone, the KRF states that the final concentration of each ion will be proportional to its mobility. Because all anions (cations) migrate in distinct zones, current continuity ensures that they migrate at the same velocity (hence the name *isotacho*phoresis), resulting in characteristic translating conductivity boundaries. Isotachophoresis in a transient manner is used as a preconcentration technique prior to capillary electrophoresis; this combination is often referred to as ITP-CE [Hirokawa, 2003]. Isotachophoresis and ITP-CE in microdevices has been described by Kaniansky et al. (2000), Vreeland et al. (2003), Wainright et al. (2002), and Xu et al. (2003).

10.2.10.3 Isoelectric Focusing

Isoelectric focusing (IEF) is another electrophoretic technique that utilizes heterogenous buffers to achieve concentration and separation [Catsimpoolas, 1976; Righetti, 1983]. Isoelectric focusing usually employs a background buffer containing carrier ampholytes (molecules that can be either negatively charged, neutral, or positively charged depending on the local pH). The pH at which an amphoteric molecule is neutral is called the isoelectric point, or pI. Under an applied electric field, the carrier ampholytes create a pH gradient along a channel or capillary. When other amphoteric sample molecules are introduced into a channel with such a stabilized pH gradient, the samples migrate until they reach the location where the pH is equal to the pI of the sample molecule. Thus IEF concentrates initially dilute amphoteric samples and separate them by isoelectric point. Because of this behavior, IEF is often used as the first dimension of multidimensional separations. IEF and multidimensional separations employing IEF have been demonstrated in microdevices by Hofmann et al. (1999), Woei et al. (2002), Li et al. (2004), Macounova et al. (2001), and Herr et al. (2003).

10.2.10.4 Temperature Gradient Focusing

Another method of sample stacking is temperature gradient focusing (TGF), which uses electrophoresis, pressure-driven flow, and electroosmosis to focus and separate samples based on electrophoretic mobility. In TGF, an axial temperature gradient applied axially along a microchannel produces a gradient in electrophoretic velocity. When opposed by a net bulk flow, charged analytes focus at points where their electrophoretic velocity and the local, area-averaged liquid velocity sum to zero. The method has been demonstrated experimentally by Ross and Locascio (2002). A review of various various electrofocusing techniques is given by Ivory (2000).

10.2.10.5 Electrothermal Flows

A fifth important class of heterogenous electrolyte electrokinetic flows are electrothermal flows. These flows are generated by electric body forces in the bulk liquid of an electrokinetic flow system with finite temperature gradients. These flows were first described by Ramos et al. (1998) and have been analyzed by Ramos et al. (1999) and Green et al. (2000a, 2000b). Work in this area is summarized in the book by Morgan and Green (2003). These researchers were interested in steady flow-streaming-like behavior observed in microfluidic systems with patterned AC electrodes. The devices were designed for dielectrophoretic particle concentration and separation. Secondary flows in these systems are generated by the coupling of AC electric fields and temperature gradients. This coupling creates body forces that can cause order 100 micron per second liquid velocities and dominate the transport of particulates in the device. Experimental validation of these flows has been presented by Green et al. (2000b) and Wang et al. (2004). The latter work used two-color micron-resolution PIV (Santiago, 1998) to independently quantify liquid and particle velocity fields.

Ramos et al. (1998) presented a linearized theory for modeling electrothermal flows. Electrothermal forces result from net charge regions in the bulk of an electrolyte with finite temperature gradients. Temperature gradients are a result of localized Joule heating in the system and affect both local electrical conductivity σ and the dielectric permittivity ε. In the Ramos model, ion density is assumed uniform and the temperature field (and therefore the conductivity and permittivity fields) is assumed known and steady. The latter assumptions imply a low value of the thermal Peclet number (Probstein, 1994) for the flow. The general body force on a volume of liquid in this system, $\bar{f}_e$ can be derived from the divergence of the Maxwell stress tensor (Melcher, 1981) and written as

$$\bar{f}_e = \rho_E \bar{E} - 0.5|\bar{E}|^2\,\nabla\varepsilon$$

Ramos et al. (1998) assume a linear expansion of the form $\bar{E} = \bar{E}_o + \bar{E}_1$, where $\bar{E}_o$ is the applied field (satisfying $\nabla \cdot \bar{E}_o = 0$) and $\bar{E}_1$ is the perturbed field, such that $|E_o| >> |E_1|$. Assuming a sinusoidal applied field of the form $\bar{E}_o(t) = \mathrm{Re}[\bar{E}_o \exp(j\omega t)]$, and substituting this linearization into an expression of the conservation of electromigration current ($\nabla \cdot (\sigma\bar{E} = 0)$, yields

$$\nabla \cdot \bar{E}_1 = \frac{-(\nabla\sigma + j\omega\nabla\varepsilon)\cdot\bar{E}_o}{\sigma + j\omega\varepsilon},$$

where higher order terms have been neglected. The (steady, nonuniform) electric charge density is then $\rho_E = \nabla\varepsilon \cdot \bar{E}_o + \varepsilon\nabla \cdot \bar{E}_1$. This charge density can be combined with the relation for $\bar{f}_e$ above to solve for motion of the liquid using the Navier–Stokes equations. Note that this model assumes steady conductivity and permittivity fields determined solely by a steady temperature field. The electric body force field is therefore uncoupled from the motion of the liquid.

10.2.10.6 Electrokinetic Flow Instabilities

Electrokinetic instabilities are a sixth interesting example of complex electrokinetic flow in heterogenous electrolyte systems. Electrokinetic instabilities (EKI) are produced by an unsteady coupling between electric fields and conductivity gradients. Lin et al. (2004) and Chen et al. (2004) present the derivation of a model for generalized electrokinetic flow that builds on the general electrohydrodynamics framework provided by Melcher (1981). This model results in a formulation of the following form:

$$\frac{\partial\sigma}{\partial t} + \mathbf{v}\cdot\nabla\sigma = \frac{1}{Ra_e}\nabla^2\sigma,$$

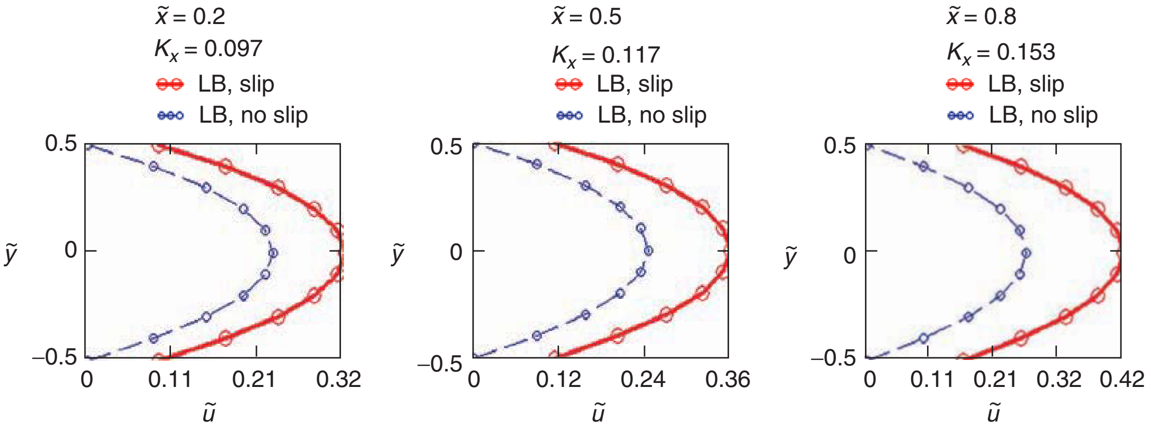
$$\nabla \cdot (\sigma\bar{E}) = 0,$$

$$\nabla \cdot \varepsilon\bar{E} = \rho_E,$$

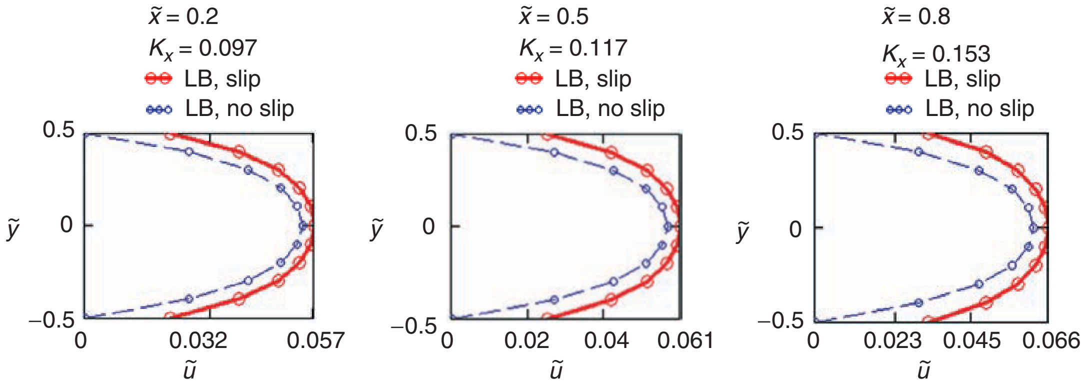
$$\nabla \cdot \mathbf{v} = 0,$$

$$\mathrm{Re}\left(\frac{\partial\mathbf{v}}{\partial t} + \mathbf{v}\cdot\nabla\mathbf{v}\right) = -\nabla p + \nabla^2\mathbf{v} + \rho_E\bar{E}.$$

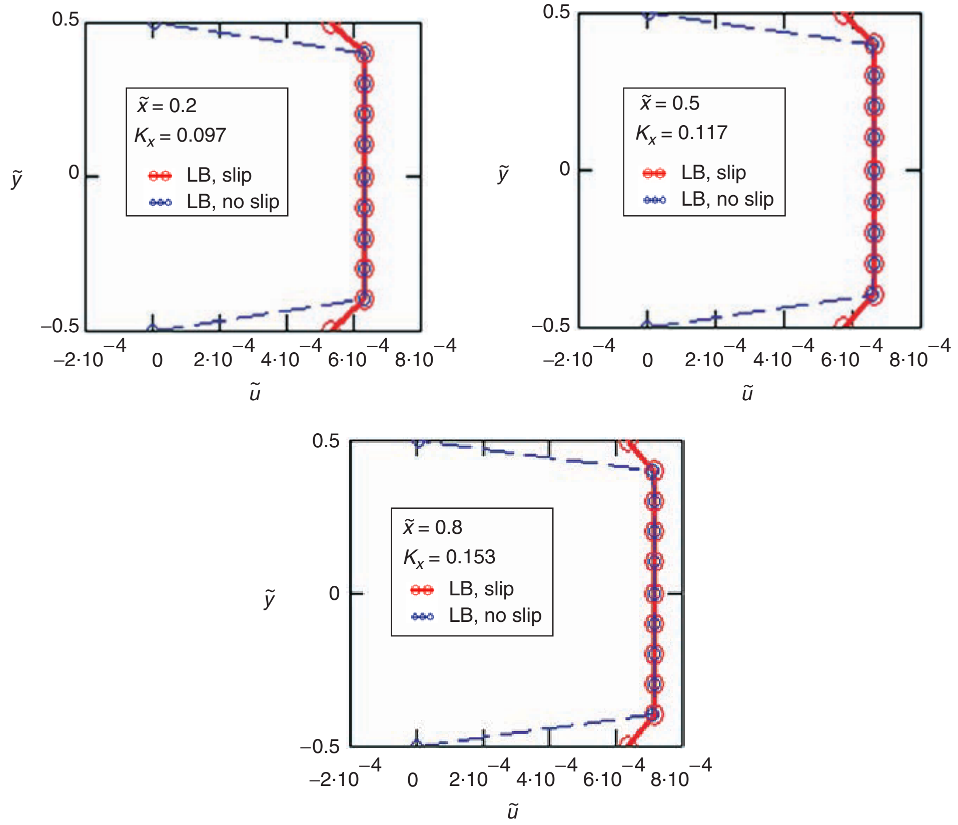
The first equation governs the development of the unsteady, nonuniform electrolyte conductivity, σ, and is derived from a summation of the charged species equations. The second equation is derived from a



COLOR FIGURE 9.5 Velocity profiles for MHD flow in a microchannel. $\mathrm{Kn_{in}} = 0.088$, $\mathrm{Kn_{out}} = 0.3$, $\mathrm{P_{in}/P_{out}} = 2.28$, $\varepsilon = H/L = 0.05$, $\alpha = 1$, $\mathrm{M} = 0.1$, $\mathrm{Ha} = 0.054$, $\mathrm{E_0} = 0$.

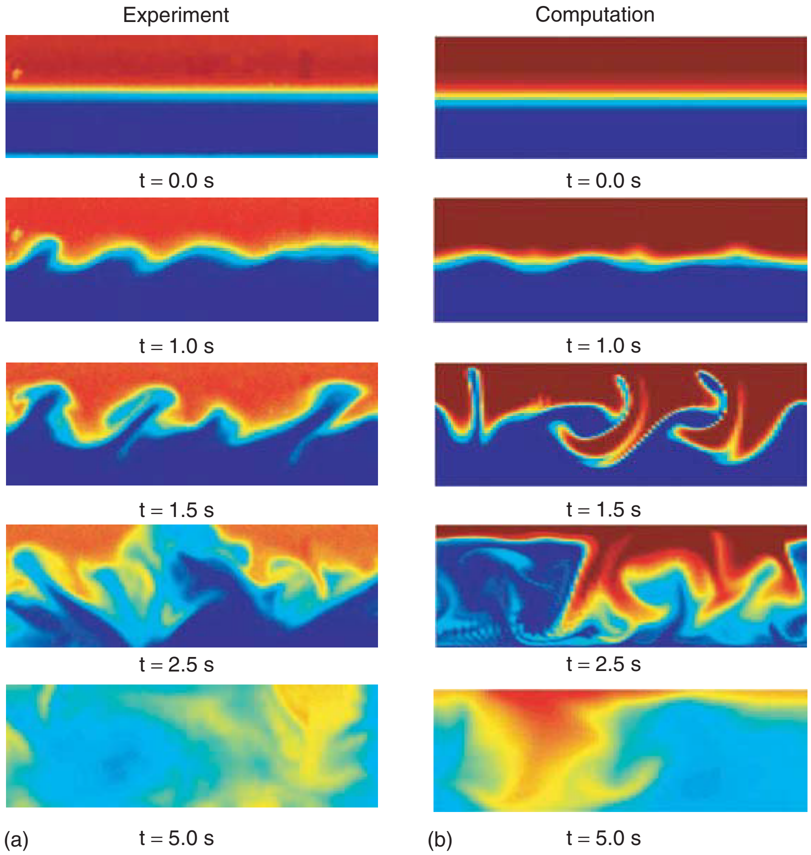


COLOR FIGURE 9.6 Velocity profiles for MHD flow in a microchannel. $\mathrm{Kn_{in}} = 0.088$, $\mathrm{Kn_{out}} = 0.3$, $\mathrm{P_{in}/P_{out}} = 2.28$, $\varepsilon = H/L = 0.05$, $\alpha = 1$, $\mathrm{M} = 0.1$, $\mathrm{Ha} = 5.4$, $\mathrm{E_0} = 0$.

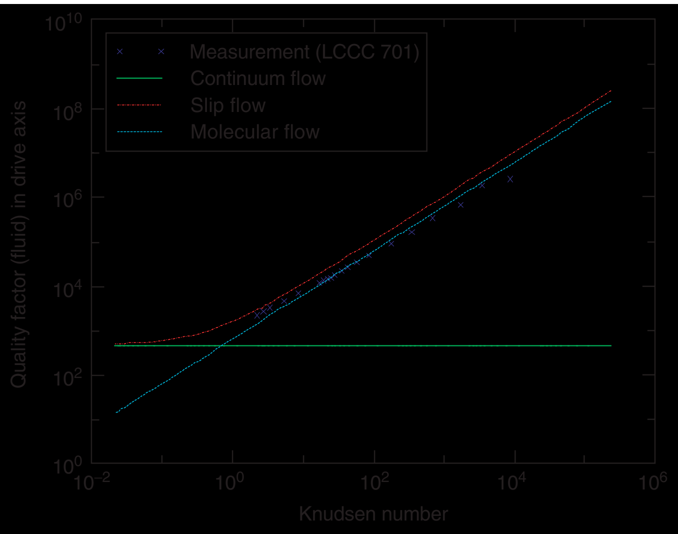


COLOR FIGURE 9.7 Velocity profiles for MHD flow in a microchannel. $Kn_{in} = 0.088$, $Kn_{out} = 0.3$, $P_{in}/P_{out} = 2.28$, $\varepsilon = H/L = 0.05$, $\alpha = 1$, $M = 0.1$, $Ha = 54$, $E_0 = 0$.

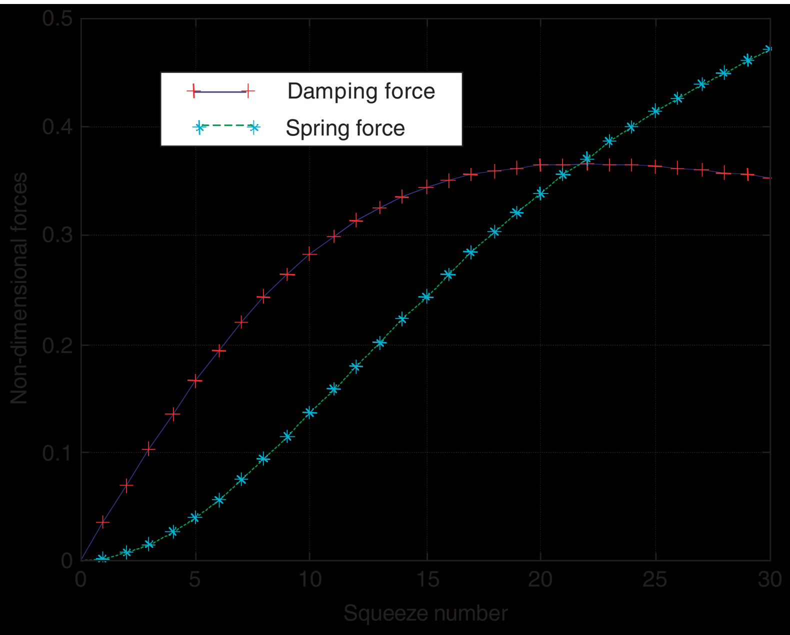
COLOR FIGURE 10.2 Blood sample cartridge using microfluidic channels. (Reprinted with permission from i-Stat, East Windsor, NJ, 2000.)



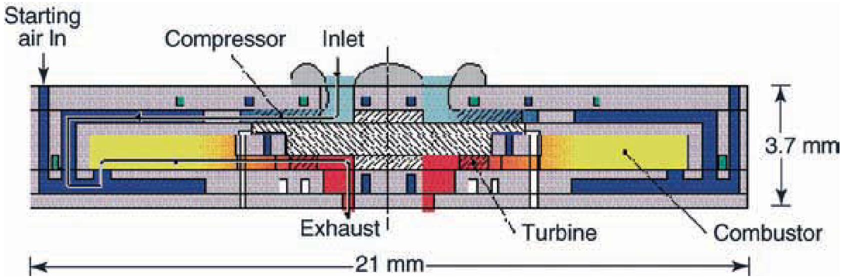
COLOR FIGURE 10.18 Time evolution of electrokinetic flow instability: (a) Experimental data of instability mixing of HEPES buffered 50 µS/cm (red) and 5 µS/cm (blue) conductivity streams [Lin et al., 2004]. At time t = 0.0 sec, a static electric field of E = 50,000 V/m is applied in the (horizontal) streamwise direction perpendicular to the initial conductivity gradients. Image area is 1 mm in the vertical direction and 3.6 mm in the streamwise direction. Channel depth (into the page) is 100 µm. Small amplitude waves quickly grow and lead to rapid stirring of the initially distinct buffer streams. (b) Reproduction of dynamics from simplified, 2-D nonlinear numerical computations. The numerical model well reproduces features of the instability observed in experiments, including wave number and time scale. Details of this model are given by Lin et al. [2004].



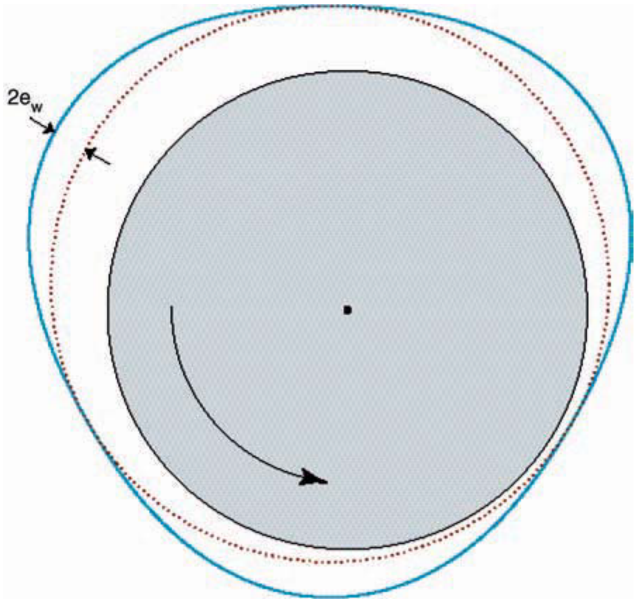
COLOR FIGURE 11.2 Theory and measurements of Couette damping in a tuning fork gyro (Kwok et al. [2005]). Note that in the high Knudsen number limit, the free molecular approximation predicts the damping more closely, but that the slip-flow model, though totally inappropriate at this high *Kn* level, is not too far from the experimental measurements.



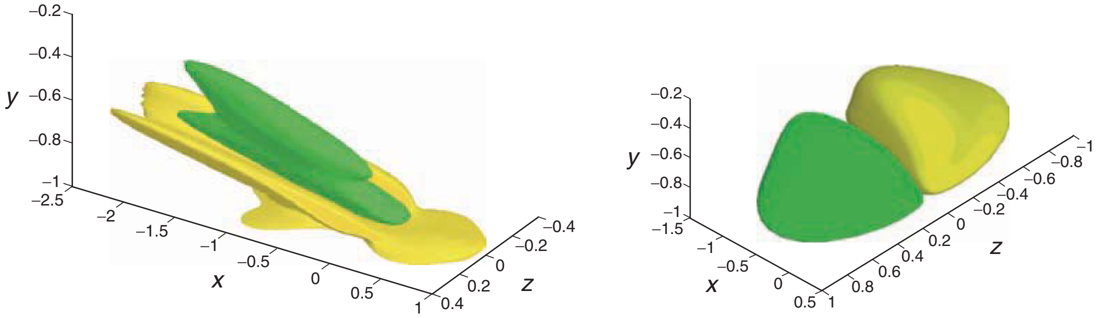
COLOR FIGURE 11.5 Solutions to the squeeze-film equation for a rectangular plate. The stiffness and damping coefficients are presented as functions of the modified squeeze number, which includes a correction due to first-order rarefaction effects [Blech, 1983; Kwok et al., 2005].



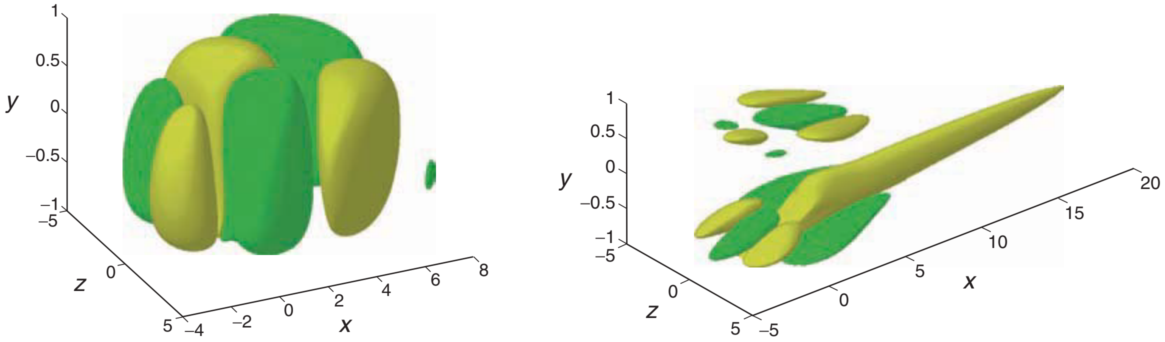
COLOR FIGURE 11.6 Schematic of the MIT Microengine, showing the air path through the compressor, combustor, and turbine. Forward and aft thrust bearings located on the centerline hold the rotor in axial equilibrium, while a journal bearing around the rotor periphery holds the rotor in radial equilibrium.



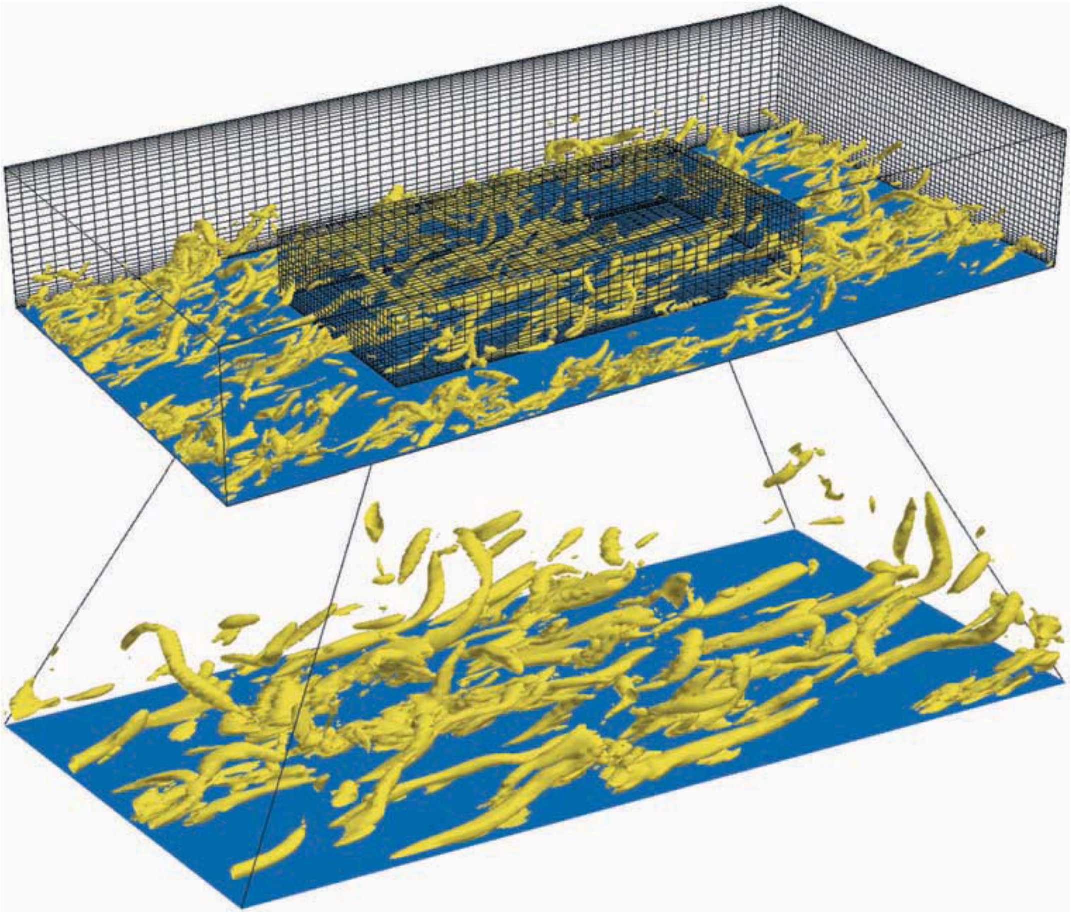
COLOR FIGURE 11.13 Geometry of a wave bearing, with the clearance greatly exaggerated for clarity [Piekos, 2000].



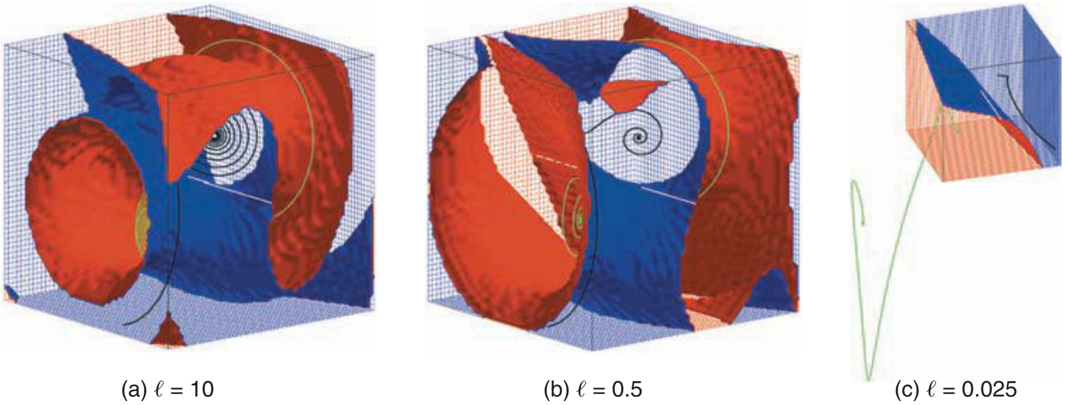
COLOR FIGURE 15.8 Localized controller gains relating the state estimate $\hat{\mathbf{x}}$ inside the domain to the control forcing $\mathbf{u}$ at the point $\{x = 0, y = -1, z = 0\}$ on the wall. Visualized are a positive and negative isosurface of the convolution kernels for (left) the wall-normal component of velocity and (right) the wall-normal component of vorticity. (Högberg, M., Bewley, T.R., and Henningson, D.S. (2003) "Linear Feedback Control and Estimation of Transition in Plane Channel Flow," *J. Fluid Mech.* **481**, pp. 149–75. Reprinted with permission from Elsevier Science.)



COLOR FIGURE 15.9 Localized estimator gains relating the measurement error $(\mathbf{y} - \hat{\mathbf{y}})$ at the point $\{x = 0, y = -1, z = 0\}$ on the wall to the estimator forcing terms $\mathbf{v}$ inside the domain. Visualized are a positive and negative isosurface of the convolution kernels for (left) the wall-normal component of velocity and (right) the wall-normal component of vorticity. (Högberg, M., Bewley, T.R., and Henningson, D.S. (2003) "Linear Feedback Control and Estimation of Transition in Plane Channel Flow," *J. Fluid Mech.* **481**, pp. 149–75. Reprinted with permission from Elsevier Science.)

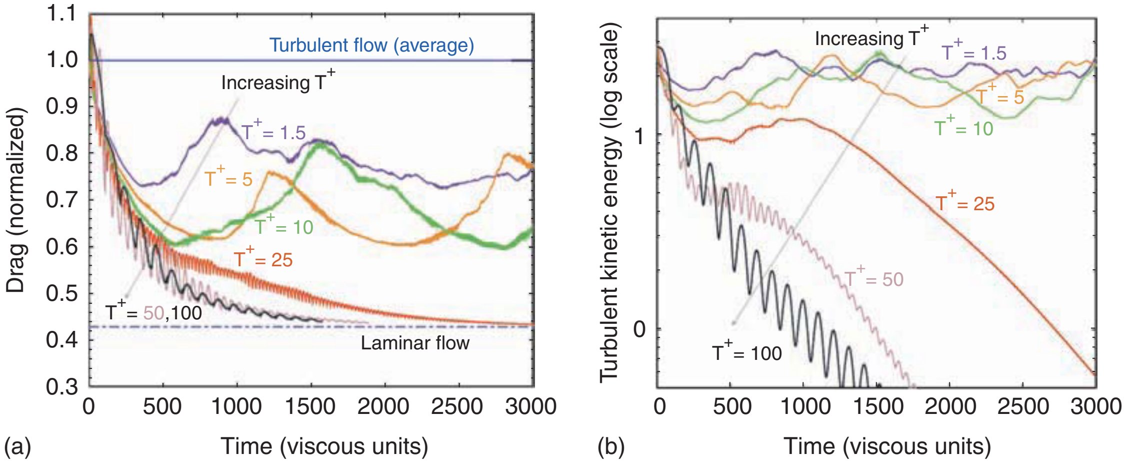


COLOR FIGURE 15.11 Visualization of the coherent structures of uncontrolled near-wall turbulence at $Re_\tau = 180$. Despite the geometric simplicity of this flow (see Figure 15.1), it is phenomenologically rich and is characterized by a large range of length scales and time scales over which energy transport and scalar mixing occur. The relevant spectra characterizing these complex nonlinear phenomena are continuous over this large range of scales, thus such flows have largely eluded accurate description via dynamic models of low state dimension. The nonlinearity, the distributed nature, and the inherent complexity of its dynamics make turbulent flow systems particularly challenging for successful application of control theory. (Simulation by Bewley, T.R., Moin, P., and Temam, R. (2001) *J. Fluid Mech.* Reprinted with permission of Cambridge University Press.)

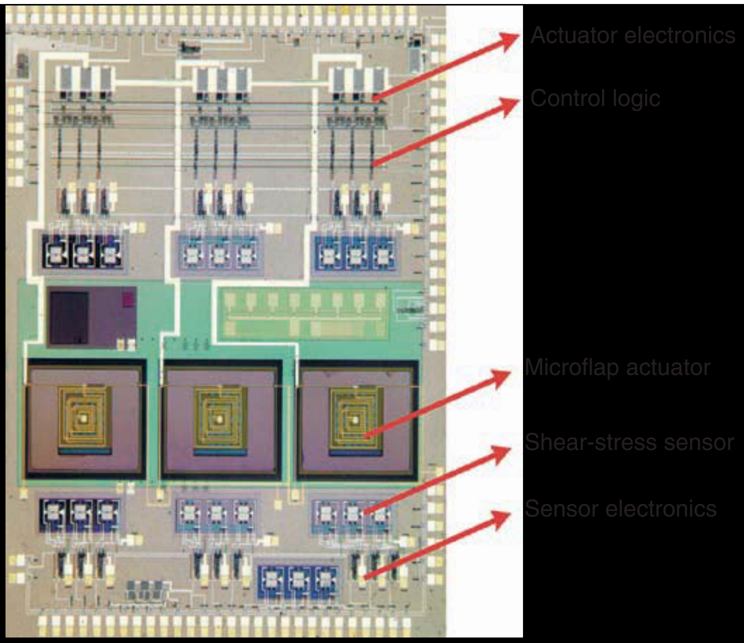


(a) $\ell = 10$ (b) $\ell = 0.5$ (c) $\ell = 0.025$

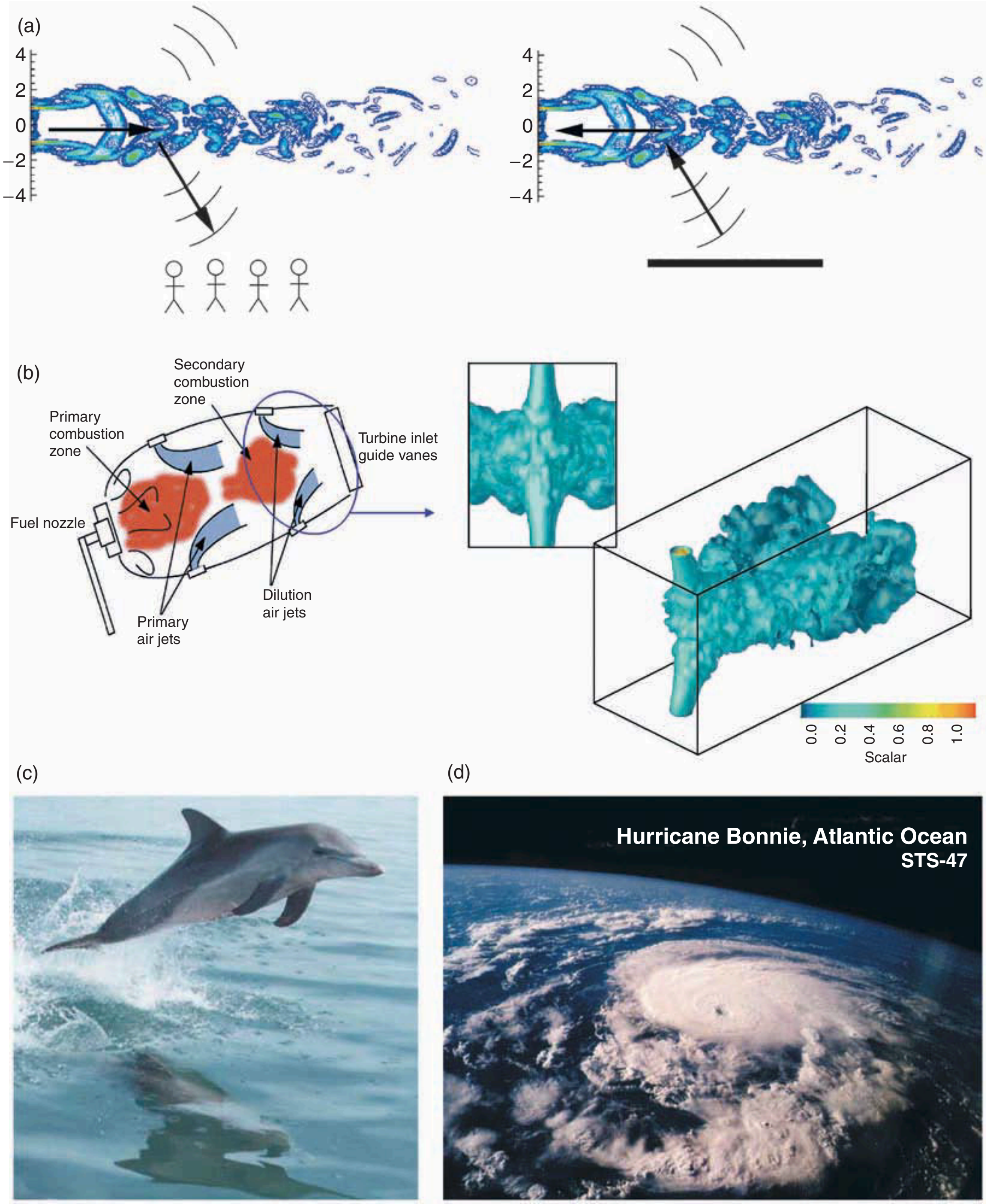
COLOR FIGURE 15.12 Example of the spectacular failure of linear control theory to stabilize a simple nonlinear chaotic convection system governed by the Lorenz equation. Plotted are the regions of attraction to the desired stationary point (blue) and to an undesired stationary point (red) in the linearly controlled nonlinear system, and typical trajectories in each region (black and green, respectively). The cubical domain illustrated is $\Omega = (-25, 25)^3$ in all subfigures. For clarity, different viewpoints are used in each subfigure. (Reprinted with permission from Bewley, T.R. (1999) *Phys. Fluids* 11, 1169–86. Copyright 1999, American Institute of Physics.)



COLOR FIGURE 15.14 Performance of optimized blowing/suction controls for formulations based on minimizing $J_o(\phi)$, case c (see Section 15.9.1.2), as a function of the optimization horizon T^+. The direct numerical simulations of turbulent channel flow reported here were conducted at $Re_\tau = 100$. For small optimization horizons ($T^+ = O(1)$, sometimes called the "suboptimal approximation"), approximately 20% drag reduction is obtained, a result that can be obtained with a variety of other approaches. For sufficiently large optimization horizons ($T^+ \geq 25$), the flow is returned to the region of stability of the laminar flow, and the flow relaminarizes with no further control effort required. No other control algorithm tested in this flow to date has achieved this result with this type of flow actuation. (From Bewley, T.R., Moin, P., and Temam, R. (2001) *J. Fluid Mech.*, to appear. Reprinted with permission of Cambridge University Press.)



COLOR FIGURE 15.20 A MEMS tile integrating sensors, actuators and control logic for distributed flow control applications. (Developed by Professors Chih-Ming Ho, UCLA, and Yu-Chong Tai, Caltech.)



COLOR FIGURE 15.22 Future interdisciplinary problems in flow control amenable to adjoint-based analysis: (a) minimization of sound radiating from a turbulent jet (simulation by Prof. Jon Freund, UCLA), (b) maximization of mixing in interacting cross-flow jets (simulation by Dr. Peter Blossey, UCSD) [Schematic of jet engine combustor is shown at left. Simulation of interacting cross-flow dilution jets, designed to keep the turbine inlet vanes cool, are visualized at right.], (c) optimization of surface compliance properties to minimize turbulent skin friction, and (d) accurate forecasting of inclement weather systems.

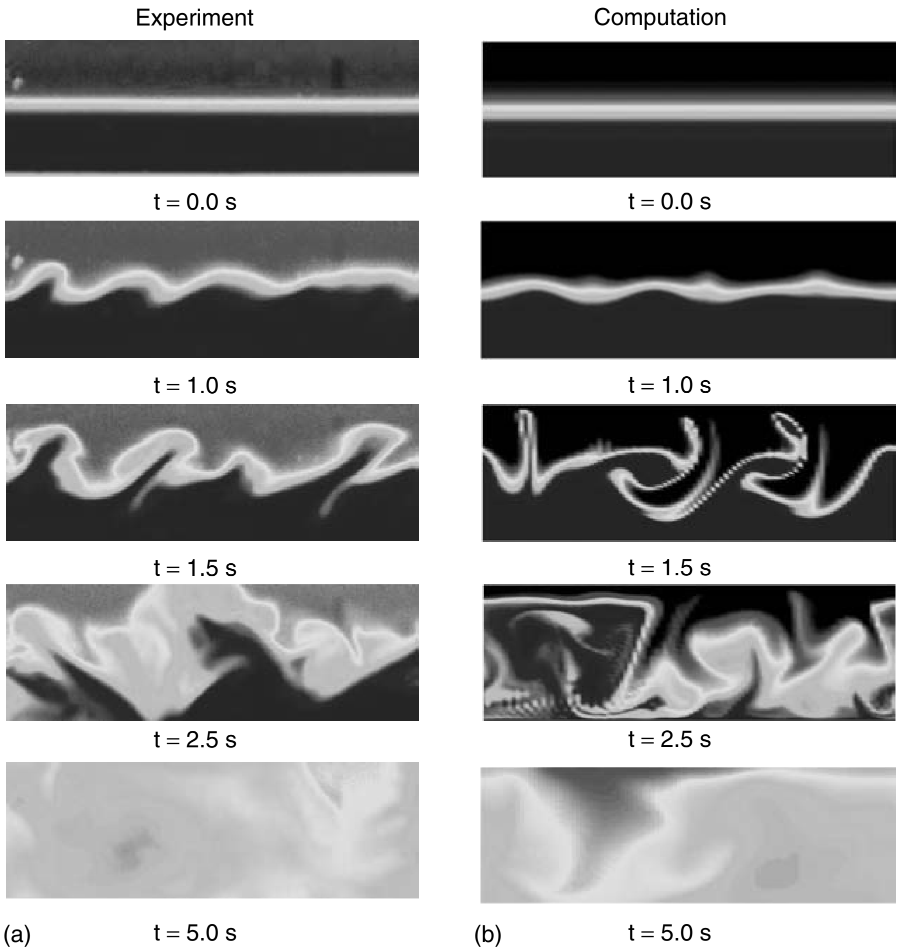


FIGURE 10.18 (**See color insert following page 10-34.**) Time evolution of electrokinetic flow instability: (a) Experimental data of instability mixing of HEPES buffered 50 μS/cm (red) and 5 μS/cm (blue) conductivity streams [Lin et al., 2004]. At time t = 0.0 sec, a static electric field of E = 50,000 V/m is applied in the (horizontal) streamwise direction perpendicular to the initial conductivity gradients. Image area is 1 mm in the vertical direction and 3.6 mm in the streamwise direction. Channel depth (into the page) is 100 μm. Small amplitude waves quickly grow and lead to rapid stirring of the initially distinct buffer streams. (b) Reproduction of dynamics from simplified, 2-D nonlinear numerical computations. The numerical model well reproduces features of the instability observed in experiments, including wave number and time scale. Details of this model are given by Lin et al. [2004].

conservation of net charge in the limit of fast charge relaxation. As discussed in detail by Lin et al. (2004), the relaxed charge assumption is consistent with the net neutrality approximation and leads to the condition that electromigration current is at all times conserved. The third equation is Gauss' law, and the last two are the Navier–Stokes equations describing fluid velocity with an electrostatic body force of the form $\rho_E\overline{E}$. Electrokinetic flow instabilities associated with electrokinetic flows with conductivity gradients arise from a coupling of these equations. This coupling results in an electric body force (per unit volume) of the form $(\varepsilon\overline{E} \cdot \nabla\sigma)\overline{E}$, which occurs in regions where local electric field is parallel to the conductivity gradient. Electrokinetic flows become unstable when the ratio of the characteristic electric body force to the viscous force in the flow exceeds a critical value [Chen et al., 2004; Lin et al., 2004]. These flows are unstable even in the limit of vanishing Reynolds number.

Electrokinetic instabilities have been experimentally demonstrated, for various geometric configurations by Oddy et al. (2001), Lin et al. (2004), and Chen et al. (2002, 2004). Figure 10.18 shows both an experimental visualization and a numerical model of a temporal instability in a microchannel with a conductivity gradient initially orthogonal to the applied electric field [Oddy, 2001]. Chen et al. (2004) show, in a slightly different geometry with much shallower channel (11 micron depth), a convective electrokinetic instability in which spatial growth of disturbances is observed. In both of these experiments threshold electric fields are observed above which the flow becomes unstable and rapid stirring and mixing occur. Together, the work of Lin et al. (2004) and Chen et al. (2004) describes the basic mechanism behind electrokinetic instabilities and identifies the critical electric Rayleigh numbers that govern the onset of the instability. Lin et al. (2004) present linear models for temporal electrokinetic instabilities, a nonlinear numerical model of the instability, and validation experiments in a long, thin microchannel structure. Chen et al. (2004) also present experimental results and describe the convective nature of the instability. The latter work identifies the electroviscous-to-electroosmotic-velocity ratio as the critical value that demarcates the boundary between convective and absolute instability.

In general, electrokinetic instabilities and flows with unsteady, nonuniform body forces due to couplings between electric fields and conductivity and permittivity gradients are directly relevant to a variety of on-chip electrokinetic systems. Such complex flow systems include field amplified sample stacking devices [Bharadwaj and Santiago, 2004; Chien, 2003; Jung et al., 2003]; low-Reynolds number micromixing [Oddy et al., 2001]; multidimensional assay systems [Herr et al., 2003]; and dielectrophoretic devices [Morgan and Green, 2003]. In general, this complex coupling of applied field and heterogenous electrolyte properties may occur in any electrokinetic system with imperfectly specified sample chemistry.

10.2.11 Practical Considerations

A few practical considerations should be considered in the design, fabrication, and operation of electrokinetic microfluidic systems. These considerations include the dimensions of the system, the choice of liquid and buffer ions, the field strengths used, and the characteristics of the flow reservoirs and interconnects. A few examples of these design issues are given here.

In the case of microchannels used to generate pressure, Equation (10.60) shows that a low liquid conductivity is essential for increasing thermodynamic efficiency of an electrokinetic pump because Joule heating is an important contributor to dissipation [Yao and Santiago, 2003a; Zhao and Liao, 2002]. In electrokinetic systems for chemical analysis, on the other hand, the need for a stable pH requires a finite buffer strength, and typical buffer strengths are in the 1–100 mM range. The need for a stable pH therefore often conflicts with a need for high fields [Bharadwaj et al., 2002] to achieve high efficiency separations because of the effects of Joule heating of the liquid.

Joule heating of the liquid in electrokinetic systems can be detrimental in two ways. First, temperature gradients within the microchannel cause a nonuniformity in the local mobility of electrophoretic particles because the local viscosity is a function of temperature. This nonuniformity in mobility results in a dispersion associated with the transport of electrophoretic species [Bosse and Arce, 2000; Grushka et al., 1989; Knox, 1988]. The second effect of Joule heating is the rise in the absolute temperature of the buffer. This temperature rise results in higher electroosmotic mobilities and higher sample diffusivities. In microchip electrophoretic separations, the effect of increased diffusivity on separation efficiency is somewhat offset by the associated decrease in separation time. In addition, the authors have found that an important limitation to the electric field magnitude in microchannel electrokinetics is that elevated temperatures and the associated decreases in gas solubility of the solution often result in the nucleation of gas bubbles in the channel. This effect of driving gas out of solution typically occurs well before the onset of boiling and can be catastrophic to the electrokinetic system because gas bubbles grow and eventually break the electrical circuit required to drive the flow. This effect can be reduced by outgassing of the solution and is, of course, a strong function of the channel geometry, buffer conductivity, and the thermal properties of the substrate material.

Another important consideration in any microfluidic device is the design and implementation of macro-to-micro fluidic interconnects. Practical implementations of fluidic interconnects span a wide range of complexity. One common practice (though rarely mentioned in publications) is to simply glue (e.g., with epoxy and by hand) trimmed plastic pipette tips or short glass tubes around the outlet port on a fluidic chip to form an end-channel reservoir. Some systems, such as those described by Gray et al.

(1999) for silicon microfluidic chips, incorporate especially microfabricated structures for integrated, low-dead-volume connections. Krulevitch (2002) describes a set of interconnects applicable to silicone rubber fluidic systems. Still other systems use Nanoport interconnect fittings commercially available from Upchurch Scientific. Fluidic interconnects are clearly an area that would benefit from an informed review of the various advantages and disadvantages of common schemes. These factors include ease of assembly, typical fabrication yield, dead volume, ability to deal with electrolytic reaction products, and pressure capacity.

Lastly, an important consideration in electrokinetic experiments is the inadvertent application and/or generation of pressure gradients in the microchannel. Probably the most common cause of this is a mismatch in the height of the fluid level at the reservoirs. Although there may not be a mismatch of fluid level at the start of an experiment, the flow rates created by electroosmotic flow may eventually create a fluid level mismatch. Also, the fluid level in each reservoir, particularly in reservoirs of 1 mm diameter or less, may be affected by electrolytic gas generation at each electrode. Because electroosmotic flow rate scales as channel diameter squared, whereas pressure-driven flow scales as channel diameter to the fourth power, this effect is greatly reduced by decreasing the characteristic channel diameter. Another common method of reducing this pressure head effect is to increase the length of the channel for a given cross-section. This length increase, of course, implies an increase in operating voltages to achieve the same flow rate. A second source of pressure gradients is a nonuniformity in the surface charge in the channel. An elegant closed-form solution for the flow in a microchannel with arbitrary axial zeta potential distribution is presented by Anderson and Idol (1985). Herr et al. (2000) visualized this effect and offered a simple analytical expression to the pressure-driven flow components associated with zeta potential gradients in fully developed channel flows.

10.3 Summary and Conclusions

In microchannels, the flow of a liquid differs fundamentally from that of a gas, primarily due to the effects of compressibility and potential rarefaction in gases. Significant differences from continuum macroscale theories have been observed. If experiments are performed with sufficient control and care in channels with dimensions of the order of tens of microns or larger, the friction factors measured in the range of accepted laminar flow behavior (i.e., $\mathrm{Re} < 2000$) agree with classical continuum hydrodynamic theory to within small or negligible differences [Sharp and Adrian, 2004], and the transition to turbulence occurs at or near the nominally accepted values for both rectangular and circular microchannels [Liu and Garimella, 2004; Sharp and Adrian, 2004].

The possibility cannot be ruled out, however, that some physical effects such as roughness or electrical charge effects are causing a deviation from conventional flow results in certain experiments. Observed differences may also be due to imperfections in the flow system of the experiment, and because imperfections may well occur in real engineering systems, it is essential to understand the sources of the observed discrepancies in order to avoid them, control them, or factor them into the designs. Measurement techniques for liquid flows are advancing quickly, both as macroscale methods are adapted to these smaller scales and as novel techniques are being developed. Further insight into phenomena present in the microscale flows, including those due to imperfections in channels or flow systems, is likely to occur rapidly given the evolving nature of the measurement techniques. Complex, nonlinear channels can be used efficiently to design for functionality.

Electrokinetics is a convenient and easily controlled method of achieving sample handling and separations on a microchip. Because the body force exerted on the liquid is typically limited to a region within a few nanometers from the wall, the resulting profiles, in the absence of imposed pressure-gradients, are often plug-like for channel dimensions greater than about 10 μm and ion concentrations greater than about 10 μM. For simple electroosmotic flows with thin EDLs, low Reynolds number, uniform surface charge, and zero imposed pressure gradients, the velocity field of these systems is well approximated by potential flow theory. This significant simplification can, in many cases, be used to predict and optimize the performance of electrokinetic systems. Further, electrokinetics can be used to generate large pressures

(>20 atm) on a microfabricated device. In principle, the handling, rapid mixing, and separation of solutes in less than 1 pL sample volumes should be possible using electrokinetic systems built with current microfabrication technologies.

References

Adamson, A.W., and Gast, A.P. (1997) "Physical Chemistry of Surfaces," 6th ed., John Wiley & Sons, Inc., New York.

Alarie, J.P., Jacobson, S.C., Culbertson, C.T., et al. (2000) "Effects of the Electric Field Distribution on Microchip Valving Performance," *Electrophoresis* **21**, pp. 100–6.

Anderson, J.L., and Idol, W.K. (1985) "Electroosmosis through Pores with Nonuniformly Charged Walls," *Chem. Eng. Commn.* **38**, pp. 93–106.

Arkilic, E.B., Schmidt, M.A., and Breuer, K.S. (1997) "Gaseous Slip Flow in Long Microchannels," *J. MEMS* **6**, pp. 167–78.

Arulanandam, S., and Li, D. (2000) "Liquid Transport in Rectangular Microchannels by Electroosmotic Pumping," *Colloid. Surface. A* **161**, pp. 89–102.

Auroux, P. A., Iossifidis, D., Reyes, D. R., and Manz, A. (2002) "Micro Total Analysis Systems: 2. Analytical Standard Operations and Applications," *Anal. Chem.* **74**, pp. 2637–52.

Baker, D.R. (1995) "Capillary Electrophoresis," in *Techniques in Analytical Chemistry Series,* John Wiley & Sons, Inc., New York.

Beckers, J.L., and Bocek, P. (2000) "Sample Stacking in Capillary Zone Electrophoresis: Principles, Advantages, and Limitations," *Electrophoresis* **21**, pp. 2747–67.

Berger, M., Castelino, J., Huang, R., Shah, M., and Austin, R.H. (2001) "Design of a Microfabricated Magnetic Cell Separator," *Electrophoresis* **22**, pp. 3883–92.

Bharadwaj, R., and Santiago, J.G. (2005) "Dynamics of Field Amplified Sample Stacking," in press, *J. Fluid Mech.*

Bharadwaj, R., Santiago, J.G., and Mohammadi, B. (2002) "Design and Optimization of On-Chip Electrophoresis," *Electrophoresis* **23**, pp. 2729–44.

Bianchi, F., Ferrigno, R., and Girault, H.H. (2000) "Finite Element Simulation of an Electroosmotic-Driven Flow Division at a T-Junction of Microscale Dimensions," *Anal. Chem.* **72**, pp. 1987–93.

Blankenstein, G., and Larsen, U.D. (1998) "Modular Concept of a Laboratory on a Chip for Chemical and Biochemical Analysis," *Biosens. Bioelectron.* **13**, pp. 427–38.

Bosse, M.A., and Arce, P. (2000) "Role of Joule Heating in Dispersive Mixing Effects in Electrophoretic Cells: Convective-Diffusive Transport Aspects," *Electrophoresis* **21**, pp. 1026–33.

Brandner, J., Fichtner, M., Schygulla, U., and Schubert, K. (2000) "Improving the Efficiency of Micro Heat Exchangers and Reactors," in *Proc. 4th Int'l. Conf. Microreaction Technology*, AIChE, 5–9 March, Atlanta, Georgia, pp. 244–49.

Branebjerg, J., Fabius, B., and Gravesen, P. (1995) "Application of Miniature Analyzers: From Microfluidic Components to TAS," in *Micro Total Analysis Systems*, A. van den Berg and P. Bergveld, eds., Kluwer Academic Publishers, Dordrecht, pp. 141–51.

Branebjerg, J., Gravesen, P., Krog, J.P., and Nielsen, C.R. (1996) "Fast Mixing by Lamination," in *Proc. 9th Annual Workshop of Micro Electro Mechanical Systems*, San Diego, California, 11–15 February, pp. 441–46.

Bridgman, P.W. (1923) "The Thermal Conductivity of Liquids Under Pressure," American Academy of Arts and Sciences 59, pp. 141–59.

Brody, J.P., Yager, P., Goldstein, R.E., and Austin, R.H. (1996) "Biotechnology at Low Reynolds Numbers," *Biophys. J.* **71**, pp. 3430–41.

Bruin, G.J.M. (2000) "Recent Developments in Electrokinetically Driven Analysis on Microfabricated Devices," *Electrophoresis* **21**, pp. 3931–51.

Brutin, D., and Tadrist, L. (2003) "Experimental Friction Factor of a Liquid Flow in Microtubes," *Phys. Fluids* **15**, pp. 653–61.

Burgi, D.S., and Chien, R.L. (1991) "Optimization of Sample Stacking for High Performance Capillary Electrophoresis," *Anal. Chem.* **63**, pp. 2042–47.

Burgreen, D., and Nakache, F.R. (1964) "Electrokinetic Flow in Ultrafine Capillary Slits," *J. Phys. Chem.* **68**, pp. 1084–91.

Castellanos, A. (1998) *Electrohydrodynamics*, New York, Springer-Verlag Wien.

Catsimpoolas, N. (1976) *Isoelectric Focusing*, New York, Academic Press.

Celata, G.P., Cumo, M., Guglielmi, M., and Zummo, G. (2002) "Experimental Investigation of Hydraulic and Single-Phase Heat Transfer in a 0.130-mm Capillary Tube," *Microscale Thermophys. Eng.* **6**, p. 85–97.

Chen, C.-H., Lin, H., Santiago, J.G., and Lele, S.K. (2005) "Convective and absolute Electrokinetic Flow Instability," with conductivity gradients *J. Fluid Mech*, **524**, pp. 263–303.

Chen, Z., Milner, T.E., Dave, D., and Nelson, J.S. (1997) "Optical Doppler Tomographic Imaging of Fluid Flow Velocity in Highly Scattering Media," *Opt. Lett.* **22**, pp. 64–66.

Chien, R. (2003) "Sample Stacking Revisited: A Personal Perspective," *Electrophoresis* **24**, pp. 486–97.

Cho, S.K., and Kim, C.-J. (2003), "Particle Separation and Concentration Control for Digital Microfluidic Systems," *Proc Sixteenth Ann. Conf. on MEMS, MEMS-03*, 19–23 January, Kyoto, Japan, pp. 686–89.

Choi, C.-H., Westin, K.J.A., and Breuer, K.S. (2003) "Apparent Slip Flows in Hydrophilic and Hydrophobic Microchannels," *Phys. Fluids* **15**, pp. 2897–2902.

Choi, S.B., Barron, R.F., and Warrington, R.O. (1991) "Fluid Flow and Heat Transfer in Microtubes," in DSC-Vol. 32, Micromechanical Sensors, Actuators and Systems, ASME Winter Annual Meeting, Atlanta, Georgia, pp. 123–34.

Chou, C.-F., Austin, R.H., Bakajin, O., Tegenfeldt, J.O., Castelino, J.A., Chan, S.S., Cox, E.C., Craighead, H., Darnton, N., Duke, T., Han, J., and Turner, S. (2000) "Sorting Biomolecules with Microdevices," *Electrophoresis* **21**, pp. 81–90.

Cui, H.-H., Silber-Li, Z.-H., and Zhu, S.-N. (2004) "Flow Characteristics of Liquids in Microtubes Driven by a High Pressure," *Phys. Fluids* **16**, pp. 1803–10.

Cummings, E.B., and Singh, A.K. (2003) "Dielectrophoresis in Microchips Containing Arrays of Insulating Posts: Theoretical and Experimental Results," *Anal. Chem.* **75**, pp. 4724–31.

Cummings, E.B., Griffiths, S.K., and Nilson, R.H. (1999) "Irrotationality of Uniform Electroosmosis," in *SPIE Conference on Microfluidic Devices and Systems II*, 20–22 September, Santa Clara, California, 3877, pp. 180–89.

Cummings, E.B., Griffiths, S.K., Nilson, R.H., et al. (2000) "Conditions for Similitude Between the Fluid Velocity and Electric Field in Electroosmotic Flow," *Anal. Chem.* **72**, pp. 2526–32.

Darbyshire, A.G., and Mullin, T. (1995) "Transition to Turbulence in Constant-Mass-Flux Pipe Flow," *J. Fluid Mech.* **289**, pp. 83–114.

Devasenathipathy, S., and Santiago, J.G. (2000) unpublished results, Stanford University, Stanford, California, October.

Devasenathipathy, S., Santiago, J.G., and Takehara, K. (2002) "Particle Tracking Techniques for Electrokinetic Microchannel Flows," *Anal. Chem.* **74**, pp. 3704–13.

Devasenathipathy, S., Santiago, J.G., Wereley, S.T., Meinhart, C. D., and Takehara, K. (2003) "Particle Imaging Techniques for Microfabricated Fluidic Systems," *Exp. Fluids* **34**, pp. 504–14.

Dutta, P., Beskok, A., and Warburton, T.C. (2002) "Electroosmotic Flow Control in Complex Microgeometries," *J. Microelectromech. Sys.* **11**, pp. 36–44.

Ermakov, S.V., Jacobson, S.C., and Ramsey, J.M. (2000) "Computer Simulations of Electrokinetic Injection Techniques in Microfluidic Devices," *Anal. Chem.* **72**, pp. 3512–17.

Everaerts, F.M., Beckers, J.L., and Verheggen, T.P.E.M. (1976) *Isotachophoresis: Theory, Instrumentation, and Applications*, Elsevier, New York.

Fiechtner, G.J., and Cummings, E.B. (2003) "Faceted Design of Channels for Low-Dispersion Electrokinetic Flows in Microfluidics Systems," *Anal. Chem.* **75,** pp. 4747–55.

Flockhart, S.M., and Dhariwal, R.S. (1998) "Experimental and Numerical Investigation into the Flow Characteristics of Channels Etched in <100> Silicon," *J. Fluids Eng.* **120**, pp. 291–95.

Fu, L.-M., Yang, R.-J., Lin, C.-H., Pan, Y.-J., and Lee, G.-B. (2004) "Electrokinetically Driven Micro Flow Cytometers with Integrated Fiber Optics for On-Line Cell/Particle Detection," *Anal. Chim. Acta* **507**, pp. 163–69.

Gad-el-Hak, M. (1999) "The Fluid Mechanics of Microdevices: The Freeman Scholar Lecture," *J. Fluids. Eng.* **121**, pp. 5–33.

Galambos, P., and Forster, F.K. (1998) "Micro-Fluidic Diffusion Coefficient Measurement," in *Micro Total Analysis Systems*, D.J. Harrison and A. van den Berg, eds., Kluwer Academic Publishers, Dordrecht, pp. 189–192.

Gan, W., Yang, L., He, Y., Zeng, R., Cervera, M.L., and d. l. Guardia, M. (2000) *Talanta* **51**, p. 667.

Ghosal, S. (2002) "Lubrication Theory for Electro-Osmotic Flow in a Microfluidic Channel of Slowly Varying Cross-Section and Wall Charge," *J. Fluid Mech.* **459**, pp. 103–28.

Ghosal, S. (2004) "Fluid Mechanics of Electroosmotic Flow and its Effects on Band Broadening in Capillary Electrophoresis," *Electrophoresis* **25**, pp. 214–28.

Glückstad, J. (2004) "Sorting Particles with Light," *Nat. Mater.* **3**, pp. 9–10.

Gray, B., Jaeggi, D., Mourlas, N., van Drieenhuizen, B., Williams, K., Maluf, N., and Kovacs, G.S. (1999) "Novel interconnection technologies for integrated microfluidic Systems," *Sensor. Actuator. A-Phys.* **77**, pp. 57–65.

Green, N.G., Ramos, A ,Gonzales, A., Castellanos, A. and Morgan, H. (2000a) "Electric-Field-Induced Fluid Flow on Microelectrodes: The Effects of Illumination," *J. Phys. D: Appl. Phys.* **33**, pp. L13–17.

Green, N.G., Ramos, A., Gonzales, A., Morgan, H., and Castellanos, A. (2000b) "Fluid Flow Induced by Non-Uniform AC Electric Fields in Electrolytes on Microelectrodes: Part 1, Experimental Measurements," *Phys. Rev. E* **61**, pp. 4011–18.

Griffiths, S.K., and Nilson, R.H. (2001) "Low Dispersion Turns and Junctions for Microchannel Systems," *Anal. Chem.* **73**, pp. 272–78.

Grushka, E., McCormick, R.M., and Kirkland, J.J. (1989) "Effect of Temperature Gradients on the Efficiency of Capillary Zone Electrophoresis Separations," *Anal. Chem.* **61**, pp. 241–46.

Hagen, G. (1839) *On the Motion of Water in Narrow Cylindrical Tubes*, (German) *Pogg. Ann.* **46**, p. 423.

Hanks, R.W., and Ruo, H.-C. (1966) "Laminar-Turbulent Transition in Ducts of Rectangular Cross Section," *I&EC Fundamentals* **5**, p. 558–61.

Harley, J.C., Huang, Y., Bau, H.H., and Zemel, J.N. (1995) "Gas Flow in Micro-channels," *J. Fluid Mech.* **284**, pp. 257–74.

Hayes, M., Kheterpal, I., and Ewing, A. (1993) "Effects of Buffer pH on Electoosmotic Flow Control by an Applied Radial Voltage for Capillary Zone Electrophoresis," *Anal. Chem.* **65**, pp. 27–31.

Henry, D.C. (1948) "The Electrophoresis of Suspended Particles: 4, The Surface Conductivity Effect," *Trans. Faraday Soc.* **44**, pp. 1021–26.

Herr, A.E., Molho, J.I., Drouvalakis, K.A., Mikkelsen, J.C.,Utz, P.J., Santiago, J.G., and Kenny, T.W. (2003) "On-Chip Coupling of Isoelectric Focusing and Free Solution Electrophoresis for Multi-Dimensional Separations," *Anal. Chem.* **75**, pp. 1180–87.

Herr, A.E., Molho, J.I., Santiago, J.G., et al. (2000) "Electroosmotic Capillary Flow with Nonuniform Zeta Potential," *Anal. Chem.* **72**, pp. 1053–57.

Hiemenz, P.C., and Rajagopalan, R. (1997) "Principles of Colloid and Surface Chemistry," 3rd ed., Marcel Dekker, Inc., New York.

Hirokawa, T., Okamoto, H., and Gas, B. (2003) "High-Sensitive Capillary Zone Electrophoresis Analysis Bb Electrokinetic Injection with Transient Isotachophoretic Preconcentration: Electrokinetic Supercharging," *Electrophoresis* **24**, pp. 498–504.

Hitt, D.L., and Lowe, M.L. (1999) "Confocal Imaging of Flows in Artificial Venular Bifurcations," *Trans. ASME J. Biomech. Eng.* **121**, pp. 170–77.

Ho, C.-M., and Tai, Y.-C. (1998) "Micro-Electro-Mechanical Systems (MEMS) and Fluid Flows," *Annu. Rev. Fluid Mech.* **30**, pp. 579–612.

Hofmann, O., Che, D.P., Cruickshank, K.A., and Muller, U.R. (1999) "Adaptation of Capillary Isoelectric Focusing to Microchannels on a Glass Chip," *Anal. Chem.* **71**, pp. 678–86.

Hsieh, S.-S., Lin, C.-Y., Huang, C.-F., and Tsai, H.-H. (2004) "Liquid Flow in a Micro-Channel," *J. Micromech. Microeng.* **14**, pp. 436–45.

Huang, T., Tsai, P., Wu, Ch., and Lee, C. (1993) "Mechanistic Studies of Electroosmotic Control at the Capillary-Solution Interface," *Anal. Chem.* **65**, pp. 2887–93.

Hunter, R.J. (1981) "Zeta Potential in Colloid Science," Academic Press, London.

Israelachvili, J.N. (1986) "Measurement of the Viscosity of Liquids in Very Thin Films," *J. Coll. Interface Sci.* **110**, pp. 263–71.

Ivory, C.F. (2000). "A Brief Review of Alternative Electrofocusing Techniques," *Separ. Sci. Technol.* **35**, pp. 1777–93.

Jacobson, S.C., Hergenroder, R., Moore, A.W. Jr., and Ramsey, J.M. (1994) "Precolumn Reactions with Electrophoretic Analysis Integrated on a Microchip," *Anal. Chem.* **66**, pp. 4127–32.

Janson, S.W., Helvajian, H., and Breuer, K. (1999) "MEMS, Microengineering and Aerospace Systems," in 30th AIAA Fluid Dyn. Conf., 28 June–1 July, Norfolk, Virginia, AIAA 99-3802.

Jiang, X.N., Zhou, Z.Y., Yao, J., Li, Y., and Ye, X.Y. (1995) "Micro-Fluid Flow in Microchannel," in *Transducers '95:Eurosensor IX*, 8th Intl. Conf. on Solid-State Sensors and Actuators, and Eurosensors IX, 25–29, 1995, in Stockholm, Sweden, pp. 317–20.

Judy, J., Maynes, D., and Webb, B.W. (2002) "Characterization of Frictional Pressure Drop for Liquid Flows through Microchannels," *Int. J. Heat Mass Transf.* **45**, pp. 3477–89.

Jung, B., Bharadwaj, R., and Santiago, J.G. (2003) "Thousand-Fold Signal Increase Using Field Amplified Sample Tacking for On-Chip Electrophoresis," *Electrophoresis* **24**, pp. 3476–83.

Kaniansky, D., Masar, M., Bielcikova, J., Ivanyi, F., Eisenbeiss, F., Stanislawski, B., Grass, B., Neyer, A., and Johnck, M. (2000) "Capillary Electrophoresis Separations on a Planar Chip with the Column-Coupling Configuration Separation Channels," *Anal. Chem.* **72**, pp. 3596–3604.

Khaledi, M.G. (1998) "High-Performance Capillary Electrophoresis," in *Chemical Analysis: A Series of Monographs on Analytical Chemistry and its Applications*, J.D. Winefordner, ed., p. 146, John Wiley & Sons, Inc., New York.

Kirby, B.J. (2004) "Zeta Potential of Microfluidic Substrates: 1. Theory, Experimental Techniques, and Effects on Separations," *Electrophoresis* **25**, pp. 187–202.

Kitahara, A., and Watanabe, A. (1984) *Electrical Phenomena at Interfaces: Fundamentals, Measurements, and Applications*, Surfactant Science Series 15, Marcel Dekker, New York.

Knox, J.H. (1988) "Thermal Effects and Band Spreading in Capillary Electro-Separation," *Chromatographia* **26**, pp. 329–37.

Kohlrausch, F. (1897) "Über Konzentrations Verschicbungen durch Electrolyse im Innern von Lösungen and Losungsgemischen," *Ann. Phys.* (Leipzig) **62**, pp. 209–39.

Koo, J., and Kleinstreuer, C. (2003) "Liquid Flow in Microchannels: Experimental Observations and Computational Analyses of Microfluidics Effects," *J. Micromech. Microeng.* **13**, pp. 568–79.

Koutsiaris, A.G., Mathiouslakis, D.S., and Tsangaris, S. (1999) "Microscope PIV for Velocity-Field Measurement of Particle Suspensions Flowing inside Glass Capillaries," *Meas. Sci. Technol.* **10**, pp. 1037–46.

Krulevitch, P., Bennett, W., Hamilton, J., Maghribi, M., and Rose, K. (2002) "Polymer-Based Packaging Platform for Hybrid Microfluidic Systems," *Biomed. Microdevices* **4**, pp. 301–8.

Landers, J.P. (1994) *Handbook of Capillary Electrophoresis*, CRC Press, Boca Raton, FL.

Lanzillotto, A.-M., Leu, T.-S., Amabile, M., and Wildes, R. (1996) "An Investigation of Microstructure and Microdynamics of Fluid Flow in MEMS," in AD-Vol. 52, Proc. of ASME Aerospace Division, Atlanta, Georgia, pp. 789–96.

Laser, D., and Santiago, J.G. (2004) "A Review of Micropumps," *J. Micromech. Microeng.* **14**, pp. R35–R64.

Lee, G.-B., Hung, C.-I., Ke, B.-J., Huang, G.-R., Hwei, B.-H., and Lai, H.-F. (2001) "Hydrodynamic Focusing for a Micromachined Flow Cytometer," *J. Fluids Eng.* **123**, pp. 672–79.

Lee, G.-B., Lin, C.-H., and Chang, G.-L. (2003) "Micro Flow Cytometers with Buried SU-8/SOG Optical Waveguides," *Sensor. Actuator. A* **103**, pp. 165–70.

Levich, V. (1962) *Physicochemical Hydrodynamics*, Prentice-Hall, Englewood Cliffs, N.J.

Li, Y., Buch, J.S., Rosenberger, F., DeVoe, D.L., and Lee, C.S. (2004) "Integration of Isoelectric Focusing with Parallel Sodium Dodecyl Sulfate Gel Electrophoresis for Multidimensional Protein Separations in a Plastic Microfluidic Network," *Anal. Chem.* **76**, pp. 742–48.

Li, Z.-X., Du, D.-X., and Guo, Z.-Y. (2003) "Experimental Study on Flow Characteristics of Liquid in Circular Microtubes," *Microscale Thermophys. Eng.* **7**, pp. 253–65.

Lichtenberg, J., Verpoorte, E., and de Rooij, N.F. (2001) "Sample Preconcentration by Field Amplification Stacking for Microchip-Based Capillary Electrophoresis," *Electrophoresis* **22**, pp. 258–71.

Lin, H., Storey, B., Oddy, M., Chen, C.-H., and Santiago, J.G. (2004) "Instability of Electrokinetic Microchannel Flows with Conductivity Gradients," *Phys. Fluids* **16**, pp. 1922–35.

Liu, D., and Garimella, S.V. (2004) "Investigation of Liquid Flow in Microchannels," *J. Thermophys. Heat Transf.* **18**, pp. 65–72.

Liu, R.H., Stremler, M.A., Sharp, K.V., Olsen, M.G., Santiago, J.G., Adrian, R.J., Aref, H., and Beebe, D.J. (2000) "Passive Mixing in a Three-Dimensional Serpentine Microchannel," *J. MEMS* **9**, pp. 190–97.

MacInnes, J., Du, X., and Allen, R. (2003) "Prediction of Electrokinetic and Pressure Flow in a Microchannel T-Junction," *Phys. Fluids* **15**, pp. 1992–2005.

Macounova, K., Cabrera, C.R., and Yager, P. (2001) "Concentration and Separation of Proteins in Microfluidic Channels on the Basis of Transverse IEF," *Anal. Chem.* **73**, pp. 1627–33.

Mala, G.M., and Li, D. (1999) "Flow Characteristics of Water in Microtubes," *Int. J. Heat Fluid Flow* **20**, pp. 142–48.

Manz, A., Effenhauser, C.S., Burggraf, N., et al. (1994) "Electroosmotic Pumping and Electrophoretic Separations for Miniaturized Chemical Analysis Systems," *J. Micromech. Microeng.* **4**, pp. 257–65.

Maynes, D., and Webb, A.R. (2002) "Velocity Profile Characterization in Sub-Millimeter Diameter Tubes Using Molecular Tagging Velocimetry," *Exp. Fluids* **32**, pp. 3–15.

Meinhart, C.D., Wereley, S.T., and Santiago, J.G. (1999) "PIV Measurements of a Microchannel Flow," *Exp. Fluids* **27**, pp. 414–19.

Melcher, J.R. (1981) *Continuum Electromechanics*. MIT Press, Boston.

Merkle, C.L., Kubota, T., and Ko, D.R.S. (1974) "An Analytical Study of the Effects of Surface Roughness on Boundary-layer Transition," AF Office of Scien. Res. Space and Missile Sys. Org., AD/A004786.

Mikkers, F.E.P., Everaerts, F.M., and Verheggen, T. (1979) "High-Performance Zone Electrophoresis," *J. Chromatogr.* **169**, pp. 11–20.

Mirowski, W., Moreland, J., Russek, S.E., and Donahue, M.J. (2004) "Integrated Microfluidic Isolation Platform for Magnetic Particle Manipulation in Biological Systems," *Appl. Phys. Lett.* **84**, pp. 1786–88.

Mohammadi, B., Molho, J.I., and Santiago, J.G. (2003) "Incomplete Sensitivities for the Design of Minimal Dispersion Fluidic Channels," *Comput. Meth. Appl. Mech. Eng.* **192**, pp. 4131–45.

Morgan, H., and Green, N.G. (2003) *AC Electrokinetics: Colloids and Nanoparticles*, Research Studies Press Ltd., Baldock, England.

Morini, G.L. (2004) "Laminar-to-Turbulent Flow Transition in Microchannels," *Microscale Thermophys. Eng.* **8**, pp. 15–30.

Myung-Suk, C., and Kwak, H.W. (2003) "Electrokinetic Flow and Electroviscous Effect in a Charged Slit-Like Microfluidic Chan with Nonlinear Poisson-Boltzmann Field," *Korea-Australia Rheol. J.* **15**, pp. 83–90.

Nguyen, N.-T., and Wereley, S. (2002) *Fundamentals and Applications of Microfluidics*, Artech House, Norwood, MA.

Novotny, E.J., and Eckert, R.E. (1974) "Rheological Properties of Viscoelastic Fluids from Continuous Flow through a Channel Approximating Infinite Parallel Plates," *Trans. Soc. Rheol.* **18**, pp. 1–26.

Obot, N.T. (2002) "Toward a Better Understanding of Friction and Heat/Mass Transfer in Microchannels: A Literature Review," *Microscale Thermophys. Eng.* **6**, pp. 155–73.

Oddy, M., and Santiago, J.G. (2004) "Alternating Electric Field Measurements of Particle Zeta-Potentials in a Microchannel," *J. Colloid Interface Sci.* **269**, pp. 192–204.

Oddy, M.H., Santiago, J.G., and Mikkelsen, J.C. (2001) "Electrokinetic Instability Micromixing," *Anal. Chem.* **73**, pp. 5822–32.

Osbourn, D.M., Weiss, D.J., and Lunte, C.E. (2000) "On-Line Preconcentration Methods for Capillary Electrophoresis," *Electrophoresis* **21**, pp. 2768–79.

Overbeek, J.T.G. (1952) "Electrochemistry of the Double Layer," in *Colloid Science*, H.R. Kruyt, ed., Elsevier, Amsterdam, pp. 115–277.

Ovryn, B. (1999) "Three-Dimensional Forward Scattering Particle Image Velocimetry in a Microscopic Field-of-View," in *Proc. 3rd Intl. Workshop PIV*, 16–18 September, Santa Barbara, California, pp. 385–93.

Papautsky, I., Brazzle, J., Ameel, T., and Frazier, A.B. (1999a) "Laminar Fluid Behavior in Microchannels Using Micropolar Fluid Theory," *Sensor. Actuator.* **73**, pp. 101–8.

Papautsky, I., Gale, B.K., Mohanty, S., Ameel, T.A., and Frazier, A.B. (1999b) "Effects of Rectangular Microchannel Aspect Ratio on Laminar Friction Constant," in *SPIE Conference on Microfluidic Devices and Systems II*, Santa Clara, California, 3877, pp. 147–58.

Patankar, N.A., and Hu, H.H. (1998) "Numerical Simulation of Electroosmotic Flow," *Anal. Chem.* 20–21 September 1999, **70**, pp. 1870–81.

Paul, P.H., Arnold, D.W., and Rakestraw, D.J. (1998a) "Electrokinetic Generation of High Pressures Using Porous Microstructures," in *Micro Total Analysis Systems*, D.J. Harrison and A. van den Berg, eds., Kluwer Academic Publishers.

Paul, P.H., Garguilo, M.G., and Rakestraw, D.J. (1998b) "Imaging of Pressure- and Electrokinetically Driven Flows Through Open Capillaries," *Anal. Chem.* Banff, Canada **70**, pp. 2459–67.

Peng, X.F., Peterson, G.P., and Wang, B.X. (1994) "Frictional Flow Characteristics of Water Flowing through Rectangular Microchannels," *Exp. Heat Transf.* **7**, pp. 249–64.

Pfahler, J., Harley, J., Bau, H., and Zemel, J. (1990a) "Liquid Transport in Micron and Submicron Channels," *Sensor. Actuator.* **A21–A23**, pp. 431–34.

Pfahler, J., Harley, J., Bau, H., and Zemel, J.N. (1991) "Gas and Liquid Flow in Small Channels," in DSC-Vol. 32, Micromechanical Sensors, Actuators and Systems, ASME Winter Annual Meeting, Atlanta, Georgia, pp. 49–59.

Pfahler, J., Harley, J., Bau, H.H., and Zemel, J. (1990b) "Liquid and Gas Transport in Small Channels," in DSC-Vol. 19, Microstructures, Sensors and Actuators, ASME Winter Annual Meeting, Dallas, Texas, pp. 149–57.

Pfund, D., Rector, D., Shekarriz, A., Popsecu, A., and Welty, J. (2000) "Pressure Drop Measurements in a Microchannel," *AIChE J* **46**, pp. 1496–1507.

Phares, D.J., and Smedley, G.T. (2004) "A Study of Laminar Flow of Polar Liquids through Circular Microtubes," *Phys. Fluids* **16**, pp. 1267–72.

Poiseuille, M. (1840, 1841) "Recherches Expérimentales Sur le Mouvement des Liquides dans les Tubes de Très Petits Diametrès," *CR Hebdomaires des Séances Acad. Sci.* **11**.

Probstein, R.F. (1994) *Physicochemical Hydrodynamics: An Introduction*, 2nd ed., John Wiley & Sons, Inc., New York.

Qu, W., Mala, G.M., and Li, D. (2000) "Pressure-Driven Water Flows in Trapezoidal Silicon Microchannels," *Int. J. Heat Mass Transf.* **43**, pp. 353–64.

Quirino, J., and Terabe, S. (1999) "Sample Stacking of Fast-Moving Anions in Capillary Zone Electrophoresis with pH-Suppressed Electroosmotic Flow," *J. Chromatogr. A* **850**, pp. 339–44.

Ramos, A., Morgan, H., Green, N.G., and Castellanos, A. (1998) "AC Electrokinetics: A Review of Forces in Microelectrode Structures," *J. Phys. D: Appl. Phys.* **21**, pp. 2338–53.

Ramos, A., Morgan, H. Green, N.G., and Castellanos, A. (1999) "AC Electric-Field-Induced Fluid Flow in Microelectrodes," *J. Colloid Interface Sci.* **21**, pp. 420–22.

Ren, L., Qu, E., and Li, D. (2001) "Interfacial Electrokinetic Effects on Liquid Flow in Microchannels," *Int. J. Heat Mass Transf.* **44**, pp. 3125–34.

Reyes, D.R., Iossifidis, D., Auroux, P.A., and Manz, A. (2002) "Micro Total Analysis Systems: 1. Introduction, Theory, and Technology," *Anal. Chem.* **74**, pp. 2623–36.

Reynolds, O. (1883) "An Experimental Investigation of the Circumstances which Determine whether the Motion of Water Will Be Direct or Sinuous, and the Law of Resistance in Parallel Channels," *Phil. Trans. Roy. Soc. London* **2**, p. 51.

Rice, C.L., and Whitehead, R. (1965) "Electrokinetic Flow in a Narrow Cylindrical Capillary," *J. Phys. Chem.* **69**, pp. 4017–24.

Righetti, P.G. (1983) *Isoelectric Focusing: Theory, Methodology, and Applications*, Amsterdam, New York.

Ross, D., and Locascio, L.E. (2004) "Microfluidic Temperature Gradient Focusing," *Anal. Chem.* **74**, pp. 2556–65.

Russel, W.B., Saville, D.A., and Schowalter, W.R. (1999) "Colloidal Dispersions," *Cambridge Monographs on Mechanics and Applied Mathematics,* G.K. Batchelor, ed., Cambridge University Press, Cambridge, United Kingdom.

Santiago, J.G. (2001) "Electroosmotic Flows in Microchannels with Finite Inertial and Pressure Forces," *Anal. Chem.* **73**, pp. 2353–65.

Santiago, J.G., Wereley, S.T., Meinhart, C.D., Beebe, D.J., and Adrian, R.J. (1998) "A Particle Image Velocimetry System for Microfluidics," *Exp. Fluids* **25**, pp. 316–19.

Saville, D. (1997) "Electrohydrodynamics: The Taylor-Melcher Leaky Dielectric Model," *Annu. Rev. Fluid Mech.* **29**, pp. 27–64.

Scales, P., Grieser, F., and Healy, T. (1992) "Electrokinetics of the Silica-Solution Interface: A Flat Plate Streaming Potential Study," *ACS J. Langmuir Surf. Colloids* **8**, pp. 965–74.

Schaller, Th., Bolin, L., Mayer, J., and Schubert, K. (1999) "Microstructure Grooves with a Width Less than 50 m Cut with Ground Hard Metal Micro End Mills," *Precision Eng.* **23**, pp. 229–35.

Schulte, T.H., Bardell, R.L., and Weigl, B.H. (2000) "On-Chip Microfluidic Sample Preparation," *J. Lab. Automat.* **5**, p. 83.

Shah, R.K., and London, A.L. (1978) "Laminar Flow Forced Convection in Ducts," in series *Adv. in Heat Transfer*, Supp. 1, Academic Press, New York.

Sharp, K.V., and Adrian, R.J. (2004) "Transition from Laminar to Turbulent Flow in Liquid Filled Microtubes," *Exp. Fluids* **36**, pp. 741–47.

Smoluchowski, M.V. (1903) Bull. Akad. Sci. Cracovie, Classe Sci. Math. Natur., **1** p. 182.

Sobhan, C.B., and Garimella, S.V. (2001) "A Comparative Analysis of Studies on Heat Transfer and Fluid Flow in Microchannels," *Microscale Thermophys.* Eng. **5**, pp. 293–311.

Sounart, T.L., and Baygents, J.C. (2001) "Electrically-Driven Fluid Motion in Channels With Streamwise Gradients of the Electrical Conductivity," *Colloid. Surface. A: Physicochem. Eng. Asp.* **195**, pp. 59–75.

Stone, H.A., Stroock, A.D., and Ajdari, A. (2004) "Engineering Flows in Small Devices: Microfluidics toward a Lab-on-a-Chip," *Annu. Rev. Fluid Mech*. **36**, pp. 381–411.

Tan, W., Fan, Z.H., Qiu, C.X., Ricco, A.J., and Gibbons, I. (2002) "Miniaturized Capillary Isoelectric Focusing in Plastic Microfluidic Devices," *Electrophoresis* **23**, pp. 3638–45.

Taylor, J.A., and Yeung, E.S. (1993) "Imaging of Hydrodynamic and Electrokinetic Flow Profiles in Capillaries," *Anal. Chem.* **65**, pp. 2928–32.

Theeuwes, F. (1987) *J. Pharm. Sci.* **64** (#12) pp. 1987–91 (1975).

Tieu, A.K., Mackenzie, M.R., and Li, E.B. (1995) "Measurements in Microscopic Flow with a Solid-state LDA," *Exp. Fluids* **19**, pp. 293–94.

Tretheway, D.C., and Meinhart, C.D. (2002) "Apparent Fluid Slip at Hydrophobic Microchannel Walls," *Phys. Fluids* **14**, pp. L9–L12.

Tretheway, D.C., and Meinhart, C.D. (2004) "A Generating Mechanism for Apparent Fluid Slip in Hydrophobic Microchannels," *Phys. Fluids* **14**, pp. L9–L12.

Tuckerman, D.B., and Pease, R.F.W. (1981) "High-Performance Heat Sinking for VLSI," *IEEE Electron Device Lett.* **EDL-2**, pp. 126–29.

Vreeland, W.N., Williams, S.J., Barron, A.E., and Sassi, A.P. (2003) "Tandem Isotachophoresis-Zone Electrophoresis via Base-Mediated Destacking for Increased Detection Sensitivity in Microfluidic Systems," *Anal. Chem.* **75**, p. 3059.

Wainright, A., Williams, S.J., Ciambrone, G., Xue, Q.F., Wei, J., and Harris, D. (2002) "Sample Pre-Concentration by Isotachophoresis in Microfluidic Devices," *J. Chromatogr. A* **979**, pp. 69–80.

Wang, D., Sigurdson, M., and Meinhart, C. D. (2004) "Experimental Analysis Of Particle and Fluid Motion in AC Electrokinetics," *Exp. Fluids*, in press.

Weigl, B.H., and Yager, P. (1999) "Microfluidic Diffusion-Based Separation and Detection," *Science* **283**, pp. 346–47.

White, F.M. (1994) *Fluid Mechanics,* 3rd ed., McGraw-Hill, Inc., New York.

White, F.M. (1991) *Viscous Fluid Flow*, 2nd ed., McGraw-Hill Series in Mechanical Engineering, J.P. Holman and J.R. Lloyd, eds., McGraw-Hill, New York.

Wilding, P., Pfahler, J., Bau, H.H., Zemel., J.N., and Kricka, L.J. (1994) "Manipulation and Flow of Biological Fluids in Straight Channels Micromachined in Silicon," *Clin. Chem.* **40**, pp. 43–47.

Woei, T., Fan, H.Z, Qiu, C.X., Ricco, A.J., and Gibbons, I. (2002) "Miniaturized Capillary Isoelectric Focusing in Plastic Microfluidic Devices," *Electrophoresis* **23**, pp. 3638–45

Wu, P., and Little, W.A. (1983) "Measurement of Friction Factors for the Flow of Gases in Very Fine Channels Used for Microminiature Joule-Thomson Refrigerators," *Cryogenics* **23**, pp. 273–77.

Xu, Z.Q., Ando, T., Nishine, T., Arai, A., and Hirokawa, T. (2003) "Electrokinetic Supercharging Preconcentration and Microchip Gel Electrophoretic Separation of Sodium Dodecyl Sulfate-Protein Complexes," *Electrophoresis* **24**, pp. 3821–27.

Yang, H., and Chien, R.-L. (2001) "Sample Stacking in Laboratory-on-a-Chip Devices," *J. Chromatogr. A* **924**, pp.155–63.

Yao, S., and Santiago, J.G. (2003a) "Porous Glass Electroosmotic Pumps: Theory," *J. Colloid Interface Sci.* **268**, pp.133–42.

Yao, S., Hertzog, D.E., Zeng, S., Mikkelsen, J.C., and Santiago, J.G. (2003b) "Porous Glass Electroosmotic Pumps: Design and Experiments," *J. Colloid Interface Sci.* **268**, pp.143–53.

Yazdanfar, S., Kulkarni, M.D., and Izatt, J.A. (1997) "High Resolution Imaging of In Vivo Cardiac Dynamics Using Color Doppler Optical Coherence Tomography," *Opt. Ex.* **1**, pp. 424–31.

Yu, D., Warrington, R., Barron, R., and Ameel, T. (1995) "An Experimental and Theoretical Investigation of Fluid Flow and Heat Transfer in Microtubes," in *Proc. of ASME/JSME Thermal Engineering Joint Conference*, 19–24 March, Maui, Hawaii, pp. 523–30.

Zeng, S., Chen, C., Mikkelsen, J.C., et al. (2000) "Fabrication and Characterization of Electrokinetic Micro Pumps," in *7th Intersoc. Conf. on Thermal and Thermomech. Phenomena in Electronic Systems*, 23–26 May, Las Vegas, Nevada.

Zhao, T.S., and Liao, Q. (2002) "Thermal Effects on Electro-Osmotic Pumping of Liquids in Microchannels," *J. Micromech. Microeng.* **12**, pp. 962–70.

Zhu, Y., and Granick, S. (2001) "Viscosity of Interfacial Water," *Phys. Rev. Lett.* **87**, pp. 096104-1–096104-4.

11

Lubrication in MEMS

Kenneth S. Breuer
Brown University

11.1 Introduction

As microengineering technology continues to advance, driven by increasingly complex and capable microfabrication and materials technologies, the need for more sophistication in MEMS design will increase. Fluid film lubrication has been a critical issue from the outset of MEMS development, particularly in the prediction and control of viscous damping in vibrating devices such as accelerometers and gyros. Much attention has been showered on the development of models for accurate prediction of viscous damping and on the development of fabrication techniques for minimizing the damping, which destroys the quality, or *Q*-factor, of a resonant system. In addition to the development and optimization of these oscillatory devices, rotating devices — micromotors, microengines, etc. — have also captured the attention of MEMS researchers since the early days of MEMS development, and there have been several demonstrations of micromotors drive by electrostatic forces [Bart et al., 1988; Mehregany et al., 1992; Nagle and Lang, 1999; and Sniegowski and Garcia, 1996].

For the most part these motors were very small, with rotors of the order of 100 microns in diameter, and although they had high rotational speeds (hundreds of thousands of revolutions per minute), their tip speeds and rotational energy, which scales with tip speed, was quite small. Tip speeds of 1 meter per second are typical. In addition, the focus of these projects was the considerable challenge of fabricating a freely moving part and integrating the drive electrodes. Lubrication and protection against wear were low priority. The demonstrated engines relied on dry-rubbing bearings in which the rotor was held in place by a bushing, but there was no design or integration of a lubrication system. The low surface speeds of these engines meant that they could operate for long times using this primitive bearing, however, failure was observed frequently due to rotor and bearing wear.

As MEMS devices become more sophisticated and have more stringent design and longevity requirements, the need for more accurate and extensive design tools for lubrication has increased. In addition, the energy density of MEMS is increasing. Devices for power generation, propulsion, and so forth are actively under development. In such devices the temperatures and stresses are stretched to the material limits. Hence, the requirement for protection of moving surfaces becomes more than a casual interest — it is critical for the success of a "power-MEMS" device.

11.1.1 Objectives and Outline

The objective of this chapter is to briefly summarize some the issues associated with lubrication in MEMS. Lubrication is a vast topic. Our focus is to review the essential features of lubrication theory and design practice, and to highlight the difficulties that arise in the design of a lubrication system for MEMS devices. A key feature of MEMS is that the fabrication, material properties, and mechanical and electrical design are all tightly interwoven and cannot be separated. For this reason, some attention is devoted to the important issue of how a successful lubrication system is influenced by manufacturing constraints and material properties.

One should always remember that MEMS is a rapidly developing, expanding, and maturing manufacturing technology. The range of geometric options, available materials, and dimensional control is continually developing and improving. This chapter focuses on the current state of the art in MEMS fabrication. For this reason this chapter favors silicon-based fabrication processes and the constraints that lithographic-based batch fabrication techniques place on lubrication system design and performance.

Examples are drawn from several sources. In the sections on translational and squeeze-film damping, examples are drawn from the extensive literature associated with accelerometer and gyro design. For rotating system lubrication, we draw heavily on the MIT Microengine project [Epstein et al., 1997], which, to our knowledge, is the only MEMS device to date with rotating elements that use a fluid film lubrication system. This device is described in some detail later in the chapter, and the analysis and examples of thrust bearings and journal bearings are drawn from that device.

11.2 Fundamental Scaling Issues

11.2.1 The Cube-Square Law

The most dominant effect that changes our intuitive appreciation of the behavior of microsystems is the so-called "cube-square law." This law states that volumes scale with the cube of the typical length scale, while areas (including surface areas) scale with the square of the length scale. Thus, as a device shrinks, surface phenomena become relatively more important than volumetric phenomena. The most important consequence of this phenomena is that the device mass and inertia become negligibly small at the micro- and nano-scales. For lubrication, this phenomena means that the volumetric loads that require support, like the weight of a rotor, quickly become negligible. As an example, consider the ratio of the weight of a microfabricated rotor (a cylinder of density ρ, diameter D, and length L) compared to the pressure (p) acting on its projected surface area. This can be expressed as a non-dimensional load parameter:

$$\zeta = \frac{\rho \pi L D^2/4}{pLD} \propto \frac{D}{p} \tag{11.1}$$

from which it can be seen that the load parameter decreases linearly as the device shrinks. For example, the load parameter due to the rotor mass for the MIT Microengine, which is a relatively large MEMS devices (measuring 4 mm in diameter and 300 microns deep), is approximately 10^{-3}. The benefits of this scaling are that orientation or the freely suspended part becomes effectively irrelevant and that unloaded operation is easy to accomplish. In addition, since the gravity loading is negligible, the primary forces that one needs to support are pressure-induced loads and in a rotating device loads due to rotational imbalance. This last load is very important and will be discussed in more detail in connection to rotating lubrication requirements. The chief disadvantage of the low natural loading is that unloaded operation is often undesirable (in hydrodynamic lubrication where a minimum eccentricity is required for journal bearing stability), and in practice, gravity loading is often used to advantage. Therefore a scheme for applying an artificial load needs to be developed. This is discussed in more detail later in the chapter.

11.2.2 Applicability of the Continuum Hypothesis

A common concern in microfluidic devices is the appropriateness of the continuum hypothesis as the device scale continues to fall. At some scale, the typical inter-molecular distances will be comparable to the device scales and the use of continuum fluid equations becomes suspect. For gases, this is measured by the Knudsen number (Kn) — the ratio of the mean free path to the typical device scale. Numerous experiments [Arkilic et al., 1997, 1993; Breuer et al., 2001] have determined that non-continuum effects become observable when Kn reaches approximately 0.1 and that continuum equations become meaningless (the "transition flow regime") at Kn of approximately 0.3. For atmospheric temperature and pressure, the mean free path of air is approximately 70 nm. Thus, atmospheric devices with features smaller than approximately 0.2 microns will be subject to non-negligible non-continuum effects. In many cases, such small dimensions are not present, and the fluidic analysis can safely use the standard Navier–Stokes equations (this is the case for the microengine).

Nevertheless, in applications where viscous damping is to be avoided (for example in high-Q resonating devices such as accelerometers or gyroscopes) the operating gaps are typically quite small (perhaps a few microns), and the gaps serve as both a physical standoff and a sense-gap where capacitive sensing is accomplished. In such examples one must work with the small dimension, and in order to minimize viscous effects, the device is packaged at low pressures where non-continuum effects are evident. For small Knudsen numbers, the Navier–Stokes equations can be used with a single modification — the boundary condition is relaxed from the standard non-slip condition to that of a slip-flow condition where the velocity at the wall is related to the Knudsen number and the gradient of velocity at the wall:

$$u_w = \lambda \frac{2-\sigma}{\sigma} \left.\frac{\partial u}{\partial y}\right|_w \tag{11.2}$$

where σ is the tangential momentum accommodation coefficient (TMAC) which varies between 0 and 1. Experimental measurements [Breuer et al., 2001] indicate that smooth native silicon has a TMAC of approximately 0.7 in contact with several commonly used gases.

Despite the fact that the slip-flow theory is valid only for low Kn, it is often used incorrectly with great success at much higher Knudsen numbers. Its adoption beyond its range of applicability stems primarily from the lack of any better approach short of solving the Bolzmann equation or Direct Simulation Monte Carlo (DSMC) computations [Beskok and Karniadakis, 1994; Cai et al., 2000]. For many simple geometries, the "extended" slip-flow theory works much better than it should and provides quite adequate results [Kwok et al., 2005]. This theory is demonstrated in the sections on Couette and squeeze-film damping later in the chapter.

11.2.3 Surface Roughness

Another peculiar feature of MEMS devices is that the surface roughness of the material used can become a significant factor in the overall device geometry. MEMS surface finishes are quite varied, ranging from atomically smooth surfaces found on polished single-crystal silicon substrates to the rough surfaces left by different etching processes. The effects of these topologies can be important in several areas for microdevice performance. Probably the most important effect is the way in which the roughness can affect structural characteristics such as crack initiation, yield strength, etc., although this will not be explored in this chapter. Secondly, the surface finish can affect fluidic phenomena such as the energy and momentum accommodation coefficient, and consequently, the momentum and heat transfer. Lastly, the surface characteristics (not only the roughness, but also the surface chemistry and affinity) can strongly affect its adhesive force. This is not treated in detail in this discussion, although it is mentioned briefly at the end of the chapter in connection with tribology issues in MEMS.

11.3 Governing Equations for Lubrication

With the proviso that the continuum hypothesis holds for micron-scale devices (perhaps with a modified boundary condition), the equations for microlubrication are identical to those used in conventional lubrication analysis and can be found in any standard lubrication textbook [Hamrock, 1984]. We present the essential results here, but the reader is referred to more complete treatments for full derivations and a detailed discussion of the appropriate limitations.

Starting with the Navier–Stokes equations, we can make a number of simplifying assumptions appropriate for lubrication problems. These are itemized here:

Inertia: The terms representing transfer of momentum due to inertia may be neglected. This arises because of the small dimensions that characterize lubrication geometries and MEMS in particular. In very high speed devices such as the MIT Microengine, inertial terms may not be as small as one might like, and corrections for inertia may be applied. However, preliminary studies suggest that these corrections are small [Piekos, 2000].

Curvature: Lubrication geometries are typically characterized by a thin fluid film with a slowly varying film thickness. The critical dimension in such systems is the film thickness, and this is assumed to be much smaller than any radius of curvature associated with the overall system. This assumption is particularly important in rotating systems where a circular journal bearing is used. Assuming that the radius of the bearing R is much larger than the typical film thickness c (i.e., $c/R \ll 1$) greatly simplifies the governing equations.

Isothermal: Because volumes are small and surface areas are large, thermal contact between the fluid and the surrounding solid is very good in a MEMS device. In addition, common MEMS materials are good thermal conductors. For both these reasons, it is safe to assume that the lubrication film is isothermal.

With these restrictions, the Navier–Stokes equations, the equation for the conservation of mass, and the equation of state for a perfect gas may be combined to yield the Reynolds equation [Reynolds, 1886], written here for two-dimensional films:

$$0 = \frac{\partial}{\partial x}\left(-\frac{\rho h^3}{12\mu}\frac{\partial p}{\partial x}\right) + \frac{\partial}{\partial y}\left(-\frac{\rho h^3}{12\mu}\frac{\partial p}{\partial y}\right) + \frac{\partial}{\partial x}\left(\frac{\rho h(u_a + u_b)}{2}\right) + \frac{\partial}{\partial y}\left(\frac{\rho h(v_a + v_b)}{2}\right)$$
$$+ \rho(w_a - w_b) - \rho u_a \frac{\partial h}{\partial x} - \rho v_a \frac{\partial h}{\partial y} + h\frac{\partial \rho}{\partial t} \tag{11.3}$$

where x and y are the coordinates in the lubrication plane: u_a etc. are the velocities of the upper and lower surfaces. An alternate and more general version may be derived [Burgdorfer, 1959] by nondimensionalization with the film length and width (l and b), the minimum clearance $h_{\min}$, a characteristic shearing velocity u_b, and a characteristic unsteady frequency ω. In addition, gas rarefaction can be incorporated for low Knudsen numbers by assuming a slip-flow wall boundary condition:

$$\frac{\partial}{\partial X}\left[(1+6K)PH^3\frac{\partial P}{\partial X}\right]+\mathrm{A}^2\frac{\partial}{\partial Y}\left[(1+6K)PH^3\frac{\partial P}{\partial Y}\right]=\Lambda\frac{\partial(PH)}{\partial X}+\sigma\frac{\partial(PH)}{\partial T} \tag{11.4}$$

where

$$\mathrm{A}=\frac{l}{b};\quad \Lambda=\frac{6\mu u_b l^2}{p_a h_{\min}^2};\quad \sigma=\frac{12\mu\omega l^2}{p_a h_{\min}^2} \tag{11.5}$$

A is the film aspect ratio, Λ is the bearing number, and σ is the squeeze number representing unsteady effects.

Solution of the Reynolds equation is straightforward, but not trivial. A chief difficulty is that gas films are notoriously unstable if they operate in the wrong parameter space. In order to determine the stability or instability of the numerically-generated solution, both the steady Reynolds equation and its unsteady counterpart need to be addressed with some accuracy. These issues are discussed more by Piekos and Breuer (1998) and others.

11.4 Couette-Flow Damping

The viscous damping of a plate oscillating in parallel motion to a substrate has been a problem of tremendous importance in MEMS devices, particularly in the development of resonating structures such as accelerometers and gyros. The problem arises because the proof mass, which may be hundreds of microns in the lateral dimension, is typically suspended above the substrate with a separation of only a few microns.

A simple analysis of Couette-flow damping for rarefied flows is easy to demonstrate by choosing a model problem of a one-dimensional proof mass (i.e., ignoring the dimension perpendicular to the plate motion). This is shown schematically in Figure 11.1.

The Navier–Stokes equations for this geometry reduce to:

$$\frac{\partial u}{\partial t}=\mu\frac{\partial^2 u}{\partial y^2} \tag{11.6}$$

in which only viscous stresses due to the velocity gradient and the unsteady terms survive. This can be solved using separation of variables and employing a slip-flow boundary condition [Arkilic and Breuer, 1993] yielding the solution the drag force experienced by the moving plate:

$$D=\frac{4\pi U^2}{\beta}\left[\frac{\sinh\beta+\sin\beta}{(\cosh\beta-\cos\beta)+D_R}\right] \tag{11.7}$$

where

$$\beta=\sqrt{\frac{\omega h^2}{\mu}} \tag{11.8}$$

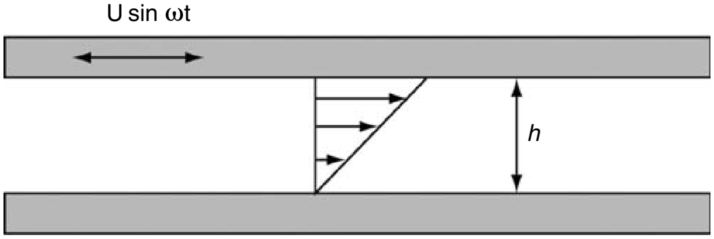


FIGURE 11.1 Schematic of Couette-flow damping geometry. The upper plate vibrates with a proscribed amplitude and frequency. For most MEMS geometries and frequencies, the unsteadiness can usually be neglected.

is a Stokes number, representing the balance between unsteady and viscous effects, and D_R is a correction due to slip flow at the wall:

$$D_R = 2Kn\beta(\sinh\beta + \sin\beta) + 2Kn^2\beta^2(\cosh\beta + \cos\beta) \tag{11.9}$$

A typical MEMS geometry might have a plate separation of one micron and an operating frequency of 10 kHz. With these parameters, the Stokes number is very small (approximately 0.1), and the flow may be considered quasi-steady to a high degree of approximation. In addition, the rarefaction effects, indicated by D_R, are also vanishingly small at atmospheric conditions.

11.4.1 Limit of Molecular Flow

Although the slip-flow solution is limited to low Knudsen numbers, the damping due to a gas at high degrees of rarefaction can be computed using a free-molecular flow approximation. In such cases the friction factor on a flat plate is given by Rohsenow and Choi (1961).

$$C_f = \sqrt{\frac{2}{\pi\gamma}}\frac{1}{M} \tag{11.10}$$

where γ is the ratio of specific heats and M is the Mach number. It is important to recognize that the damping (and Q) in this case is provided, not only by the flow in the gap, but also by the flow above the vibrating plate. However, it is unlikely that the fluid damping provides the dominant source of damping at such extremely low pressures. More likely, damping derived from the structure (e.g., flexing of the support tethers, non-elastic strain at material interfaces, etc.) will take over as the dominant energy-loss mechanism. Kwok et al. (2005) compared the continuum, slip-flow, and free molecular flow models for Couette damping with data obtained by measuring the "ring down" of a tuning fork gyroscope fabricated by Draper Laboratories. Figure 11.2 shows the measurements and theory confirming the functional behavior of the damping as the pressure drops (Kn increases) and the unexpected accuracy of these rather simple models. Although the trends are well-predicted, the absolute value of the Q-factor is still in error by a factor of two, suggesting that more detailed computations are still of interest.

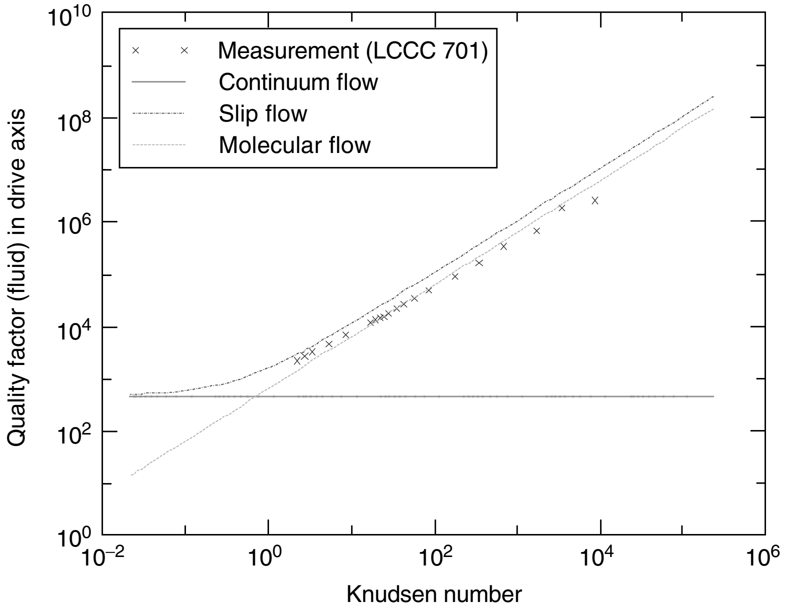


FIGURE 11.2 (**See color insert following page 10-34.**) Theory and measurements of Couette damping in a tuning fork gyro (Kwok et al. [2005]). Note that in the high Knudsen number limit, the free molecular approximation predicts the damping more closely, but that the slip-flow model, though totally inappropriate at this high Kn level, is not too far from the experimental measurements.

11.5 Squeeze-Film Damping

Squeeze-film damping arises when the gap size changes in an oscillatory manner squeezing the trapped fluid (Figure 11.3). Fluid, usually air, is trapped between the vibrating proof mass and the substrate resulting in a squeeze film, which can significantly reduce the quality factor of the resonator. In some cases this damping is desirable, but as with the case of Couette-flow damping, it is often parasitic, and the MEMS designer tries to minimize its effects and maximize the resonant *Q*-factor of the device. Common methods for alleviating squeeze-film effects are to fabricate breathing holes ("chimneys") throughout the proof mass which relieve the build up of pressure and to package the device at low pressure. Both of these solutions have drawbacks. The introduction of vent holes reduces the vibrating mass, necessitating an even larger structure, while the low-pressure packaging adds considerable complexity to the overall device development and cost. Figure 11.4 illustrates a high-performance tuning fork gyroscope fabricated by Draper Laboratories.

11.5.1 Derivation of Governing Equations

The analysis of the squeeze-film damping is presented in the following section. The Reynolds equations may be used as the starting point. However, a particularly elegant and complete solution was published by Blech (1983) for the case of the continuum flow and was extended by Kwok et al. (2005) to the case of slip-flow and flows films with vent holes. This analysis is summarized here.

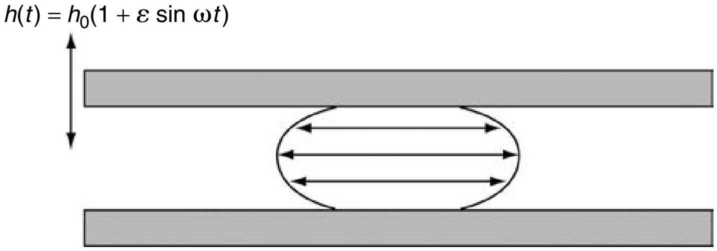


FIGURE 11.3 Schematic of squeeze-film damping between parallel plates. As with Couette damping, for most practical embodiments of MEMS, the damping is quasi-steady.

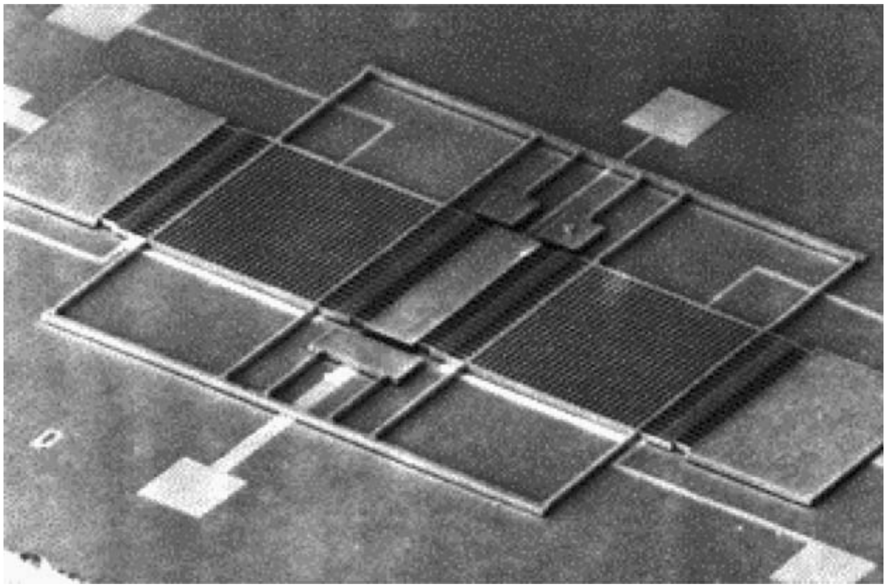


FIGURE 11.4 Photograph of a typical microfabricated vibrating proof mass used in a high-performance tuning fork gyroscope. (Reprinted with permission of M. Weinberg at Draper Laboratories.)

The Navier–Stokes equations are written for the case of a parallel plate vibrating sinusoidally in a proscribed manner in the vertical direction. If we assume that the motion, subsequent pressure, and velocity perturbations are small, a perturbation analysis yields the classical squeeze-film equation derived by Blech, with an additional term due to the rarefaction:

$$\frac{\partial}{\partial X}\left(\Psi H^3 \frac{\partial \Psi}{\partial X}\right) + \frac{\partial}{\partial X}\left(6KH^2 \frac{\partial \Psi}{\partial X}\right) = \sigma \frac{\partial(\Psi H)}{\partial T} \tag{11.11}$$

where the variables have been non-dimensionalized, so that H represents the film gap, normalized by the nominal film gap $H = h(x, y, t)/h_o$; Ψ is the pressure, normalized by ambient pressure $\Psi = P(x, y, t)/P_0$; X and Y are the coordinates, normalized by the characteristic plate geometry $X = x/L_x$, $Y = y/L_y$; T is time, normalized by the vibration frequency $T = \omega t$; and the squeeze number σ is defined as before:

$$\sigma = \frac{12\mu\omega L_x^2}{P_0 h_0^2} \tag{11.12}$$

Assuming small amplitude, harmonic forcing of the gap $H = 1 + \varepsilon \sin T$, and a harmonic response of the pressure, we can derive a pair of coupled equations describing the in-phase (Ψ_0) and out-of-phase (Ψ_1) pressure distributions in the gap representing stiffness and damping coefficients, respectively:

$$\frac{\partial^2 \Psi_0}{\partial x^2} + \frac{\sigma}{1 + 6K}\Psi_1 + \frac{\sigma}{1 + 6K} = 0$$

$$\frac{\partial^2 \Psi_1}{\partial x^2} + \frac{\sigma}{1 + 6K}\Psi_0 = 0 \tag{11.13}$$

Note that these equations represent the standard conditions with the adoption of a modified squeeze number, $\sigma_m \equiv \sigma/(1 + 6K)$. The solutions are achieved either by manual substitution of Fourier sine and cosine series or by direct numerical solution. The results for rectangular plates with no vent holes are shown in Figure 11.5.

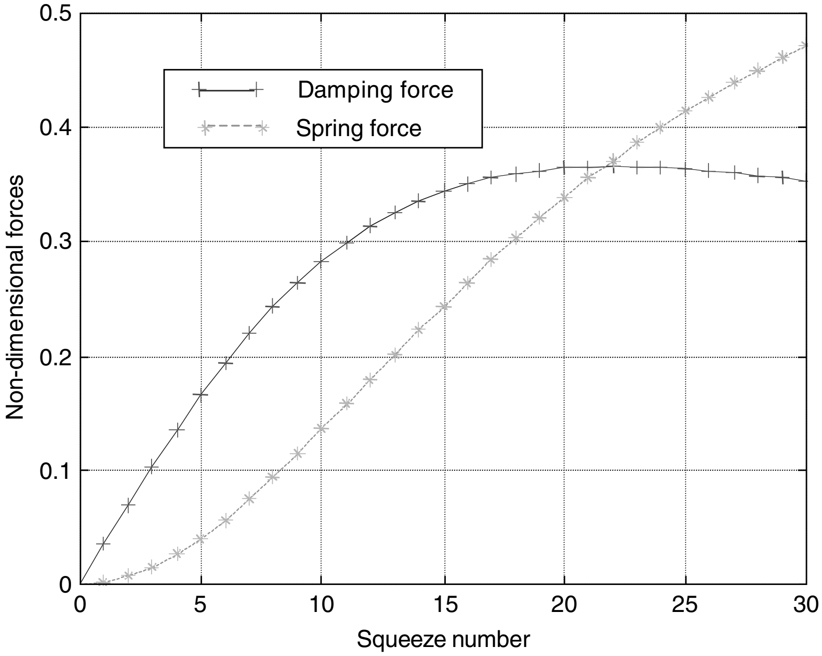


FIGURE 11.5 (**See color insert following page 10-34.**) Solutions to the squeeze-film equation for a rectangular plate. The stiffness and damping coefficients are presented as functions of the modified squeeze number, which includes a correction due to first-order rarefaction effects [Blech, 1983; Kwok et al., 2005].

11.5.2 Effects of Vent Holes

The equations as previously derived are made more useful by extending them to account for the presence of vent holes in the vibrating proof mass. In such cases the boundary condition at each vent hole is no longer atmospheric pressure ($\Psi_0 = \Psi_1 = 0$), but rather an elevated pressure proscribed by the pressure drop through the "chimney" which vents the squeeze film to the ambient. Kwok et al. (2005) demonstrate that this can be incorporated into the previous model (in the limit of low squeeze number) by a modified boundary condition for the squeeze-film equations for Ψ_0:

$$\Psi = \left[\frac{32\,\dfrac{t}{L_x}\left(\dfrac{h_0}{L_x}\right)^3\left(1-\left(\dfrac{L_h}{L_x}\right)^2\right)}{12\left(\dfrac{L_h}{L_x}\right)^4}\right]\left[\frac{1}{1+8\,\dfrac{\lambda}{L_h}}\right]\sigma \qquad (11.14)$$

This boundary condition has three components: a geometric component dependent on the plate thickness t, length L_x, hole size L_h, and nominal gap size h_0; a rarefaction component (here based on the hole size); and a time-dependent component — the squeeze number σ. Note that as the thickness of the plate decreases and the chimney pressure drop falls, the boundary condition approaches zero. Similarly, as the open area fraction of the plate increases (more venting), the boundary condition approaches that of the ambient. This boundary condition can be applied at the chimney locations and can accurately simulate the squeeze-film damping of perforated micromachined plates.

11.5.3 Reduced-Order Models for Complex Geometries

Most devices of practical interest have geometries that are too complex to enable full numerical simulation of the kind described previously. Reduced-order models are of great value in such cases. Many such models have been developed, including those based on acousto-electric models [Veijola et al., 1995]. In the case of squeeze-film damping in the limit of low squeeze numbers, such models reduce to solution of a resistor network that models the pressure drops associated with each segment of the squeeze film. This is effectively a finite-element approach to the problem. Instead of modeling a large number of elements, as is generally the case in a numerical solution, a relatively small number of discrete elements can be used, if higher-order solutions can be employed to connect each element together. Kwok et al. (2005) demonstrate this approach and model the damping associated with an inclined plate with vent holes. More complex numerical solution techniques based on boundary integral techniques have also been presented [Aluru and White, 1998; Kanapka and White, 1999] providing a good balance between solution fidelity and required computing power.

11.6 Lubrication in Rotating Devices

Rotating MEMS devices bring a new level of complexity to MEMS fabrication and to the lubrication considerations. As discussed in the introduction, many rotors and motors have been demonstrated with dry-rubbing bearings, and the success of these devices is due to the low surface speeds of the rotors. However, as the surface speed increases in order to get high power densities, the dry rubbing bearings are no longer an option, and true lubrication systems need to be considered. An example of "Power-MEMS" development is provided by a project initiated at the Massachusetts Institute of Technology in 1995 to demonstrate a fully functional microfabricated gas turbine engine [Epstein et al., 1997]. The baseline engine, illustrated in Figure 11.6, consists of a centrifugal compressor, fuel injectors (hydrogen is the initial fuel, although hydrocarbons are planned for later configurations), a combustor operating at 1600 K, and a radial inflow turbine. The device is constructed from single crystal silicon and is fabricated by extensive and complex fabrication of multiple silicon wafers that are fusion bonded in a stack to form the complete

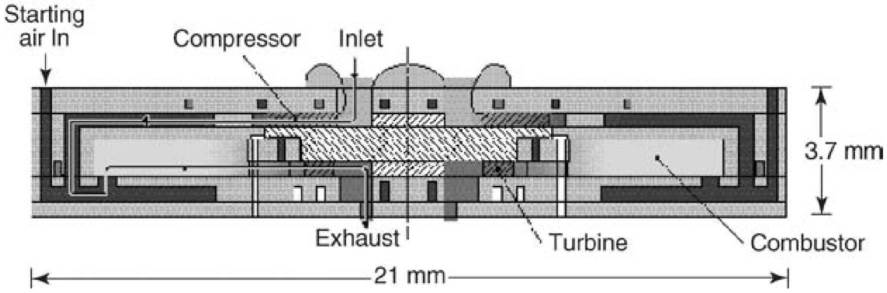


FIGURE 11.6 (**See color insert following page 10-34.**) Schematic of the MIT Microengine, showing the air path through the compressor, combustor, and turbine. Forward and aft thrust bearings located on the centerline hold the rotor in axial equilibrium, while a journal bearing around the rotor periphery holds the rotor in radial equilibrium.

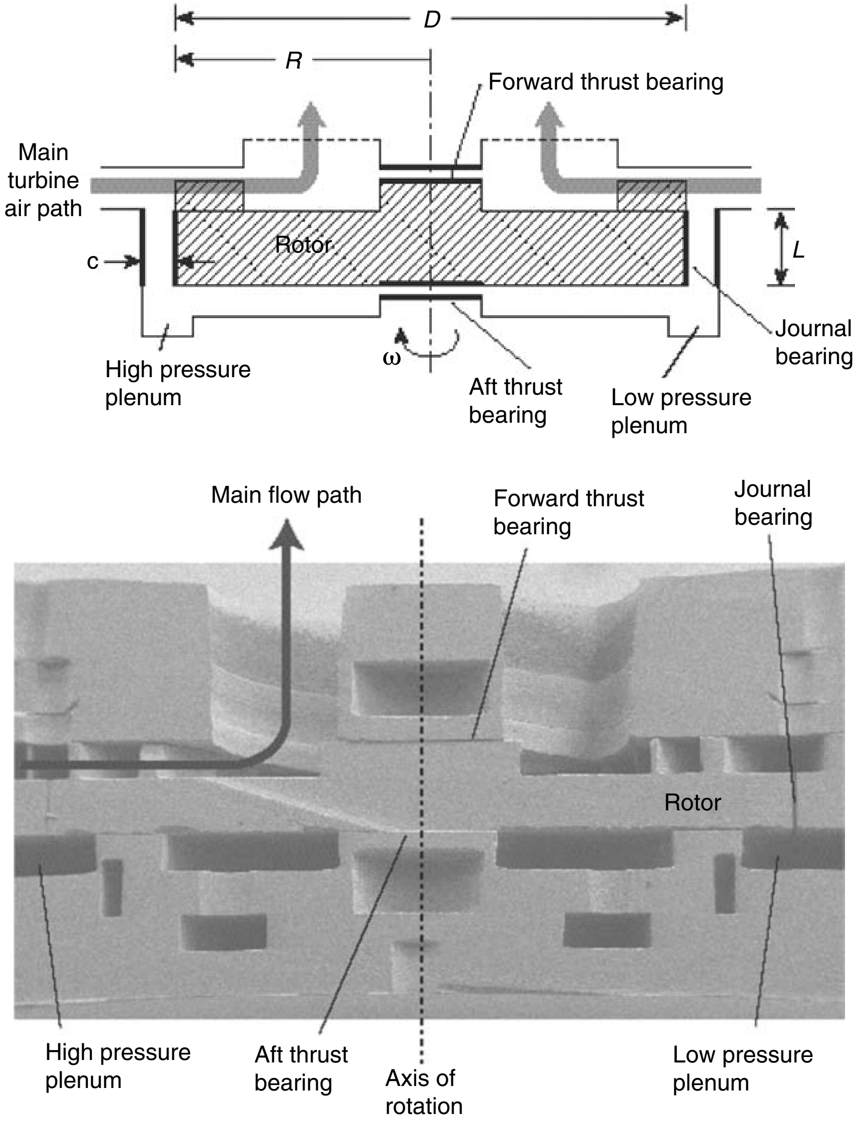


FIGURE 11.7 Illustrating schematic and corresponding SEM of a typical microfabricated rotor, supported by axial thrust bearings and a radial journal bearing.

device. An electrostatic induction generator may also be mounted on a shroud above the compressor to produce electric power instead of thrust [Nagle and Lang, 1999]. The baseline MIT Microengine has at its core a "stepped" rotor consisting of a compressor with an 8 mm diameter and a journal bearing and turbine with a diameter of 6 mm. The rotor spins at 1.2 million r/min.

Key to the successful realization of such a device is the ability to spin a silicon rotor at high speed in a controlled and sustained manner. The key to spinning a rotor at such high speeds is the demonstration of efficient lubrication between the rotating and stationary structures. The lubrication system needs to be simple enough to be fabricated but with sufficient performance and robustness to be of practical use in the development program and in future devices. Figure 11.7 illustrates a microbearing rig that was fabricated to develop this technology. The core of the rotating machinery has been implemented but without the substantial complications of the thermal environment that the full engine brings. The rig consists of a radial inflow turbine mounted on a rotor and embedded inside two thrust bearings that provide axial support. A journal bearing located around the disk periphery provides radial support for the disk as it rotates.

11.7 Constraints on MEMS Bearing Geometries

11.7.1 Device Aspect Ratio

Perhaps the most restrictive aspect of microbearing design is that MEMS devices are limited to rather shallow etches, resulting in devices with low aspect ratio. Even with the advent of deep reactive ion etchers (DRIE) in which the ion etching cycle is interleaved with a polymer passivation step, the maximum practical etch depth that can be achieved while maintaining dimensional control is about 500 microns. Even this has an etch time of about nine hours, which makes its adoption a very costly decision. In comparison, typical rotor dimensions are a few millimeters. The result is that microbearings are characterized by very low aspect ratios (Length/Depth, or *L*/*D*). In the case of the MIT microturbine test rig, the journal bearing is nominally 300 microns deep while the rotor is 4 mm in diameter, yielding an aspect ratio of 0.075. To put this in perspective, commonly available design charts [Wilcox, 1972] present data for values of *L*/*D* as low as 0.5 or perhaps 0.25. Prior to this work there was no data for lower *L*/*D*. The implications of the low aspect ratio bearings are that the task becomes supporting a disk rather than a shaft.

The low aspect ratio bearings do not have terrible performance by any standard. The key features of the low *L*/*D* bearings are:

The load capacity is reduced compared to conventional designs. This is because the fluid leaking out of the ends relieves any tendency for the bearing to build up a pressure distribution. For a given geometry and speed, a short bearing supports a lower load per unit length than its longer counterpart.

The bearing acts as an incompressible bearing over a wide range of operation. Pressure rises, which might lead to gas compressibility, are minimized by the flow leaking out of the short bearing. Incompressible behavior (without the usual fluid cavitation that is commonly assumed in incompressible liquid bearings) can be observed to relatively high speeds and eccentricities.

11.7.2 Minimum Etchable Clearance

It is reasonable to question why one could not fabricate a 300 micron long “shaft”, but with a much smaller diameter, to greatly enhance the *L*/*D*. For example, a shaft with a diameter of 300 microns would result in a reasonable value for *L*/*D*. This raises the second major constraint on bearing design by current microfabrication technologies — that of the minimum etchable clearance.

In the current microengine manufacturing process, the bearing and rotor combination is defined by a single deep and narrow etch, currently 300 microns deep and about 12 microns in width. No foreseeable advance in fabrication technology will make it possible to significantly reduce the minimum etchable clearance, and this has considerable implications for bearing design. In particular, if one were to fabricate a bearing with a diameter of 300 microns in an attempt to improve the *L*/*D* ratio, the result would be a bearing with a clearance to radius *c*/*R* of 12/300, or 0.04. For a fluid bearing, this is two orders of magnitude above conventional bearings and has several detrimental implications.

The most severe implication is the impact on the dynamic stability of the bearing. The non-dimensional mass of the rotor depends on (*c*/*R*) raised to the fifth power [Piekos, 2000]. Bringing the bearing into the center of the disk and raising the *c*/*R* by a factor of 10 results in a mass parameter increasing by a factor of 10^5. This increase in effective mass has severe implications for the stability of the bearing.

These reasons and others not enumerated here make the implementation of an inner-radius bearing less attractive. Therefore, the constraint of small *L/D* is unassailable as long as one requires that the microdevice be fabricated in situ. If one were to imagine a change in the fabrication process such that the rotor and bearing could be fabricated separately and subsequently assembled reliably, this situation would be quite different. In such an event, the bearing gap is not constrained by the minimum etch dimension of the fabrication process, and almost any "conventional" bearing geometry could be considered and would probably be superior in performance to the bearings discussed here. Such fabrication could be considered for a "one-off" device, but does not appear feasible for mass production, which relies on the monolithic fabrication of the parts. Lastly, the risk of contamination during assembly — a common concern for all precision-machined MEMS — effectively rules out piece-by-piece manufacture and assembly and constrains the bearing geometry as described.

11.8 Thrust Bearings

Thrust bearings support any axial loads generated by rotating devices such as turbines, engines, or motors. Current fabrication techniques require that the axis of rotation in MEMS devices lie normal to the lithographic plane. This lends a significant advantage in the design and operation of thrust bearings because the area available for the thrust bearing is relatively large as defined by lithography, while the weight of the rotating elements will be typically small due to the cube-square law and the low thicknesses of microfabricated parts. For these reasons, thrust bearings are one area of microlubrication where solutions abound and problems are relatively easily dealt with.

Two thrust bearing options exist: (a) hydrostatic (externally-pressured) thrust bearings, in which the fluid is fed from a high-pressure source to a lubrication film, and (b) hydrodynamic, where the supporting pressure is generated by a viscous pump fabricated on the surface of the thrust bearing itself (see Figure 11.9). Hydrostatic bearings are easy to operate and relatively easy to fabricate. These have been successfully demonstrated in the MIT Microengine program [Frechette et al., 2005; Liu et al., 2003]. The thrust bearing in Figure 11.8 shows an scanning electron micrograph (SEM) of the fabricated device cut though the middle to reveal the plenum, restrictor holes, and the bearing lubrication gap, which is approximately 1 micron wide. Key to the successful operation of hydrostatic thrust bearings is the accurate

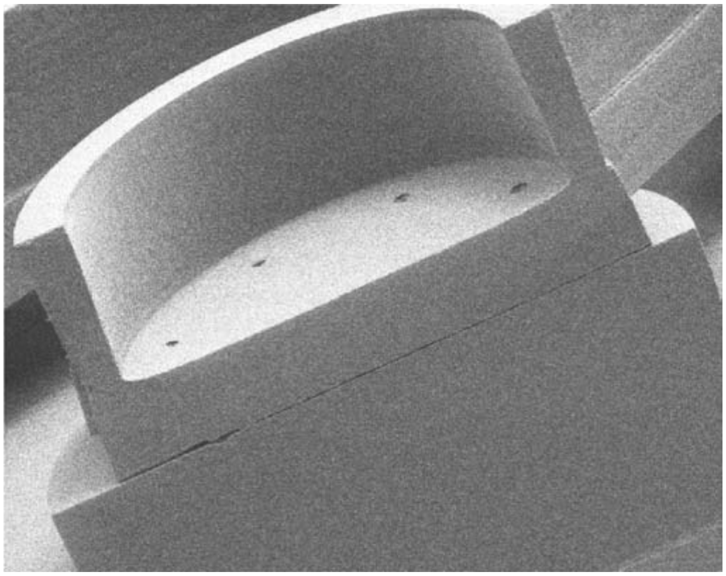


FIGURE 11.8 Close-up cutaway view of micro thrust bearing showing the pressure plenum (on top), the feed-holes, and the bearing gap (faintly visible). (SEM reprinted with permission of Lin et al. [1999].)

manufacture of the restrictor holes, maintenance of sharp edges at the restrictor exit, and careful control of the dimension of the lubrication film. In an initial fabrication run, the restrictor holes were fabricated 2 microns larger than specified. While the bearing operated, its performance was well below its design peak because of the off-design restrictor size. Current specifications of the fabrication protocols control the restrictor size carefully, ensuring close to optimal operation.

Hydrodynamic or spiral groove bearings (SGBs), illustrated in Figure 11.9, were first analyzed in detail forty years ago [Muijderman, 1966] but have not received much attention due to their low load capacity compared to hydrostatic thrust bearings and due to complex manufacturing requirements.

SGBs operate by using the rotor motion against a series of spiral grooves etched in the bearing to viscously pump fluid into the lubrication gap. This process creates a high-pressure cushion on which the rotor can ride. The devices typically have relatively low load capacity, which has limited their use in macroscopic applications. The load capacity becomes more than adequate at microscales due to favorable cube-square

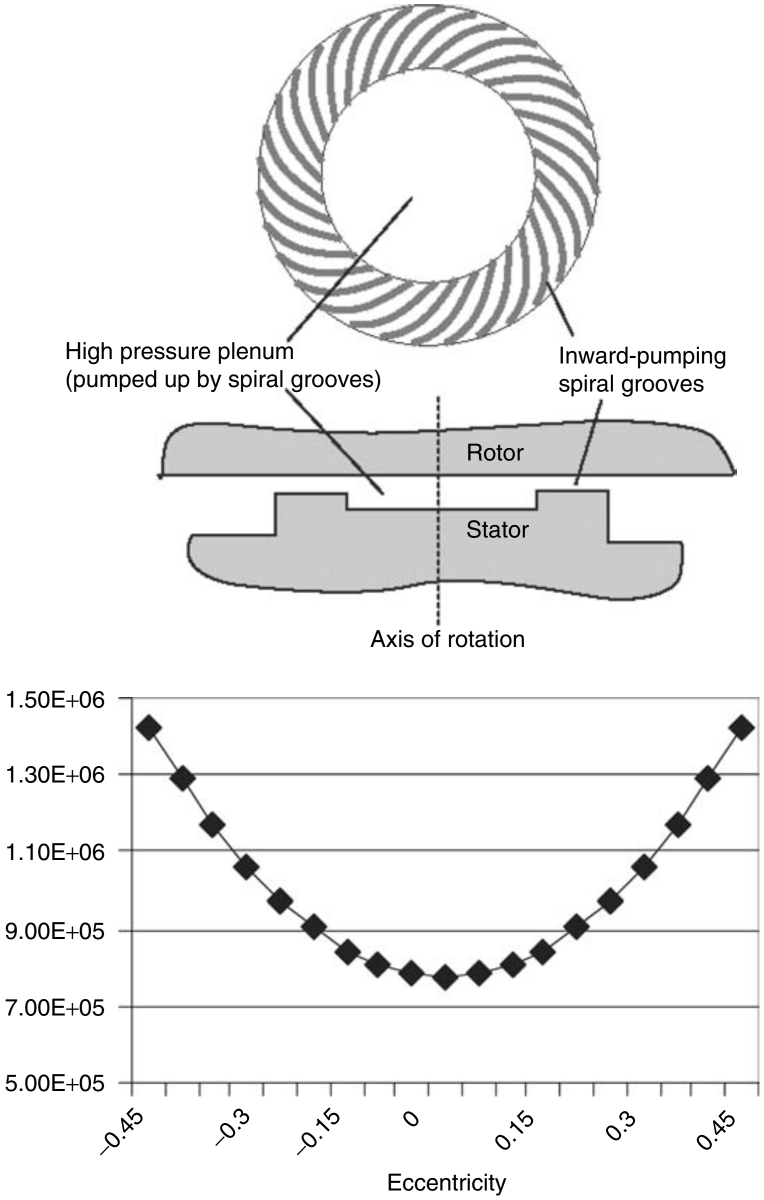


FIGURE 11.9 Schematic of hydrodynamic thrust bearings and predicted performance (stiffness in N/m vs. axial eccentricity) for a typical spiral groove thrust bearing for use in a high speed MEMS rotor.

scaling. Thus, they gain considerable advantage when compared with conventional hydrostatic thrust bearings as the scale and Reynolds number decreases. In addition, the fabrication of the multitude of shallow spiral features, which is an expensive task for a traditional SGB, is ideally suited for lithographic fabrication technologies such as MEMS.

Figure 11.9 illustrates the bearing stiffness for a particular single-point design for the MIT microrotor rotating at design speed (2.4 million r/min) and supported by matched forward and aft spiral groove bearings. The stiffness at full speed is quite impressive and superior to comparable hydrostatic bearings, but the SGB do suffer at lower speeds since the bearing stiffness is roughly proportional to rotational speed. For this design, the lift-off speed (the speed at which the film can support the weight of the rotor and the pressure distribution associated with the turbine flow) is only a few thousand r/min, and the dry rubbing endured during startup will be minimal. SGBs also have the strong advantage that the two matched spiral groove bearings, forward and aft, naturally balance each other with no supply pressures to maintain or adjust, and the removal of the thrust bearing plena and restrictor holes considerably simplifies the overall device fabrication. This simplification allows for the use of two fewer wafers in the wafer-bonded stack, which is a considerable advantage from the perspective of manufacturing process cost and yield. A hybrid bearing consisting of both hydrostatic and hydrodynamic bearings has been recently successfully demonstrated [Wong et al., 2004] up to a speed of approximately 450,000 r/min.

11.9 Journal Bearings

Journal bearings, which are used to support radial loads in a rotating machine, have somewhat unusual requirements in MEMS. These requirements derive from the extremely shallow structures that are currently fabricated. Rotating devices tend to be disk-shaped, and their corresponding journal bearings are characterized by very low aspect ratios which are defined as the ratio of the bearing height to its radius. In addition, the minimum etchable gap allowed by current fabrication techniques results in a paradoxically large bearing clearance — a 10 micron gap over a 2 mm radius rotor, or a c/R of 1/200. This bearing clearance is large in comparison to conventional journal bearings, which typically have c/R ratios that are smaller by a factor of 10.

Journal bearings can operate in two distinct modes: hydrodynamic and hydrostatic. Typically any operating condition will contain aspects of both modes. These modes are discussed in the following sections.

11.9.1 Hydrodynamic Operation

Hydrodynamic operation occurs when the rotor is forced to operate at an eccentric position in the bearing housing. As a result, a pressure distribution develops in the gap to balance the viscous stresses that arise due to the rotor motion. This pressure distribution supports the rotor statically against the applied force and dynamically to suppress random excursions of the rotor due to vibration, etc. Hydrodynamic operation has the advantage of requiring no external supply of lubrication fluid. However, it has two distinct drawbacks: it requires a means to load the rotor to an eccentric position, and insufficient eccentricity results in instability (the so-called "fractional speed whirl") and likely failure. Both of these issues are particularly difficult in the case of MEMS bearings.

11.9.1.1 Static Journal Bearing Behavior

Figure 11.10 shows the static behavior of a MEMS journal bearing. This figure presents the load capacity ζ and the accompanying attitude angle (the angle between the applied load and the eccentricity vector) as functions of the bearing number and the operating eccentricity. The geometry considered here is for a low-aspect ratio bearing ($L/D = 0.075$) typical of a deep reactive ion etched rotor such as the MIT microengine. The bearing number is defined as:

$$\Lambda = \frac{6\mu\omega}{p}\left[\frac{R}{c}\right]^2 \tag{11.15}$$

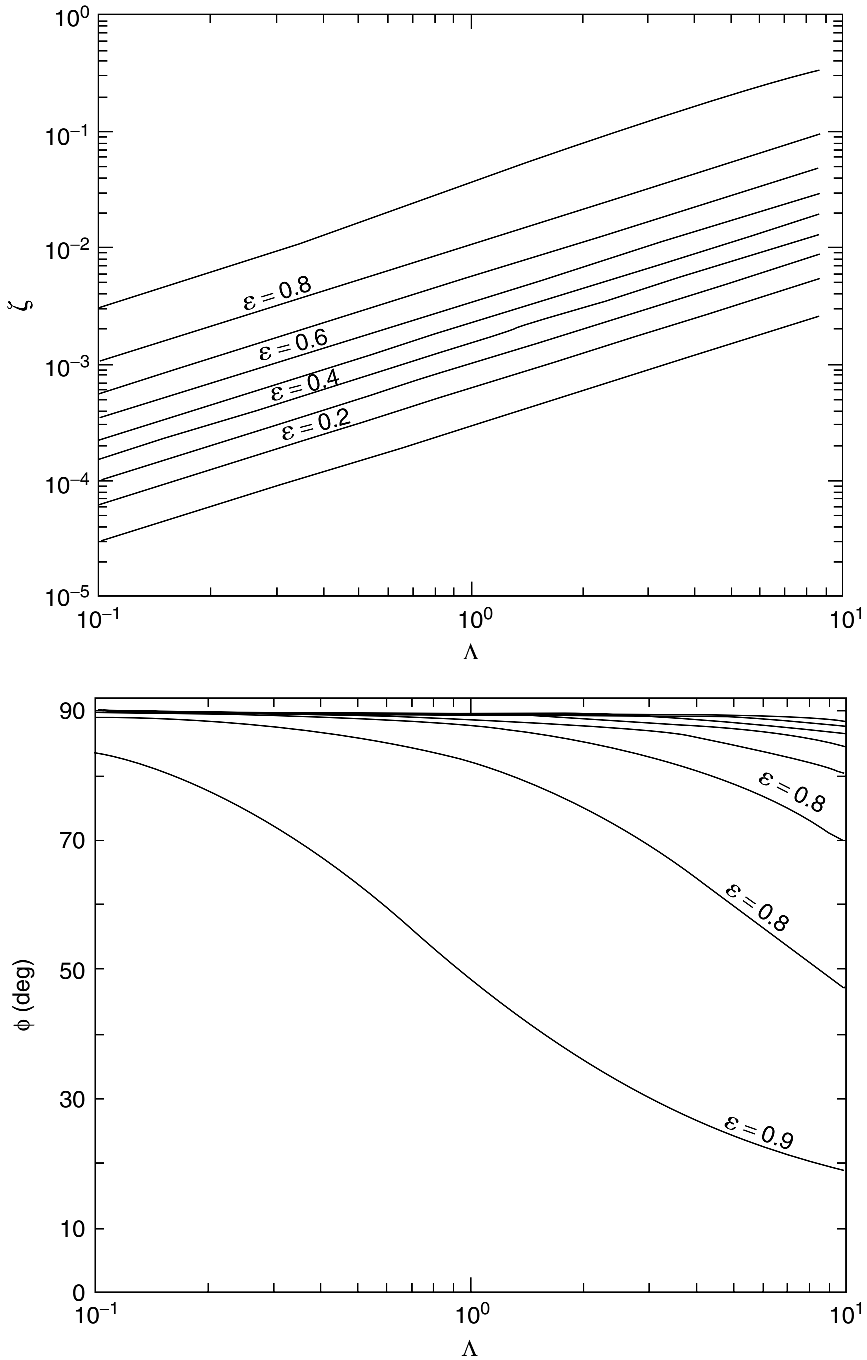


FIGURE 11.10 Static performance (eccentricity and attitude angle vs. bearing number) for a journal bearing with $L/D = 0.075$. Notice that the load lines are almost constant (linear), indicating the absence of compressibility effects. This is also indicated by the attitude angle, which remains close to 90 degrees except at very high eccentricities [Piekos and Breuer, 1998].

where μ is the fluid viscosity, ω the rotation rate, p the ambient pressure, and R/c the ratio of the radius to clearance. For a given bearing geometry, Λ can be interpreted as operating speed.

Several aspects of these results should be noted. The load capacity is quite small when compared with bearings of higher L/D. This is because for very short bearings, the applied load simply squeezes the fluid out of the bearing ends, and consequently it is difficult to develop any significant restoring force. The same mechanism is responsible for the load lines being straight. Straight load lines indicate that little compressibility of the fluid is taking place, which usually results in a "saturation" of the load parameter at higher values of the bearing number. Again, this is because any tendency to compress the gas is alleviated by the fluid venting at the bearing edge. The behavior of the attitude angle, which maintains a high angle (close to $\pi/2$) over a wide range of bearing numbers and eccentricities, illustrates this point. This value of the attitude angle corresponds to the analytic behavior of a Full-Sommerfeld incompressible short bearing [Orr, 1999]. This value is a good approximation for such short bearings at low to moderate eccentricities when the eccentricity remains below approximately 0.6. Below 0.6, compressibility finally becomes important. This incompressible behavior is much more extensive than conventional gas bearings of higher aspect ratio and has profound ramifications, particularly with respect to the dynamic properties of the system.

11.9.1.2 Journal Bearing Stability

The stability of a hydrodynamic journal bearing has long been recognized as troublesome and is foreshadowed by the static behavior shown in Figure 11.10. The high attitude angle suggests that the bearing spring stiffness is dominated by cross stiffness as opposed to direct stiffness. Thus, any perturbation to the rotor will result in its motion perpendicular to the applied force. If this reaction is not damped, the rotor will enter a whirling motion. This is precisely what is observed, and gas bearings are notorious for their susceptibility to fractional-speed whirl. The instability is suppressed by the generation of more damping and increased direct stiffness, both of which are obtained by increasing the loading and the static eccentricity of the rotor.

Figure 11.11 shows a somewhat unusual presentation of the stability boundaries for a low-aspect ratio MEMS journal bearing. The vertical axis shows the non-dimensional mass of the rotor $\overline{M}$ which is defined as:

$$\overline{M} = \frac{mp}{72L\mu^2}\left[\frac{c}{R}\right]^5 \tag{11.16}$$

This is the "mass" which appears in the non-dimensionalized equations of motion for the rotor and it is fixed for a given geometry. Close inspection of Figure 11.11 indicates that $\overline{M}$ does changes very slightly with speed. This is because of the elastic expansion of the rotor due to centrifugal forces, the variation in the ambient pressure at different speeds, and temperature effects on viscosity. The horizontal axis of Figure 11.11 shows the bearing number, which can be interpreted as speed, for a fixed bearing geometry. The contours on the graph represent the stability boundary at fixed eccentricity. Stable operation lies above each line. For a fixed $\overline{M}$ at low bearing number (i.e., speed), a minimum eccentricity must be

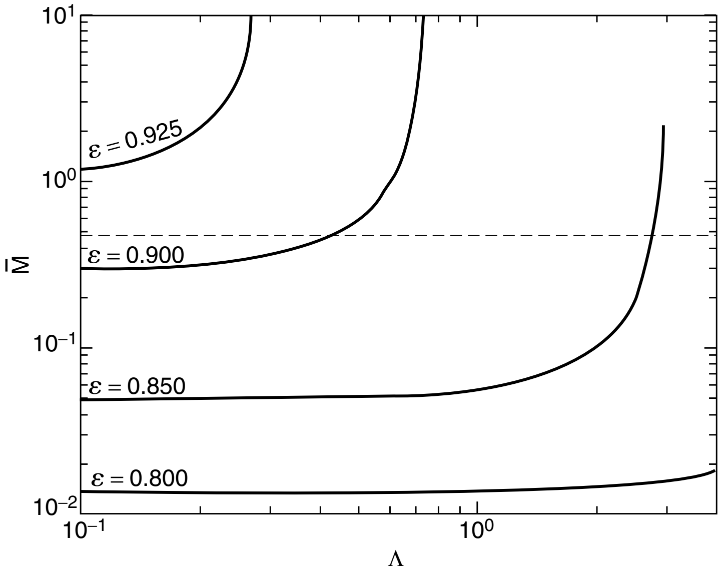


FIGURE 11.11 Stability boundaries for a typical microbearing plotted vs. bearing number (speed for a fixed geometry). The dotted line represents an operating line for a microbearing which has almost constant $\overline{M}$ (varying only due to centrifugal expansion of the rotor at high speeds [Piekos, 2000]).

obtained to ensure stability. As the speed increases, this minimum eccentricity remains almost constant (the lines are horizontal) until a particular speed at which the lines break upward, and the minimum eccentricity required for stability starts to drop as indicated by Figure 11.10, as Λ increases, the load required to maintain a fixed eccentricity increases linearly due to the stiffening of the hydrodynamic bearing. The key feature of this chart is that the minimum eccentricities are very high and suggest that stable operation requires running very close to the wall. This is troublesome. The high eccentricities are driven by high values of the mass parameter $\overline{M}$ which is due to the relatively high value of the clearance-to-radius ratio (c/R) and the short length L. The low aspect ratio (L/D) also contributes to high minimum eccentricities. At high speeds, the problem becomes less severe, because the high speed allows the bearing to generate sufficient direct stiffness. Even at these points, the eccentricity is very high and might not be manageable in practical operation.

Orr (1999) has demonstrated on a scaled-up experimental rig that matches the microengine geometry that stable high-eccentricity operation is possible for extended periods of time. His experiments achieved 46,000 r/min which, when translated to the equivalent speed at the microscale, correspond to approximately 1.6 million r/min. In order to accomplish this high eccentricity operation, he noted that the rotor system must (a) have very good axial thrust bearings to control axial and tipping modes of the rotor system, and (b) be well-balanced. A rotor with imbalance of more than a few percent could not be started from rest. Piekos (2000) also explored the tolerance of the microbearing system to imbalance and found that it was surprisingly robust to imbalance of several percent. His computations were achieved assuming that the rotor was at full speed and then carefully subjected to imbalance. In practice, the imbalance will exist at rest, and the rotor is stable at full speed but unable to accelerate to that point. This "operating line" issue is discussed in more detail by Savoulides et al. (2001) who explored several options for accelerating microbearings from rest under both hydrodynamic and hydrostatic modes of operation.

Figure 11.12 illustrates a convenient summary of the trade-offs for design of a hydrodynamic MEMS bearing. This figure presents the variation of the low-speed minimum eccentricity asymptote, or worst-case eccentricity, as a function of the mass parameter $\overline{M}$ and other geometric factors (L/D, clearance, c,

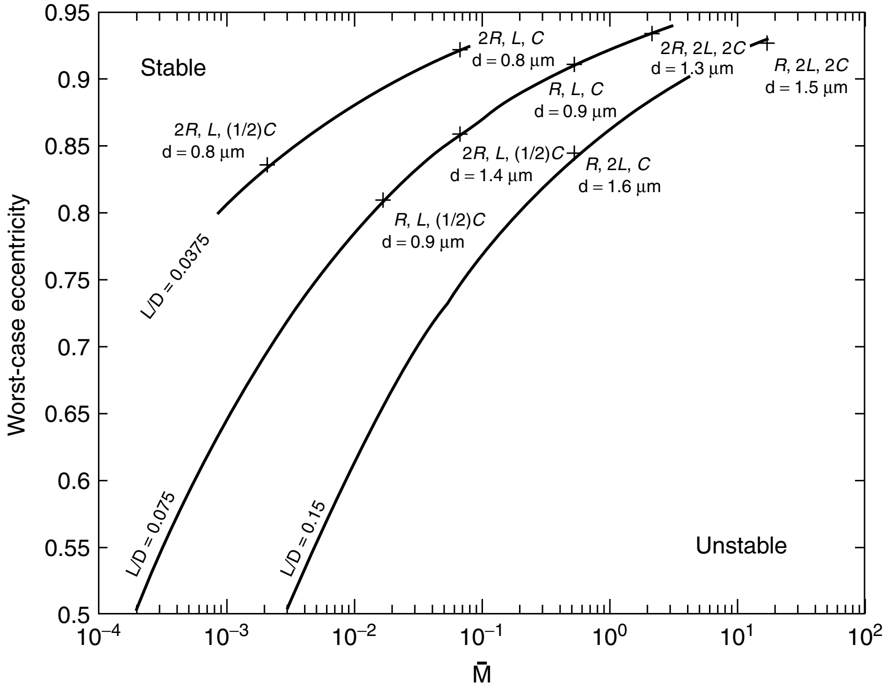


FIGURE 11.12 Tradeoff chart for microbearing design. For a given length-to-diameter (L/D) and a given $\overline{M}$, the worst-case (i.e., low-speed) eccentricity is shown for a variety of geometric perturbations. In general, lower eccentricities are preferred. (Reprinted with permission from Piekos [2000].)

etc.). Notice that the worst-case eccentricity improves as the *L/D* increases and the $\overline{M}$ decreases. However, the physical running distance from the wall is actually increased slightly by running at a higher eccentricity with a larger bearing gap. In all cases, the stable eccentricity is alarmingly high, and other alternatives need to be sought for simpler stable operation.

11.9.2 Advanced Journal Bearing Designs

One prospect for further improvement in the journal bearing performance is the incorporation of wave bearings [Dimofte, 1995] as illustrated in Figure 11.13. These bearing geometries suppress the subsynchronous whirl due to the excitation of multi-synchronous pressure perturbations imposed by the bearing geometry. The geometric complexity of the wave bearing is no problem for lithographic manufacturing processes that are used for MEMS. This alleviates many of the reservations and costs that might inhibit their adoption. Because the MEMS constraint is the minimum gap dimension, the wave bearing in a MEMS machine can be implemented only by selectively enlarging the bearing gap. Piekos (2000) analyzed the performance of the wave bearing for the microengine geometry and found (Figure 11.14) that while the load capacity is diminished, the stability is enhanced and the load required to maintain stable operation (i.e., to achieve the minimum stable eccentricity) is reduced considerably with the introduction of a wave geometry. In microbearings the load capacity is usually sufficient, and the wave bearing is attractive as a stabilizing mechanism.

Rotor imbalance, which is increasingly becoming a first-order issue, can only be contained with excess load capacity, and this tradeoff is not clear. The adoption and testing of wave bearing geometries are scheduled to be explored as part of our development program.

11.9.3 Side Pressurization

Due to the small mass of the rotor in a MEMS device, any eccentricity required to enable stable hydrodynamic operation must be applied using some other means. Typically, this requires the use of a pressure distribution introduced around the circumference of the bearing. This pressure distribution loads the bearing preferentially to one side. Such a scheme is illustrated in Figure 11.15 for the MIT Microengine.

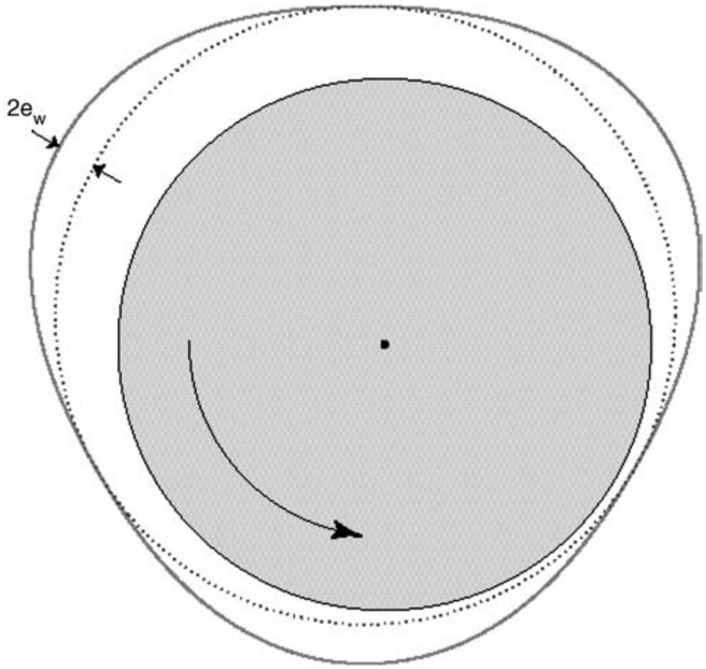


FIGURE 11.13 **(See color insert following page 10-34.)** Geometry of a wave bearing, with the clearance greatly exaggerated for clarity [Piekos, 2000].

The aft side of the rotor is divided into two plena isolated by seals. Each plenum can be separately pressurized. The pressure in each plenum forces an axial flow through the journal bearing to the forward side (which is assumed to be at a uniform pressure), and thus establishes two differing pressure distributions on the high- and low-pressure sides of the rotor. As a result, the axial flow through the bearing generates a hydrostatic stiffness mechanism and an associated hydrostatic critical frequency. These results are discussed in the following section.

11.9.4 Hydrostatic Operation

Although hydrodynamically lubricated bearings with low aspect ratio are predicted to operate successfully and have been demonstrated on a scaled-up level [Orr, 1999], there are a number of issues that make

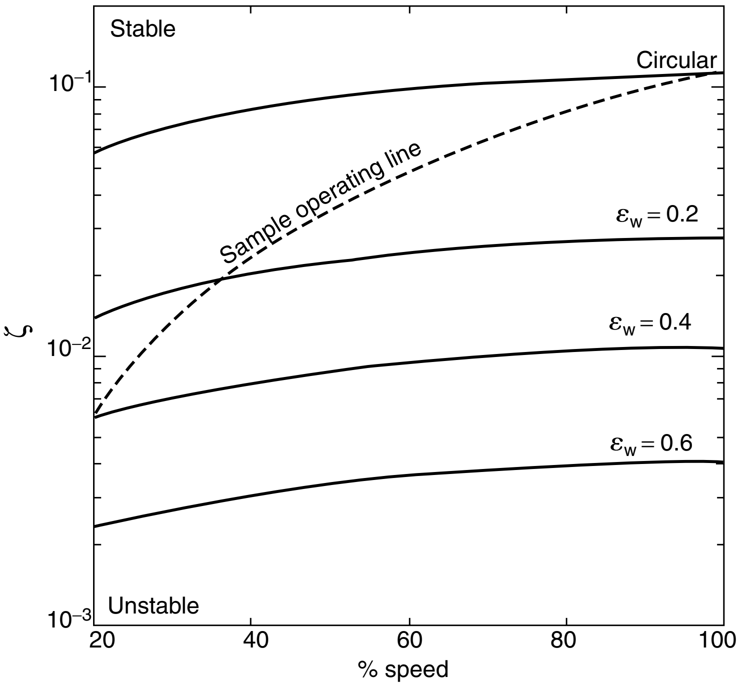


FIGURE 11.14 Effect of wave bearing amplitude on journal bearing stability as a function of rotor speed. The dotted line shows a typical operating line for a microengine [Piekos, 2000].

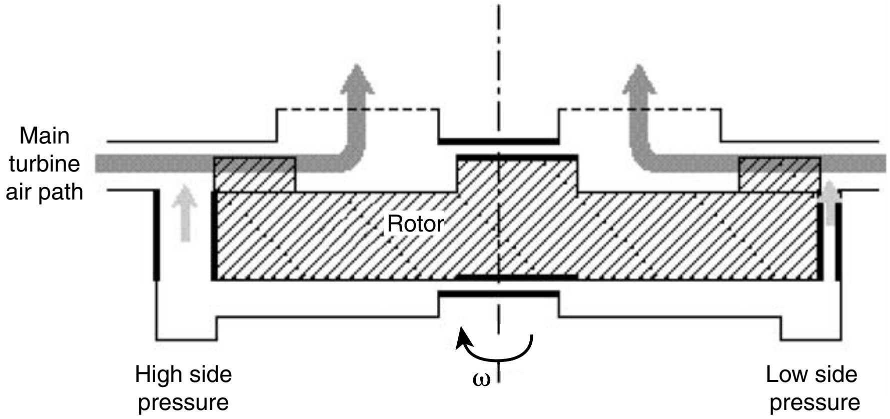


FIGURE 11.15 Schematic of the pressure-loading scheme used in the microengine to provide a side load to the rotor during hydrodynamic operation. The side load is developed by applying a differential pressure to the two plena located on the aft side of the rotor.

them undesirable in a practical MEMS rotor system. The primary difficulty is that, in order to satisfy the requirements of sub-synchronous stability, the rotor needs to operate at very high eccentricity (made unavoidable due to the low aspect ratio of the journal). For a MEMS device this means operating 1–2 microns from the wall. This is hard to control, particularly with the limited available instrumentation for MEMS devices. An alternative mode of operation is to use a hydrostatic lubrication system. In this mode, fluid is forced from a high-pressure source through a series of restrictors, all of which impart a fixed resistance. The fluid then flows through the lubrication passage (the bearing gap). If the rotor moves to one side, the restrictor and lubrication film act as a pressure divider such that the pressure in the lubrication film rises, forcing the rotor back towards the center of the bearing. The advantages of using hydrostatic lubrication in MEMS devices are that:

The rotor tends to operate near the center of the housing, and small clearances are avoided. This is safer, more tolerant of any motion induced by rotor imbalance, and results in lower viscous resistance.

Because the hydrostatic system is a zero-eccentricity based system, no position information about the rotor is needed. This greatly simplifies instrumentation requirements.

There are significant disadvantages to a hydrostatic system, including:

Pressurized air needs to be supplied to the bearing. This requires supply channels, which complicate the fabrication process and come with a system cost: the high-pressure air must come from somewhere. In a turbomachinery application, bleed air from the compressor could be used.

Since the bearing gaps are relatively large due to minimum etchable dimensions previously discussed, the mass flow through hydrostatic systems can be substantial and might be impractical in anything but demonstration experiments.

Fabrication constraints make the manufacture of effective flow restrictors very difficult. Flow restrictors need to have very well controlled dimensions, sharp edges, and other specific geometric features. Only the simplest restrictors can be implemented without undue cost and effort, severely limiting the hydrostatic design.

Orr (1999) demonstrated a novel method for achieving hydrostatic lubrication for journal bearings with low aspect ratio. The mechanism relies on the small pressure differences that exist between the forward and aft sides of the rotor. The mechanism also relies on the flow resistance to the pressure differences being small enough such that an axial flow will ensue for a short bearing of the kind seen in MEMS devices.

As the flow enters the bearing channel, boundary layers develop along the wall eventually merging to form the fully developed lubrication film. This boundary layer development (Figure 11.16) acts as an inherent restrictor. If the rotor moves off the centerline and disturbs the axisymmetric symmetry of the flow, a restoring force is generated. This source of hydrostatic stiffness supports the bearing at zero eccentricity

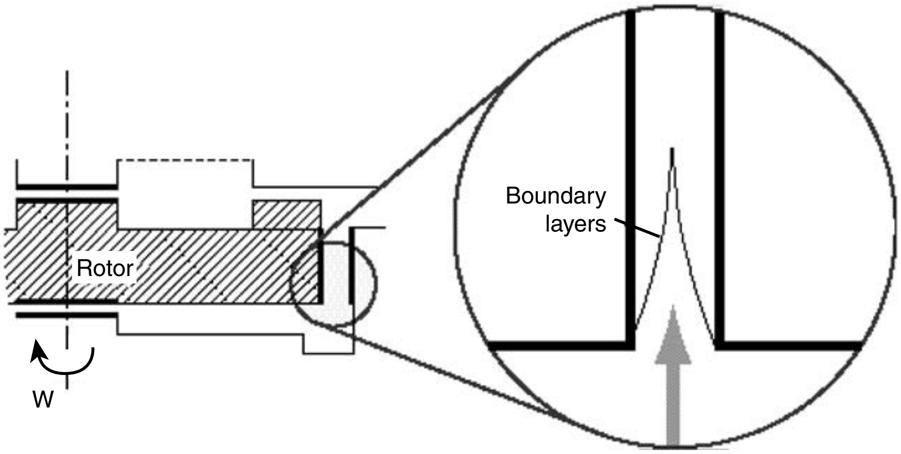


FIGURE 11.16 Schematic illustrating the origin of the axial-through-flow hydrostatic mechanism.

and is effective even when the rotor is not moving. The conventional inherent restriction of the flow entering the lubrication channel also enhances the stiffness. The stiffness coupled with the rotor mass defines a natural frequency which was measured by Orr (1999). The presence of this frequency led to the discovery of the axial-through-flow mechanism. Simple theory [Orr, 1999] was also able to predict the frequency in a scaled-up experimental facility with reasonable accuracy (Figure 11.17).

There is a severe gap in our ability to accurately predict and account for all the hydrostatic lubrication phenomena in a real microrotor. Experiments conducted at the microscale [Frechette et al., 2005] demonstrate successful operation at high speeds (1.4 million r/min) despite theoretical predictions of failure. Experimental measurements suggest that the natural frequency is higher than predicted by the simple axial-through-flow theory of Piekos (2000) and that the damping is sufficient to operate at critical speed ratios (rotor frequency, scaled by the natural frequency of the hydrostatic system) greater than 10. Conventional analysis [Orr, 1999] suggests that the instability occurs at critical speed ratios of 2. This discrepancy suggests that the real bearing exhibits significantly higher damping than is accounted for by the theory, perhaps

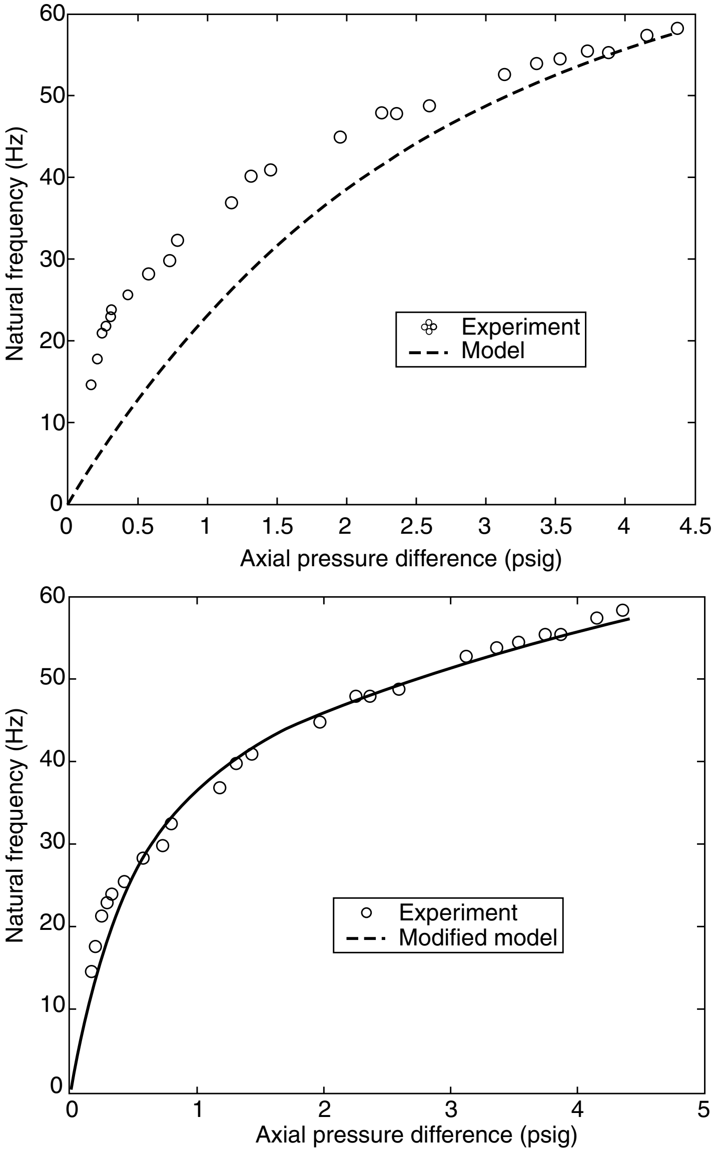


FIGURE 11.17 Prediction and measurement of natural frequency associated with axial-through-flow in a low-aspect ratio microbearing. The left-hand frame shows the measurements (from a 26:1 scaled-up experimental rig) along with the theoretical predictions based on the assumed geometry. The right-hand frame shows the same measurements compared with the same model but using slightly modified geometric parameters. [Orr, 1999].

deriving from the turbine which drives the rotor or some other source of fluid damping not yet considered. The resolution of these discrepancies need more attention and will be aided greatly by improved models and more detailed measurements of the microrotor in operation.

11.10 Fabrication Issues

A key challenge to the successful operation of a high-speed microbearing is the accurate fabrication of bearing geometries. Two aspects of this challenge need to be considered: the need to hold design tolerances in any given fabrication process, and the ability to manufacture multiple devices with good uniformity in a single fabrication run.

The issue of achieving design tolerances is a matter of process maturity. The attention paid to the maintenance of tight tolerances and small details is the hallmark of a well-established fabrication process. The microengine process is very complex and continually advances the state of the art in micromachining complexity. Almost any fabrication run that results in a freely rotating turbine should be considered a manufacturing triumph. From the standpoint of the success of the system, we must have much more stringent manufacturing requirements. The bearing designs are sensitive to critical dimensions such as the bearing-rotor gap and the size of restrictor holes for hydrostatic injectors. The failure to hold these dimensions within a specified tolerance can make the difference between a device that operates with a lubricated film and one that grinds the rotor and stator surface until failure. The very first version of the microbearing rig ran in this mode with occasional demonstrations of lubricated operation. Subsequent designs and builds have paid attention to dimensional accuracy, and the fabrication protocols are quite mature so that this precision is ensured from one build to the next.

11.10.1 Cross-Wafer Uniformity

Typically, multiple microengines are fabricated in parallel on a single silicon wafer. In addition to the accurate manufacture of critical dimensions on a single microengine die, the importance of manufacturing uniformity from one die to the next on a single wafer is vital. Manufacturing unity is a major obstacle to device yield. It is very common for a given process to exhibit cross-wafer variations. For example, a shallow plasma etch into a silicon substrate might show a variance of as much as 10% from one side of the wafer to the other because of variations in the plasma that are intrinsic to the fabrication tool. All fabrication processes will exhibit such variations, and any microfabrication process needs to identify and accommodate these variations. Should the variations be unacceptably large, either the fabrication tool needs to be improved, or a different processing path needs to be considered. This need is a common driver throughout both the MEMS and microelectronics industry. This industry also desires greater process uniformity as feature size diminishes and processing moves to larger and larger wafers.

As previously discussed, there are several critical etches that need to be controlled to a high degree of precision for microbearing design. The difficulty in maintaining cross-wafer uniformity results in some operational devices on the wafer (typically from the center of the wafer, where the process was initially honed to precision). Many devices from the wafer edge are out of specification and will not operate satisfactorily. At this stage, most of the uniformity issues have been addressed. However, two items are still troublesome. The deep reactive ion etcher being a relatively new tool exhibits a fairly significant variation in etch rate between the center and edge of a wafer. This variation results in a gradient in etch depth that is particularly severe on devices lying on the wafer periphery (3 microns variation across a 4 mm rotor wheel). This gradient contributes to a mass imbalance of as much as 25% of the bearing gap, rendering the bearing inoperable at high speed. The imbalance force increases with the square of the rotational speed.

The second continuing difficulty is that of front-to-back mask alignment during fabrication. It is common during the fabrication process for a single silicon wafer to be patterned on both the front and back surfaces. For example, the rotor has the turbine blades patterned from one side and the bearing gap patterned from the other side. Any slight misalignment between the lithography on the front and back surfaces of the wafer will result in an offset of front and back features which, as with the etch-depth gradient, leads to a rotor imbalance. Currently, mask alignment of critical features, primarily the rotor blades and the bearing gap, must be maintained to within 0.5 microns or better in order to ensure operable rotors from every die on the wafer. This is an extremely tight, but achievable, tolerance, and work continues to improve the alignment even further and to improve process design to minimize imbalance.

11.10.2 Deep Etch Uniformity

The last issue for fabrication precision is that of deep etch uniformity. Any high-speed bearing depends critically on the straightness and parallelism of the sidewalls that constitute the bearing and rotor surfaces. This is particularly true for hydrodynamic operation at high eccentricity. In the drive to generate deep trenches so that the bearing aspect ratio is minimized, the quality of the bearing etch is often compromised. These two issues, the etch depth and the etch quality, constantly pull against each other. Their relative advantages need to be weighed against each other in any final design.

Figure 11.18 shows typical non-uniform etch profiles for DRIE. This figure illustrates two common phenomena: etch bow, where the trench widens in the middle, and etch taper, where the trench widens (usually) at the bottom. The effects of these non-uniformities have been analyzed computationally [Piekos and Breuer, 2002]. As one might expect, the static performance of the bearing (load capacity) is degraded by the blow-out, particularly in the case of the tapered bearing where fluid pressure cannot accumulate in the gap but rather leaks out the enlarged end. The bowed bearing, because of its concave curvature, tends to hold the pressure more successfully, and the loss in load capacity is typically less severe. As mentioned earlier, load capacity is less of an issue in microbearings, and it is the effect on hydrodynamic stability that is of most interest. Figure 11.19 summarizes the effects of bow and taper on hydrodynamic operation. This figure shows the minimum stable eccentricity as a function of bearing number (i.e., speed, for a fixed bearing geometry) for different levels of bow and taper. The effects of taper are most severe, and considerable effort has been placed in the fabrication process design to minimize bearing taper.

11.10.3 Material Properties

One of the key benefits realized at the microscale is the improvement in strength-related material properties. This is particularly true in silicon-based MEMS where the baseline structural material is single-crystal silicon, which can be fabricated to have very good mechanical properties. The strength of brittle materials is controlled primarily by flaws and to some extent by grain boundaries, both of which become smaller or non-existent in a single-crystal material with surfaces defined by microfabrication processes. The device size becomes comparable with the flaw distribution such that the incidence of "super-strong"

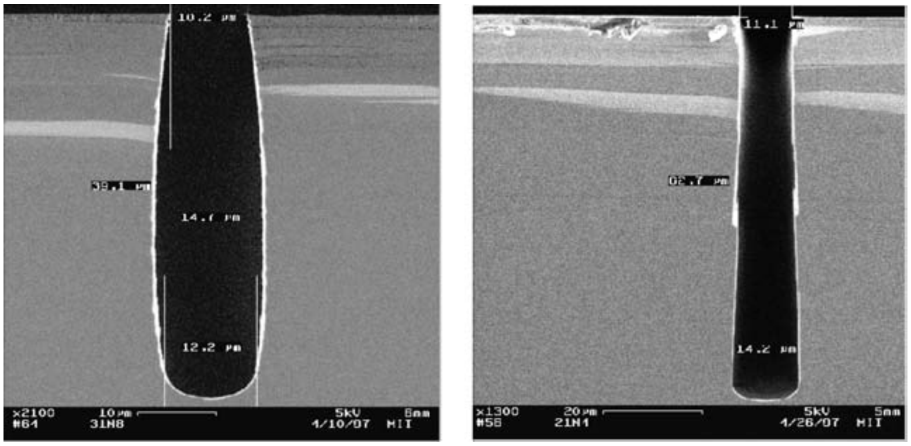


FIGURE 11.18 SEMs of bearing etches, illustrating typical manufacturing non-uniformities. The left-hand SEM shows an etch with a bow in the center. The right-hand SEM shows an etch with a taper. (SEM reprinted with permission of A. Ayon.)

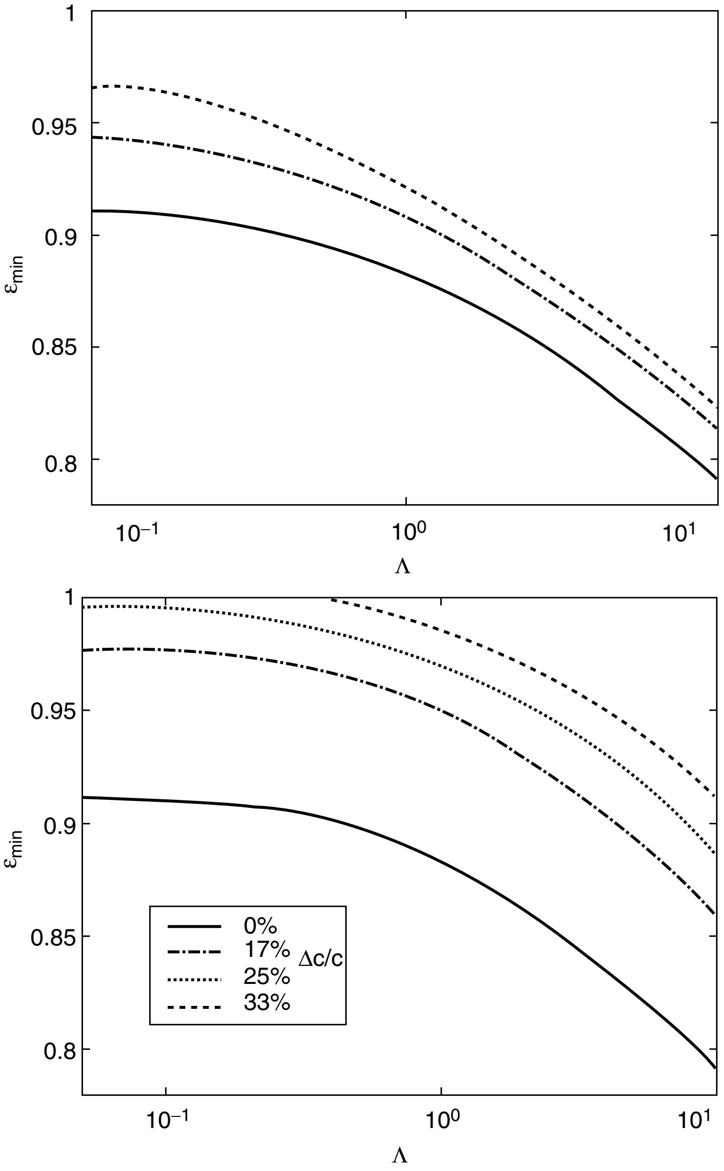


FIGURE 11.19 Degradation of hydrodynamic stability due to bow (left frame) and taper (right frame), as indicated by the minimum eccentricity required for hydrodynamic stability at a given bearing number (speed) [Piekos and Breuer, 2002].

devices increases in microscale systems. Silicon is a light material with a density (2330 kg/m^3) lower than that of aluminum (2700 kg/m^3). The strength-to-weight ratio of silicon micromachined structures is unparalleled, which is a key for high-speed rotating machinery. Despite its high specific strength silicon is a very brittle material. For a high speed rotating system, such as a turbine, this can be problematic since an impact or touchdown at any appreciable speed is likely to result in a catastrophic failure rather than an elastic rebound or more benign plastic deformation. Figure 11.20 shows a photograph of a microturbine rotor that crashed during a high-speed test run. The importance of robust bearings is emphasized because the material is extremely unforgiving.

11.11 Tribology and Wear

When lubrication fails, tribology and wear become important as the focus shifts from the prevention of contact to the mitigation of its effects. Tribology has been a subject of technical and industrial importance

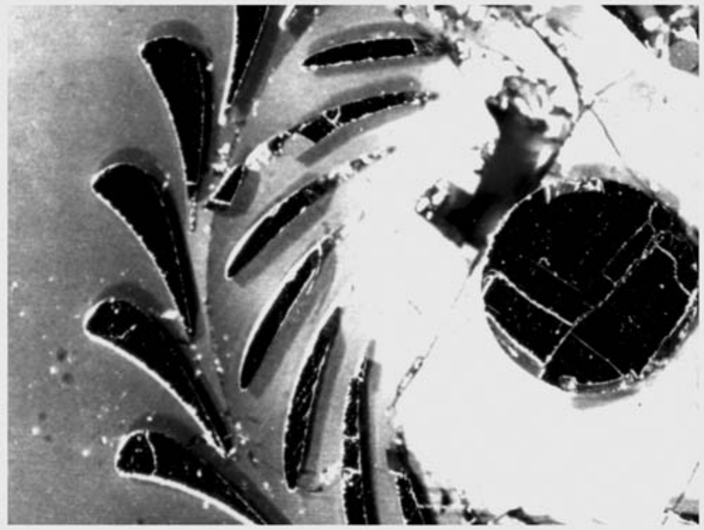


FIGURE 11.20 Photograph of the remains of a silicon rotor after experiencing a high-speed crash. The instant fracture of the rotor (largely along crystallographic planes) is visible.

since the industrial revolution, and a certain level of accomplishment was achieved that allows the design and operation of complex machinery in difficult environments. With the advent of microengineering and the development of the atomic force microscope, the field has defined a new set of problems and has witnessed a rebirth of focus on the micro- and nano-scale processes associated with friction and wear. This chapter does not address the progress in micro- and nano-tribology, however, a few general comments are made that are valuable in practical MEMS devices.

11.11.1 Stiction

A common problem associated with the failure of MEMS devices is that of stiction in which two surfaces touch and stick together due to the high surface energies. The problem is exacerbated by the use of wet-etching during fabrication and by the relatively smooth surfaces (and thus high contact areas) associated with MEMS materials. Many lubricants have been used to mitigate the problem. Self-assembling monolayers (SAMs), such as perfluoro-decyl-trichlorosilane, coat the surface with a monolayer of a long molecule that adheres to the surface at one end and is hydrophobic at the other end thus preventing stiction. Other lubricants are also under investigation such as Fomblin-Zdol, which is used in the hard-disk industry to protect the disk surface during head crashes and can be vapor-deposited during manufacture. Other remedies include the intentional design of textured surfaces, or standoffs, to prevent large areas coming into direct contact.

11.11.2 The Tribology of Silicon

Since most MEMS devices are silicon-based, the tribology of silicon has received considerable attention in the past several years. Silicon is not a very desirable bearing material [Gardos 2001] and exhibits high wear rates, high coefficients of friction, and poor stiction characteristics. Surface treatments (e.g., silicon carbide, and fluorocarbon (Teflon™-like) materials) and appropriate design have improved matters considerably. Most MEMS have not yet been designed with significant surface motions that require either supporting lubrication films or protective coatings. This is likely to change as fabrication processes enable more complex devices with higher power-densities and more challenging lubrication and wear requirements. The lifetime and reliability requirements of MEMS are becoming more severe because devices are being developed for space, medical, and national defense applications. All of these applications have unique and stringent requirements for predicting lifetime, wear, and failure, and so this oft-neglected corner of the industry will receive the attention it deserves.

11.12 Conclusions

In this chapter, we have focused on some of the key issues that face the design, manufacture, and operation of microdevices that need to operate with minimal friction or wear. Many of the issues in Couette and squeeze-film lubrication have been successfully addressed. Fundamental surface models (such as gas accommodation coefficients) are still unknown, and reliable prediction methods are just becoming available. Stiction issues remain problematic although they are managed in an engineering manner. Increased understanding of surface science is needed to solve this problem.

Rotating MEMS devices need much more development before they can be reliably manufactured and used. Even with dramatic advances in lubricant and surface treatment technologies, high-speed operation will likely require gas bearings which have always offered high-speed, low wear operation with the attendant cost of a narrow and treacherous window of stable operation. Many of the commonly held assumptions and design rules that have guided previous fluid film bearings in conventional machinery have been revisited due to the consequences of scaling and the current limitations in microfabrication technology.

Future research needs to focus attention in many areas. On the manufacturing side, the single largest obstacle to trouble-free production of gas bearings for high-speed rotors and shafts is the issue of precision microfabrication. Macroscopic systems, with typical scale of 1 meter, need precision manufacturing in places with sub-millimeter tolerances. Microdevices with typical dimensions of 1 mm will need tolerances of 1 micron or less. The ability to manufacture with such precision will require much improved understanding of micromachining technologies such as etching and deposition so that cross-device and cross-wafer uniformities can be improved.

At the level of lubrication technologies, the new parameter regimes that are exposed by microfabricated systems (very low aspect ratios, relatively large clearances, insignificant inertial properties, etc.) need to be further explored and understood. Despite the low Reynolds numbers, inertial losses are critically important for hydrostatic lubrication mechanisms and need to be better understood and predicted. Similarly, the coupled fluid-structure interactions at high eccentricities and the interactions between hydrodynamic and hydrostatic mechanisms need to be more fully explored. Fundamental issues of fluid and solid physics need to be addressed as the scale continues to shrink. Gas surface interactions, momentum and energy accommodation phenomena, and the effects of surface contamination (whether deliberate or accidental) need to be rigorously studied so that the macroscopic behavior can be predicted with some certainty. These issues will become more important as manufacturing scales decrease further and as continuum assumptions become more problematic.

Acknowledgments

I am indebted to my several friends and collaborators over the years with whom I have been fortunate to work in this exciting area. In particular, the contributions of Errol Arkilic, Stu Jacobson, Fred Ehrich, Alan Epstein, Luc Frechette, Peter Kwok, C.-C. Lin, D.J. Orr, Ed Piekos, Nick Savoulides, Marc Weinberg, and C.-W. Wong are gratefully acknowledged. Some of the contents of this chapter are drawn from a similar paper presented at a NIST-sponsored workshop and published by Kluwer Press [Hsu, 2001].

References

Aluru, N.R., and White, J. (1998) "A Fast Integral Equation Technique for Analysis of Microflow Sensors Based on Drag Force Calculations," *International Conference on Modeling and Simulation of Microsystems, Semiconductors, Sensors and Actuators* pp. 283–6, April, Santa Clara, CA.

Arkilic, E.B., Schmidt, M.A., and Breuer, K.S. (1997) "Gaseous Slip Flow in Long Microchannels," *J. MicroElectroMech. Syst.* **6**(2), pp. 167–78.

Arkilic, E., and Breuer, K.S. (1993) "Gaseous Flow in Small Channels," *AIAA 93-3270, AIAA Shear Flow Conference*, July, Orlando, FL.

Bart, S.F., Lorber, T.A., Howe, R.T., Lang, J.H., and Schlecht, M.F. (1988) "Design Considerations for Micromachined Electric Actuators," *Sensor. Actuator.*, vol. **14**, pp. 269–92.

Beskok, A., and Karniadakis, G.E. (1994) "Simulation of Heat and Momentum Transfer in Complex Micro-Geometries," *J. Thermophys. Heat Transf.* **8**(4), pp. 647–55.

Blech, J.J. (1983) "On Isothermal Squeeze Films," *J. Lubr. Technol.*, vol. **105**, pp. 615–20.

Breuer, K.S., Arkilic, E.B., and Schmidt, M.A. (2001) "Slip Flow and Tangential Momentum Accommodation in Silicon Micromachined Channels," *J. Fluid Mech.* **437**, pp. 29–44.

Burgdorfer, A. (1959) "The Influence of the Molecular Mean Free Path on the Performance of Hydrodynamics Gas Lubricated Bearing," *J. Basic Eng.*, vol. **81**, pp. 94–9.

Cai, C.-P., Boyd, I.D., Fan, J., and Candler, G.V. (2000) "Direct Simulation Methods for Low-Speed Microchannel Flows," *J. Thermophys. Heat Transf.*, vol. **14**, pp. 368–78.

Dimofte, F. (1995) "Wave Journal Bearing with Compressible Lubricant. Part 1: The Wave Bearing Concept and a Comparison to the Plain Circular Bearing," *Tribol. Trans.* vol. **38**, pp. 153–60.

Epstein, A.H., Senturia, S.D., Al-Midani, O., Anathasuresh, G., Ayon, A., Breuer, K., Chen, K.-S., Ehrich, F.E., Esteve, E., Frechette, L., Gauba, G., Ghodssi, R., Groshenry, C., Jacobson, S., Kerrebrock, J.L., Lang, J.H., Lin, C.-C., London, A., Lopata, J., Mehra, A., Mur Miranda, J.O., Nagle, S., Orr, D.J., Piekos, E.S., Schmidt, M.A., Shirley, G., Spearing, S.M., Tan, C.S., Tzeng, Y.-S., and Waitz, I.A. (1997) "Micro-Heat Engines, Gas Turbines and Rocket Engines - the MIT Microengine Project," *AIAA Paper 97-1773* Snowmass, CO, June.

Fréchette, L.G., Jacobson, S.A., Breuer, K.S., Ehrich, F.F., Ghodssi, R., Khanna, R., Wei Wong, C., Zhang, X., Schmidt, M.A., and Epstein, A.H. (2005) "High-Speed Microfabricated Silicon Turbomachinery and Fluid Film Bearings," *J. MicroElectroMech. Syst.* **14**(1), pp. 141–52.

Gardos, M. (2002) *Nanotribology*, S. Hsu, ed., Kluwer Academic Press, New York.

Hamrock, B.J. (1984) *Fundam. Fluid Film Lubr.*, McGraw-Hill, New York.

Hsu, S., ed. (2001) *Nanotribology*, Kluwer Press, New York.

Kwok, P., Weinberg, M., and Breuer, K.S. (2005) "Fluid Effects in Vibrating Micro-Machined Structures" *J. MicroElectroMech. Syst.* (in press)

Lin, C.-C., Ghodssi, R., Ayon, A.A., Chen, D-Z., Jacobson, S., Breuer, K.S., Epstein, A.H., and Schmidt, M.A. (1999) "Fabrication and Characterization of Micro Turbine/Bearing Rig," *Proceedings, MEMS99*, January, ***IEEE***, Orlando, FL.

Liu, L.X., Teo, C.J., Epstein, A.H., and Spakovszky, Z.S. (2003) "Hydrostatic Gas Journal Bearings for Micro-Turbomachinery," *Proceedings of ASME-DEC'03*, September, Chicago.

Mehregany, M., Senturia, S.D., and Lang, J.H. (1992) "Measurement of Wear in Polysilicon Micromotors," *IEEE Trans. Electron Devices*, vol. **39**, pp. 1136–43.

Muijderman, E.A. *Spiral Groove Bearings*, Springer Verlag, New York.

Nagle, S.F., and Lang, J.H. (1999) "A Micro-Scale Electric-Induction Machine for a Micro Gas-Turbine Generator;" *Presented at the 27th Meeting of the Electrostatics Society of America*, June, Boston, MA.

Orr, D.J. (1999) Macro-Scale Investigation of High Speed Gas Bearings for MEMS Devices, Ph.D. thesis, Department of Aeronautics and Astronautics, MIT Cambridge, MA.

Piekos, E.S. (2000) Numerical Simulations of Gas-Lubricated Journal Bearings for Microfabricated Machines, Ph.D. thesis, Department of Aeronautics and Astronautics, MIT Cambridge, MA.

Piekos, E.S., and Breuer, K.S. (1999) "Pseudo-Spectral Orbit Simulation of Non-Ideal Gas-Lubricated Journal Bearings for Microfabricated Turbomachines," *J. Tribol.*, vol. **121**, pp. 604–9.

Piekos, E.S., and Breuer, K.S. (2002) "Manufacturing Effects in microfabricated gas bearings: axially varying clearance" *J. Tribol.* **124**, pp. 815–21.

Reynolds, O. (1886) "On the Theory of Lubrication and Its Application to Mr. Beauchamp Tower's Experiments, Including an Experimental Determination of the Viscosity of Olive Oil," *Royal Society, Phil. Trans.*, pt. 1.

Rohsenow, W.M., and Choi, H.Y. (1961) *Heat, Mass, and Momentum Transfer*, Prentice-Hall, Inc., Englewood Cliffs, NJ.

Savoulides, N., Breuer, K., Jacobson, S., and Ehrich, F. (2001) "Low Order Models for Very Short Hybrid Gas Bearings," *J. Tribol.* **123**, pp. 368–75.

Sniegowski, J., and Garcia, E. (1996) "Surface-Micromachined Geartrains Driven by an On-Chip Electrostatic Microengine," *IEEE Electron Device Lett.*, vol. **17**, p. 366.

Veijola, T., Kuisma, H., Lahdenpera, J., and Ryhanen, T. (1995) "Equivalent-Circuit Model of the Squeezed Gas Film in a Silicon Accelerometer," *Sensor. Actuator.*, vol. **A48**, pp. 239–48.

Wilcox, D.F., ed. (1972) *Design of Gas Bearings* M.T.I. Inc., Latham, NY.

Wong, C.W., Zhang, X., Jacobson, S.A., and Epstein, A. (2004) "A Self-Acting Gas Thrust Bearing for High-Speed Microrotors," *J. MicroElectroMech. Syst.* **13**(2), pp. 158–64.

12

Physics of Thin Liquid Films

Alexander Oron
Technion — Israel Institute of Technology

12.1 Introduction

Various aspects of fluid mechanics in microelectromechanical systems (MEMS), such as flows in microconfigurations, flow transducers, and flow control by microsystems were reviewed by Ho and Tai (1998). However, the issue of thin liquid films and their dynamics in the context of microelectromechanical systems was not included in the scope of that important work. This chapter intends to fill this gap.

Thin liquid films are encountered in a variety of phenomena and technological applications [Myers, 1998]. On large scale they emerge in geophysics as gravity currents under water or as lava flows [Huppert and Simpson, 1980; Huppert, 1982]. On the engineering scale liquid films serve in heat and mass transfer processes to control fluxes and to protect surfaces, and their various applications arise in paints, coatings, and adhesives. They also occur in foams [Schramm, 1994; Prud'homme and Khan, 1996], emulsions [Ivanov, 1988; Edwards et al., 1991], and detergency [Adamson, 1990]. In biological applications they appear as membranes, as linings of mammalian lungs [Grotberg, 1994], or as tear films in the eye [Sharma and Ruckenstein, 1986]. On the microscale in MEMS, thin liquid films are used to produce insulating coating of solid surfaces, to form stable liquid bridges at specified locations, to create networks of microchannels on patterned microchips [Herminghaus et al., 1999, 2000], and to design fluid microreactors [Ichimura et al., 2000].

The presence of the deformable interface between the liquid and the ambient (normally gaseous, but possibly also another liquid) phases engenders various kinds of dynamics driven by one or usually several physical factors simultaneously. Liquid films may undergo, spontaneously or under the influence of external factors, diverse profound changes in their shapes. These changes are related to different kinds of instability that might interact. Such interactions might lead either to a mutual enhancement or quenching of each other, so that the overall film dynamics may be rather complex. The film can rupture when its local depth vanishes and dewet the solid, leading to holes in the liquid that expose the substrate to the ambient gas. In this case, the continuous character of the film changes if droplets of liquid are detached from the film. Changes in structure might occur in flows with contact lines leading to wavy fronts, fingered patterns, or rivulets. Liquid films might be isothermal or subjected to the influence of a temperature field which normally alters their dynamics. Liquid films might also undergo phase changes, such as mass loss by evaporation, mass gain by condensation or solidification. Liquid films exhibit many fascinating examples of behavior, and some of them are presented below.

Sharma and Reiter (1996) studied experimentally the process of spontaneous dewetting of thin (less than 60 nanometers thick) polystyrene films on various coated silicon wafers and found a wealth of types of pattern formation. Different stages of dewetting identified in their experiments were: (1) rupture of the film and emergence of holes; (2) expansion of the holes, their coalescence and formation of polygonal cellular pattern where most of the liquid gathers in the ridges (see also [Reiter, 1998]); (3) disintegration of liquid ridges into isolated and ultimately spherical drops. Also fingering instability of the hole rim during hole expansion was observed on low wettability coatings, which resulted in the emergence of separate drops (see also [Elbaum and Lipson, 1994]). The growth rate of the initial disturbance, the time of rupture, the number density of holes, and the size of the polygons depend *only* on the solid substrate and is independent of the coating. The contact angle, which strongly depends on the choice of the coating layer, affects the generation of droplets via the fingering instability, which is faster for larger contact angles. Also, the size of spherical drops forming as a result of the breakup of the liquid ridges depends on the contact angle. Figure 12.1 reproduced from Sharma and Reiter (1996) displays the final pattern established by a 45 nanometers thick liquid polysterene film on a coated silicon wafer, where the average contact angle for this combination of solid and liquid was about 22°. The pattern presented in Figure 12.1 consists of spherical droplets of various sizes aligned along the polygonal structure obtained as a result of the evolution of an initially uniform film that went through all of the stages previously mentioned. Problems associated with dewetting of solid surfaces by liquid films are discussed in the section on isothermal films.

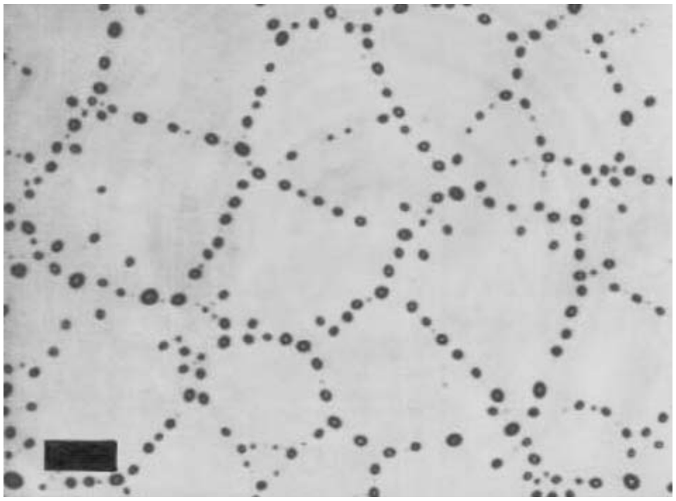


FIGURE 12.1 Photograph of the final polygonal pattern of spherical droplets for an initially flat polysterene film of the mean thickness of 45 nanometers on a coated silicon wafer. For reference, the length of the bar is 70 microns. (Reprinted with permission from Sharma and Reiter (1996).)

Thiele et al. (1998) carried out experiments with thin (≈10 nanometers) volatile films consisting of a collagen solution in acetic acid in various conditions of ambient humidity. For low ambient humidity and thus a high evaporation rate, the pattern of very small holes along with several large ones was observed suggesting that the former emerged because of the polar interactions with the short residence time, while the latter nucleated because of defects, like dust particles or imperfections of the substrate. However, when the humidity was high and therefore the evaporation rate was low, the pattern of a homogeneous polygonal network with large spacing was found. The large size of the polygons was explained by the long residence time, during which holes created by nucleation were able to open up. The effects associated with evaporation are discussed in the section dedicated to phase changes.

Figure 12.2 reproduced from Herminghaus et al. (2000) displays the patterns generated by condensation of water on a spatially heterogeneous solid poorly wetted (hydrophobic) silicone rubber substrate

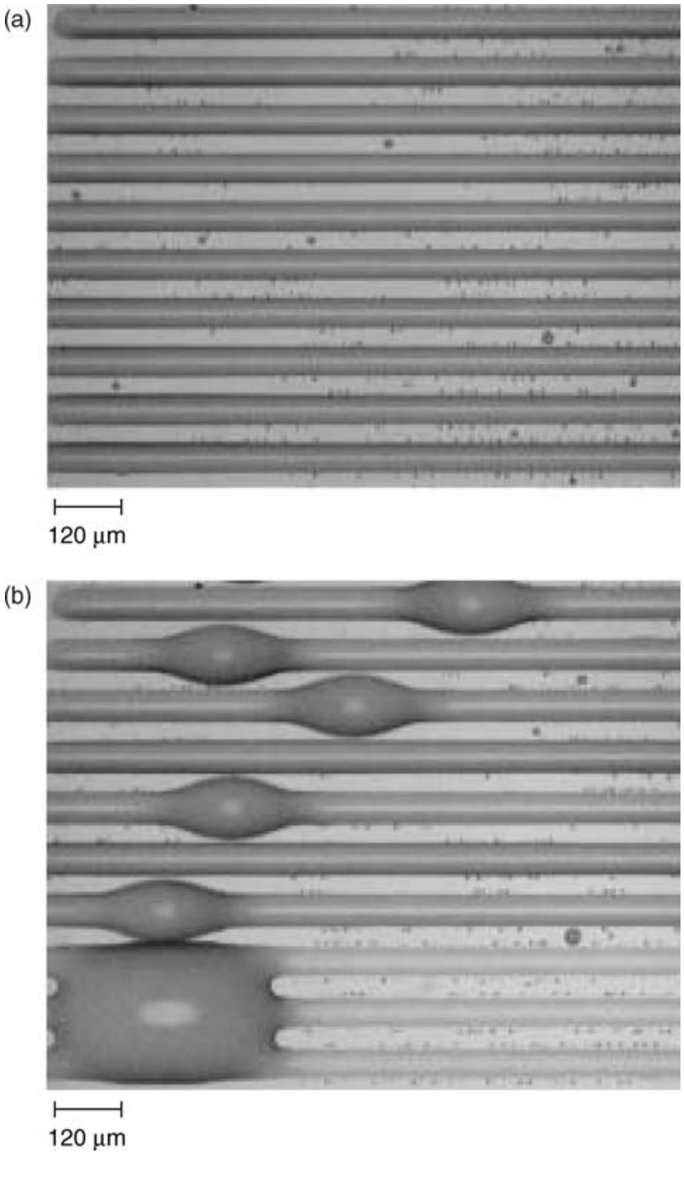


FIGURE 12.2 Water condensation on hydrophilic (magnesium fluoride) stripes of an elsewhere hydrophobic (silicone rubber) substrate. (a) Low-condensate-volume regime: the parallel channels of condensed water have a constant cross-section and a small contact angle. Several droplets are also seen on the hydrophobic domains of the substrate. (b) High-condensate-volume regime: some of the liquid channels develop a single drop, when the contact angle exceeds a certain critical value. If two drops are in a close proximity, they merge to form a microbridge between the two neighboring microchannels. (Reprinted with permission from Herminghaus et al. (2000).)

with well wetted (hydrophilic) magnesium fluoride stripes. Figure 12.2(a) shows the intermediate stage of water condensation, in which the liquid phase forms parallel microchannels with a certain contact angle θ at the edges of the stripes where the liquid appears to be pinned. The cross-section of these microchannels is position-independent and represents a circular cap. Several drops of the condensate between the hydrophilic stripes are also observed in Figure 12.2(a). When the process of water condensation proceeds further the contact angle θ increases beyond a certain value, and the microchannels undergo morphological change. As a result, droplets emerge on some of the microchannels, one per channel, as seen in Figure 12.2(b). When such droplets develop near each other, they merge to create a bridge as seen in the bottom left corner of Figure 12.2(b). These fundamental phenomena can guide the liquid into the desired location(s) on the substrate with a specially designed wettability pattern [Herminghaus et al., 2000].

Figure 12.3 reproduced from Herminghaus et al. (2000) shows the time evolution (from top to bottom) of deposition of water condensing onto curved wettable patches of different width in its corners. This width increases from channel (a) to channel (e). When water condenses and gradually fills the channels, the behavior of the liquid depends solely on the width of the corner. If the latter is large, as in the case (e), the drop develops in the corner. However, if it is small as in the case (a), the uniform channel configuration becomes unstable and a droplet develops in the straight part before it occurs in the corner. In the intermediate range (b)–(d) the corner first develops a structure similar to the case (a), but it suddenly and discontinuously moves into the corner when a critical value of capillary pressure is attained. In such a configuration the contact area of the liquid with the hydrophilic patch of the substrate is maximized. If the corners are sufficiently close to each other, two droplets will merge to produce a microbridge between the two neighboring microchannels similar to what is shown in Figure 12.2.

VanHook et al. (1997) carried out experiments to study the thermocapillary convection produced by variations of surface tension in bilayer systems containing silicone oil films of 0.007 to 0.027 cm thick and overlying gas gap. Figure 12.4 reproduced from VanHook et al. (1997) shows some of their representative results. The dominant feature for thinner films was the emergence of a drained spot, Figure 12.4(a), or the emergence of a localized elevated structure with a peak touching the upper lid, Figure 12.4(b). The drained spot may contain in certain circumstances isolated droplets trapped inside it. All these are manifestations of the long-wave instability of the film. This and other effects will be discussed in the section dedicated to thermal effects. However, along with long-wave features of the system other phenomena were observed for generally thicker films. Figure 12.4(c, d) show these short-wave phenomena (i.e., whose length scale is comparable with the mean thickness of the layer) displayed the formation of hexagons or a combination of hexagons with the emergence of a dry spot. Mathematical treatment of such short-wave phenomena will not be considered here. Rupture of thin liquid films and following it growth of the dry spot are frequent features associated with various physical mechanisms. Figure 12.5 reproduced from VanHook et al. (1997) shows such an evolution of a local depression to the stage of film rupture and later to the growth and saturation of the dry spot driven by the thermocapillary effect (see the section on thermal effects).

A low-cost high-yield passive alignment method, known as controlled collapse chip connection or as a C4 process, was designed [Goldmann, 1969] and used in optoelectronic packaging, where alignment accuracies at the submicron level are required [Wale and Edge, 1990; Lin et al., 1995]. Such precision alignment techniques, as illustrated in Figure 12.6 reproduced from Salalha et al. (2000), are employed for coupling fibers and wave guides to devices such as lasers and photo detectors, and are being actively developed and improved. These techniques use molten solder and are based on the restoring force arising from surface tension that drives the misaligned solder joint to become well-aligned and minimizes the total interfacial energy of the system. The final well-aligned configuration is then fixed by cooling down and solidifying the solder. Figure 12.6(a) presents such a misaligned layout of the two chips with four solder joints seen as dark circles. The misalignment is illustrated by the position of the center of the cross with respect to the point between the four squares. Figure 12.6(b) is a close-up of the area designated by the large circle on Figure 12.6(a) and presents the initial position of the misaligned system, which moves through the intermediate state, Figure 12.6(c), to its final well-aligned position, Figure 12.6(d). Figure 12.6(e) displays the final cross-section of the solder between the two chips. Note that

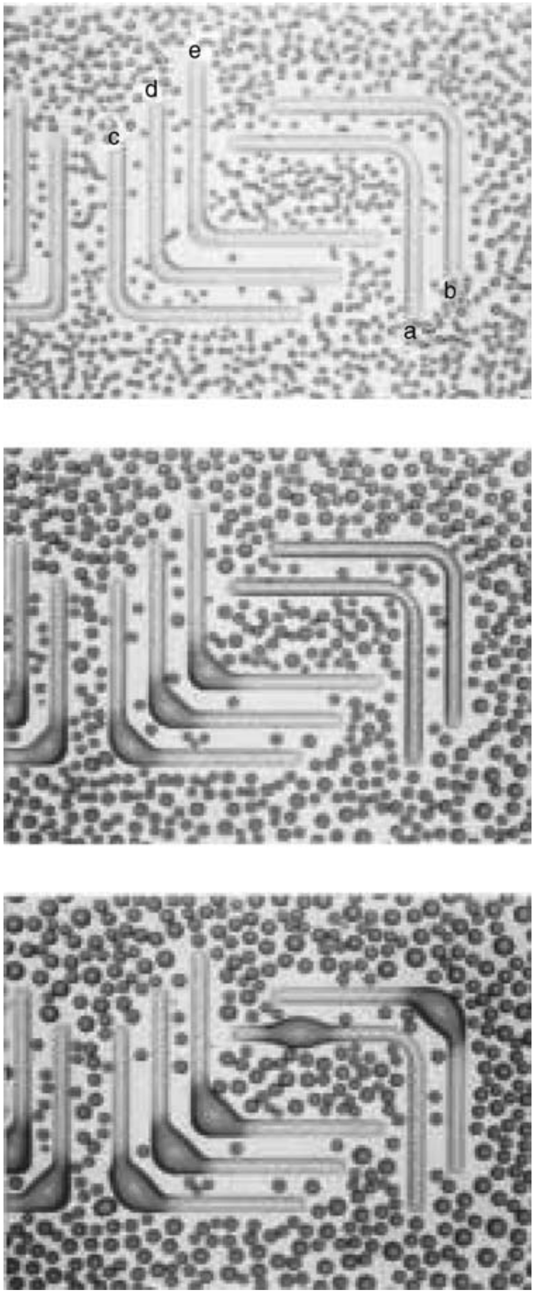


FIGURE 12.3 Deposition of water onto a patterned surface with hydrophilic microchannels with corners. The width of the channel in the corner region increases from channel (a) to channel (e). Time and therefore the volume of the condensate increase from top to bottom. When a microchannel undergoes a morphological change of its shape, the drop moves to the corner to maximize the contact area with the hydrophilic part of the substrate. (Reprinted with permission from Herminghaus et al. (2000).)

the alignment process presented here takes place in the time range of a second. A similar mechanism, by which a hard contact lens centers itself over the cornea in a human eye, was discussed by Moriarty and Terrill (1996).

The centrifugal spinning of volatile solutions is a convenient and efficient means of coating planar solids with thin films. This process, known as spin coating, has been widely used in many technological processes, such as deposition of dielectric layers onto silicon wafers in the microelectronic industry, formation of ultrathin antireflective coatings for deep UV lithography, and others. Two important stages of

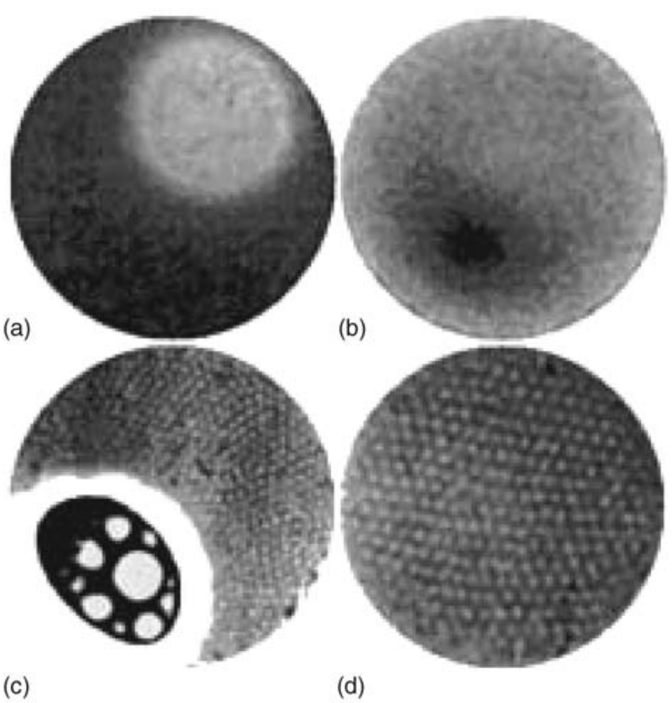


FIGURE 12.4 Infrared images of various states as seen in the experiments. The temperature increases with increasing brightness, so warm depression regions are white (except in (c)) and cool elevated regions are dark. Each image has its own brightness, so temperatures in different images cannot be compared. (a) A localized depression (dry spot) with a helium gas layer and $d = 0.025$ cm. (b) A localized elevation (high spot) with an air gas layer and $d = 0.037$ cm. (c) A dry spot with hexagons in the surrounding region and $d = 0.025$ cm. (d) Hexagons with an air gas layer and $d = 0.045$ cm. For more detail refer to the source. (Reprinted with permission from VanHook et al. (1997).)

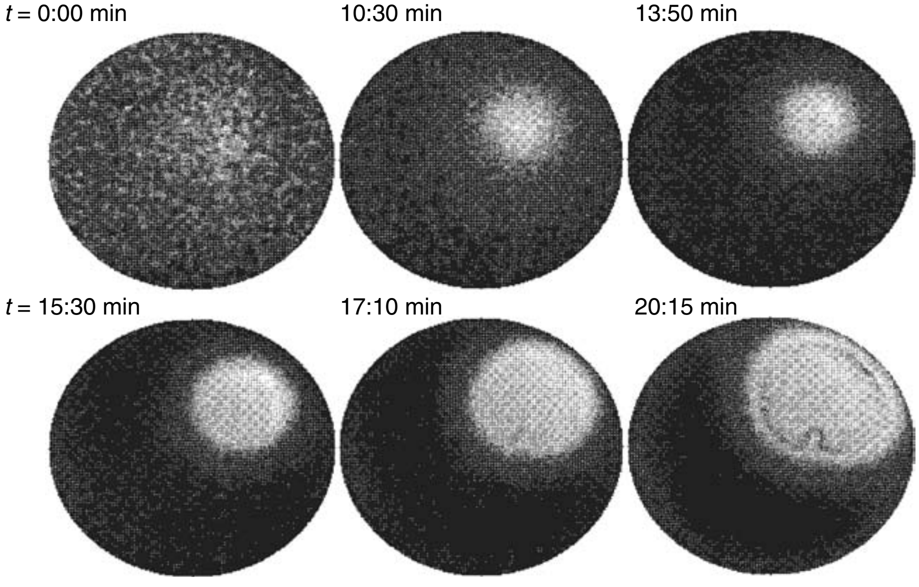


FIGURE 12.5 The evolution of a localized depression and formation of a dry spot in silicone oil of depth $d = 0.0267 \pm 0.0008$ cm and helium in the gas layer. At $t = 0$ (an arbitrary starting point) there is negligible deformation of the interface. The liquid layer begins to form a localized depression (the white circle), and in 15 minutes the interface has ruptured ($h_{\min} \to 0$) and formed a dry spot. The dry spot continues to grow for several more minutes before saturating. Bright (dark) regions are hot (cool) because they are closer (farther) to (from) the heater. All images have the same intensity scaling. (Reprinted with permission from VanHook et al. (1997).)

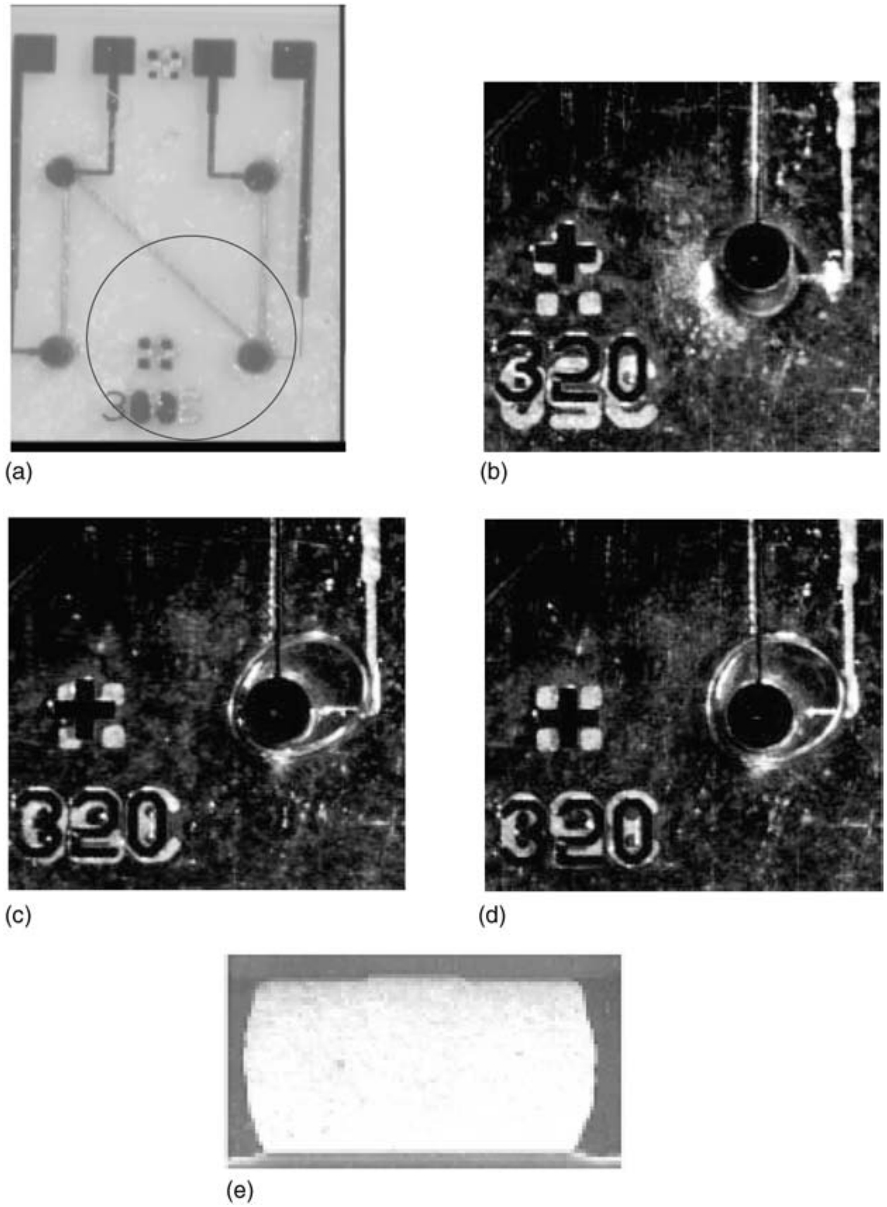


FIGURE 12.6 Photographs of C4 bonding based on self-alignment mechanism. (a) Layout of the chip (4 mm by 4 mm) which consists of four solder joints made of 63Sn37Pb. The upper chip is not aligned with the lower one, as can be seen from the position of the upper cross relative to four squares at the lower chip. Initial misalignment is 150 microns. (b) An enlarged picture of one of the solder joints at the initial moment. (c) An intermediate stage. (d) The final position. (e) A side view showing the cross-section of the solder joint at the final stage. (Reprinted with permission from Salalha et al. (2000).)

the process are usually considered. The first stage occurs shortly after the liquid volume is delivered to the disk surface rotating usually at the speed of 1000–10,000 r/min. At the beginning of this stage the liquid film is relatively thick (usually greater than 500 microns). The film thins mainly because of radial drainage under the influence of centrifugal forces. Inertial forces are important and can lead to the appearance of instabilities of the spinning film. The second stage occurs when the film has thinned to the point where inertia is no longer important (film thickness usually less than 100 microns), and the flow slows down considerably, but deformations of the fluid interface may still be present because of the instabilities that appeared during the first stage. The film continues to thin mainly because of solvent evaporation until the solvent becomes depleted, and the film solidifies and ceases to flow. Such problems are discussed in the sections on isothermal films and phase changes.

Numerous applications relevant for MEMS involve the dynamics of liquid films or drops. This area is in constant progress and new exciting developments are often reported in the literature. Knight et al. (1998) describes a new method of enhancement and control of nanoscale fluid jets. They demonstrated this method with a design of a continuous-flow mixer capable of mixing flow rates of nanoliters per second within the time scale of 10 microseconds. Such a mixer can be useful in nanofabrication techniques and serve as an essential part of a microreactor built on a chip.

Spatially controlled changes in the chemical structure of a solid substrate can guide a deposited liquid along the substrate. Ichimura et al. (2000) reported their experimental results showing the possibility of reversible guidance of liquid motion by light irradiation of a photoresponsive solid substrate. Asymmetric irradiation of the solid surface with blue light led to movement of a 2 microliter olive oil droplet with a typical speed of 35 microns/sec. A similar irradiation with a homogeneous blue light stopped the movement of the droplet completely. The speed of the droplet and the direction of its movement were adjustable to the conditions of such irradiation. The phenomenon described has a potential applicability in design of microreactors and microchips.

Schaeffer et al. (2000) proposed a new technique of creating and replicating lateral structures in films on submicron length scales. This technique is based on the fact that lateral gradients of the electric field applied in the vicinity of the film interface induce variations of surface tension and thus lead to the electrocapillary effect. The electrocapillary effect is similar to the thermocapillary effect previously mentioned and is addressed more thoroughly in the section on thermal effects. The electrocapillary effect triggering the electrocapillary instability of the film results in formation of ordered patterns on the film interface and focusing of the interfacial troughs and peaks in the desired locations following the master pattern of the electrodes. Schaeffer et al. (2000) reported the replication of patterns of lateral dimensions of order 140 nanometers while employing this technique. A complete investigation of the electrocapillary instability of thin liquid films has not yet appeared in the literature. Lee and Kim (2000) presented a liquid micromotor and liquid–metal droplets rotating along a microchannel loop driven by continuous electrowetting (CEW) phenomenon based on the electrocapillary effect. They identified and developed key technologies to design, manufacture, and test the first MEMS devices employing CEW.

A mathematical treatment of this and other phenomena must consider that the interface of the film lying or flowing on a solid surface is partially or entirely a free boundary whose configuration evolving both temporally and spatially must be determined as an integral part of the solution of the governing equations. This renders the problem too difficult and often almost intractable analytically, which might lead researchers to rely on computing only. Computing also becomes complicated because of the free-boundary character of the problem which requires a careful design of adequate numerical methods.

Another property of such mathematical problems is their strong inherent nonlinearity, which is present in both governing equations and boundary conditions. This nonlinearity of the problem presents another complexity. Consideration of coupled phenomena, such as those previously mentioned, requires compact description of simultaneous instabilities that interact in intricate ways. This compact form must be tractable and, at the same time, complex enough to retain the main features of the problem at hand.

The most appropriate analytical method of dealing with the above complexities is to analyze only long scale phenomena, in which the characteristic lateral length scales are much larger than the average film thickness, the flow-field and temperature variations along the film are much more gradual than those normal to it, and the time variations are slow. Similar theories arise in a variety of areas of classical physics: shallow-water theory for water waves, lubrication theory in viscous flows, slender-body theory in aerodynamics, and in dynamics of jets [e.g., Yarin, 1993]. In all of these examples, a geometrical disparity is used to practically separate the variables and to simplify the analysis. In thin viscous films, most rupture and instability phenomena occur on long scales, and a long-wave approach explained later is very useful.

The long-wave theory approach is based on the asymptotic reduction of the governing equations and boundary conditions to a simplified system, which consists often, but not always, of a single nonlinear partial differential equation formulated in terms of the local thickness of the film varying in time and space. The rest of the unknowns (i.e., the fluid velocity, pressure, temperature, etc.) are determined via functionals of the solution of this differential equation usually called *evolution* equation. The notorious complexity of a free-boundary problem thus is removed. The corresponding penalty is, however, the presence of the strong nonlinearity in the evolution equation(s) and the higher-order spatial derivatives (usually up to the fourth) appearing there. A simplified linear stability analysis of the problem can be carried out based on the resulting evolution equation. A weakly nonlinear analysis of the problem is also possible through that equation. However, the fully nonlinear analysis that allows one to study finite-amplitude deformations of the film interface must be performed numerically. Numerical solution of the evolution equation is incomparably less difficult than that of the original, free-boundary problem.

Several encouraging verifications of the long-wave theory versus the experimental results have appeared in the literature. Burelbach et al. (1990) carried out a series of experiments in an attempt to check the long-wave theory of Tan et al. (1990) for steady thermocapillary flows induced by non-uniform heating of the solid substrate. The measured steady shapes were favorably tested against theoretical predictions for layers less than 1 mm thick under moderate heating conditions. However, the relative error was large for conditions near rupture, where the long-wave theory is formally invalid [Burelbach et al., 1988], but in all other cases the predicted and measured values of the minimal film thickness agreed within 20%. The theory (see Equation (3.6) of [Tan et al., 1990]) also predicts rupture when the parameter L exceeds a certain critical value and predicts steady patterns otherwise. Experimental results (see Figure 1 of [Burelbach et al., 1990]) show that L is an excellent qualitative indicator of whether the film ruptures.

VanHook et al. (1995, 1997) performed experiments on the onset of the long-wavelength instability in thin layers of silicone oil of varying thickness, aspect ratios, and transverse temperature gradients across the layer. A formation of "dry spots" at randomly varying locations was found above the critical temperature difference across the layer in qualitative agreement with corresponding numerical simulations. The experimental support for the theoretical results is discussed in various sections of this chapter.

Another test for the validity of an asymptotic theory, such as the long-wave theory presented here, is the comparison between the numerical solutions for the full free-boundary problem in its original form and the solutions obtained for the corresponding long-wave evolution equations. Due to the difficulty of carrying out direct numerical simulations previously discussed, the number of such comparative studies is quite limited. Krishnamoorthy et al. (1995) performed a full-scale direct numerical simulation of the governing equations to study the rupture of thin liquid films because of thermocapillarity and found very good qualitative agreement with the results arising from the solution of the corresponding evolution equation, except for times prior to rupture. Oron (2000b) found even better agreement at rupture between his results and the direct simulations of the Navier–Stokes equations of Krishnamoorthy et al. (1995). There has been a long debate in the literature about the validity of fingered structures of the film interface often arising from the solution of the evolution equations and whether they are artifacts of the asymptotic reduction applied. Direct solution of the Navier–Stokes equations [Krishnamoorthy et al., 1995] provides convincing evidence supporting the validity of the evolution equations even in the domain where some assumptions leading to their derivation are violated.

The analysis of thin liquid films has progressed significantly in recent years. In the review article by Oron et al. (1997) such analyses were unified into a simple framework in which the special cases naturally emerged. In this chapter the physics of thin liquid films is reviewed with emphasis on the phenomena of considerable interest for MEMS. The theory of drop spreading, despite its importance, is not included here. Refer to other reviews [de Gennes, 1985; Leger and Joanny, 1992; Oron et al., 1997] for more detailed information.

The general evolution equation describing the general dynamics of thin liquid films is derived following Oron et al. (1997) and is discussed in the next section. The topic addressed in the second section is isothermal films, where the physical effects discussed are viscous, surface tension, gravity, and centrifugal forces along with van der Waals interactions. The third section examines the influence of thermal effects on the dynamics of liquid films. The fourth section considers the dynamics of liquid films undergoing phase changes, such as evaporation and condensation.

12.2 The Evolution Equation for a Liquid Film on a Solid Surface

We now describe the long-wave approach and apply it to a flow of a viscous liquid in a film. The film is supported below by a solid horizontal plate and is bounded above by an interface separating the liquid and a passive gas and slowly evolving in space and time, as given by its equation $z = h\ (x, y, t)$. Assume the possibility of external interfacial forces Π with the components $\{\Pi_3, \Pi_1, \Pi_2\}$ in the normal and tangential to the film surface directions, respectively, determined by the vectors

$$\mathbf{n} = \frac{\{-h_x, -h_y, 1\}}{\sqrt{1 + h_x^2 + h_y^2}}, \quad \mathbf{t}_1 = \frac{\{1, 0, h_x\}}{\sqrt{1 + h_x^2}}, \quad \mathbf{t}_2 = \frac{\{0, 1, h_y\}}{\sqrt{1 + h_y^2}}. \tag{12.1}$$

The components of the vectors $\mathbf{n}$, $\mathbf{t}_1$, $\mathbf{t}_2$ in Equation (12.1) are specified in the order of x-, y-, and z- directions, where x and y are the spatial coordinates in the given solid plane and z is normal to the latter and directed across the film. The presence of a conservative body force determined by the potential ϕ acting on the liquid phase, such as gravity, centrifugal, or van der Waals force, is accounted for as well. We note that the vectors $\mathbf{t}_1$, $\mathbf{t}_2$ are not orthogonal, but it is sufficient for our later application that $(\mathbf{n}, \mathbf{t}_1)$ and $(\mathbf{n}, \mathbf{t}_2)$ constitute pairs of orthogonal vectors. The letter subscripts denote the partial derivatives with respect to the corresponding variable.

The liquid considered in this work is assumed to be a simple Newtonian incompressible viscous fluid whose dynamics are well described by the Navier–Stokes and mass conservation equations, provided that the length scales characteristic for the flow domain are within the continuum range exceeding several molecular spacings. The mass conservation and Navier–Stokes equations for such a liquid in three dimensions have the form

$$\begin{aligned} u_x + v_y + w_z &= 0, \\ \rho(u_t + uu_x + vu_y + wu_z) &= -p_x + \mu(u_{xx} + u_{yy} + u_{zz}) - \phi_x, \\ \rho(v_t + uv_x + vv_y + wv_z) &= -p_y + \mu(v_{xx} + v_{yy} + v_{zz}) - \phi_y, \\ \rho(w_t + uw_x + vw_y + ww_z) &= -p_z + \mu(w_{xx} + w_{yy} + w_{zz}) - \phi_z, \end{aligned} \tag{12.2}$$

where ρ, μ are, respectively, the density and kinematic viscosity of the liquid; u, v, w are the respective components of the fluid velocity vector $\mathbf{v}$ in the directions x, y, z; t is time; and p is pressure.

The classical boundary conditions between the liquid and the solid surface supporting it are those of no-penetration $w = 0$ and no-slip $u = 0$, $v = 0$. These conditions are appropriate for the continuous films to be considered. Problems with a contact line, where the liquid on a solid surface spreads or recedes will not be examined in this chapter. The reader interested in this topic is referred to the review papers by de Gennes (1985), Leger and Joanny (1992), and Oron et al. (1997).

The boundary conditions at the solid surface are therefore

$$w = 0, \quad u = 0, \quad v = 0 \qquad \text{at } z = 0. \tag{12.3}$$

At the film surface $z = h(x, y, t)$ the boundary conditions are formulated in the vector form [e.g., Wehausen and Laitone, 1960]:

$$h_t + \mathbf{v} \cdot \nabla^* h \sim w = 0, \tag{12.4a}$$

$$\mathbf{T} \cdot \mathbf{n} = -2\widetilde{H}\sigma\mathbf{n} + \nabla_s\sigma + \Pi, \tag{12.4b}$$

where $\mathbf{T}$ is the stress tensor of the liquid, Π is the prescribed forcing at the interface, $\widetilde{H}$ is the mean curvature of the interface determined from

$$2\widetilde{H} = \nabla^* \cdot \mathbf{n} = -\frac{h_{xx}(1 + h_y^2) + h_{yy}(1 + h_x^2) - 2h_x h_y h_{xy}}{(1 + h_x^2 + h_y^2)^{3/2}}, \tag{12.5}$$

$\nabla^* = (\partial/\partial x, \partial/\partial y, \partial/\partial z)$ is the gradient operator and ∇_s is the surface gradient with respect to the interface $z = h(x, y, t)$. Note that in Equation (12.4) the "dot" represents both the inner product of two vectors and the product of a tensor and a vector, respectively.

Equation (12.4a) is the kinematic boundary condition formulated in the absence of interfacial mass transfer and represents the balance between the normal component of the liquid velocity at the interface and the velocity of the interface itself. An appropriate change should be made in Equation (12.4a) to accommodate the phenomena of evaporation or condensation (see the section on phase changes). Equation (12.4b), which constitutes the balance of interfacial stresses in the absence of interfacial mass transfer, has three components. The physical meaning of its two tangential components is that the shear stress at the interface is balanced by the sum of the respective Π_i, $i = 1, 2$ and the surface gradient of surface tension σ. The normal component of Equation (12.4b) states that the difference between the normal interfacial stress and Π_3 exhibits a jump equal to the product of twice the mean curvature of the film interface and surface tension. This jump is known in the literature as the capillary pressure. When the external force Π is zero, and the fluid has zero viscosity or the fluid is static $v = 0$, then $\mathbf{T} \cdot \mathbf{n} \cdot \mathbf{n} = -p$, and Equation (12.4b) reduces to the well-known Young–Laplace equation. This equation describes, for instance, the excess pressure in an air bubble gauged to the external pressure, as twice the surface tension divided by the bubble radius (see e.g., [Landau and Lifshitz, 1987]). The subsequent derivations closely follow those made by Oron et al. (1997) when explicitly extended into three dimensions.

Projecting Equation (12.4b) onto the directions $\mathbf{n}$, $\mathbf{t}_1$, $\mathbf{t}_2$, respectively, yields

$$-p + \frac{2\mu[u_x(h_x^2 - 1) + v_y(h_y^2 - 1) + h_x h_y(u_y + v_x) - h_x(u_z + w_x) - h_y(v_z + w_y)]}{1 + h_x^2 + h_y^2} = 2\widetilde{H}\sigma + \Pi_3,$$

$$\mu[(u_z + w_x)(1 - h_x^2) - (v_z + w_y)h_x h_y - (u_y + v_x)h_y - 2(u_x - w_z)h_x] = \left(\Pi_1 + \frac{\partial\sigma}{\partial x}\right)(1 + h_x^2 + h_y^2)^{1/2},$$

$$\mu[-(u_z + w_x)h_x h_y + (v_z + w_y)(1 - h_y^2) - (u_y + v_x)h_x - 2(v_y - w_z)h_y] = \left(\Pi_2 + \frac{\partial\sigma}{\partial y}\right)(1 + h_x^2 + h_y^2)^{1/2}. \tag{12.6}$$

Let us now introduce scales appropriate for thin films where the transverse length scale is much smaller than the lateral ones. Assume length scales in the lateral directions, x and y, to be defined by wavelength λ of the interfacial disturbance on a film of mean thickness d. The film is referred to as *thin* film if the interfacial distortions are much longer than the mean film thickness, that is,

$$\varepsilon = \frac{d}{\lambda} \ll 1. \tag{12.7}$$

The z-coordinate (normal to the solid substrate) is normalized with respect to d, while the coordinates x, y are scaled with λ or equivalently d/ε. Thus the dimensionless z-coordinate is defined as

$$\varsigma = \frac{z}{d}, \tag{12.8a}$$

while the dimensionless x- and y-coordinates are given by

$$\xi = \frac{\varepsilon x}{d}, \quad \eta = \frac{\varepsilon y}{d}. \tag{12.8b}$$

It is assumed that in the new spatial variables no rapid variations occur as $\varepsilon \to 0$, then

$$\frac{\partial}{\partial \xi}, \frac{\partial}{\partial \eta}, \frac{\partial}{\partial \varsigma} = O(1). \tag{12.8c}$$

If the lateral components of the velocity field u, v are assumed to be of order one and U_0 denotes the characteristic velocity of the problem, the dimensionless fluid velocities in the x- and y- directions are defined as

$$U = \frac{u}{U_0}, \quad V = \frac{v}{U_0}. \tag{12.8d}$$

Then the continuity Equation (12.2) requires that the z-component of the velocity field w is small, and the dimensionless fluid velocity in the z-direction is defined as

$$W = \frac{w}{\varepsilon U_0} \tag{12.8e}$$

We stress that the characteristic velocity U_0 is not specified here for the sake of generality. The freedom of choosing this value is thus given to the user. We just note one of the possible choices but not the unique one $U_0 = \mu/\rho d$, which is known in the literature as a "viscous velocity."

Time is scaled in the units of λ/U_0, so that the asymptotically long-time behavior of the film is considered. The dimensionless time is therefore defined via

$$\tau = \frac{\varepsilon U_0 t}{d} \tag{12.8f}$$

Finally, because of the assumed slow lateral variation of the film interface, one expects locally parallel flow in the liquid, so that the pressure gradient is balanced with the viscous stress $p_x \propto \mu u_{zz}$, and the dimensionless interfacial stresses, body-force potential and pressure are defined, respectively, as

$$(\Pi_1, \Pi_2, \Pi_3) = \frac{d}{\mu U_0} \, (\hat{\Pi}_1, \hat{\Pi}_2, \varepsilon\hat{\Pi}_3), (\Phi, \mathrm{P}) = \frac{\varepsilon d}{\mu U_0} \, (\phi, p). \tag{12.8g}$$

Notice that pressure is asymptotically large similar to the situation arising in the lubrication effect [Schlichting, 1968].

If all these dimensionless variables are substituted into the governing system of Equations (12.2)–(12.5), the following scaled system is obtained:

$$U_\xi + V_\eta + W_\varsigma = 0, \tag{12.9a}$$

$$\varepsilon R(U_\tau + UU_\xi + VU_\eta + WU_\varsigma) = -P_\xi + U_{\varsigma\varsigma} + \varepsilon^2(U_{\xi\xi} + U_{\eta\eta}) - \Phi_\xi, \tag{12.9b}$$

$$\varepsilon R(V_\tau + UV_\xi + VV_\eta + WV_\varsigma) = -P_\eta + V_{\varsigma\varsigma} + \varepsilon^2(V_{\xi\xi} + V_{\eta\eta}) - \Phi_\eta, \tag{12.9c}$$

$$\varepsilon^3 R(W_\tau + UW_\xi + VW_\eta + WW_\varsigma) = -P_\varsigma + \varepsilon^2 W_{\varsigma\varsigma} + \varepsilon^4(W_{\xi\xi} + W_{\eta\eta}) - \Phi_\varsigma. \tag{12.9d}$$

At $\varsigma = 0$:

$$W = 0, \quad U = 0, \quad V = 0. \tag{12.10}$$

At $\varsigma = H$:

$$W = H_\tau + UH_\xi + VH_\eta, \tag{12.11a}$$

$$\frac{2\varepsilon^2[U_\xi(\varepsilon^2 H_\xi^2 - 1) + V_\eta(\varepsilon^2 H_\eta^2 - 1) + \varepsilon^2 H_\xi H_\eta(U_\eta + V_\xi) - H_\xi(U_\varsigma + W_\xi) - H_\eta(V_\varsigma + W_\eta)]}{1 + \varepsilon^2 (H_\xi^2 + H_\eta^2)}$$

$$= P + \hat{\Pi}_3 + \frac{\bar{S}\varepsilon^3[H_{\xi\xi}(1 + \varepsilon^2 H_\eta^2) + H_{\eta\eta}(1 + \varepsilon^2 H_\xi^2) - 2\varepsilon^2 H_\xi H_\eta H_{\xi\eta}]}{[1 + \varepsilon^2(H_\xi^2 + H_\eta^2)]^{3/2}}, \tag{12.11b}$$

$$\begin{gathered}(U_\varsigma + \varepsilon^2 W_\xi)(1 - \varepsilon^2 H_\xi^2) - \varepsilon^2(V_\varsigma + \varepsilon^2 W_\eta)H_\xi H_\eta - \varepsilon^2(U_\eta + V_\xi)H_\eta - 2\varepsilon^2(U_\xi - W_\varsigma) H_\xi \\ = (\hat{\Pi}_1 + \Sigma_\xi)[1 + \varepsilon^2(H_\xi^2 + H_\eta^2)]^{1/2},\end{gathered} \tag{12.11c}$$

$$\begin{aligned}(V_\varsigma + \varepsilon^2 W_\eta)(1 - \varepsilon^2 H_\eta^2) - \varepsilon^2(U_\varsigma + \varepsilon^2 W_\xi)H_\xi H_\eta - \varepsilon^2(U_\eta + V_\xi)H_\xi - 2\varepsilon^2(V_\eta - W_\varsigma)H_\eta \\ = (\hat{\Pi}_2 + \Sigma_\eta)[1 + \varepsilon^2\,(H_\xi^2 + H_\eta^2)]^{1/2}.\end{aligned} \tag{12.11d}$$

Here $H = h/d$ is the dimensionless thickness of the film and $\Sigma = \varepsilon\sigma/\mu U_0$ is the dimensionless surface tension normalized with respect to its characteristic value. The Reynolds number R and the inverse capillary number $\bar{S}$ are defined by

$$R = \frac{U_0 d\rho}{\mu}, \quad \bar{S} = \frac{\sigma}{U_0\mu}. \tag{12.12}$$

The continuity Equation (12.9a) is now integrated in ς across the film from 0 to H (ξ, η, τ), and Equations (12.10) and (12.11a) are used along with integration by parts to obtain

$$H_\tau + \frac{\partial}{\partial\xi}\int_0^H U d\varsigma + \frac{\partial}{\partial\eta}\int_0^H V d\varsigma = 0. \tag{12.13}$$

Equation (12.13) is a more convenient form of the kinematic condition because only two of three components of the fluid velocity field appear explicitly. It also warrants conservation of mass in a domain with a deflecting upper boundary.

The solution of the governing Equations (12.2)–(12.5) is sought in the form of expansion of the dependent variables into asymptotic series in powers of the small parameter ε:

$$\begin{aligned}U &= U^{(0)} + \varepsilon U^{(1)} + \varepsilon^2 U^{(2)} + \cdots, \quad V = V^{(0)} + \varepsilon V^{(1)} + \varepsilon^2 V^{(2)} + \cdots, \\ W &= W^{(0)} + \varepsilon W^{(1)} + \varepsilon^2 W^{(2)} + \cdots, \quad P = P^{(0)} + \varepsilon P^{(1)} + \varepsilon^2 P^{(2)} + \cdots.\end{aligned} \tag{12.14}$$

One way to approximate the solution of the governing system is to assume that R, $\bar{S} = O(1)$ as $\varepsilon \to 0$. Under this assumption the inertial terms, measured by εR, are one order of magnitude smaller than the dominant viscous terms, consistent with the local-parallel-flow assumption. The surface tension terms, measured by $\bar{S}\varepsilon^3$, are two orders of magnitude smaller and would be lost. It is essential to retain surface-tension effects at leading order, so it is assumed that capillary effects are strong relative to those of viscosity and

$$\bar{S} = S\varepsilon^{-3}. \tag{12.15}$$

It is then assumed that R, $S = O(1)$, as $\varepsilon \to 0$.

Equations (12.14) and (12.15) are substituted into Equations (12.9)–(12.11) and (12.13), and the resulting equations are sorted with respect to the powers of ε. At leading order in ε the governing system becomes, after omitting the superscript "zero" in $U^{(0)}$, $V^{(0)}$, $W^{(0)}$, and $P^{(0)}$,

$$U_{\varsigma\varsigma} = (P + \Phi)_\xi, \tag{12.16a}$$

$$V_{\varsigma\varsigma} = (P + \Phi)_\eta, \tag{12.16b}$$

$$(P + \Phi)_\varsigma = 0, \tag{12.16c}$$

$$H_\tau + UH_\xi + VH_\eta - W = 0, \tag{12.16d}$$

$$U_\xi + V_\eta + W_\varsigma = 0 \tag{12.16e}$$

with the boundary conditions at $\varsigma = 0$:

$$W = 0, \quad U = 0, \quad V = 0, \tag{12.17}$$

and at $\varsigma = H$:

$$\begin{aligned}P &= -\hat{\Pi}_3 - S(H_{\xi\xi} + H_{\eta\eta}), \\ U_\varsigma &= \hat{\Pi}_1 + \Sigma_\xi, \\ V_\varsigma &= \hat{\Pi}_2 + \Sigma_\eta.\end{aligned} \tag{12.18}$$

We note here that Equations (12.16)–(12.18) are linear with respect to the variables U, V, W, P. The only nonlinearity of this problem is associated, as seen from Equation (12.19) in conjunction with the kinematic condition Equation (12.16d), with the local film thickness $H(\xi, \eta, \tau)$. Solving Equations (12.16)–(12.18) yields

$$\begin{aligned} U &= \left[\frac{1}{2}\varsigma^2 - H\varsigma\right](\Phi - \hat{\Pi}_3|_{\varsigma=H} - S\nabla^2 H)_\xi + \varsigma(\hat{\Pi}_1 + \Sigma_\xi), \\ V &= \left[\frac{1}{2}\varsigma^2 - H\varsigma\right](\Phi - \hat{\Pi}_3|_{\varsigma=H} - S\nabla^2 H)_\eta + \varsigma(\hat{\Pi}_2 + \Sigma_\eta), \\ W &= -\int_0^\varsigma (U_\xi + V_\eta)d\varsigma, \quad P = -\hat{\Pi}_3|_{\varsigma=H} - S\nabla^2 H. \end{aligned} \tag{12.19}$$

If Equation (12.19) is substituted into the mass conservation Equation (12.13), one obtains the appropriate evolution equation for the interface,

$$H_\tau + \frac{1}{2}\nabla \cdot [H^2(\hat{\Pi}^\star + \nabla\Sigma)] + \frac{1}{3}\nabla \cdot \{H^3[\nabla(\hat{\Pi}_3 - \Phi|_{\varsigma=H}) + S\nabla\nabla^2 H]\} = 0, \tag{12.20}$$

where $\hat{\Pi}^\star = (\hat{\Pi}_1, \hat{\Pi}_2)$ is the tangential projection of the dimensionless vector $\hat{\Pi}$, $\nabla \equiv (\partial/\partial\xi, \partial/\partial\eta)$ and $\nabla^2 \equiv \partial^2/\partial\xi^2 + \partial^2/\partial\eta^2$.

In two dimensions ($\partial/\partial\eta = 0$) this evolution equation reduces to

$$H_\tau + \frac{1}{2}[H^2(\hat{\Pi}_1 + \Sigma_\xi)]_\xi + \frac{1}{3}\{H^3[(\hat{\Pi}_3 - \Phi|_{\varsigma=H})_\xi + SH_{\xi\xi\xi}]\}_\xi = 0. \tag{12.21}$$

In these equations the location of the film interface $H = H(\xi, \eta, \tau)$ is unknown and is determined from the solution of the corresponding partial differential equation. When such a solution is obtained, the components of the velocity and the pressure fields can be determined from Equation (12.19).

The physical significance of the terms becomes apparent when Equations (12.20) and (12.21) are written in the original dimensional variables:

$$\mu h_t + \frac{1}{2}\overline{\nabla} \cdot [h^2(\Pi^\star + \overline{\nabla}\sigma)] + \frac{1}{3}\overline{\nabla} \cdot \{h^3[\overline{\nabla}(\Pi_3 - \phi|_{z=h}) + \sigma\overline{\nabla}\,\overline{\nabla}^2 h]\} = 0, \tag{12.22}$$

with $\overline{\nabla} \equiv (\partial/\partial x, \partial/\partial y)$, $\overline{\nabla}^2 \equiv (\partial^2/\partial x^2 + \partial^2/\partial y^2)$ and

$$\mu h_t + \frac{1}{2}[h^2(\Pi_1 + \sigma_x)]_x + \frac{1}{3}\{h^3[(\Pi_3 - \phi\,|_{z=h})_x + \sigma h_{xxx}]\}_x = 0. \tag{12.23}$$

The first term in Equations (12.22) and (12.23) represents the effect of viscous damping, while the next ones account, respectively, for the effects of the imposed tangential interfacial stress, non-uniformity of surface tension, the imposed normal interfacial stress, body forces, and surface tension on the dynamics of the film.

In the following examples, two- and three-dimensional cases are examined. Unless specified, only disturbances periodic in x and y are discussed. Thus, λ is the wavelength of these disturbances, and $2\pi d/\lambda$ is the corresponding dimensionless wavenumber. In accordance with this, Equations (12.20)–(12.23) are normally solved with periodic boundary conditions. These equations whether in two or three dimensions are of fourth order in each of the spatial variables, and therefore four boundary conditions are needed to define a well-posed mathematical problem. These four boundary conditions imply periodicity of the solution H and its first, second, and third derivatives with respect to the corresponding spatial variable. At the same time, Equations (12.20)–(12.23) are of first order in time, thus one initial condition is needed to complete the well-posed statement of the problem. This initial condition representing the location of the film interface at $t = 0$ or $\tau = 0$ is usually taken as a small-amplitude random or sinusoidal disturbance on top of the uniform state given by $H = 1$. In two dimensions it can be written by

$$H(\tau = 0, \xi) = 1 + \delta\sin(\xi + \varphi) \quad \text{or} \quad H(\tau = 0, \xi) = 1 + \delta\,\mathrm{rand}(\xi), \tag{12.24}$$

where $\delta \ll 1$, φ is a phase, and rand(ξ) is a random function uniformly distributed in the interval $(-1, 1)$. An extension of Equation (12.24) can be obtained in the three-dimensional case.

12.3 Isothermal Films

We now examine the dynamics of films whose temperature remains unchanged and phase changes do not occur.

12.3.1 Constant Surface Tension and Gravity

Consider the simplest case in which the film is supported from below by a solid surface and subjected to the influence of gravity and constant surface tension. In this case one has $\Sigma_\xi = \Sigma_\eta = \hat{\Pi}_1 = \hat{\Pi}_2 = \hat{\Pi}_3 = 0$ and $\Phi = G\varsigma$, so that in two dimensions Equation (12.21) becomes

$$H_\tau - \frac{1}{3}G(H^3H_\xi)_\xi + \frac{1}{3}S(H^3H_{\xi\xi\xi})_\xi = 0, \tag{12.25a}$$

where G is the unit-order positive gravity number

$$G = \frac{\rho g d^2}{\mu U_0}.$$

The second term of Equation (12.25a) accounts for the influence of gravity, while the third one describes the effect of the capillary forces. The dimensional version of Equation (12.25a) is obtained from Equation (12.23) as

$$\mu h_t - \frac{1}{3}\rho g(h^3h_x)_x + \frac{1}{3}\sigma(h^3h_{xxx})_x = 0. \tag{12.25b}$$

In the absence of surface tension Equation (12.25b) is a nonlinear (forward) diffusion equation so that one can envision that no disturbance to $h = d$ experiences growth in time. Surface tension acts through a fourth-order (forward) dissipation term only enhancing stabilization of the interface, so that no instabilities would occur in the case described by Equation (12.25b) for $G > 0$.

To formally assess these intuitive observations one can investigate the stability properties of the uniform film $h = d$ perturbing it by a small disturbance h'periodic in x (i.e. $h = d + h'$ with $h' \ll d$). Substituting this into Equation (12.25b) and linearizing it with respect to h', one obtains the linear-stability equation for the uniform state $h = d$. Since this equation has coefficients independent of t and x, one can seek separable solutions of the form

$$h' = h_0' \exp(ikx + \omega t), \quad h_0' = \text{const},$$

which constitute a complete set of "normal modes" and can be used to represent any disturbance by means of the Fourier series. Here k is the wavenumber of the disturbance in the x direction. If these normal modes are substituted into the linear-stability equation, one obtains the following characteristic equation for ω:

$$\mu\omega = -\frac{1}{3d}(\rho g d^2 + \sigma a^2)a^2, \tag{12.26}$$

where $a = kd$ is the non-dimensional wavenumber and ω is the growth rate of the perturbation. In general, the amplitude of the perturbation will decay if the real part of the growth rate Re(ω) is negative, and will grow if Re(ω) is positive. Purely imaginary values of ω will correspond to translation along the x-axis and give rise to traveling-wave solutions. Finally, zero values of Re(ω) will correspond to neutral perturbations.

Two remarks are now in order. First, the linear stability analysis is carried out here in the dimensional form, but it could be done in the same way in the dimensionless form when its starting point would be Equation (12.25a). Second, the linear stability analysis is carried out here in the two-dimensional case. The same can be done in the three-dimensional case with respect to the normal modes

$$h' = h'_0 \exp(ik_x x + ik_y y + \omega t), \quad h'_0 = \text{const},$$

where k_x, k_y are, respectively, the wavenumbers in the x and y directions. As in the physical problem at hand, the symmetry is such that the spatial variables x and y are interchangeable and the characteristic equation for ω will be identical to Equation (12.26), but now $k = (k_x^2 + k_y^2)^{1/2}$ is the total wavenumber of the disturbance.

Equation (12.26) describes the rate of film leveling since $\omega < 0$ for any value of the dimensionless wavenumber a and the rest of the parameters. If at time $t = 0$ a small bump is imposed on the interface, Equation (12.26) describes how it will relax to zero and the interface will return to $h = d$.

The overall rate of film leveling can be estimated by the maximal value of the growth (decay in the case at hand) rate ω, as given by Equation (12.26). If the lateral size of the film is L, the fastest decaying mode is the longest available one so that its wavenumber is $k = 2\pi/L$. Thus the rate of disturbance decay is given by

$$\omega_m = -\frac{4\pi^2 d^3}{3\mu L^2}\left(\rho g + \frac{4\pi^2\sigma}{L^2}\right),$$

so the amplitude of the disturbance will reach the value of, say a thousandth of the initial amplitude at the time of $t = (\ln 0.001)/\omega_m$. However, this is only an estimate based on the linear stability analysis, and the effect of nonlinearities on the rate of film leveling can be found only from the solution of Equation (12.25).

Equations (12.25a, b) with the obvious change in the sign of the gravity term in each of these also apply to the case of a film on the underside of a plate. This case is known in the literature as the Rayleigh–Taylor instability [Chandrasekhar, 1961] of a thin viscous layer. To study the stability properties of such a system one replaces g by $-g$ in Equation (12.26) and finds that

$$\mu\omega = \frac{1}{3d}(\rho|g|d^2 - \sigma a^2)a^2. \tag{12.27}$$

The film is linearly unstable if

$$a^2 < a_c^2 \equiv \frac{\rho|g|d^2}{\sigma} \equiv Bo,$$

that is, if the perturbations are so long that the nondimensional wavenumber is smaller than the square root of the Bond number Bo, which measures the relative importance of gravity and capillary effects. The value of a_c is often called the (dimensionless) cutoff wavenumber for neutral stability. The cutoff wavenumber is defined in a way that all perturbations with the wavenumber larger than a_c are damped, while those with the wavenumber smaller than a_c are amplified.

We point out that Equations (12.25) constitute the valid limit to the governing set of equations and boundary conditions when the Bond number Bo is asymptotically small. This follows from the relationships $G = \varepsilon Bo\overline{S}$, $G = O(1)$, and the large value of $\overline{S}$, as assumed in Equation (12.15).

The case of Rayleigh–Taylor instability was studied by Yiantsios and Higgins (1989, 1991) for a thin film of a light fluid atop the plate and overlain by a large body of a heavy fluid, and by Oron and Rosenau (1992) for a thin liquid film on the underside of a plane. It was found that evolution of an interfacial disturbance of small amplitude leads to rupture of the film, that is, at certain location(s) the local thickness of the film is driven to zero.

The dimensionless wavenumber of the fastest growing mode is determined for a film of an infinite lateral extent from Equation (12.27) as $a = \sqrt{Bo/2}$, and its growth rate is determined from Equation (12.27) as

$$\omega_m = \frac{\rho^2 g^2 d^3}{12\mu\sigma}.$$

Thus the time of film rupture can be estimated by $t = (\ln d/h'_0)/\omega_m$.

Yiantsios and Higgins (1989) showed that Equation (12.25a) with $G < 0$ admits several steady solutions. These consist of various numbers of sinusoidal drops separated by "dry" spots of zero film thickness, as shown in Figure 8 in Yiantsios and Higgins (1989). The examination of an appropriate free energy functional [Yiantsios and Higgins, 1989] suggests that multi-drop states are energetically less preferred than a one-drop state. These analytical results were partially confirmed by numerical simulations. As found in the long-time limit, the solutions can asymptotically approach multi-humped states with different amplitudes and spacings. This suggests that terminal states depend upon the choice of initial data [Yiantsios and Higgins, 1989]. If the overlying semi-infinite fluid phase is more viscous than the thin liquid film, the process of the film rupture slows down in comparison with the single-fluid case.

Note that Equation (12.25a) with $G < 0$ was also derived and studied by Hammond (1983) in the context of capillary instability of a thin liquid film on the inner side of a cylindrical surface when gravity was neglected. The gravitational term was due to the destabilizing effect of the capillary forces arising from longitudinal (along the axis of the cylinder) disturbances. Hammond (1983) also showed that the film ruptures, but the process of rupture is infinitely long.

The three-dimensional version of the problem of the Rayleigh–Taylor instability was considered by Fermigier et al. (1992) using the weakly nonlinear analysis. Formation of patterns of different symmetries and transition between these patterns were experimentally studied. Axially symmetric cells and hexagons were preferred. Droplet detachment was observed at the final stage of the experiment as a manifestation of a film rupture. The growth of an axisymmetric drop is shown in Figure 5 in Fermigier et al. (1992). A theoretical study of the Rayleigh–Taylor instability in an extended geometry [Fermigier et al., 1992] on the basis of the long-wave equation showed the tendency of the hexagonal structures to emerge as a preferred pattern in agreement with their own experimental observations.

Saturation of the Rayleigh–Taylor instability of a thin liquid film, and therefore prevention of its rupture by an imposed advection in the longitudinal (parallel to the interface) direction, is discussed by Babchin et al. (1983). Similarly, capillary instability of an annular film saturates because of a through flow [Frenkel et al., 1987].

Stillwagon and Larson (1988) considered the problem of a film leveling under the action of capillary force on a substrate with topography given by $z = \lambda(x)$. Using the approach previously described, they derived the evolution equation that for the case of zero gravity reads

$$\mu h_\tau + \frac{1}{3}\sigma[h^3\,(h + \lambda)_{xxx}]_x = 0 \qquad (12.28)$$

Numerical solutions of Equation (12.28) showed a good agreement with their own experimental data. At short times there is film deplanarization because of the emergence of capillary humps, but these relax at longer times.

12.3.2 van der Waals Forces and Constant Surface Tension

Because of very small typical length scales of MEMS applications (and particularly of liquid film thickness) that go down into the range of fractions of a micrometer, new physics related mainly to intermolecular forces is considered. These fundamental types of forces acting on interatomic or intermolecular distances can affect the dynamics of macroscopic thin liquid films. Some of them, like weak and strong interactions, are short-range (i.e., much beyond the validity limits of continuum theory considered here). Others, like electromagnetic and gravitational forces, are of a long range and will be thus of a great importance for the subject of the current review.

Israelachvili (1992) presents a classification of electromagnetic forces into three categories. The first category consists of purely electrostatic forces arising from the Coulomb interaction. These forces include interactions between charges, dipoles, etc. The second category consists of polarization forces that stem from the dipole moments induced in totally neutral particles by the electric fields associated with other neighboring particles and permanent dipoles. These forces include interactions in a solvent medium. The third category consists of forces of quantum mechanics origin. Such forces lead to chemical bonding and to repulsive steric interactions. Among these forces is the force which acts, similar to the gravitational force, between all kinds of particles whether charged or neutral. This force is called "dispersion force" or "London force." The origin of the dispersion force is explained by the following consideration: in an electrically neutral particle whose time-averaged dipole moment vanishes, an instantaneous dipole moment does not vanish according to time-varying relative distribution of negative and positive charges. Such an instantaneous dipole moment gives rise to a dipole moment in the neighboring neutral particles, and the interaction between these dipoles induces the force with a non-vanishing time-averaged value. These dispersion forces are long-range forces acting at the distances from several angstroms to several hundred angstroms. They play, as we see later, a very important role in the dynamics of ultrathin liquid films whose average thickness is in this range and in various phenomena such as wetting and adhesion. The dispersion forces can be either attractive or repulsive affecting the properties of good or poor wetting of solids by liquids. The presence of other bodies alters the dispersion interaction between the molecules, thus the dispersion force is strictly non-additive. As shown in Table 6.3 of Israelachvili (1992), the dispersion force constitutes in many cases, except for highly polar water molecules, the main contribution to the total intermolecular force called van der Waals force. Various types of potentials describing the forces acting between molecules were reviewed by Israelachvili (1992).

Dzyaloshinskii et al. (1959) developed a theory for van der Waals interactions in which an integral representation is given for the excess Helmholtz free energy of the layer as functions of the frequency-dependent dielectric properties of the materials in the layered system.

The potential ϕ of the van der Waals forces is frequently specified in terms of the excess intermolecular free energy ΔG. These two values are related each to other via

$$\phi = \frac{\partial \Delta G}{\partial h}. \tag{12.29}$$

It follows in this case from Equation (12.22) in the 3-D case and Equation (12.23) in the 2-D case that the film is unstable to infinitesimal disturbances only if

$$\frac{\partial \phi}{\partial h} < 0 \quad \text{or} \quad \text{equivalently} \; \frac{\partial^2 \Delta G}{\partial h^2} < 0. \tag{12.30}$$

It follows from Equation (12.30) that the film is unstable only if the potential ϕ has a decreasing branch or ΔG displays a negative curvature, both as functions of the film thickness h.

In the special case of an apolar film with parallel boundaries and non-retarded forces,

$$\phi = \phi_r + A'h^{-3}/6\pi, \tag{12.31a}$$

where ϕ_r is an additive reference value for the body-force potential omitted hereafter and A' is the dimensional Hamaker constant [Dzyaloshinskii et al., 1959]. When $A' > 0$, there is negative disjoining pressure (referred to sometimes as conjoining pressure), and a corresponding attraction of the two interfaces (solid–liquid and liquid–gas) toward each other causes the instability of the flat state of the film surface and eventually its breakup. When the disjoining pressure is positive $A' < 0$ the interfaces repel each other, and the flat state of the film surface is energetically preferred.

The literature provides various forms for the potential ϕ accounting for more complex physical situations. Mitlin (1993), Mitlin and Petviashvili (1994), Khanna and Sharma (1997), and others used the 6–12 Lennart-Jones potential for van der Waals interactions between the solid and the apolar liquid

$$\phi = A_3'h^{-3} - A_9'h^{-9} \tag{12.31b}$$

with positive dimensional Hamaker coefficients A_j'. In this case the two interfaces of the film are mutually attracting when the separation distance is relatively large. This drives the instability of the flat state of the film surface. On the other hand, the two interfaces of the film are mutually repelling when the separation distance is relatively short. This leads to a final saturation of the amplitude of the interfacial undulation.

If the solid substrate is coated with a layer of thickness δ, the potential of the intermolecular pairwise interactions between the solid, coating, passive air, and apolar liquid phases is given by [Bankoff, 1990; Hirasaki, 1991; Sharma and Reiter, 1996; Khanna et al., 1996; Oron and Bankoff, 1999]

$$\phi = A_3' h^{-3} + \hat{A}_3'(h + \delta)^{-3}, \tag{12.31c}$$

where $A_3' = (A_{LL}' - A_{cL}')/6\pi$, $\hat{A}_3' = (A_{sL}' - A_{cL}')/6\pi$ with A_{ij}' being the Hamaker constant related to the interaction between the phases *i* and *j*, $A_{ij}' = A_{ii}'^{1/2} A_{jj}'^{1/2}$ [Israelachvili, 1992], and subscripts *s*, *c*, and *L* corresponding, respectively, to solid, coating, and liquid phases.

Oron and Bankoff (1999) derived the potential topologically similar to the Lennart-Jones potential Equation (12.31a) but with different exponents

$$\phi = A_3' h^{-3} - A_4' h^{-4} \tag{12.31d}$$

to model the simultaneous action of the attractive ($A_3' > 0$) long-range and repulsive ($A_4' > 0$) (relatively) short-range van der Waals interactions and their influence on the dynamics of the film. To obtain the potential Equation (12.31d), Equation (12.31a) was expanded into the Taylor series in h under assumption of $\delta \ll d$ with$\hat{A}_3' > 0, A_3' + \hat{A}_3' > 0$, and only two leading terms of this expansion were kept. Thus the coefficients $A_3' A_4'$ are specified by the properties of the three phases. The potential of the form Equation (12.31d) is also appropriate for liquid films on a rough solid substrate [Teletzke et al., 1987; Mitlin, 2000].

A combination of long-range apolar (van der Waals) and shorter-range polar intermolecular interactions gives rise to the generalized disjoining pressure expressed by the potential

$$\phi = A_3' h^{-3} - S^p \exp(-h/\lambda)/\lambda, \tag{12.31e}$$

where S^p, λ are dimensional constants [Williams, 1981; Sharma and Jameel, 1993; Jameel and Sharma, 1994; Paulsen et al., 1996; Sharma and Khanna, 1998; and others] that are, respectively, the strength of the polar interaction and its decay length λ called the correlation length for polar interaction. The polar component of the potential is repulsive if $S^p > 0$ and is attractive if $S^p < 0$. Sharma and Jameel (1993) classified films with polar and apolar components into four groups: type I systems with both polar and apolar attractive forces ($A_3' > 0$, $S^p < 0$), type II systems with apolar attractions and polar repulsions($A_3' > 0$, $S^p > 0$), type III systems with both polar and apolar repulsions ($A_3' < 0, S^p > 0$), and type IV systems with apolar repulsions and polar attractions ($A_3' < 0$, $S^p < 0$). Films of type I are always unstable and their dynamics are in many ways similar to that of apolar films described by the potential Equation (12.31a), while those of type III are always stable. Films of type II and IV display ranges of stability and instability according to the sign of the derivative $\partial\phi/\partial h$. See the instability criterion Equation (12.30).

12.3.2.1 Homogeneous Substrates

Scheludko (1967) observed experimentally spontaneous breakup of ultrathin, static films and proposed that negative disjoining pressure is responsible. He also used linear stability analysis to calculate a critical thickness of the film below which breakup occurs, while neglecting the presence of electric double layers. Since then a great deal of scientific activity has focused on the phenomenon.

The dynamics of ultrathin liquid films and the process of dewetting of solid surfaces have attracted a special interest during the last decade. Progress and development of both experimental techniques such as ellipsometry, X-ray reflectometry, and atomic force microscopy (AFM), and computational techniques along with the availability and affordability of fast computers helped to advance the study of the pertinent phenomena. The main interest is centered about the pattern formation and the quest for the dominant mechanisms driving the film evolution. In the context of the latter issue the polemics are ongoing between the two candidates, namely thin film instability arising from the interaction between the intermolecular and capillary forces called sometimes in the literature "spinodal dewetting" or "a spinodal mode," and nucleation of holes from impurities or defects. It should be noted that most if not all of the experiments with dewetting recorded in the literature were carried out on liquid polymer films, while the theory is currently available for simple Newtonian liquids. The reasons for using polymer films in terms of controllability of the experiments were discussed by Sharma and Reiter (1996) and Reiter et al. (1999b).

Bischof et al. (1996) performed experiments on ultra-thin (≈40 nanometers) metal (gold, copper, and nickel) films on a fused silica substrate irradiated by a laser and turned into the liquid phase. Isolated holes, coalesced holes, and the typical rims surrounding them were observed. Little humps were found in the center of many holes, and the mechanism of heterogeneous hole nucleation was suggested to be responsible for formation of these. However, along with this mechanism, growing film surface deformations were detected, and thus the mechanism of spinodal dewetting is also in effect. The characteristic size of film surface deformations is well-correlated with the wavelength of the most amplified linear mode proportional to d^2. Similar conclusions about the dominance of the nucleation mechanism were drawn later by Jacobs et al. (1998). Experimental evidences of spinodal dewetting were given by Brochard-Wyart and Daillant (1990), Reiter (1992), Sharma and Reiter (1996), Xie et al. (1998), Reiter et al. (1999b, 2000), and others. Reiter et al. (2000) showed for the first time that the spinodal length and time scales are consistent with the results of their experiments. Independent molecular dynamics simulations [Koplik and Banavar, 2000] support the spinodal character of dewetting.

Khanna et al. (2000) presented the first real time experimental observation of the pattern formation in thin unstable polydimethylsiloxane (PDMS) films placed on a coated silicon wafer and bounded by aqueous surfactant solutions. The process of film disintegration ("self-destruction") was described by the following sequence of stages: self-organization of the pattern and selective amplification of the interfacial disturbance, breakup of the film and formation of isolated circular holes, lateral expansion of the holes and emergence of long liquid ridges, and lastly breakup of the ridges into droplets standing on an equilibrium film plateau and ripening of the droplet structure.

Muller-Buschbaum et al. (1997) studied the process of dewetting of thin polysterene films on silicon wafers covered with an oxide layer of different thicknesses and observed the emergence of "nano-dewetting structures" inside the dewetted areas. These structures in the form of troughs of about 70 nanometers in diameter confirmed that the dewetted areas were neither completely dry nor covered with a flat ultrathin layer of the liquid. Such patterns were detected along with micrometer-size drops usually observed in similar situations on top of oxide layers that were 24 angstroms thick but were not present on thinner oxide layers where only drops emerged. The dependence of the mean drop size as well as the trough diameter on the initial thickness of the film was in agreement with theoretical predictions based on the assumption of spinodal dewetting [Muller-Buschbaum et al., 1997].

Consider now a film under the influence of van der Waals forces and constant surface tension only, so that $\Pi_1 = \Pi_2 = \Pi_3 = \sigma_x = \sigma_y = 0$. As we see shortly the planar film is unstable when $A' > 0$ and stable when $A' < 0$. In two dimensions Equation (12.23) in the case at hand becomes [Williams and Davis, 1982]

$$\mu h_t + \frac{1}{6\pi} A'(h^{-1}h_x)_x + \frac{1}{3}\sigma(h^3 h_{xxx})_x = 0. \tag{12.32a}$$

Its dimensionless version reads

$$H_\tau + A(H^{-1}H_\xi)_\xi + \frac{1}{3}S(H^3 H_{\xi\xi\xi})_\xi = 0, \tag{12.32b}$$

where

$$A = \frac{\varepsilon A'}{6\pi\rho\nu^2 d}$$

is the scaled dimensionless Hamaker constant. Here the characteristic velocity was chosen as $U_0 = \nu/d$.

If Equation (12.32a) is linearized around $h = d$ the following characteristic equation for ω

$$\mu\omega = \left(\frac{a}{d}\right)^2\left(\frac{A'}{6\pi d} - \frac{1}{3}\sigma d a^2\right) \tag{12.33a}$$

is obtained. It follows from Equation (12.33a) that there is instability for $A' > 0$, driven by the long-range molecular forces, and stabilization is due to surface tension. The cutoff wavenumber a_c is given then by

$$a_c = \frac{1}{d}\left(\frac{A'}{2\pi\sigma}\right)^{1/2}, \tag{12.33b}$$

which reflects that an initially corrugated interface has its thin regions thinned further by van der Waals forces while surface tension cuts off the small scales. Instability is possible only if $0 < a < a_c$, as seen by combining Equations (12.33a) and (12.33b):

$$\mu\omega = \frac{\sigma a^2}{3d}\,(a_c^2 - a^2). \tag{12.34}$$

Similar results were obtained in the linear stability analysis presented by Jain and Ruckenstein (1974). On the periodic infinite domain of wavelength $\lambda = 2\pi/k$, the linearized theory predicts that the film is always unstable since all wave numbers are available to the system. In an experimental situation the film resides in a container of finite width, say L. The solution obtained from the linear stability theory for $0 \leqslant \xi \leqslant L$ would show that only perturbations of the non-dimensional wavenumber lower than a_c, see Equation (12.34), and those of small enough wavelength that "fit" in the box (i.e., $\lambda < L$) are unstable. Hence no instability would occur by this estimate if $2\pi d/L > a_c$. It is inappropriate to seek a "global" critical thickness from the theory but only a critical thickness for a given experiment, since the condition depends on the system size L.

The evolution of the film interface as described by Equation (12.32) with periodic boundary conditions and an initial linearly unstable perturbation of the uniform state leads to the rupture of the film in a finite (non-dimensional) time τ_R [Williams and Davis, 1982]. This rupture manifests itself by the fact that at a certain time the local thickness of the film becomes zero. The time of rupture of the film of an infinite lateral extent can be estimated from the linear stability theory by

$$t_R = \frac{48\pi^2 d^5\sigma}{A'}\ln\left(\frac{d}{h_0'}\right).$$

However, the rate of film thinning, measured as the rate of decrease of the minimal thickness of the film, explosively increases with time and becomes much larger than the disturbance growth rate given by Equation (12.33a) according to the linear theory. This phenomenon was found numerically from the solution of Equation (12.32b) [Williams and Davis, 1982] and analytically by weakly nonlinear theory [Sharma and Ruckenstein, 1986; Hwang et al., 1993]. Hwang et al. (1997) studied the three-dimensional version of this problem using the natural extension of Equation (12.32b). They confirmed film rupture and found that it occurs pointwise and not along a line. Moreover, the rupture time in the three-dimensional case is shorter than in the two-dimensional case.

Burelbach et al. (1988) used numerical analysis to show that, in a certain time range near the rupture point, surface tension has a minor effect, and therefore the local behavior of the interface is governed by the backward diffusion equation

$$H_\tau + A(H^{-1}H_\xi)_\xi = 0. \tag{12.35}$$

Looking for separable solutions for Equation (12.35) in the form $H(\xi, \tau) = T(\tau)\,X(\xi)$, Oron et al. (1997) used the known temporal asymptotics [Burelbach et al., 1988] and found that [also, Rosenau, 1995]

$$H(\xi, \tau) = A\frac{b^2}{2}(\tau_R - \tau)\sec^2\left(\frac{b\xi}{2}\right), \tag{12.36}$$

where τ_R is the time of rupture and b is the constant which should be determined from the matching with the far-from-rupture solution. The minimal thickness of the film close to the rupture point is therefore expected to decrease linearly with time. This allows the long-wave analysis to be extrapolated closer to the point where adsorbed layers and moving contact lines appear. However, the solution Equation (12.35) is not expected to be valid very close to the rupture point, where the film progresses toward rupture and the fluid velocities diverge. Recently, the existence of infinite sets of similarity solutions in which both van der Waals and surface tension forces are equally important near rupture was shown [Zhang and Lister, 1999; Witelski and Bernoff, 1999, 2000]. These solutions have the same form in both two-dimensional and axisymmetric cases

$$H(\xi,\tau) = (\tau_R - \tau)^{1/5}\, g[\xi(\tau_R - \tau)^{-2/5}], \tag{12.37}$$

where g is a function to be determined. Among this infinite set of self-similar solutions the fundamental solution stable to linear perturbations was identified as the only asymptotic behavior observed in the direct numerical solution of Equation (12.32b) [Witelski and Bernoff, 1999; Zhang and Lister, 1999]. It is described by the function g the least oscillatory one among the possible solutions of the corresponding ordinary differential equation. The point rupture is the preferred mode of film rupture in three dimensions [Witelski and Bernoff, 2000].

Several authors [Kheshgi and Scriven, 1991; Mitlin, 1993; Sharma and Jameel, 1993; Jameel and Sharma, 1994; Mitlin and Petviashvili, 1994; Oron and Bankoff, 1999] have considered the dynamics of thin liquid films in the process of dewetting a solid surface. The effects important for a meaningful description of the process are gravity, capillarity, and if necessary, the use of a generalized disjoining pressure, which contains a sum of intermolecular attractive and repulsive potentials. The generalized disjoining pressure of the Mie type is destabilizing (attractive) or stabilizing (repulsive) for the film of a larger (smaller) thickness, still within the range of several hundreds of angstroms [Israelachvili, 1992] where van der Waals interactions are effective. Equations (12.21) and (12.23) can be rewritten in the situation considered, respectively, in the form

$$H_\tau - \frac{1}{3}[H^3(GH - SH_{\xi\xi} + \Phi)_\xi]_\xi = 0, \tag{12.38a}$$

$$\mu h_t - \frac{1}{3}[h^3(\rho g h - \sigma h_{xx} + \phi)_x]_x = 0. \tag{12.38b}$$

Linearizing Equation (12.38b) around $h = d$, one obtains

$$\mu\omega = -\frac{1}{3}a^2 d\left(\rho g + \frac{\partial\phi}{\partial h}d + \frac{\sigma a^2}{d^2}\right). \tag{12.39}$$

It follows from Equation (12.39) that the necessary condition for linear instability is

$$\frac{\partial\phi}{\partial h}d < -\rho g, \tag{12.40}$$

that is, the destabilizing effect of the van der Waals force has to be stronger than the leveling effect of gravity.

Kheshgi and Scriven (1991) studied the evolution of the film using Equation (12.38a) with the potential Equation (12.31a) and found that smaller disturbances decay because of the presence of gravity leveling, while larger ones grow and lead to film rupture propelled by van der Waals force. Mitlin (1993) and Mitlin and Petviashvili (1994) discussed possible stationary states for the late stage of solid-surface dewetting with the potential Equation (12.31b) and drew the formal analogy between the latter and the Cahn theory of spinodal decomposition [Cahn, 1961]. Sharma and Jameel (1993) and Jameel and Sharma (1994) followed the film evolution as described by Equations (12.38) and (12.31e) with no gravity ($G = 0$) and concluded that thicker films break up, while thinner ones undergo “morphological phase separation” that manifests itself in creation of steady structures of drops separated by ultra-thin practically flat liquid films (holes). Similar patterns of morphological phase separation were also observed by Oron and Bankoff (1999) in their study of the dynamics of thin spots near film breakup. Figure 2 in Oron and Bankoff (1999) shows typical steady-state solutions for Equation (12.38a) with the potential Equation (12.31d) and $G = 0$ for different sets of parameters.

Khanna and Sharma (1998) used the Lennart-Jones potential Equation (12.31b) to study the three-dimensional dynamics of an apolar liquid film on a solid substrate. Their investigation based on the dimensionless evolution equation

$$H_\tau - \frac{1}{3}\nabla \cdot (H^3\nabla\Phi) + \frac{1}{3}\, S\nabla \cdot (H^3\nabla\nabla^2 H) = 0 \tag{12.41}$$

showed that in the case of $A'_9 d^6 \ll A'_3$ the corresponding film evolution displays the formation of steep holes. These holes are axisymmetric when the size of the periodic domain slightly exceeds the critical wavelength. However, they are non-axisymmetric with uneven rims surrounding the holes for larger domains.

Sharma and Khanna (1998) studied the film dynamics governed by Equation (12.41) with the potential Equation (12.31e) that engenders short-range polar repulsion, intermediate-range van der Waals attraction, and long-range polar repulsion. The linear and weakly nonlinear analyses fail to predict the structure of the emerging patterns. The former, however, can successfully predict the length scale of the resulting pattern. Two characteristic morphologically different patterns were found and in both of them the true dewetting does not occur. A microfilm covering the solid surface emerges and persists instead. The first pattern is typical for the films whose thickness is closer to the upper critical thickness. In this case the film undergoes the stages of reorganization into a pattern of a length scale corresponding to the fastest growing linear mode, emergence of circular holes with rims uneven in height, coalescence of the holes, and slow evolution into circular drops standing on top of a flat microfilm. The second pattern typical for relatively thin films of initial thickness near the lower critical thickness does not exhibit formation of circular holes and instead produces droplets that tend to be circular subject to the capillary forces. This type of a film evolution seems to be less frequent but was also observed in the experiments of Xie et al. (1998). The flat microfilm covering the substrate emerges after the formation of isolated drops. Finally, a stable state that consists of a single circular drop standing on a flat equilibrium film is reached. In the intermediate range of the initial film thickness, the patterns consisting of holes, ridges, and drops coexist when the number of each of these depends on the initial film thickness. As will be discussed later, all kinds of structures that contain holes, drops, and ridges may coexist on heterogeneous substrates [Konnur et al., 2000].

Sharma et al. (2000) attributed the type of film dewetting to the relative position of the average thickness of the film d and the location of the minimum of the function $\partial\phi/\partial h$. When the film is thicker than the thickness corresponding to the minimum of $\partial\phi/\partial h$, the film dewets by formation of holes. In the opposite case, dewetting sets in by formation of liquid ridges which break up further into droplets. In either case, ripening of the droplet structure takes place, and larger droplets grow at the expense of smaller ones.

Oron (2000c) studied the evolution of a film on a coated solid substrate as described by Equation (12.41) with the potential Equation (12.31d) given in dimensionless form as $\Phi = A^3H^{-3} - A_4H^{-4}$, where A_3 and A_4 are positive non-dimensional Hamaker constants. As noted previously, this potential acts as long-range van der Waals attraction and short-range repulsion, both apolar. The evolution of a small-amplitude disturbance of a flat initial state $H = 1$ leads to self-organization of the surface, emergence of holes, their expansion, coalescence, and formation of polygonal network of liquid ridges on top of the essentially flat microlayer. Later the liquid ridges break up into isolated drops and ridges that pump their liquid by means of the capillary forces into the largest drop making the latter bigger and more circular. The existence of a "thick" microlayer facilitates a relatively free liquid flow along the coated substrate and the accumulation of the liquid in an isolated drop standing on a plateau minimizing the free energy of the system. Finally, a steady state is reached, where a circular drop persists when standing on a flat equilibrium film, as seen in Figure 12.7. The film evolution described follows the typical sequence of events as described in the experiments by Khanna et al. (2000).

Reiter et al. (1999a) carried out theoretical and experimental studies of the dynamics of films on wettable solid surfaces and in contact with an ambient phase of varying physicochemical composition. By exchanging the ambient phase it is possible to vary the total Hamaker constant of the system and even to change its sign, thus turning the initially stable configuration into the unstable one. Experiments with PDMS films on a silicon wafer with alternating air and water ambient phases provide an example of such

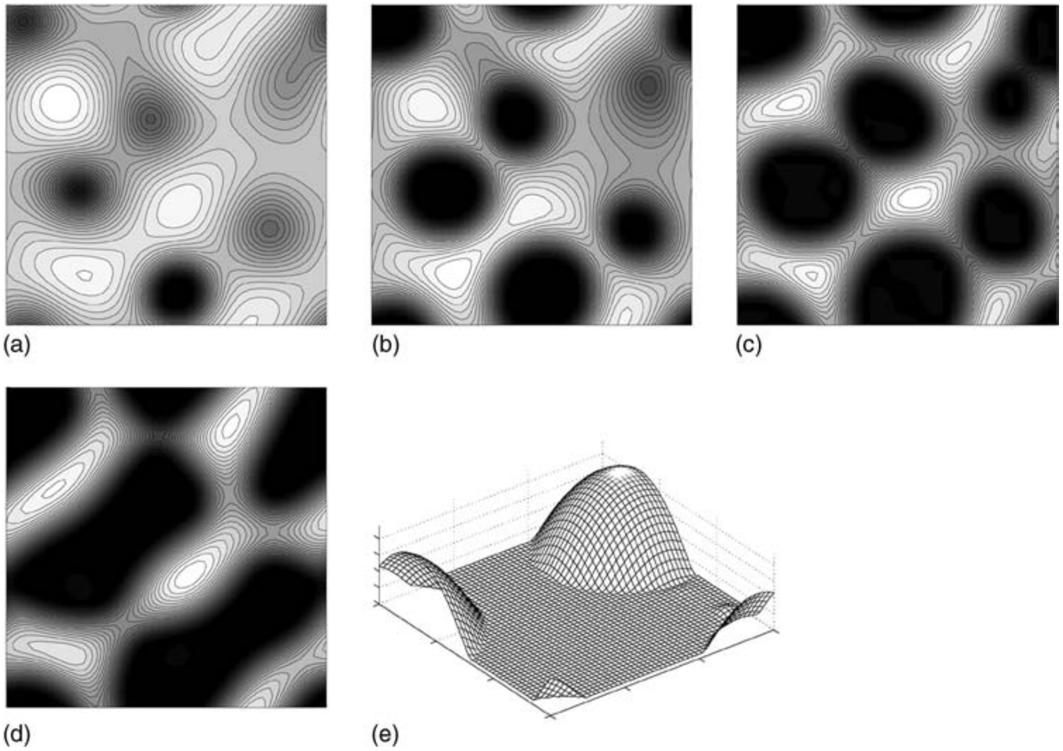


FIGURE 12.7 The stages of evolution of a non-evaporating film as described by Equation (12.41) with the potential Equation (12.31d). The first four consecutive snapshots are given in the form of a contour plot, while the last one is in the form of surface plot. Each image has its own brightness, so the film thickness in different images cannot be compared. A polygonal network of liquid ridges qualitatively similar to the experimental observations made by Sharma and Reiter (1996) is seen in the snapshots (b)–(d.) Bright and dark shades correspond to elevations and depressions, respectively. (Reprinted with permission from Oron (2000c).)

a system [Reiter et al., 1999a, b]. When in contact with air, the film remained flat and did not exhibit any evidence of instability. However, while in contact with water instability sets in, and the film, whose initial thickness ranged between 30 and 110 nanometers, finally reached the state in which small droplets stood on top of a thin wetting layer. This phenomenon was studied theoretically [Reiter et al., 1999a] using a three-dimensional evolution Equation (12.41) with the potential topologically similar to that of Equation (12.31b). Qualitative agreement between theory and experiments was quite good. However, as noted by Reiter et al. (1999a), even quantitative agreement between the two could be achieved but for "unexpectedly high effective Hamaker constant." The reason for that is still unclear.

12.3.2.2 Heterogeneous Substrates

A study of the dynamics of thin liquid films on a heterogeneous substrate can be motivated by the presence of dust particles or other impurities, oxidized or rough patches, or varying chemical composition leading to non-uniform wettability properties of the solid surface underlying the film. These and other types of heterogeneity of the substrate may be present unintentionally or created deliberately to achieve a certain goal.

The governing equation studied in this context is Equation (12.41). In contrast with the case of the homogeneous substrate where the potential of the intermolecular forces depends solely on the film thickness $\Phi = \Phi(H)$, in the current case the potential explicitly depends on the lateral spatial coordinates. This dependence enters the equations via spatial variation of the Hamaker coefficients.

A series of papers [Lenz and Lipowsky, 1998; Herminghaus et al., 1999, 2000; Gau et al., 1999; Lenz, 1999; Lipowsky et al., 2000] examined the morphological transitions of liquid layers on heterogeneous structured substrates. Lenz and Lipowsky (1998) showed by minimization of the total interfacial free energy that for a domain containing a hydrophilic patch confined between the hydrophobic ones, three

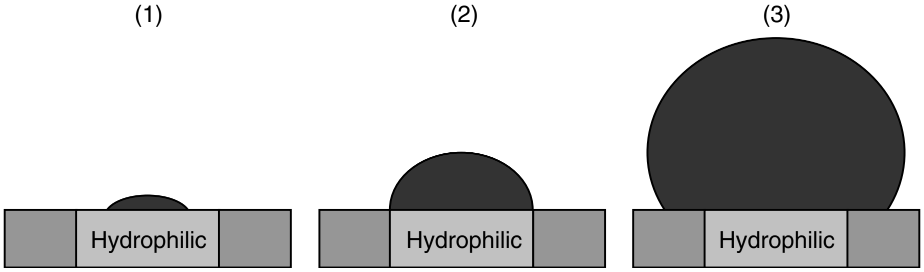


FIGURE 12.8 Equilibrium states of the droplets on a heterogeneous substrate that consists of alternating hydrophilic and hydrophobic patches. These equilibria depend on the droplet volume.

different regimes depending on the volume of the droplet are possible. Figure 12.8 demonstrates these regimes. In the regimes (1) and (3), the respective contact angles are prescribed a priori by the phases chosen and satisfy the Young equation. The regime (2) is characterized by the droplet volume and the contact angle spanning over the range between the respective values of regimes (1) and (3). In the limiting case of perfectly wettable hydrophilic and non-wettable hydrophobic patches, the regime (2) is only possible. In the case of a two-dimensional square periodic lattice of N circular hydrophilic patches surrounded by hydrophobic domains, the equilibrium state for a low total liquid volume consists of N identical droplets, all of them covering their own hydrophilic patch similar to regime (1) for the case of an isolated patch. As the total volume of the liquid increases, the droplets grow and the system undergoes transition to a heterogeneous equilibrium state that consists of one large drop and $N - 1$ small identical drops. More complex heterogeneous states were unstable [Gau et al., 1999]. If the total volume of the liquid increases beyond a certain value, a third equilibrium state that represents a single completely wetting layer covering the whole system becomes possible. The transition to this equilibrium state is possible from either of the aforementioned states. For striped periodic domains, all of the three equilibria states found in the previous case persist. However, a new kind of transition from the homogeneous state to the film state exists here. This transition consists of the stages where identical droplets span over several hydrophilic patches and the hydrophobic ones in between.

Gau et al. (1999) performed a series of experiments with liquid microchannels created by hydrophilic stripes of about 40 microns wide and further condensation of water onto the substrate. When the total amount of condensed water was low, the microchannels had a shape of cylindrical caps of a constant cross-section with a small contact angle θ between the liquid and the solid. However, when the total volume of water exceeded a certain value, the straight channels underwent instability, which led to the formation of a single bulge on each of the stripes. Moreover, when the bulges on two neighboring channels were in close proximity, they merged to form a big drop or microbridge between the channels. Gau et al. (1999) found theoretically that the cylindrical cap configuration with the contact angle θ is linearly stable for $\theta < 90°$ and unstable to long-wave disturbances for $\theta > 90°$, provided that the wavelength of the disturbance is sufficiently large

$$\lambda > \lambda_c = \left[\frac{\pi/2}{\theta^2 - (\pi/2)^2}\right]^{1/2} \frac{\theta}{\sin\theta} a_\gamma,$$

where a_γ is the width of the hydrophilic stripe. The presence of this instability disallows the emergence of long homogeneous liquid channels with a contact angle larger than 90°. The onset of the instability occurs at $\theta = 90°$, and the wavelength of the critical disturbance is infinite. This explains the formation of a single bulge on the microchannel [Herminghaus et al., 2000]. The precise shape of the configuration of liquid microchannels with bulges was numerically calculated by Gau et al. (1999) using minimization of the total free energy. A very good agreement was found between the experimental and theoretical results.

Konnur et al. (2000) and Kargupta et al. (2000) studied the three-dimensional dynamics of liquid crystal films using Equation (12.41) with the potential Equation (12.31e) with different sets of fixed positive values a_3, S^p, λ on the patches of the substrate. They reported a new mechanism of film instability associated with the substrate heterogeneity. This mechanism is driven by the pressure gradient generated by the spatial variation of ϕ and directed from the less to the more wettable domains on the solid. The potential Equation (12.31e) employed by Konnur et al. (2000) prescribes instability for both relatively thin and thick films, while films in the intermediate range are stable. They found that the presence of heterogeneity is able to destabilize even spinodally stable films, speed up the rupture process of the film, and produce spatially complex and locally ordered patterns. Destabilization of spinodally stable films arises even when the heterogeneous patch is much smaller than the spinodal length scale determined as the wavelength of the fastest growing linearly unstable disturbance. The true rupture can occur for spinodally stable films if the local thickness of the film is reduced by the heterogeneous mechanism to the value where the spinodal instability condition is met, and both of the mechanisms propel the film to rupture. The evolution of an initially flat film typically exhibits such morphological patterns as: a lack of surface deformations prior to the formation of a hole, emergence of a non-growing hole on a perfectly wetted substrate or in a spinodally stable film, formation of a "castle-moat" pattern with a central drop surrounded by a ring-like depression or hole, and formation of locally ordered structures with alternating depressions and rims [Konnur et al., 2000; Kargupta et al., 2000]. The heterogeneous mechanism was strong for relatively thick films, and its time scale was several orders of magnitude lower than that of the spinodal mechanism. Kargupta et al. (2000) considered also the two-dimensional dynamics of the film on a substrate with a heterogeneous patch of varying size. They found that the presence of heterogeneity always causes the emergence of local interfacial depression, which can evolve into film rupture when the length of the patch becomes sufficiently large. The rupture time rapidly decreases when the patch length increases beyond the critical length and becomes independent of the patch length when the latter is large. Kargupta et al. (2001) also considered drying of thin isothermal liquid films on heterogeneous substrates. They found that the rate of dewetting can be increased by evaporation, and the latter induces the formation of a large number of ring-like patterns containing satellite holes. Theoretical results of Kargupta and Sharma (2001) were recently confirmed experimentally when the pattern size is larger than the spinodal wavelength on a homogeneous surface [Sehgal et al., 2002]. Brusch et al. (2002) studied the process of dewetting two-dimensional films with the diffuse interface on a heterogeneous substrate with a sinusoidal modulation of the disjoining pressure via the investigation of possible steady states. Scenarios of the emergence of both pinning and coarsening patterns were discussed. They found that pinning is possible when the heterogeneity is of a larger periodicity than that of the critical dewetting mode. Large domains of coexistence of both types of patterns were also found. Patterning of thin liquid films by templating on heterogeneous substrates was investigated by Kargupta and Sharma (2002a, b, c), (2003); and Sharma et al. (2003).

12.3.2.3 Flow on a Rotating Disc

Reisfeld et al. (1991) considered the isothermal, axisymmetric flow of an incompressible viscous liquid on a horizontal rotating disk. Cylindrical polar coordinates r, θ, z are used in the frame of reference rotating with the disk. The film interface is located at $z = h(r, t)$. In the coordinate system chosen, outward unit normal vector **n** and unit tangent vector **t** are

$$\mathbf{n} = \frac{(-h_r, 0, 1)}{(1 + h_\mathrm{r}^2)^{1/2}}, \quad \mathbf{t} = \frac{(1, 0, h_r)}{(1 + h_\mathrm{r}^2)^{1/2}}.$$

The hydrodynamic equations analogous to Equation (12.2), taking into account both the centrifugal forces and Coriolis acceleration, are written in the vector form as

$$\nabla \cdot \mathrm{v} = 0, \quad \rho[\mathrm{v}_t + (\mathrm{v} \cdot \nabla)\mathrm{v}] = -\nabla p + \mu\nabla^2\mathrm{v} - \rho[\mathbf{g} + 2\omega \times \mathrm{v} + \omega \times \omega \times \mathrm{v}],$$

where ω is the angular-velocity vector with the components $(0, 0, \overline{\varpi})$. The boundary conditions are given by Equation (12.4) formulated in cylindrical polar coordinates with $\nabla_s\sigma = 0$ and $\Pi = 0$.

The characteristic length scale in the horizontal direction is chosen as the radius of the rotating disk $\overline{R}$ and the velocity scale is taken as $U_0 = \rho\overline{\varpi}^2\overline{R}d^2/\mu$. A small parameter ε is defined in accord with Equation (12.7) as $\varepsilon = d/\overline{R}$. The dimensionless parameters of the problem are the Reynolds number R as given in

Equation (12.12), the scaled inverse capillary number S given by Equations (12.12) and (12.15), and the Froude number F,

$$F = \frac{U_0}{(gd)^{1/2}}.$$

Using the procedures previously outlined, one obtains at leading order the following evolution equation

$$H_\tau + \frac{1}{3r}\left\{r^2H^3 + SrH^3\left[\frac{1}{r}(rH_r)_r\right]_r\right\}_r = 0. \tag{12.42}$$

The terms describing the effect of inertia and gravity appeared in the terms of first order in ε and thus were omitted. However, they may be retained to investigate the dynamics of the rotating film in the first phase of the process, including inertia and amplification of kinematic waves [Reisfeld et al., 1991]. Equation (12.42) models the combined effect of capillary forces and centrifugal drainage, neither of which describes any kind of instability.

For most spin coating applications, S is very small and the corresponding term may be neglected, although it may be very important in planarization studies where the leveling of liquid films on rough surfaces is investigated. Equation (12.42) can be thus simplified

$$H_\tau + \frac{1}{3r}(r^2H^3)_r = 0. \tag{12.43}$$

This simplified equation can then be used for further analysis. Looking for flat basic states $H = H(\tau)$, Equation (12.43) is reduced to the ordinary differential equation which is to be solved with the initial condition $H(0) = 1$. The film thins because of centrifugal drainage according to the solution

$$H(\tau) = \left(1 + \frac{4}{3}\tau\right)^{-1/2},$$

which predicts a decrease of the thickness to zero at the infinite time. The cases where inertia was taken into account were considered in [Reisfeld et al., 1991] where linear stability analysis of flat base states was given.

Stillwagon and Larson (1990) considered the spin coating process and leveling of a non-volatile liquid film over an axisymmetric, uneven solid substrate. For a given local dimensionless height of the substrate $\lambda(r)$, their equation derived from the Cartesian version valid for capillary leveling of a film in a trench resembles Equation (12.42) and reads

$$H_\tau + \frac{1}{3\hat{r}}[\alpha\,\hat{r}^2H^3 + \hat{S}\hat{r}H^3(H_{\xi\xi\xi} + \lambda_{\xi\xi\xi})]_\xi = 0, \tag{12.44}$$

where $\hat{r}$, ξ are, respectively, the radial coordinate and the radial distance from the trench, both scaled with the trench width. α is the ratio of the width and the location of the trench. Equation (12.44) can be further simplified under assumption that the width of the trench is small compared to its radial position and can be brought to the form

$$H_\tau + \frac{1}{3}[H^3 + \Omega^{-2}H^3(H_{\xi\xi\xi} + \lambda_{\xi\xi\xi})]_\xi = 0, \tag{12.45}$$

where Ω^2 is the ratio between the centrifugal and capillary forces. Stillwagon and Larson (1990) calculated quasi-steady-state solutions close to the trench solving the time-independent version for Equation (12.44)

$$H^3(H_{\xi\xi\xi} + \lambda_{\xi\xi\xi}) + \Omega^2H^3 = \Omega^2, \tag{12.46}$$

where the right-hand-side term arises from the condition of uniformity of the film far from the trench. Experiments with liquid films reported in Stillwagon and Larson (1990) demonstrate quantitative agreement between measured film profiles and those obtained from Equation (12.46).

Wu et al. (1999) and Wu and Chou (1999) used Equation (12.46) to study the degree of planarization for periodic uneven substrates expressed as the ratio between the amplitude of the deformed film interface and the average thickness of the film. They showed that this value is independent of Ω and slightly varies with the trench spacing for large Ω. This value decreases with the increase of spacing for small fixed values of Ω.

Chou and Wu (2000) studied the effect of air shear on the process of film planarization. Similar to the case considered in Equation (2.31) in Section IIC of Oron et al. (1997), where the term proportional to the imposed shear stress multiplied by hh_x arises in the evolution equation, air shear produces the advective term proportional to H^2, which has to be added to the expressions in the square brackets of the left-hand side of Equations (12.44) and (12.45). Corresponding additional terms will appear in Equation (12.46). Chou and Wu (2000) studied such an extended Equation (12.46) and found that the shear stress enhances the amplitude of the film interface, and thus opposes film planarization during spin coating for both isolated and periodic features of the substrate.

Peurrung and Graves (1993) considered three-dimensional quasi-steady states in spin coating over topography using the natural extension of Equation (12.45) into three dimensions. Their theoretical and experimental results agree qualitatively, both showing the emergence of wake-like structures at the downstream side of the protrusion with crests extending along each of the corners and the depression near the center.

12.4 Thermal Effects

One of the best known fluid flows under the influence of heat transfer is the buoyancy or Rayleigh convection [Chandrasekhar, 1961] of a stagnant liquid layer lying on a horizontal solid surface triggered by heating from below and a subsequent establishing of unstable density stratification. This convection sets in when the temperature difference across the layer exceeds a certain critical value, which is proportional among other physical parameters of the system to the third power of the layer thickness d. Due to the fact that the range of very small values of the film thickness is of a major interest in the context of MEMS, the Rayleigh effect is much weaker than the thermocapillary or Marangoni effect addressed next. The latter scales with the first power of d in contrast with d^3 in the case of the Rayleigh effect.

12.4.1 Thermocapillarity, Surface Tension, and Gravity

The thermocapillary or Marangoni effect (e.g., see [Davis, 1987] accounts for the emergence of interfacial shear stresses because of the variation of surface tension with temperature ϑ, $\sigma = \sigma(\vartheta)$, which is, in most cases, monotonically decreasing. Such a shear stress is mathematically expressed by $\nabla_s \sigma$ [Edwards et al., 1991]. In order to incorporate the thermocapillary effect into the equations, one needs to add an energy equation and the appropriate boundary conditions related to heat transfer to the governing system Equations (12.2)–(12.4).

The energy equation in three dimensions and the boundary conditions have the form

$$\rho c(\vartheta_t + u\vartheta_x + v\vartheta_y + w\vartheta_z) = k_{th}(\vartheta_{xx} + \vartheta_{yy} + \vartheta_{zz}) + \dot{q}, \tag{12.47}$$

$$\vartheta = \vartheta_0 \quad \text{at } z = 0, \tag{12.48a}$$

$$k_{th}\mathbf{n} \cdot \nabla\vartheta + \alpha_{th}(\vartheta - \vartheta_\infty) = 0 \quad \text{at } z = h(x, y, t) \tag{12.48b}$$

Here c is the specific heat of the fluid, k_{th} is its thermal conductivity, ϑ_0 is the temperature of the rigid substrate assumed to be uniform, and $\dot{q}$ is the rate of internal heat generation. The boundary condition Equation (12.48b) is Newton`s cooling law, and α_{th} is the heat-transfer coefficient describing the rate of heat transfer from the liquid to the ambient gas phase held at the constant temperature ϑ_∞.

Turning to the two-dimensional case, scaling the temperature by

$$\Theta = \frac{\vartheta - \vartheta_\infty}{\vartheta_0 - \vartheta_\infty} \tag{12.49}$$

and substituting scales Equation (12.8) into Equations (12.47) and (12.48) yields

$$\varepsilon RP(\Theta_\tau + U\Theta_\xi + W\Theta_\varsigma) = \varepsilon^2\Theta_{\xi\xi} + \Theta_{\xi\xi} + 2Qf(\varsigma), \tag{12.50}$$

$$\Theta = 1 \quad \text{at } \varsigma = 0, \tag{12.51a}$$

$$\Theta_\varsigma - \varepsilon^2\Theta_\xi \mathrm{H}_\xi + B\Theta(1 + \varepsilon^2 H_\xi^2)^{1/2} = 0 \quad \text{at } \varsigma = H, \tag{12.51b}$$

where P and B are, respectively, the Prandtl and Biot numbers, Q is the dimensionless measure of the rate of internal energy generation defined by

$$P = \frac{\rho c v}{k_{th}}, \quad B = \frac{\alpha_{th} d}{k_{th}}, \quad Q = \frac{\dot{q} d^2}{2k_{th}\vartheta_r}, \tag{12.52}$$

where ϑ_r is the reference temperature chosen as $\vartheta_r = \vartheta_0 - \vartheta_\infty$ if $\vartheta_0 > \vartheta_\infty$ and as $\vartheta_r = \vartheta_0$ if $\vartheta_0 = \vartheta_\infty$. Furthermore, $f(\varsigma)$ expresses the dependence of the rate of internal energy generation on the vertical coordinate ς.

Begin first with the case of no internal heat generation $\dot{q} = 0$ leading to $Q = 0$. Expand the temperature Θ in a perturbation series in ε along with the expansions Equations (12.14), and substitute these into the system given by Equations (12.50) and (12.51). Assume again that $R = O(1)$ and let $P, B = O(1)$, so that the convective terms in Equation (12.50) are delayed to next order, that is, declaring that conduction in the liquid is dominant, and the conductive heat flux at the interface balances the heat loss to the environment.

At leading order in ε the governing system for $\Theta^{(0)}$ consists of condition Equation (12.51a),

$$\Theta_{\varsigma\varsigma} = 0, \tag{12.53}$$

and

$$\Theta_\varsigma + B\Theta = 0 \quad \text{at } \varsigma = H, \tag{12.54}$$

where the superscript "zero" has been dropped. The solution to this system is

$$\Theta = 1 - \frac{B\varsigma}{1 + BH} \quad \text{and} \quad \Theta_i = \frac{1}{1 + BH}, \tag{12.55}$$

where $\Theta_i = \Theta(\tau, \xi)$ is the surface temperature in order to substitute it into Equation (12.21).

It is now required to determine the thermocapillary stress Σ_ξ. By the chain rule

$$\Sigma_\xi = M\left(\frac{d\Sigma}{d\Theta}\right)(\Theta_\xi + H_\xi\Theta_\varsigma) \equiv -M\frac{\gamma(H)H_\xi}{(1 + BH)^2}, \tag{12.56a}$$

where

$$\begin{gathered}\gamma(H) = -(d\Sigma/d\Theta)_{\Theta=\Theta_i}, \\ M = \frac{\Delta\sigma}{\mu U_0}\end{gathered} \tag{12.56b}$$

is the Marangoni number, and the sign change is inserted because $d\Sigma/d\Theta$ is negative for most common materials. Here $\Delta\sigma$ is the change of surface tension over the temperature domain between the characteristic temperatures, usually ϑ_∞ and ϑ_0. To be more precise, if $\vartheta_\infty < \vartheta_0$ (heating at the bottom of the layer), then $\Delta\sigma > 0$ for standard fluid pairs with surface tension decreasing with temperature. For heating at the interface side, $\vartheta_\infty > \vartheta_0$ and $\Delta\sigma < 0$. The shear stress condition, Equation (12.18b), has at leading order in ε the form

$$U_\varsigma + M\frac{\gamma(H)H_\xi}{(1+BH)^2} = 0 \quad \text{at } \varsigma = H. \tag{12.57}$$

Thus, in this case $\hat{\Pi}_1 = \hat{\Pi}_3 = \Phi = 0$, Equation (12.21) becomes

$$H_\tau + \frac{1}{2}MB\left[\frac{H^2\gamma(H)H_\xi}{(1+BH)^2}\right]_\xi + \frac{1}{3}S(H^3H_{\xi\xi\xi})_\xi = 0. \tag{12.58}$$

If gravity forces are to be included, $\Phi = g\varsigma$, Equation (12.58) becomes

$$H_\tau + \left\{\left[\frac{1}{2}MB\frac{H^2\gamma(H)}{(1+BH)^2} - \frac{1}{3}GH^3\right]H_\xi\right\}_\xi + \frac{1}{3}S(H^3H_{\xi\xi\xi})_\xi = 0. \tag{12.59}$$

For the most ubiquitous case in which surface tension is a linearly-decreasing function of temperature $\sigma = \sigma(\vartheta)$, the value $(d\Sigma/d\Theta) = const$ and $\gamma(H) = 1$. Equations (12.58) and (12.59) with $\gamma(H) = 1$ appeared in Davis (1983) for $B \ll 1$ and in Kopbosynov and Pukhnachev (1986); Bankoff and Davis (1987); Burelbach et al. (1988); Oron and Rosenau (1992); Deissler and Oron (1992); VanHook et al. (1995); and Oron (2000b).

For $B \ll 1$, Equation (12.59) in dimensional form becomes

$$\mu h_t + \frac{\alpha_{th}\Delta\sigma}{2k_{th}}(h^2h_x)_x - \frac{1}{3}\rho g(h^3h_x)_x + \frac{1}{3}\sigma(h^3h_{xxx})_x = 0. \tag{12.60}$$

Linearization of Equation (12.60) around the state $h = d$ yields the characteristic equation

$$\mu\omega = \left(\frac{\alpha_{th}\Delta\sigma}{2k_{th}} - \frac{1}{3}\rho g d - \frac{\sigma}{3d}a^2\right)a^2. \tag{12.61}$$

Equation (12.61) shows that if $g > 0$ (gravity acting towards the base of the film), gravity has a stabilizing effect (similar to that described in the section on isothermal films), while thermocapillarity has a destabilizing effect on the interface. Equation (12.61) shows that the gravitational stabilization is enhanced with the thickness of the film. The dimensionless cutoff wavenumber a_c is given in this case by

$$a_c = \left(\frac{3}{2}B\frac{\Delta\sigma}{\sigma} - Bo\right)^{1/2}. \tag{12.62}$$

Thermocapillary destabilization of a film can be explained by examining the behavior of an initially deformed interface in the linear temperature field produced by the heat transfer at the interface. The depression lies in the region of higher temperature than its neighbors. Therefore, if surface tension is a decreasing function of temperature, interfacial stresses proportional to the surface gradient of the surface tension [e.g., Levich, 1962; Landau and Lifshitz, 1987] drive the interfacial liquid away from it. Because the liquid is viscous, it is dragged away from the depression causing it to deepen further. Hydrostatic and capillary forces cannot prevent this deepening, and the film proceeds to zero thickness (ruptures) at some location.

Studies of Equation (12.58) with $\gamma(H) = 1$ [Oron and Rosenau, 1992] reveal that evolution of a small-amplitude initial data usually results in rupture of the film qualitatively similar to that displayed in Figure 11 of Oron and Rosenau (1992). The three-dimensional version of Equation (12.58) was studied by VanHook et al. (1995), and the results were tested against their experiments. The existence of a ruptured region predicted by Equation (12.58) was qualitatively confirmed by the experiment. However, the theoretical predictions of the instability threshold were about 50% higher than the experimental data. Becerril et al. (1998) recently addressed this discrepancy in terms of side-walls effects and deflected initial interface shapes.

VanHook et al. (1997) developed a "two-layer" theory modeling the dynamics of systems containing superposed layers of a liquid and a passive gas confined between two horizontal rigid differentially heated surfaces. This approach takes into consideration the change in the temperature profile in the air due to deformation of the interface. The two-layer setting leads to the thermal problem containing Equation (12.53) formulated in each layer with the boundary conditions of temperature and heat flux continuity at the liquid–gas interface. Two new parameters arise from the solution of this thermal problem

$$\eta = \frac{k_{th,g}d}{k_{th}d_g}, \quad F = \frac{(d/d_g) - \eta}{1 + \eta}, \tag{12.63}$$

where d_g is the thickness of the gas layer and $k_{th,g}$ is its thermal conductivity. The parameter η replaces the Biot number in the "one-layer" model described above.

In the case at hand the dimensionless interfacial temperature is

$$\Theta_i = 1 - \frac{H}{1 + F - FH}, \tag{12.64}$$

and the corresponding dimensionless evolution equation in the standard case of $\gamma(H) = 1$ in two dimensions is obtained in the form [VanHook et al., 1997]

$$H_\tau + \left\{\left[\frac{1}{2}M(1 + F)\frac{H^2}{(1 + F - FH)^2} - \frac{1}{3}GH^3\right]H_\xi\right\}_\xi + \frac{1}{3}S(H^3H_{\xi\xi\xi})_\xi = 0. \tag{12.65}$$

Equation (12.65) reduces to the "one-layer" model Equation (12.59) when $F = -\eta/(1 + \eta)$ that corresponds to $d/d_g \to 0$. For one-layer systems the parameter F is always non-positive, while for two-layer systems F is usually positive. For both of these cases $F > -1$. The solutions of Equation (12.65) were of two distinct types, namely "dry spots" that represent rupture at the bottom solid surface, see Figure 12.4(a), and "high spots" that represent rupture at the top solid surface by the elevated film interface, see Figure 12.4(b). Dry spots emerge for $F < 1/2$, while high spots form when $F > 1/2$. The transition between these two different kinds of solutions, which depend on the value of the Bond number Bo and the initial condition, occurs in the vicinity of $F = 1/2$ [VanHook et al., 1997]. As in the "one-layer" theory reviewed earlier, steady non-ruptured states of the system were not found. The experimental results of VanHook et al. (1997) qualitatively agree for certain liquid depths with a "two-layer" model.

In the case of "negative gravity," $g < 0$, that is, when the film is on the underside of the solid plane, the Rayleigh–Taylor instability (heavy fluid overlying light fluid) enhances the thermocapillary instability and broadens the band of linearly unstable modes:

$$a_c = \left(\frac{3}{2}B\frac{\Delta\sigma}{\sigma} + Bo\right)^{1/2}. \tag{12.66}$$

Stabilization of the Rayleigh–Taylor instability by thermocapillarity was investigated [Oron and Rosenau, 1992; Deissler and Oron, 1992] for two- and three-dimensional cases, respectively. They found that negative thermocapillarity, that is, with $\Delta\sigma < 0$, corresponding to heating at the interface side or cooling at the rigid bottom, in conjunction with surface tension can lead to saturation of the Rayleigh–Taylor instability and to formation of steady drops. The experimental confirmation of such saturation was recently obtained by Burgess et al. (2001).

We now turn to the case where the $\dot{q}$-term is present in Equation (12.47). Its presence stands for the effect of internal energy generation, which might be induced by irradiation of the film and further absorption of the radiation energy within the non-scattering liquid phase. In this context the solid substrate is assumed to be black, that is, absorbing all radiation penetrating through the liquid film. Oron and Peles (1998) considered the simplified case of spatially uniform energy absorption, $f(\varsigma) \equiv 1$. In this situation the solution of the thermal problem Equations (12.50) and (12.51) is given by

$$\Theta = \Theta_0\left(1 - \frac{B\varsigma}{1 + BH}\right) + Q\left(-\varsigma^2 + \frac{BH^2 + 2H}{1 + BH}\varsigma\right) \quad \text{and} \quad \Theta_i = \frac{B}{1 + BH}(\Theta_0 + QH^2), \tag{12.67}$$

where $\Theta_0 = 1$ if $\vartheta_0 = \vartheta_\infty$, and $\Theta_0 = 0$ if $\vartheta_0 = \vartheta_\infty$. The corresponding evolution equation of the film obtained upon calculation of the corresponding value of Σ_ξ using Equation (12.67) and its substitution into Equation (12.21) with gravity neglected and $\gamma(H) = 1$ is

$$H_\tau + \frac{1}{2}M\left[H^2 \frac{B\Theta_0 - Q(BH^2 + 2H)}{(1 + BH)^2} H_\xi\right]_\xi + \frac{1}{3}S(H^3H_{\xi\xi\xi})_\xi = 0. \qquad (12.68)$$

The main result found by Oron and Peles (1998) was that internal heat generation stabilizes the interface via the thermocapillary effect associated with it. In the simplest case where the temperature of the solid is equal to the saturation temperature $\vartheta_0 = \vartheta_\infty$ (i.e., $\Theta_0 = 0$). This follows directly from the fact that the interfacial temperature Θ_i is an increasing function of the film thickness H. The effect of stabilization becomes apparent because at leading order the heat transfer in a thin liquid film is one-dimensional across it, and the energy input from absorption of radiation energy in the thicker part of the film is greater than in its thinner part. Thus, the interfacial temperature at the depression is lower than at the crest of the interface, and the thermocapillary stress drives the liquid into the depression promoting stabilization of the interface. All this is different from the standard case discussed in the literature where internal energy generation is absent. In the latter case of $Q = 0$ and $\Theta_w = 1$, the interfacial temperature decreases with H, and instability of the spatially uniform state of the interface is thus triggered. When the internal heat generation is sufficiently large, $Q \geqslant 1$, the film becomes unconditionally stable. When the film is linearly unstable, the range of unstable modes narrows with the increase of Q for a fixed value of B [Oron and Peles, 1998].

Oron (2000a) considered thin liquid films with an optically smooth non-reflective deformed interface irradiated with monochromatic beam of a specified wavelength λ. The intensity of radiation i_λ of such a beam normally impinging on the optically smooth non-reflective interface was shown in the absence of emission by the irradiated liquid phase to decay exponentially with the distance from the flat liquid surface [Siegel and Howell, 1992]

$$i_\lambda(z) = i_\lambda(z_0)e^{-K_\lambda(z_0 - z)}, \qquad (12.69)$$

where z_0 is the location of the film surface and K_λ is the extinction coefficient of the given liquid assumed to be constant. The extinction coefficient is a property of the medium and in general varies with its temperature, pressure, and the wavelength of the incident radiation.

Equation (12.69) is often referred as to Bouguer's [Siegel and Howell, 1992] or Beer`s law. The attenuation of the radiation intensity is associated with the absorption and scattering of energy. The extinction coefficient is in general represented as a sum of the absorption and scattering components $K_\lambda = \mathrm{a}_\lambda + \mathrm{a}_{s,\lambda}$. In the case of a vanishing scattering $a_{s,\lambda} = 0$, $K_\lambda = a_\lambda$, and the optical thickness κ_λ of a liquid film of a uniform thickness d can be defined as

$$\kappa_\lambda \equiv a_\lambda d = d/L_m,$$

where L_m is the mean penetration length of the incident radiation by the wave of the wavelength λ. Assuming that the film is non-scattering, the solid surface underneath is non-reflecting and the intensity of absorbed radiation is equal to the intensity of internal heat sources, the latter is expressed [Oron, 2004a] by

$$\dot{q}(z) = \dot{q}\exp[-a_\lambda(h - z)],$$

where h represents the location of the interface and $\dot{q} = i_\lambda(z_0)a_\lambda$ is the constant representing the rate of energy absorption at the film interface. Note that the intensity of the heat sources $\dot{q}(z)$ varies also with x, y, t when $a_\lambda \neq 0$. By comparing the value of the optical thickness of the film with unity, one can examine the following limiting cases: (a) if $\kappa_\lambda \ll 1$, the radiation passes through the film, and such a film is called optically thin or transparent; (b) if $\kappa_\lambda \gg 1$, the radiation penetrates only into a very thin boundary layer adjacent to the film interface, and in this case the film is called optically thick or opaque.

Solving the thermal problem given by Equations (12.50) and (12.51) with $f(\varsigma) = \exp[-\beta\,(H - \varsigma)]$, constant a_λ, and $\beta = a_\lambda\, d$ a non-dimensional attenuation coefficient, yields the interfacial temperature in the form

$$\Theta_i = \frac{B}{1 + BH}\left[\Theta_0 + \frac{2Q}{\beta^2}(e^{-\beta H} + \beta H - 1)\right]. \tag{12.70}$$

The corresponding evolution equation of the film obtained upon calculation of the term Σ_ξ based on Equation (12.70) and its substitution into Equation (12.21) with gravity neglected and $\gamma(H) = 1$ is

$$H_\tau + \frac{M}{2}\left\{\frac{H^2}{(1 + BH)^2}\left[B\Theta_0 - \frac{2Q}{\beta^2}((\beta + BH)(1 - e^{-\beta H}) - \beta BHe^{-\beta H})\right]H_\xi\right\}_\xi + \frac{S}{3}(H^3\, H_{\xi\xi\xi})_\xi = 0. \tag{12.71}$$

As in the case of uniform heat generation, Oron (2000a) found for irradiated films following the Bouguer's law that an increase of the radiation intensity leads to stabilization of the interface because of the appropriate change in the profile of the interfacial temperature. In the presence of heating across the film, $\Theta_0 = 1$, there exists a critical value of $Q = Q_c$ depending on the Biot number B, such that the film becomes linearly stable when $Q > Q_c$, and remains linearly unstable when $Q < Q_c$, albeit the rate of the disturbance growth slows down in comparison with the case of $Q = O$. This critical value Q_c is obtained from the linear stability analysis as

$$Q_c = \frac{\beta^2}{2}[(1 + \beta B^{-1})(1 - e^{-\beta}) - \beta e^{-\beta}]^{-1}. \tag{12.72}$$

This critical value tends to its limiting value of $Q_c = B/(2 + B)$ for optically thin films $\beta \ll 1$, which is the limit of the spatially uniform absorption, and to $Q_c = \beta B/2$ for optically thick films, $\beta \gg 1$.

It was experimentally discovered that dilute aqueous solutions of long-chain alcohols exhibit non-monotonic dependence of surface tension on temperature [Legros et al., 1984; Legros, 1986]. This dependence can be approximated quite well by the quadratic polynomial

$$\sigma(\vartheta) = \delta(\vartheta - \vartheta_m)^2, \tag{12.73a}$$

where δ is constant and ϑ_m is the temperature corresponding to the minimal surface tension. In this case,

$$\gamma(H) \propto \left(\frac{\vartheta_m - \vartheta_\infty}{\vartheta_0 - \vartheta_\infty} - \frac{1}{1 + BH}\right). \tag{12.73b}$$

The instability (called QM instability) arising from the variation of surface tension given by Equation (12.73a) was studied by Oron and Rosenau (1994). In contrast with the case of the standard thermocapillary instability described by Equation (12.58) with $\gamma(H) = 1$, evolution of QM instability may result in a non-ruptured steady state. Figure 4 in Oron and Rosenau (1994) displays such a state along with the streamlines of the flow field obtained from solving Equation (12.58) with $\gamma(H)$ given by Equation (12.73b). The intersections of the Θ_0-line with the film interface in Figure 4 of Oron and Rosenau (1994) correspond to the locations of the minimal surface tension. The existence of these creates surface shear stresses acting in opposite directions, as shown by the arrows on the graph, and leads to film stabilization.

12.4.2 Liquid Film on a Thick Substrate

The methods described in the previous sections can be easily implemented in the case of a liquid film lying on top of a solid slab of thickness that is small compared to the characteristic wavelength of the interfacial disturbance [Oron et al., 1996]. In this case the thermal conduction equation in the solid has to be solved simultaneously with the energy equation in the liquid. This coupled thermal problem is written at leading order in ε as

$$\Theta_{w,\varsigma\varsigma} = 0, \quad -\frac{d_w}{d} \leqslant \varsigma \leqslant 0, \tag{12.74}$$
$$\Theta_{\varsigma\varsigma} = 0, \quad 0 \leqslant \varsigma \leqslant H$$

with the boundary conditions

$$\Theta_w = \Theta, \quad -k_{th,w}\Theta_{w,\varsigma} = -k_{th}\Theta_{\varsigma} \quad \text{at } \varsigma = 0,$$
$$\Theta_w = 1 \quad \text{at } \varsigma = -\frac{d_w}{d}, \tag{12.75}$$

where Θ_{w} and d_w/d are the dimensionless temperature the thickness of the solid slab scaled with d, and $k_{th,w}$ is its thermal conductivity. The upper equations in Equation (12.75) express the conditions of continuity of both the temperature and heat flux at the solid–liquid boundary, while the lower equation in Equation (12.75) prescribes a uniform temperature at the bottom of the solid substrate. The last boundary condition is taken at the film interface and at leading order in ε, it is given by Equation (12.54). Appropriate extension has to be made in the case of a volatile liquid (see the section on phase changes).

Solution of Equations (12.74) and (12.75) results in

$$\Theta = 1 - \frac{B(\bar{\kappa}\varsigma + d_w/d)}{\bar{\kappa}(1 + BH) + Bd_w/d}, \quad \Theta_w = 1 - \frac{B(\varsigma + d_w/d)}{\bar{\kappa}(1 + BH) + Bd_w/d} \tag{12.76a}$$

with $\bar{\kappa} = \kappa_{th,w}/k_{th}$, which implies the interfacial temperature in the form

$$\Theta_i = \left[1 + B\left(H + \frac{d_w/k_{th,w}}{d/k_{th}}\right)\right]^{-1}. \tag{12.76b}$$

Comparing the expressions for the interfacial temperatures Θ_i, as given by Equations (12.55) and (12.76b), in addition to the thermal resistance due to each conduction and convection at the interface in the former case, the latter contains a thermal resistance owing to conduction in the solid. The evolution equation, analogous to Equation (12.58), will have the same form except for the change in the denominator of the second term containing an additional additive term

$$a \equiv \frac{d_w/k_{th,w}}{d/k_{th}} \tag{12.77}$$

This additional term represents the ratio between the values of the thermal conductive resistance of the solid and the liquid.

Using Equations (12.76) one can derive the expressions for the temperatures along the gas–liquid (GL) and solid–liquid (SL) interfaces: $\Theta_{GL} \equiv \Theta(\varsigma = H)$ and $\Theta_{SL} \equiv \Theta_w(\varsigma = 0)$. When the film ruptures (i.e., $H = 0$), the values for Θ_{GL} and Θ_{SL} are equal if

$$a = \frac{d_w}{\bar{\kappa}d} \neq 0 \tag{12.78}$$

However, the temperature singularity $\Theta_{GL} \neq \Theta_{\mathrm{SL}}$ emerges at the rupture point when $a = 0$. Equation (12.78) is the sufficient condition to be satisfied in order to relieve this singularity [Oron et al., 1996]. If it is satisfied

$$\lim_{H\to 0}\lim_{a\to 0}\Theta_{GL} = \lim_{H\to 0}\lim_{a\to 0}\Theta_{SL} = 0,$$

and the singularity is removed. The problematic case of $a \to 0$ materializes if the substrate is of a negligible thickness.

12.5 Change of Phase: Evaporation and Condensation

12.5.1 Interfacial Conditions

We now consider the case of an evaporating (condensing) thin film of a simple liquid lying on a heated (cooled) plane surface held at constant temperature ϑ_0 which is higher (lower) than the saturation temperature at the given vapor pressure. It is assumed that the speed of vapor particles is sufficiently low, so that the vapor can be considered an incompressible fluid.

The boundary conditions appropriate for phase transformation at the film interface $z = h$ are now formulated. The mass conservation equation at the interface is given by the balance between the liquid and vapor fluxes through the interface

$$j = \rho_\upsilon(\mathbf{v}_\upsilon - \mathbf{v}_i) \cdot \mathbf{n} = \rho_f(\mathbf{v}_f - \mathbf{v}_i) \cdot \mathbf{n}, \tag{12.79a}$$

where j is the mass flux due to evaporation; ρ_υ and ρ_f are, respectively, the densities of the vapor and the liquid; $\mathbf{v}_\upsilon$ and $\mathbf{v}_f$ are the vapor and liquid velocities at $z = h$; and $\mathbf{v}_i$ is the velocity of the interface. Equation (12.79a) provides the relationship between the normal components of the vapor and liquid velocities at the interface. The tangential components of both of the velocity fields are equal at the interface:

$$(\mathbf{v}_f - \mathbf{v}_\upsilon) \cdot \mathbf{t}_m = 0, \quad m = 1, 2. \tag{12.79b}$$

The boundary condition that expresses the stress balance and extends Equation (12.4b) to the case of phase transformation reads [Delhaye, 1974; Burelbach et al., 1988]

$$j(\mathbf{v}_f - \mathbf{v}_\upsilon) - (\mathbf{T} - \mathbf{T}_\upsilon) \cdot \mathbf{n} = 2\widetilde{\mathrm{H}}\sigma(\vartheta)\mathbf{n} - \nabla_s\sigma, \tag{12.80a}$$

where $\mathbf{T}_\upsilon$ is the stress tensor in the vapor phase and temperature dependence of surface tension is accounted for.

The energy balance at $z = h$ is given by [Delhaye, 1974; Burelbach et al., 1988]

$$j\left(L + \frac{1}{2}\upsilon^2_{\upsilon,n} - \frac{1}{2}\upsilon^2_{f,n}\right) + (k_{th}\nabla\vartheta - k_{th,\upsilon}\nabla\vartheta_\upsilon) \cdot \mathbf{n} + 2\mu(\mathbf{e}_f \cdot \mathbf{n}) \cdot \mathbf{v}_{f,r} - 2\mu_\upsilon(\mathbf{e}_\upsilon \cdot \mathbf{n}) \cdot \mathbf{v}_{\upsilon,r} = 0, \tag{12.80b}$$

where L is the latent heat of vaporization per unit mass; $k_{th,\upsilon}$, μ_υ, ϑ_υ are, respectively, the thermal conductivity, viscosity, and the temperature of the vapor; $\mathbf{v}_{\upsilon,r} = \mathbf{v}_\upsilon - \mathbf{v}_i$, $\mathbf{v}_{f,r} = \mathbf{v}_f - \mathbf{v}_i$ are the vapor and liquid velocities relative to the interface, respectively; $\upsilon_{\upsilon,n} = \mathbf{v}_{\upsilon,r} \cdot \mathbf{n}$, $\mathbf{v}_{f,n} = \mathbf{v}_{f,r} \cdot \mathbf{n}$ are the normal components of the latter; and $\mathbf{e}_f$, $\mathbf{e}_\upsilon$ are the rate-of-deformation tensors in the liquid and the vapor, respectively. In Equation (12.80b) the first term represents the contribution of the latent heat, the combination of the second and the third terms represents the interfacial jump in the momentum flux, the combination of the fourth and the fifth terms represents the jump in the conductive heat flux at both sides of the interface, while the combination of the last two terms is associated with the viscous dissipation of energy at both sides of the interface.

Since $\rho_\upsilon/\rho_f \ll 1$, typically of order 10^{-3}, it follows from Equation (12.79a) that the magnitude of the normal velocity of the vapor relative to the interface is much greater than that of the liquid. Hence, the phase transformation causes large accelerations of the vapor at the interface where the back reaction, called the vapor recoil, represents a force exerted on the interface. During evaporation (condensation) the troughs of the deformed interface are closer to the hot (cold) plate than the crests, so they have greater evaporation (condensation) rates j. The dynamic pressure at the vapor side of the interface is much larger than that at the liquid side,

$$\rho_\upsilon\upsilon^2_{\upsilon,n} = \frac{j^2}{\rho_\upsilon} \gg \rho_f\upsilon^2_{f,n} = \frac{j^2}{\rho_f}. \tag{12.81}$$

Momentum fluxes are thus greater in the troughs than at the crests of surface waves. Vapor recoil is a destabilizing factor for the interface dynamics for both evaporation ($j > 0$) and condensation ($j < 0$) [Burelbach et al., 1988]. Scaled with j^2, see Equation (12.84), the vapor recoil is only important for applications where very high mass fluxes are involved.

Vapor recoil generally exerts a reactive downward pressure on a horizontal evaporating film. Bankoff (1961) introduced the effect of vapor recoil in the analysis of the film boiling. In this analysis the liquid overlays the vapor layer generated by boiling and leads to the Rayleigh–Taylor instability of an evaporating liquid–vapor interface above a hot horizontal wall. In this case the vapor recoil stabilizes the film boiling because the reactive force is greater for the wave crests approaching the wall than for the troughs.

To obtain a closure for the system of governing equations and boundary conditions, an equation relating the dependence of the interfacial temperature ϑ_i and the local pressure in the vapor phase is added [Plesset and Prosperetti, 1976; Palmer, 1976; Sadhal and Plesset, 1979]. Its linearized form is

$$\widetilde{K}j = \vartheta_i - \vartheta_s \equiv \Delta\vartheta_i, \tag{12.82}$$

where

$$\widetilde{K} = \frac{\vartheta_s^{3/2}}{\hat{\alpha}\rho_v L}\left(\frac{2\pi R_v}{M_w}\right)^{1/2},$$

ϑ_s is the absolute saturation temperature, $\hat{\alpha}$ is the accommodation coefficient, R_v is the universal gas constant, and M_w is the molecular weight of the vapor [Palmer, 1976; Plesset and Prosperetti, 1976; Burelbach et al., 1988]. Note that the absolute saturation temperature ϑ_s serves now as the reference temperature instead of ϑ_∞ in the normalization, Equation (12.49). When $\Delta\vartheta_i = 0$, the phases are in thermal equilibrium with each other, and in order for net mass transport to take place, a vapor pressure driving force must exist, given for ideal gases by kinetic theory [Schrage, 1953]. The latter is represented in the linear approximation by the parameter $\widetilde{K}$ [Burelbach et al., 1988]. Departure from ideal behavior is addressed in the parameter $\widetilde{K}$ by the presence of an accommodation coefficient $\hat{\alpha}$ depending on interface/molecule orientation and steric effects which represents the probability of a vapor molecule sticking upon hitting the liquid–vapor interface.

The set of the boundary conditions Equations (12.80) can be simplified to what is known as a "one-sided" model for evaporation or condensation [Burelbach et al., 1988] in which the dynamics of the liquid are decoupled from those of the vapor. This simplification is possible because of the assumption of smallness of density, viscosity, and thermal conductivity of the vapor with respect to the respective properties of the liquid. The vapor dynamics are ignored in the one-sided model, and only the mass conservation and the effect of vapor recoil stand for the presence of the vapor phase.

The energy balance Equation (12.80b) becomes

$$-k_{th}\nabla\vartheta \cdot \mathbf{n} = j\left(L + \frac{1}{2}\frac{j^2}{\rho_v^2}\right), \tag{12.83}$$

suggesting that the heat flux conducted to the interface in the liquid is converted to latent heat of evaporation and the kinetic energy of vapor particles.

The stress balance at the interface Equation (12.80a) is reduced and now rewritten explicitly for the components of the normal and tangential stresses as

$$\begin{gathered} -\frac{j^2}{\rho_v} - \mathbf{T}\cdot\mathbf{n}\cdot\mathbf{n} = 2\widetilde{H}\sigma(\vartheta), \\ \mathbf{T}\cdot\mathbf{n}\cdot\mathbf{t} = \nabla_s\sigma\cdot\mathbf{t}. \end{gathered} \tag{12.84}$$

In Equation (12.84) the j^2-term stands for the contribution of vapor recoil. Finally, the remaining boundary conditions Equations (12.79) and (12.82) are unchanged.

The procedure of asymptotic expansions outlined in the beginning of this chapter is used again to derive the pertinent evolution equation. The dimensionless mass balance Equation (12.13) is modified by the presence of the non-dimensional evaporative mass flux, $J = jdL/k_{th}(\vartheta_0 - \vartheta_s)$

$$EJ = (-H_\tau - UH_\xi - VH_\eta + W)(1 + H_\xi^2)^{-1/2}, \tag{12.85a}$$

or at leading order of approximation

$$H_\tau + Q_\xi^{(x)} + Q_\eta^{(y)} + EJ = 0, \tag{12.85b}$$

where $Q^{(x)}(\xi, \eta, \tau) = \int_0^H U d\varsigma$, $Q^{(y)}(\xi, \eta, \tau) = \int_0^H V d\varsigma$ are the components of the scaled volumetric flow rate per unit width parallel to the wall. The parameter E in Equation (12.85) is an evaporation number

$$E = \frac{k_{th}(\vartheta_0 - \vartheta_s)}{\rho\nu L},$$

which represents the ratio of the viscous time scale $t_\nu = d^2/\nu$ to the evaporative time scale, $t_e = \rho d^2 L/k_{th}(\vartheta_0 - \vartheta_s)$ [Burelbach et al., 1988].

The dimensionless versions of Equations (12.82) and (12.83) are:

$$\begin{aligned} KJ &= \Theta \quad \text{at } \varsigma = H, \\ \Theta_\varsigma &= -J \quad \text{at } \varsigma = H, \end{aligned} \tag{12.86}$$

where

$$K = \tilde{K}\frac{k_{th}}{dL}.$$

In the lower equation in Equation (12.86) the kinetic energy term is neglected. For details refer to Burelbach et al. (1988). Equations (12.18), (12.19), (12.53), and (12.86) pose the problem whose solution is substituted into Equation (12.85b) to obtain the sought evolution equation. The general dimensionless evolution Equation (12.21) will then contain an additional term EJ, which arises from the mass flux because of evaporation and condensation now expressed via the local film thickness H.

A different approach to theoretically describe the rate of evaporative flux j in the *isothermal* case is known in the literature [Sharma, 1998; Padmakar et al., 1999]. This approach is based on the extended Kelvin equation that accounts for the local interfacial curvature and the disjoining and conjoining pressures, both entering the resulting expression for the evaporative mass flux j. It was shown by Padmakar et al. (1999) that their evaporation model admits the emergence of a flat adsorbed layer remaining in equilibrium with the ambient vapor phase, and thus in this state the evaporation rate from the film vanishes. This adsorbed layer, however, is usually several molecular spacings thick, which is beyond the resolution of continuum theory.

12.5.2 Evaporation/Condensation Only

We first consider the case of an evaporating or condensing thin liquid layer lying on a rigid plane held at constant temperature. Mass loss or gain is retained, while all other effects are neglected.

Solving first Equation (12.53) along with boundary conditions Equations (12.51a) and (12.86) and eliminating the mass flux J from the latter yields the dimensionless temperature field and the evaporative mass flux through the interface

$$\Theta = 1 - \frac{\varsigma}{H + K}, \quad J = \frac{1}{H + K}. \tag{12.87}$$

An initially flat interface will remain flat as evaporation or condensation proceeds. If surface tension, thermocapillary, and convective thermal effects are negligible (i.e., $M = S = \varepsilon RP = 0$), it will give rise to a scaled evolution equation of the form

$$H_\tau + \frac{\overline{E}}{H + K} = 0, \tag{12.88}$$

where $\overline{E} = \varepsilon^{-1}E$, positive in the evaporative case and negative in the condensing one. K, the scaled interfacial thermal resistance, is equivalent to the inverse Biot number B^{-1}. On the physical grounds, $K \neq 0$ represents a temperature jump from the liquid surface temperature to the uniform temperature of the saturated vapor ϑ_s. This jump drives the mass transfer. The conductive resistance of the liquid film is proportional to H, and the total thermal resistance, assuming infinite thermal conductivity of the solid, is given by $(H + K)^{-1}$. For a specified temperature difference $\vartheta_0 - \vartheta_s$ Equation (12.88) represents a volumetric balance whose solution, subject to the initial condition $H\,(\tau = 0) = 1$, is

$$H = -K + [(K + 1)^2 - 2\overline{E}\tau]^{1/2}. \tag{12.89}$$

In the case of evaporation $\overline{E} > 0$ and when $K \neq 0$, the film vanishes in a finite time $\tau_e = (2K + 1)/2\overline{E}$, and the rate of disappearance of the film at $\tau = \tau_e$ is finite

$$\frac{dH}{d\tau}\bigg|_{\tau = \tau_e} = -\frac{\overline{E}}{K}.$$

For $K \neq 0$, the value of $dH/d\tau$ remains finite, because as the film thins the interface temperature ϑ_i, nominally at its saturation value ϑ_s, increases to the wall temperature. If $K = 0$ however, the problem becomes singular. In this case the thermal resistance vanishes, and the mass flux will increase indefinitely if a finite temperature difference $\vartheta_0 - \vartheta_s$ is sustained. The speed of the interface at rupture becomes infinite as well.

Burelbach et al. (1988) showed that the interfacial thermal resistance $K = 10$ for a 10 nanometers thick water film. Since K is inversely proportional to the initial film thickness, $K \approx 1$ for $d = 100$ nanometers, so that $H/K \approx 1$ at this point. However, $H/K \approx 10^{-1}$ at $d = 30$ nanometers, so that the resistance to conduction is small compared to the interfacial transport resistance. Shortly after, van der Waals forces become appreciable.

12.5.3 Evaporation/Condensation, Vapor Recoil, Capillarity, and Thermocapillarity

The dimensionless vapor recoil gives an additional normal stress at the interface determined by the j^2-term in Equation (12.84), $\hat{\Pi}_3 = -\frac{3}{2}\overline{E}^2 D^{-1} J^2$, where D is a unit-order scaled ratio between the vapor and liquid densities

$$D = \frac{3}{2}\varepsilon^{-3}\,\frac{\rho_v}{\rho}.$$

This stress can be calculated using Equation (12.87). The resulting scaled evolution equation for an evaporating film on an isothermal horizontal surface neglecting the thermocapillary effect and body forces is obtained using the combination of Equations (12.21) and (12.88) with $\Pi_1 = 0$, $\Sigma_\xi = 0$ [Burelbach et al., 1988]:

$$H_\tau + \frac{\overline{E}}{H + K} + \left[\overline{E}^2 D^{-1}\left(\frac{H}{H + K}\right)^3 H_\xi\right]_\xi + \frac{1}{3}S(H^3 H_{\xi\xi\xi})_\xi = 0. \tag{12.90}$$

Since usually $t_e \gg t_v$, $\overline{E}$ can be a small number and can be used as an expansion parameter for slow evaporation compared to the non-evaporating base state [Burelbach et al., 1988] appropriate to very thin evaporating films.

Taking into account van der Waals forces and thermocapillarity, the complete evolution equation for a thin heated or cooled film on a horizontal plane surface was given by Burelbach et al. (1988) in the form

$$H_\tau + \frac{\bar{E}}{H+K} + \left\{\left[AH^{-1} + \bar{E}^2D^{-1}\left(\frac{H}{H+K}\right)^3 + K\overline{M}P^{-1}\left(\frac{H}{H+K}\right)^2\right]H_\xi\right\}_\xi + \frac{1}{3}S(H^3H_{\xi\xi\xi})_\xi = 0 \quad (12.91)$$

with $\overline{M} = \varepsilon M$. Here the first term represents the rate of volumetric change; the second one the mass loss/gain; the third, fourth, and fifth ones the attractive van der Waals, vapor recoil, and thermocapillary terms, all destabilizing; while the sixth term describes the stabilizing capillary force. This was the first full statement of the possible competition among various stabilizing and destabilizing effects on a horizontal plate, with scaling making them present at the same order. Other effects such as gravity may be included in Equation (12.91). Joo et al. (1991) extended the work to an evaporating (condensing) liquid film draining down a heated (cooled) inclined plate.

Oron and Bankoff (1999) studied the two-dimensional dynamics of an evaporating ultrathin film on a coated solid surface when the potential Equation (12.31d) was used. Three different types of the evolution of a volatile film were identified. One type is related to low evaporation rates associated with relatively small $\bar{E} > 0$ when holes covered by a liquid microlayer emerge, and the expansion of such holes is governed mainly by the action of the attractive molecular forces. These forces impart the squeeze effect to the film and, as a result of this, the liquid flows away from the hole. In this stage the role of evaporation is secondary. Figure 12.9 displays such an evolution of a volatile liquid film. Following the nucleation of the hole and during the process of surface dewetting, one can identify the formation of a large ridge, or drop, on either side of the trough. The former grows during the evolution of the film until the drops at both ends of the periodic domain collide. A further recession of the walls of the dry spot leads to the formation of a single large drop that flattens and ultimately disappears, according to Equation (12.89). The stages of the film evolution shown in Figure 12.9(a) are very similar to that sketched in Figure 3 of Elbaum and Lipson (1995). This type of evolution also resembles the results obtained by Padmakar et al. (1998) for the isothermal film subject to hydrophobic interactions and to evaporation driven by the difference between the equilibrium vapor pressure and the pressure in the vapor phase. Such films thin uniformly to a critical thickness and then spontaneously to dewet the solid substrate by the formation of growing dry spots when the solid was partially wetted. In the completely wetted case, thin liquid films evolved to an array of islands that disappeared by evaporation to a thin equilibrium flat film. Two other regimes corresponding to intermediate and high evaporation rates were discussed in Oron and Bankoff (1999).

An important phenomenon was found in the last stage of the evolution of an evaporating film where the latter finally disappears by evaporation: prior to that the film equilibrates, so that its disappearance is practically uniform in space. The film equilibration is caused by the "reservoir effect," which is driven by the difference in disjoining pressures and manifests itself by feeding the liquid from the large drops into the ultrathin film that bridges between them.

Oron and Bankoff (2001) studied the dynamics of condensing thin films on a horizontal coated solid surface. In the case of a relatively fast condensation, where the initial depression of the interface rapidly fills up because of the enhanced mass gain there, the film equilibrates and grows uniformly in space according to Equation (12.89). Note that $\bar{E} < 0$. When condensation is relatively slow, the evolution of the film exhibits several distinct stages. The first stage, dominated by attractive van der Waals forces, leads to the opening of a hole covered by a microlayer, as shown in the first three snapshots of Figure 12.10(a). This is accompanied with continuous condensation with the highest rate of mass gain attained in the microlayer region corresponding to the smallest thickness H in Equation (12.87). However, opposite to the evaporative case [Oron and Bankoff, 1999], where the "reservoir effect" arising from the difference between the disjoining pressures causes feeding of the liquid from the large drops into the microlayer and film equilibration, in the condensing case the excess liquid is driven from the microlayer into the large drops. This effect is referred to as the "reversed reservoir effect." The thickness of the microlayer remains nearly constant because of local mass gain by condensation compensating for the impact of the reverse reservoir effect. The first stage of the film evolution terminates in the situation where the size of the hole

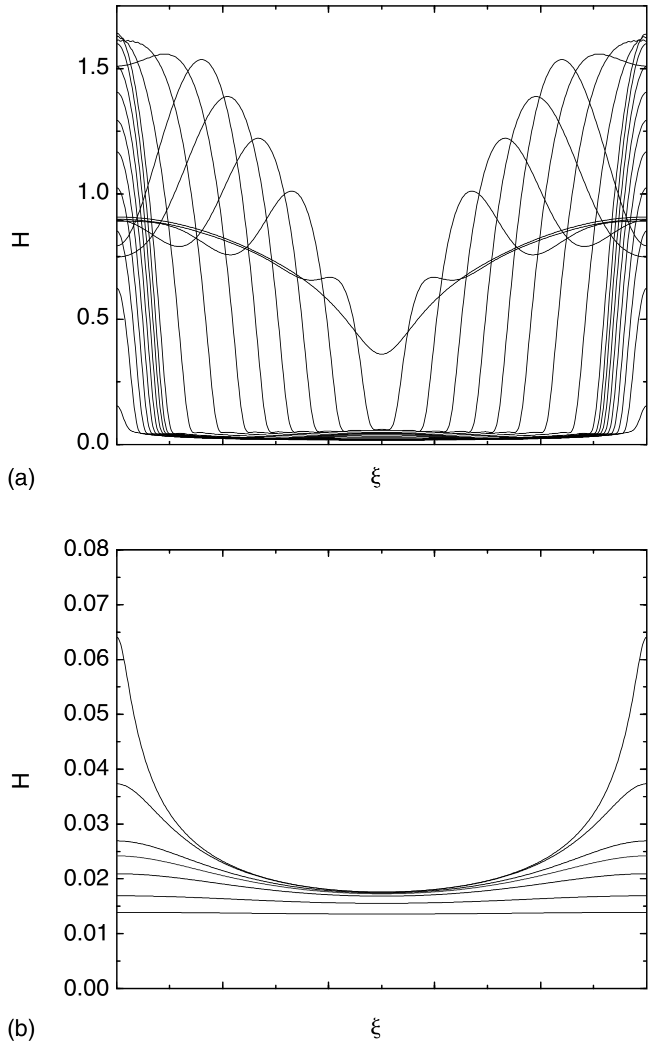


FIGURE 12.9 The evolution of a slowly evaporating film: (a) the initial and intermediate stages of the film evolution, and (b) the final stage of the evolution. The curves in both graphs correspond to the interfacial shapes in consecutive times (not necessarily equidistant). (Reprinted with permission from Oron and Bankoff (1999).)

is the largest. The receding of the drops stops due to the increase of the drop curvature and buildup of the capillary pressure that comes to balance with the squeeze effect of the attractive van der Waals forces. From this moment the hole closes driven by condensation, as shown in Figure 12.10(a, b). Once the hole closes, the depression fills up rapidly, the amplitude of the interfacial disturbance decreases, and the film tends to flatten out. The film then grows uniformly in space following the solution Equation (12.89) with negative $\overline{E}$.

Oron (2000c) studied the three-dimensional evolution of an evaporating film on a coated solid surface subject to the potential Equation (12.31d). The main stages of the evolution repeat those mentioned previously in the case of a non-volatile film in the section on isothermal films, except for the stage of disappearance accompanied by the reservoir effect. Because of the reservoir effect, the minimal film thickness decreases very slowly during the stage of film equilibration.

12.5.4 Flow on a Rotating Disc

Reisfeld et al. (1991) considered the axisymmetric flow of an incompressible viscous volatile liquid on a horizontal, rotating disk. The liquid was assumed to evaporate because of the difference between the

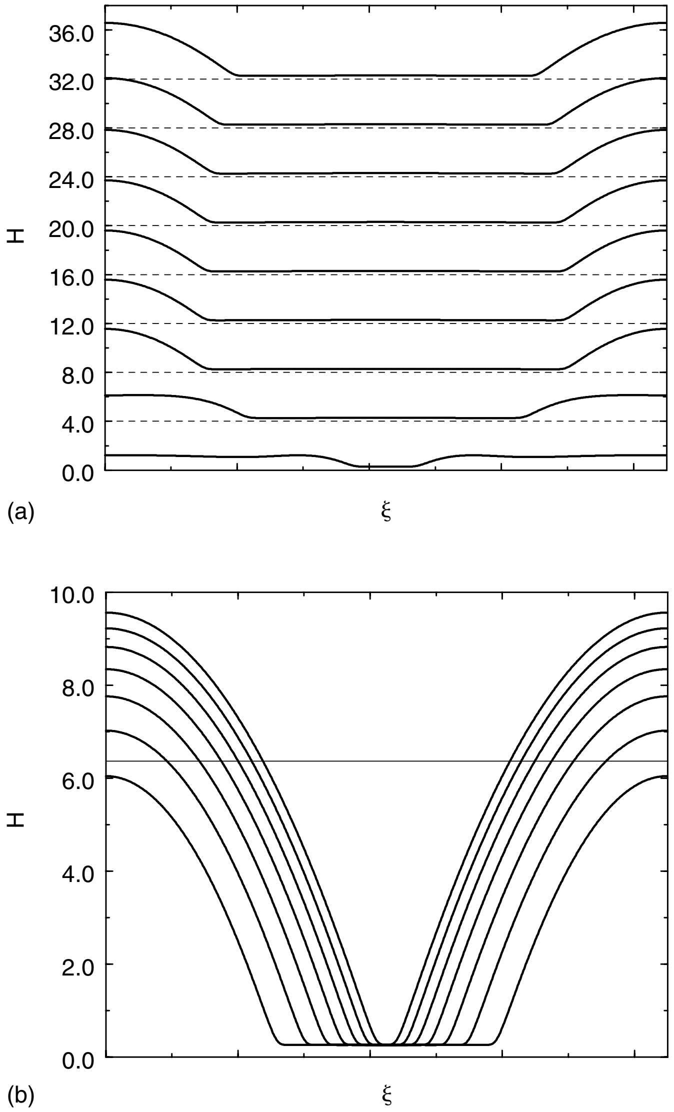


FIGURE 12.10 The evolution of a slowly condensing film on the horizontal plane. (a) The curves from the bottom to the top correspond to consecutive times (not necessarily equidistant). (b) The curves from the left to the right correspond to consecutive times (not necessarily equidistant). The flat curve corresponds to the interface at a certain time after which the film grows uniformly in space according to Equation (12.89). In the graph (a), the dashed lines represent the location of the solid substrate $H = 0$. (Reprinted with permission from Oron and Bankoff (2000).)

vapor pressures of the solvent species at the fluid–vapor interface and in the gas phase. This situation is analogous to the phase two of spin coating process.

The analysis is similar to what is done in the section on isothermal films, but now with an additional parameter describing the process of evaporation, which for a prescribed evaporative mass flux j is defined as

$$E = \frac{3j}{2\varepsilon\rho U_0}.$$

Using the procedures outlined in the section on isothermal films, one obtains at leading order the following evolution equation

$$H_\tau + \frac{2}{3}E + \frac{1}{3}r\left\{r^2H^3 + SrH^3\left[\frac{1}{r}(rH_r)_r\right]_r\right\}_r = 0. \tag{12.92}$$

Equation (12.92) models the combined effect of local mass loss, capillary forces and centrifugal drainage, none of which describe any kind of instability.

For most spin coating applications S is very small, and the corresponding term may be neglected, although it may be very important in planarization studies where the leveling of liquid films on rough surfaces is investigated. Therefore, Equation (12.92) can be simplified

$$H_\tau + \frac{2}{3}E + \frac{1}{3}r(r^2H^3)_r = 0. \tag{12.93}$$

This simplified equation can then be used for further analysis. Looking for flat basic states $H = H(\tau)$, Equation (12.93) is reduced to the ordinary differential equation which is to be solved with the initial condition $H(0) = 1$. In the case of $E > 0$, both evaporation and drainage cause thinning of the layer. Equation (12.93) describes the evolution in which the film thins monotonically to zero thickness in a finite time in contrast with an infinite thinning time by centrifugal drainage only. Explicit expressions for $H(\tau)$ and for the time of film disappearance are given in Reisfeld et al. (1991). In the condensing case $E < 0$ drainage competes with condensation to thin the film. Initially the film thins due to drainage until the rate of mass gain because of condensation balances the rate of mass loss by drainage. At this point the film interface reaches its steady location $H = |E|^{1/3}$. The cases where inertia is taken into account are considered in Reisfeld et al. (1991), where linear stability analysis of flat base states is given.

Experiments with volatile rotating liquid films [Stillwagon and Larson, 1990] showed that the final stage of film leveling was affected by an evaporative shrinkage of the films. Therefore, they suggested separating the analysis of the evolution of evaporating spinning films into two stages with fluid flow dominating the first stage and solvent evaporation dominating the second one [Stillwagon and Larson, 1992].

12.6 Closing Remarks

In this chapter the physics of thin liquid films is reviewed and various examples of their dynamics relevant for MEMS are presented, some of them with reference to the corresponding experimental results. The examples discussed examine isothermal, non-isothermal with no phase changes, and evaporating and condensing films under the influence of surface tension, gravity, van der Waals, and centrifugal forces. The long-wave theory has been proven to be a powerful tool for the research of the dynamics of thin liquid films.

However, there exist several optional approaches suitable for a study of the dynamics of thin liquid films. Direct numerical simulation of the hydrodynamic equations (Navier–Stokes and continuity) [Scardovelli and Zaleski, 1999] mentioned briefly in the introduction represents one of these options. A variety of methods were developed to carry out such simulations: techniques based on Finite Elements Method (FEM) [Ho and Patera, 1990; Salamon et al., 1994; Krishnamoorthy et al., 1995; Tsai and Yue, 1996; Ramaswamy et al., 1997], techniques based on the boundary-integral method [Pozrikidis, 1992, 1997; Newhouse and Pozrikidis, 1992; Boos and Thess, 1999], surface tracking technique [Yiantsios and Higgins, 1989], and others. Another optional approach is that of molecular dynamics (MD) simulations [Allen and Tildesley, 1987; Koplik and Banavar, 1995, 2000]. Refer directly to these works for more detail.

A new approach treating the film interface as a diffuse rather than a sharp one, as presented in this chapter, was recently developed [Pismen and Pomeau, 2000] and applied to various physical situations [Pomeau, 2001; Pismen, 2001; Bestehorn and Neuffer, 2001; Thiele et al., 2001a, b; 2002a, b; 2003].

Lastly, new frontiers in the investigation of the dynamics of thin liquid films were recently discussed in the special issue of "European Physical Journal E, Vol. 12(3), 2003". An attempt was made to bridge between numerous theoretical and experimental results in order to explain the main mechanism(s) liable to rupture of a film. Open questions, controversial approaches, and contradictory conclusions were all in the focus of the discussion [Ziherl and Zumer, 2003; van Effenterre and Valignat, 2003; Morariu et al., 2003; Kaya and Jérôme, 2003; Bollinne et al., 2003; Sharma, A., 2003; Thiele, 2003; Stöckelhuber, 2003; Richardson et al., 2003; Müller-Buschbaum, 2003; Green and Ganesan, 2003; Oron, 2003; Manghi and Aubouy, 2003; Reiter, 2003].

Acknowledgments

It is a pleasure to express my gratitude to A. Sharma, G. Reiter, and U. Thiele for reading the manuscript and sharing their comments and thoughts with me. A. Sharma, S. Herminghaus, M. F. Schatz, and E. Zussman are acknowledged for providing me with their experimental results used in this chapter.

References

Adamson, A.W. (1990) *Physical Chemistry of Surfaces*, Wiley, New York.

Allen, M.P., and Tildesley, D.J. (1987) *Computer Simulation of Liquids*, Clarendon, Oxford.

Babchin, A.J., Frenkel, A.L., Levich, B.G., and Sivashinsky, G.I. (1983) "Nonlinear Saturation of Rayleigh–Taylor Instability," *Phys. Fluids* **26**, pp. 3159–61.

Bankoff, S.G. (1961) "Taylor Instability of an Evaporating Plane Interface," *AIChE J.* **7**, pp. 485–7.

Bankoff, S.G. (1990) "Dynamics and Stability of Thin Heated Films," *J. Heat Transfer Trans. ASME* **112**, pp. 538–46.

Bankoff, S.G., and Davis, S.H. (1987) "Stability of Thin Films," *Physicochem. Hydrodyn.* **9**, pp. 5–7.

Becerril, R., Van Hook, S.J., and Swift, J.B. (1998) "The Influence of Interface Profile on the Onset of Long-Wavelength Marangoni Convection," *Phys. Fluids* **10**, pp. 3230–2.

Bestehorn, M., and Neuffer, K. (2001) "Surface Patterns of Laterally Extended Thin Liquid Films in Three Dimensions," *Phys. Rev. Lett.* **87**, pp. 046101-1–046101-4.

Bischof, J., Scherer, D., Herminghaus, S., and Leiderer, P. (1996) "Dewetting Modes of Thin Metallic Films: Nucleation of Holes and Spinodal Dewetting," *Phys. Rev. Lett.* **77**, pp. 1536–9.

Bollinne, C., Cuenot, S., Nysten, B., and Jonas, A.M. (2003) "Spinodal-Like Dewetting of Thermodynamically-Stable Thin Polymer Films," *Eur. Phys. J.* E **12**, pp. 389–96.

Boos, W., and Thess, A. (1999) "Cascade of Structures in Long-Wavelength Marangoni Instability," *Phys. Fluids* **11**, pp. 1484–94.

Brochard-Wyart, F., and Daillant, J. (1990) "Drying of Solids Wetted by Thin Liquid Films," *Can. J. Phys.* **68**, pp. 1084–8.

Brusch, L., Kuhne, H., Thiele, U., and Bar, M. (2002) "Dewetting of Thin Films on Heterogeneous Substrates: Pinning Versus Coarsening," *Phys. Rev.* E **66**, pp. 011602-1–011602-5.

Burelbach, J.P., Bankoff, S.G., and Davis, S.H. (1988) "Nonlinear Stability of Evaporating/Condensing Liquid Films," *J. Fluid Mech.* **195**, pp. 463–94.

Burelbach, J.P., Bankoff, S.G., and Davis, S.H. (1990) "Steady Thermocapillary Flows of Thin Liquid Layers. II. Experiment," *Phys. Fluids A* **2**, pp. 322–33.

Burgess, J.M., Juel, A., McCormick, W.D., Swift, J.B., and Swinney, H.L. (2001) "Suppression of Dripping from a Ceiling," *Phys. Rev. Lett.* **86,** pp. 1203–6.

Cahn, J.W. (1961) "On Spinodal Decomposition," *Acta Metall.* **9**, pp. 795–801.

Chandrasekhar, S. (1961) *Hydrodynamic and Hydromagnetic Stability*, Clarendon, Oxford.

Chou, F.-C., and Wu, P.-Y. (2000) "Effect of Air Shear on Film Planarization During Spin Coating," *J. Electrochem. Soc.* **147**, pp. 699–705.

Davis, S.H. (1983) "Rupture of Thin Films," in *Waves on Fluid Interfaces*, R.E. Meyer, ed., Academic Press, New York, pp. 291–302.

Davis, S.H. (1987) "Thermocapillary Instabilities," *Annu. Rev. Fluid Mech.* **19**, pp. 403–35.

Deissler, R.J., and Oron, A. (1992) "Stable Localized Patterns in Thin Liquid Films," *Phys. Rev. Lett.* **68**, pp. 2948–51.

Delhaye, J.M. (1974) "Jump Conditions and Entropy Sources in Two-Phase Systems. Local Instant Formulation," *Int. J. Multiphase Flow* **1**, pp. 395–409.

Dzyaloshinskii, I.E., Lifshitz, E.M., and Pitaevskii, L.P. (1959) *Zh Eksp Teor Fiz* **37**, pp. 229–41; [(1960) "Van der Waals Forces in Liquid Films," *Sov. Phys. JETP* **10**, pp. 161–170].

Edwards, D.A., Brenner, H., and Wasan, D.T. (1991) *Interfacial Transport Processes and Rheology*, Butterworth-Heinemann, Boston.

Elbaum, M., and Lipson, S.G. (1994) "How Does a Thin Wetted Film Dry Up?," *Phys. Rev. Lett.* **72**, pp. 3562–5.

Elbaum, M., and Lipson, S.G. (1995) "Pattern Formation in the Evaporation of Thin Liquid Films," *Israel J. Chem.* **35**, pp. 27–32.

Fermigier, M., Limat, L., Wesfreid, J.E., Boudinet, P., and Quilliet, C. (1992) "Two-Dimensional Patterns in Rayleigh–Taylor Instability of a Thin Layer," *J. Fluid Mech.* **236**, pp. 349–83.

Frenkel, A.L., Babchin, A.J., Levich, B.G., Shlang, T., and Sivashinsky, G.I. (1987) "Annular Flow Can Keep Unstable Flow from Breakup: Nonlinear Saturation of Capillary Instability," *J. Colloid Interface Sci.* **115**, pp. 225–33.

Gau, H., Herminghaus, S., Lenz, P., and Lipowsky, R. (1999) "Liquid Morphologies on Structured Surfaces: from Microchannels to Microchips," *Science* **283**, pp. 46–9.

de Gennes, P.G. (1985) "Wetting: Statics and Dynamics," *Rev. Mod. Phys.* **57**, pp. 827–63.

Goldmann, L.S. (1969) "Geometric Optimization of Controlled Collapse Interconnections," *IBM J. Res. Dev.* **13**.

Green, P.F., and Ganesan, V. (2003) "Dewetting of Polymeric Films: Unresolved Issues," *Eur. Phys. J. E* **12**, pp. 449–54.

Grotberg, J.B. (1994) "Pulmonary Flow and Transport Phenomena," *Annu. Rev. Fluid Mech.* **26**, pp. 529–71.

Hammond, P.S. (1983) "Nonlinear Adjustment of a Thin Annular Film of Viscous Fluid Surrounding a Thread of Another within a Circular Cylindrical Pipe," *J. Fluid Mech.* **137**, pp. 363–84.

Herminghaus, S., Fery, A., Schlagowski, S., Jacobs, K., Seeman, R., Gau, H., Moench, W., and Pompe, T. (2000) "Liquid Microstructures at Solid Interfaces," *J. Phys. Condens. Matter* **12**, pp. A57–A74.

Herminghaus, S., Gau, H., and Moench, W. (1999) "Artificial Liquid Microstructures," *Adv. Mater.* **11**, pp. 1393–5.

Hirasaki, G.J. (1991) "Thin Films and Fundamentals of Wetting Phenomena" in *Interfacial Phenomena in Petroleum Recovery*, N.R. Morrow, ed., Marcel Dekker, New York.

Ho, C.-M., and Tai, Y.-C. (1998) "Micro-Electro-Mechanical Systems (MEMS) and Fluid Flows," *Annu. Rev. Fluid Mech.* **30**, pp. 579–612.

Ho, L.-W., and Patera, A.T. (1990) "A Legendre Spectral Element Method for Simulation of Unsteady Incompressible Viscous Free-Surface Flow," *Comp. Meth. Appl. Mech. Eng.* **80**, pp. 355–66.

Huppert, H.E. (1982) "The Propagation of Two-Dimensional and Axisymmetric Viscous Gravity Currents over a Rigid Horizontal Surface," *J. Fluid Mech.* **121**, pp. 43–58.

Huppert, H.E., and Simpson, J.E. (1980) "The Slumping of Gravity Currents," *J. Fluid Mech.* **99**, pp. 785–99.

Hwang, C.-C., Chang, S.-H., and Chen, J.-L. (1993) "On the Rupture Process of Thin Liquid Film," *J. Colloid Interface Sci.* **159**, pp. 184–8.

Hwang, C.-C., Lin, C.-K., and Uen, W.-Y. (1997) "A Nonlinear Three-Dimensional Rupture Theory of Thin Liquid Films," *J. Colloid Interface Sci.* **190**, pp. 250–2.

Ichimura, K., Oh, S.-K., and Nakagawa, M. (2000) "Light-Driven Motion of Liquids on a Photoresistive Surface," *Science* **288**, pp. 1624–6.

Israelachvili, J.N. (1992) *Intermolecular and Surface Forces*, 2nd ed., Academic Press, London.

Ivanov, I.B. (1988) *Thin Liquid Films: Fundamentals and Applications*, Marcel Dekker, New York.

Jacobs, K., Herminghaus, S., and Mecke, K.R. (1998) "Thin Liquid Polymer Films Rupture via Defects," *Langmuir* **14**, pp. 965–9.

Jain, R.K., and Ruckenstein, E. (1974) "Spontaneous Rupture of Thin Liquid Films," *J. Chem. Soc. Faraday Trans. II* **70**, pp. 132–47.

Jameel, A.T., and Sharma, A. (1994) "Morphological Phase Separation in Thin Liquid Films," *J. Colloid Interface Sci.* **164**, pp. 416–27.

Joo, S.W., Davis, S.H., and Bankoff, S.G. (1991) "Long-Wave Instabilities of Heated Films: Two Dimensional Theory of Uniform Layers," *J. Fluid Mech.* **230**, pp. 117–46.

Kargupta, K., Konnur, R., and Sharma, A. (2000) "Instability and Pattern Formation in Thin Liquid Films on Chemically Heterogeneous Substrates," *Langmuir* **16**, pp. 10243–53.

Kargupta, K., Konnur, R., and Sharma, A. (2001) “Spontaneous Dewetting and Ordered Patterns in Evaporating Thin Liquid Films on Homogeneous and Heterogeneous Substrates,” *Langmuir* **17**, pp. 1294–305.

Kargupta, K., and Sharma, A. (2001) “Templating of Thin Films Induced by Dewetting on Patterned Surfaces,” *Phys. Rev. Lett.* **86**, pp. 4536–9.

Kargupta, K., and Sharma, A. (2002a) “Dewetting of Thin Films on Periodic Physically and Chemically Patterned Surfaces,” *Langmuir* **16**, pp. 1893–903.

Kargupta, K., and Sharma, A. (2002b) “Morphological Self-Organization by Dewetting in Thin Films on Chemically Patterned Substrates” *J. Chem. Phys.* **116**, pp. 3042–51.

Kargupta, K., and Sharma, A. (2002c) “Creation of Ordered Patterns by Dewetting of Thin Films on Homogeneous and Heterogeneous Substrates,” *J. Colloid Interface Sci.* **245**, pp. 99–115.

Kargupta, K., and Sharma, A. (2003) “Mesopatterning of Thin Liquid Films by Templating on Chemically Patterned Complex Substrates,” *Langmuir* **19**, pp. 5153–63.

Kaya, H., and Jérôme, B. (2003) “Unstable Thin Films: New Questions,” *Eur. Phys. J. E* **12**, pp. 383–8.

Khanna, R., Jameel, A.T., and Sharma, A. (1996) “Stability and Breakup of Thin Polar Films on Coated Substrates: Relationship to Macroscopic Parameters of Wetting,” *Ind. Eng. Chem. Res.* **35**, pp. 3081–92.

Khanna, R., and Sharma, A. (1998) “Pattern Formation in Spontaneous Dewetting of Thin Apolar Films,” *J. Colloid Interface Sci.* **195**, pp. 42–50.

Khanna, R., Sharma, A., and Reiter, G. (2000) “The ABC of Pattern Evolution in Self-Destruction of Thin Polymer Films,” *Euro. Phys. J. E* **2**, pp. 1–9.

Kheshgi, H.S., and Scriven, L.E. (1991) “Dewetting: Nucleation and Growth of Dry Regions,” *Chem. Eng. Sci.* **46**, pp. 519–26.

Knight, J.B., Vishwanath, A., Brody, J.P., and Austin, R.H. (1998) “Hydrodynamic Focusing on a Silicon Chip: Mixing Nanoliters in Microseconds,” *Phys. Rev. Lett.* **80**, pp. 3863–6.

Konnur, R., Kargupta, K., and Sharma, A. (2000) “Instability and Morphology of Thin Liquid Films on Chemically Heterogeneous Substrates,” *Phys. Rev. Lett.* **84**, pp. 931–4.

Kopbosynov, B.K., and Pukhnachev, V.V. (1986) “Thermocapillary Flow in Thin Liquid Films,” *Fluid Mech. Sov. Res.* **15**, pp. 95–106.

Koplik, J., and Banavar, J.R. (1995) “Continuum Deductions from Molecular Hydrodynamics” *Annu. Rev. Fluid Mech.* **27**, pp. 257–92.

Koplik, J., and Banavar, J.R. (2000) “Molecular Simulations of Dewetting,” *Phys. Rev. Lett.* **84**, pp. 4401–4.

Krishnamoorthy, S., Ramaswamy, B., and Joo, S.W. (1995) “Spontaneous Rupture of Thin Liquid Films Due to Thermocapillarity: A Full-Scale Direct Numerical Simulation,” *Phys. Fluids A* **7**, pp. 2291–3.

Landau, L.D., and Lifshitz, E.M. (1987) *Fluid Mechanics*, Pergamon, Oxford.

Lee, J., and Kim, C.-J. (2000) “Surface-Tension-Driven Microactuation Based on Continuous Electrowetting,” *J. Microelectromech. Syst.* **9**, pp. 171–80.

Leger, L., and Joanny, J.F. (1992) “Liquid Spreading,” *Rep. Prog. Phys.* **55**, pp. 431–86.

Legros, J.C. (1986) “Problems Related to Non-Linear Variations of Surface Tension,” *Acta Astron.* **13**, pp. 697–703.

Legros, J.C., Limbourg-Fontaine, M.C., and Petre, G. (1984) “Influence of Surface Tension Minimum as a Function of Temperature on the Marangoni Convection,” *Acta Astron.* **14**, pp. 143–7.

Lenz, P. (1999) “Wetting Phenomena on Structured Surfaces,” *Adv. Mater.* **11**, pp. 1531–3.

Lenz, P., and Lipowsky, R. (1998) “Morphological Transitions of Wetting Layers on Structured Surfaces,” *Phys. Rev. Lett.* **80**, pp. 1920–3.

Levich, V.G. (1962) *Physicochemical Hydrodynamics*, Prentice-Hall, Englewood Cliffs.

Lin, W., Patra, S.K., and Lee, Y.C. (1995) “Design of Solder Joints for Self-Aligned Optoelectronic Assemblies,” *IEEE Trans. Compon., Packag. Manuf. Technol. B* **18**, pp. 543–51.

Lipowsky, R., Lenz, P., and Swain, P.S. (2000) “Wetting and Dewetting of Structured and Imprinted Surfaces,” *Colloid. Surface. A* **161**, pp. 3–22.

Manghi, M., and Aubouy, M. (2003) “Short Commentary: Theoretical Considerations on Concentrated Polymer Interfaces: an Attempt to Explain Dewetting of Ultra-Thin Films,” *Eur. Phys. J. E* **12**, pp. 459–63.

Mitlin, V.S. (1993) "Dewetting of Solid Surface: Analogy with Spinodal Decomposition," *J. Colloid Interface Sci.* **156**, pp. 491–7.

Mitlin, V.S. (2000) "Dewetting Revisited: New Asymptotics of the Film Stability Diagram and the Metastable Regime of Nucleation and Growth of Dry Zones," *J. Colloid Interface Sci.* **227**, pp. 371–9.

Mitlin, V.S. and Petviashvili, N.V. (1994) "Nonlinear Dynamics of Dewetting: Kinetically Stable Structures," *Phys. Lett. A* **192**, pp. 323–6.

Morariu, M.D., Schäffer, E., and Steiner, U. (2003) "Capillary Instabilities by Fluctuation Induced Forces," *Eur. Phys. J. E* **12**, pp. 375–82.

Moriarty, J.A., and Terrill, E.L. (1996) "Mathematical Modeling of the Motion of Hard Contact Lenses," *Eur. J. Appl. Math.* **7**, pp. 575–94.

Müller-Buschbaum, P. (2003) "Influence of Surface Cleaning on Dewetting of Thin Polystyrene Films," *Eur. Phys. J. E* **12**, pp. 443–8.

Muller-Buschbaum, P., Vanhoorne, P., Scheumann, V., and Stamm, M. (1997) "Observation of Nano-Dewetting Structures," *Europhys. Lett.* **40**, pp. 655–60.

Myers, T.G. (1998) "Thin Films with High Surface Tension," *SIAM Rev.* **40**, pp. 441–62.

Newhouse, L.A., and Pozrikidis, C. (1992) "The Capillary Instability of Annular Layers and Liquid Threads," *J. Fluid Mech.* **242**, pp. 193–209.

Oron, A. (2000a) "Nonlinear Dynamics of Irradiated Thin Volatile Liquid Films," *Phys. Fluids* **12**, pp. 29–41.

Oron, A. (2000b) "Nonlinear Dynamics of Three-Dimensional Long-Wave Marangoni Instability in Thin Liquid Films," *Phys. Fluids* **12**, pp. 1633–45.

Oron, A. (2000c) "Three-Dimensional Nonlinear Dynamics of Thin Liquid Films," *Phys. Rev. Lett.* **85**, pp. 2108–11.

Oron, A. (2003) "Short Commentary: Theory of Thin Liquid Films: Some Questions and Challenges," *Eur. Phys. J. E* **12**, pp. 455–8.

Oron, A., and Bankoff, S.G. (1999) "Dewetting of a Heated Surface by an Evaporating Liquid Film under Conjoining/Disjoining Pressures," *J. Colloid Interface Sci.* **218**, pp. 152–66.

Oron, A., and Bankoff, S.G. (2001) "Dynamics of a Condensing Liquid Film under Conjoining/Disjoining Pressures," *Phys. Fluids* **13**, pp. 1107–17.

Oron, A., Bankoff, S.G., and Davis, S.H. (1996) "Thermal Singularities in Film Rupture," *Phys. Fluids A* **8**, pp. 3433–5.

Oron, A., Davis, S.H., and Bankoff, S.G. (1997) "Long-Scale Evolution of Thin Liquid Films," *Rev. Mod. Phys.* **68**, pp. 931–80.

Oron, A., and Peles, Y. (1998) "Stabilization of Thin Liquid Films by Internal Heat Generation," *Phys. Fluids A* **10**, pp. 537–9.

Oron, A., and Rosenau, P. (1992) "Formation of Patterns Induced by Thermocapillarity and Gravity," *J. Phys. II France* **2**, pp. 131–46.

Oron, A., and Rosenau, P. (1994) "On a Nonlinear Thermocapillary Effect in Thin Liquid Layers," *J. Fluid Mech.* **273**, pp. 361–74.

Padmakar, A.S., Kargupta, K., and Sharma, A. (1999) "Stability and Dewetting of Evaporating Thin Water Films on Partially and Completely Wettable Substrates," *J. Chem. Phys.* **110**, pp.1735–44.

Palmer, H.J. (1976) "The Hydrodynamic Stability of Rapidly Evaporating Liquids at Reduced Pressure," *J. Fluid Mech.* **75**, pp. 487–511.

Paulsen, F.G., Pan, R., Bousfield, D.W., and Thompson, E.V. (1996) "The Dynamics of Bubble/Particle Attachment and the Application of Two Disjoining Film Rupture Models to Flotation," *J. Colloid Interface Sci.* **178**, pp. 400–10.

Peurrung, L.M., and Graves, D.B. (1993) "Spin Coating over Topography," *IEEE Trans. Semicond. Manufact.* **6**, pp. 72–6.

Pismen, L.M. (2001) "Nonlocal Diffuse Interface Theory of Thin Films and the Moving Contact Line," *Phys. Rev. E* **64**, pp. 021603-1–021603-9.

Pismen, L.M., and Pomeau, Y. (2000) "Disjoining Potential and Spreading of Thin Liquid Layers in the Diffuse-Interface Model Coupled to Hydrodynamics," *Phys. Rev. E* **62**, pp. 2480–92.

Plesset, M.S., and Prosperetti, A. (1976) “Flow of Vapor in a Liquid Enclosure,” *J. Fluid Mech.* **78**, pp. 433–44.

Pomeau, Y. (2001) “Moving Contact Line,” *Journal de Physique IV* **11**, pp. 199–212.

Pozrikidis, C. (1992) *Boundary Integral and Singularity Methods for Linearized Viscous Flow*, Cambridge University, Cambridge.

Pozrikidis, C. (1997) “Numerical Studies of Singularity Formation at Free Surfaces and Fluid Interfaces in Two-Dimensional Stokes Flow,” *J. Fluid Mech.* **331**, pp. 145–67.

Prud’homme, R.K., and Khan, S.A., eds. (1996) *Foams: Theory, Measurements and Applications*, Marcel Dekker, New York.

Ramaswamy, B., Krishnamoorthy, S., and Joo, S.W. (1997) “Three-Dimensional Simulation of Instabilities and Rivulet Formation in Heated Falling Films,” *J. Comp. Phys.* **131**, pp. 70–88.

Reisfeld, B., Bankoff, S.G., and Davis, S.H. (1991) “The Dynamics and Stability of Thin Liquid Films During Spin-Coating. I. Films with Constant Rates of Evaporation or Absorption. II. Films with Unit-Order and Large Peclet Numbers,” *J. Appl. Phys.* **70**, pp. 5258–77.

Reiter, G. (1992) “Dewetting of Thin Polymer Films,” *Phys. Rev. Lett.* **68**, pp. 75–8.

Reiter, G. (1998) “The Artistic Side of Intermolecular Forces,” *Science* **282**, pp. 888–9.

Reiter, G., Sharma, A., Casoli, A., David, M.-O., Khanna, R., and Auroy, P. (1999a) “Destabilizing Effect of Long-Range Forces in Thin Liquid Films on *Wettable* Substrates,” *Europhys. Lett.* **46**, pp. 512–8.

Reiter, G., Sharma, A., Casoli, A., David, M.-O., Khanna, R., and Auroy, P. (1999b) “Thin Film Instability Induced by Long-Range Forces,” *Langmuir* **15**, pp. 2551–8.

Reiter, G., Khanna, R., and Sharma, A. (2000) “Enhanced Instability in Thin Liquid Films by Improved Compatibility,” *Phys. Rev. Lett.* **85**, pp. 1432–5.

Reiter, G. (2003) “Summary and Conclusions: Progress in Our Understanding of Instabilities in Thin Films,” *Eur. Phys. J. E* **12**, pp. 465–8.

Richardson, H., Carelli, C., Keddie, J.L., and Sferrazza, M. (2003) “Structural Relaxation of Spin-Cast Glassy Polymer Thin Films as a Possible Factor in Dewetting,” *Eur. Phys. J. E* **12**, pp. 437–41.

Rosenau, P. (1995) “Fast and Superfast Diffusion Processes,” *Phys. Rev. Lett.* **74**, pp. 1056–9.

Sadhal, S.S., and Plesset, M.S (1979) “Effect of Solid Properties and Contact Angle in Dropwise Condensation and Evaporation,” *J. Heat Transf.* **101**, pp. 48–54.

Salalha, W., Zussman, E., Meltser, M., and Kaldor, S. (2000) “Prediction of Yield for Flip-Chip Packaging,” *Proc. 10th International CIRP Design Seminar,* pp. 259–63, Haifa, Israel.

Salamon, T.R., Armstrong, R.C., and Brown, R.A. (1994) “Traveling Waves on Vertical Films: Numerical Analysis Using the Finite Elements Method,” *Phys. Fluids A* **6**, pp. 2202–20.

Scardovelli, R., and Zaleski, S. (1999) “Direct Numerical Simulation of Free-Surface and Interfacial Flow,” *Ann. Rev. Fluid Mech.* **31**, pp. 567–603.

Schaeffer, E., Thurn-Albrecht, T., Russell, T.P., and Steiner, U. (2000) “Electrically Induced Structure Formation and Pattern Transfer,” *Nature* **403**, pp. 874–7.

Scheludko, A.D. (1967) “Thin Liquid Films,” *Adv. Coll. Interface Sci.* **1**, pp. 391–464.

Schlichting, H. (1968) *Boundary-Layer Theory*, McGraw-Hill, New York.

Schrage, R.W. (1953) *A Theoretical Study of Interphase Mass Transfer*, Columbia University, New York.

Schramm, L.A., ed. (1994) *Foams: Fundamentals and Applications in the Petroleum Industry*, American Chemical Society, Washington.

Sehgal, A., Ferreiro, V., Douglas, J.F., Amis, E.J., and Karim, A. (2002) “Pattern-Directed Dewetting of Ultrathin Polymer Films,” *Langmuir* **18**, pp. 7041–8.

Sharma, A. (1998) “Equilibrium and Dynamics of Evaporating or Condensing Thin Fluid Domains: Thin Film Stability and Heterogeneous Nucleation,” *Langmuir* **14**, pp. 4915–28.

Sharma, A. (2003) “Many Paths to Dewetting of Thin Films: Anatomy and Physiology of Surface Instability,” *Eur. Phys. J. E* **12**, pp. 397–408.

Sharma, A., and Jameel, A.T. (1993) “Nonlinear Stability, Rupture, and Morphological Phase Separation of Thin Fluid Films on Apolar and Polar Substrates,” *J. Colloid Interface Sci.* **161**, pp. 190–208.

Sharma, A., and Khanna, R. (1998) "Pattern Formation in Unstable Thin Liquid Films," *Phys. Rev. Lett.* **81**, pp. 3463–6.

Sharma, A., Konnur, R., and Kargupta, K. (2003) "Thin Liquid Films on Chemically Heterogeneous Substrates: Self-Organization, Dynamics and Patterns in Systems Displaying a Secondary Minimum," *Physica A* **318**, pp. 262–78.

Sharma, A., Konnur, R., Khanna, R., and Reiter, G. (2000) "Morphological Pathways of Pattern Evolution and Dewetting in Thin Liquid Films," in *Emulsions, Foams and Thin Films,* K.L. Mittal, ed., pp. 211–32, Marcel Dekker, New York.

Sharma, A., and Reiter, G. (1996) "Instability of Thin Polymer Films on Coated Substrates: Rupture, Dewetting, and Drop Formation," *J. Colloid Interface Sci.* **178**, pp. 383–99.

Sharma, A., and Ruckenstein, E. (1986), "An Analytical Nonlinear Theory of Thin Film Rupture and Its Application to Wetting Films," *J. Colloid Interface Sci.* **113**, pp. 456–79.

Siegel, R., and Howell, J.R. (1992) *Thermal Radiation Heat Transfer*, Hemisphere, Washington.

Stillwagon, L.E., and Larson, R.G. (1988) "Fundamentals of Topographic Substrate Leveling," *J. Appl. Phys.* **63**, pp. 5251–8.

Stillwagon, L.E., and Larson, R.G. (1990) "Leveling of Thin Films over Uneven Substrates during Spin Coating," *Phys. Fluids A* **2**, pp. 1937–44.

Stillwagon, L.E., and Larson, R.G. (1992) "Planarization during Spin Coating," *Phys. Fluids A* **4**, pp. 895–903.

Stöckelhuber, K.W. (2003) "Stability and Rupture of Aqueous Wetting Films," *Eur. Phys. J. E* **12**, pp. 431–5.

Tan, M.J., Bankoff, S.G., and Davis, S.H. (1990) "Steady Thermocapillary Flows of Thin Liquid Layers," *Phys. Fluids A* **2**, pp. 313–21.

Teletzke, G.F., Davis, H.T., and Scriven, L.E. (1987) "How Liquids Spread on Solids," *Chem. Eng. Comm.* **55**, pp. 41–82.

Thiele, U. (2003) "Open Questions and Promising New Fields in Dewetting," *Eur. Phys. J. E* **12**, pp. 409–16.

Thiele, U., Brusch, L., Bestehorn, M., and Bar, M. (2003) "Modelling Thin-Film Dewetting on Structured Substrates and Templates: Bifurcation Analysis and Numerical Simulations," *Eur. Phys. J. E* **11**, pp. 255–71.

Thiele, U., Neuffer, K., Bestehorn, M., Pomeau, Y., Velarde, M.G. (2002a) "Sliding Drops on an Inclined Plane," *Colloids and Surfaces A — Physicochemical and Engineering Aspects.* **206**, pp. 87–104.

Thiele, U., Neuffer, K., Pomeau, Y., and Velarde, M.G. (2002b) "On the Importance of Nucleation Solutions for the Rupture of Thin Liquid Films," *Colloids and Surfaces A — Physicochemical and Engineering Aspects.* **206**, pp. 135–55.

Thiele, U., Mertig, M., and Pompe, W. (1998) "Dewetting of an Evaporating Thin Liquid Film: Heterogeneous Nucleation and Surface Instability," *Phys. Rev. Lett.* **80**, pp. 2869–71.

Thiele, U., Velarde, M.G., Neuffer, K., Bestehorn, M., and Pomeau, Y. (2001b) "Sliding Drops in the Diffuse Interface Model Coupled to Hydrodynamics," *Phys. Rev. E* **64**, pp. 061601-1–061601-12.

Thiele, U., Velarde, M.G., Neuffer, K., and Pomeau, Y. (2001a) "Film Rupture in the Diffuse Interface Model Coupled to Hydrodynamics," *Phys. Rev. E* **64**, pp. 031602-1–031602-14.

Tsai, W.T., and Yue, D.K.P. (1996) "Computation of Nonlinear Free-Surface Flow," *Annu. Rev. Fluid Mech.* **28**, pp. 249–78.

van Effenterre, D., and Valignat, M.P. (2003) "Stability of Thin Nematic Films," *European Physical Journal E* **12**, pp. 367–72.

VanHook, S.J., Schatz, M.F., McCormick, W.D., Swift, J.B., and Swinney, H.L. (1995) "Long-Wavelength Instability in Surface-Tension-Driven Benard Convection," *Phys. Rev. Lett.* **75**, pp. 4397–400.

VanHook, S.J., Schatz, M.F., Swift, J.B., McCormick, W.D., and Swinney, H.L. (1997) "Long-Wavelength Instability in Surface-Tension-Driven Benard Convection: Experiment and Theory," *J. Fluid Mech.* **345**, pp. 45–78.

Wale, M.J., and Edge, C. (1990) "Self-Aligned, Flip-Chip Assembly of Photonic Devices with Electrical and Optical Connections," *IEEE Trans. Comp. Hybrids Manufact. Technol.* **13**, pp. 780–6.

Wehausen, J.V., and Laitone, E.V. (1960) "Surface Waves" in *Encyclopedia of Physics*, S. Flugge, ed., Springer-Verlag, Berlin, vol.IX, Fluid Dynamics III, pp. 446–778.

Williams, M.B. (1981) Nonlinear Theory of Film Rupture, Ph.D. thesis, Johns Hopkins University.

Williams, M.B., and Davis, S.H. (1982) "Nonlinear Theory of Film Rupture," *J. Colloid Interface Sci.* **90**, pp. 220–8.

Witelski, T.P., and Bernoff, A.J. (1999) "Stability of Self-Similar Solutions for van der Waals Driven Thin Film Rupture," *Phys. Fluids* **11**, pp. 2443–5.

Witelski, T.P., and Bernoff, A.J. (2000) "Dynamics of Three-Dimensional Thin Film Rupture," *Physica D* **147**, pp. 155–76.

Wu, P.-Y., and Chou, F.-C. (1999) "Complete Analytical Solutions of Film Planarization during Spin Coating," *J. Electrochem. Soc.* **146**, pp. 3819–26.

Wu, P.-Y., Chou, F.-C., and Gong, S.-C. (1999) "Analytical Solutions of Film Planarization for Periodic Features," *J. Appl. Phys.* **86**, pp. 4657–9.

Xie, R., Karim, A., Douglas, J.F., Han, C.C., and Weiss, R.A. (1998) "Spinodal Dewetting of Thin Polymer Films," *Phys. Rev. Lett.* **81**, pp. 1251–4.

Yarin, A.L. (1993) *Free Liquid Jets and Films*, Longman, New York.

Yiantsios, S.G., and Higgins, B.G. (1989) "Rayleigh–Taylor Instability in Thin Viscous Films," *Phys. Fluids A* **1**, pp. 1484–501.

Yiantsios, S.G., and Higgins, B.G. (1991) "Rupture of Thin Films: Nonlinear Stability Analysis," *J. Colloid Interface Sci.* **147**, pp. 341–50.

Zhang, W.W., and Lister, J.R. (1999) "Similarity Solutions for van der Waals Rupture of a Thin Film on a Solid Substrate," *Phys. Fluids* **11**, pp. 2454–62.

Ziherl, P., and Zumer, S. (2003) "Morphology and Structure of Thin Liquid-Crystalline Films at Nematic-Isotropic Transition," *Eur. Phys. J. E* **12**, pp. 361–6.

13

Bubble/Drop Transport in Microchannels

Hsueh-Chia Chang

University of Notre Dame

13.1 Introduction

Many microdevices involve fluid flows. Microducts, micronozzles, micropumps, microturbines, and microvalves are examples of small devices with gas or liquid flow. Designing similar devices for two-phase flows is desirable, and one can envision many attractive applications, if microreactors and microlaboratories could include immiscible liquid–liquid and gas–liquid systems. Miniature evaporative and distillation units, bubble generators, multiphase extraction and separation units, and many other conventional multiphase chemical processes could be fabricated at microscales. Efficient multiphase heat exchangers could be designed for microelectromechanical systems (MEMS) devices to minimize joule or frictional heating effects. Even for the current generation of microlaboratories using electrokinetic flow, multiphase flow has many advantages. Drops of organic samples could be transported by flowing electrolytes, thus extending the electrokinetic concept to a broader class of samples. Gas bubbles could be used as spacers for samples in a channel or act as a piston to produce pressure-driven flow on top of the electrokinetic flow. Flow valves and pumps that employ air bubbles, like those in the ink reservoirs of ink jet printers, are already being tested for microchannels. Drug-delivery and diagnostic devices involving colloids, molecules, and biological cells are also active areas of research.

Before multiphase flow in microchannels becomes a reality, several fundamental problems that arise from the small dimension of the channels must be solved. Most of these problems originate from the large curvature of the interface between two phases in these small channels. Furthermore, the menisci along the channel often have opposite curvatures that give rise to large capillary pressure drops of opposite signs. This makes it difficult to sustain a pressure gradient in the same direction along the channel. Another related problem concerns three-phase contact lines that can exist at these menisci. Contact-line resistance is often negligible in macroscopic flows. The contact-line region, defined by intermolecular and capillary forces, is small compared to the macroscopic length scales. However, in microchannels, the contact-line region is comparable in dimension to the channel size. As a result, the large stress in that region (the classical contact-line logarithm stress singularity) can dominate the total viscous dissipation [Kalliadasis and Chang, 1994; Veretennikov et al., 1998; Indeikina and Chang, 1999]. Hence, it is inadvisable to have contact lines in microchannels unless one is prepared to apply enormous pressure or electric potential driving forces. One fluid should wet the channel or capillary walls while the other is dispersed in the form of bubbles. Due to the small channel dimension, the bubbles usually have a free radius larger than the channel radius — it is typically difficult to generate colloid-size bubbles smaller than the channel. This chapter addresses several fundamental issues in the transport of these "large" bubbles and suggests the most realistic and attainable conditions for such multiphase microfluidic flows.

13.2 Fundamentals

Schematics of a bubble immersed in a wetting liquid within a capillary of radius R are shown in Figure 13.1. The dimensionless coordinate r is scaled by the capillary radius R. If the bubble is not translating, the capillary pressure drop across the bubble cap is of order σ/R, where s is the interfacial tension. In contrast, the pressure drop necessary to drive a liquid slug of length l at speed U in the same channel is of order $Ul\,\mu/R^2$. Hence, the slug length l scales as RCa^{-1} where $Ca = \mu U/\sigma$ is the capillary number. In microchannels, Ca ranges from 10^{-8} to 10^{-4} (for aqueous solutions moving at 10^{-4} to 1 mm/sec), thus the equivalent slug length l is many orders of magnitude higher than R. Equivalently, the capillary pressure across the static meniscus can drive a liquid slug of length R at the astronomically large dimensionless speed of $Ca = 1$. For electrokinetic flow, such speeds can be achieved only by an electric field of more than 104 V/cm. The capillary pressure across a static meniscus in a capillary, sometimes called the invasion pressure, is the

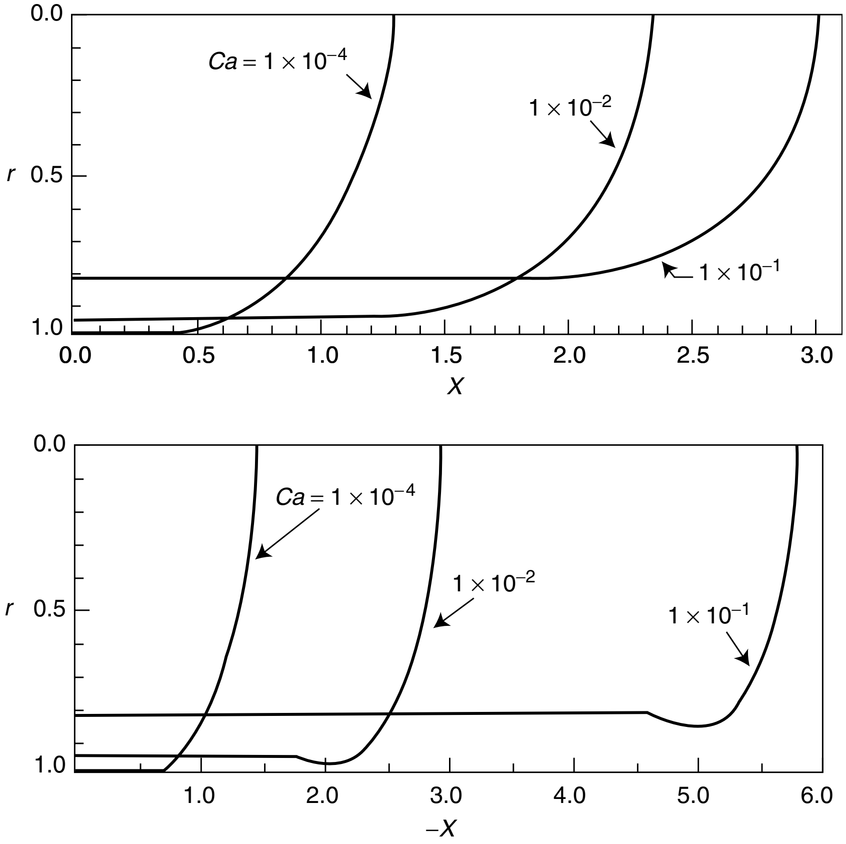


FIGURE 13.1 Front and back profiles for very long bubbles.

required pressure to insert a meniscus in the capillary. After the bubble is set into motion, the required pressure to sustain its motion is less than σ/R but is still significant.

The thickness of the wetting film around a moving bubble in a capillary and the pressure drop across the wetting film were first studied by Bretherton (1961). For capillary radii R smaller than the capillary length $(\sigma/\Delta\rho g)^{1/2}$, which is about 1 mm for aqueous solutions, buoyancy effects are negligible, and the bubble is axisymmetrically placed within the capillary. The flat annular film around the bubble allows only unidirectional longitudinal flow. This lubrication limit stipulates that the pressure be constant across the film and determined by the local interfacial curvature, the sum of the axial and azimuthal curvatures of the axisymmetric bubble. Pressure variation is only in the longitudinal direction. For pressure-driven mobile bubbles, the flat annular film at the middle of the drop indicates that no pressure gradient is present and that there is no flow in the film. Liquid flow only occurs at the transition regions near the caps where the film is no longer flat in the longitudinal direction. Near the front cap, the azimuthal curvature decreases behind the tip, and the resulting capillary pressure gradient drives fluid into the annular film. The reverse happens near the back cap to pick up the stagnant liquid laid down by the front cap. Unlike the usual symmetric Stokes flow, the flow around the two caps are not mirror images of each other in this free-surface problem. If they were reflectively symmetric, the net pressure drop across the bubble would be zero, which is impossible for a translating bubble. The same negative bulk pressure gradient results in pressure-driven liquid flow before and after the bubble. The capillary pressure gradients at the two caps are in opposite directions relative to this bulk gradient. As a result, the two caps are not mirror images of each other, and the capillary pressure across the back cap must be smaller than that at the front cap.

Simple scaling arguments determine the pressure drop across the bubble and the thickness of the surrounding film. The leading order estimate of capillary pressure drops at both caps is identical at $2\sigma/R$, and the axial curvature at the tips is $1/R$. The axial curvature at the surrounding annular film is d^2h/dx^2, where h is the interfacial thickness measured from the capillary wall and x is the longitudinal direction. (The azimuthal curvature gradient scales as h_x and is negligible compared to the axial curvature gradient h_{xxx} in the short transition region.) Balancing the axial curvature d^2h/dx^2 to $1/R$ reveals that the ratio of the length of the transition region scales as the square root of the film thickness, with both lengths small compared to the capillary radius R. The pressure gradient in the transition regions provided by the capillary pressure drives a liquid flow at the speed U of the bubble. Balancing the viscous dissipation estimate $\mu U/h^2$ with dp/dx and using the above scalings for each quantity, we conclude that the ratio of h^2 to the transition length x is of order Ca. Reconciling this with the relative scalings imposed by curvature matching, we obtain the classical Bretherton scalings — the film thickness scales as $RCa^{2/3}$ while the transition regions near the cap are of the order of $RCa^{1/3}$ long, with $\mathrm{Ca} << 1$. The total viscous dissipation due to the flow at the caps is the integral of μ times the normal gradient of the flow field at the wall over the transition length. This is the capillary pressure required to balance the dissipation. Using the previous scaling, this capillary pressure is of the order $(\sigma/R)Ca^{2/3}$. Due to the asymmetry of the two caps, this capillary pressure is different at the two caps. The difference in the pressure drop across the two caps is then of the order $(\sigma/R)Ca^{2/3}$.

Using this new estimate for the pressure drop, we conclude that the equivalent slug length l scales as $RCa^{-1/3}$. Equivalently, in a train of translating bubbles spaced by continuous liquid slugs, the pressure drop across each bubble roughly corresponds to a liquid slug of length $RCa^{-1/3}$, or 10 to 1000 times the capillary radius. Hence, the pressure drop required to drive most bubble trains occurs at the bubble caps. Even without contact-line resistance, pressure-driven multiphase transport in microchannels is expected to require orders of magnitude higher pressure drops. In the next section, we estimate this pressure drop with and without Marangoni traction introduced by surfactants, and we sketch the effects of drop viscosity and noncircular capillaries. It is unlikely that we can achieve pressure-driven multiphase flow under realistic conditions. The following section shows that electrokinetically driven multiphase flow is achievable and demonstrates that bubble speed can reach as high as the electrokinetic speed of pure liquids. Such flows occur under very specific conditions, which are described in some detail. We conclude with some conjectures on other multiphase microfluidics.

13.3 The Bretherton Problem for Pressure-Driven Bubble/Drop Transport

The previous scaling arguments can be made more precise with matched asymptotics. Using a local Cartesian coordinate for the thin-film region, the usual lubrication analysis yields the following longitudinal velocity profile:

$$u(y) = \frac{-\sigma}{\mu}\frac{\partial^3 h}{\partial x^3}\left(\frac{y^2}{2} - yh\right) \tag{13.1}$$

The normal coordinate y is measured from the capillary wall. The pressure p is independent of y and is equal to $-\sigma h_{xx}$, the axial curvature of the film. Hence, integrating over the film thickness, one obtains the flow rate $q = \left(\frac{\sigma h^3}{3\mu}\right)\frac{\partial^3 h}{\partial x^3}$. The cubic power dependence arises from the parabolic profile of $u(y)$ in Equation (13.1). Mass balance over the entire film cross section yields:

$$\frac{\partial h}{\partial t} = -\frac{\partial q}{\partial x} \tag{13.2}$$

In a frame moving with the bubble speed U, the time derivative is converted into $-Uh_x$ in the moving frame. Integrating from the flat-film region where the third derivative vanishes into the transition region yields:

$$3\left(\frac{\mu U}{\sigma}\right)(h - h_\infty) = h^3 h_{xx} \tag{13.3}$$

Scaling h by the unknown flat-film thickness h_∞ and scaling the x coordinate by $h_\infty/(3Ca)^{1/3}$, we obtain the Bretherton equation:

$$H_{XXX} = \frac{H-1}{H^3} \tag{13.4}$$

This nonlinear equation for H describes the transition regions of both caps. However, the front one corresponds to $X \to \infty$, while the back one corresponds to $X \to -\infty$. The two asymptotic behaviors are not identical, indicating that the two caps are not mirror images. Nevertheless, as H blows up in both infinities, its third derivative must vanish according to Equation (13.4), and one expects quadratic blowup in both directions. These quadratic asymptotes must then be matched to the outer cap solutions. As h blows up, viscous effects become negligible and the outer caps are, to the leading order, just static solutions of the Laplace–Young equation. Without gravitational effects, these axisymmetric solutions are just spherical caps of radius R that make quadratic contact with the wall.

Linearizing about $H = 1$, the behavior away from the flat film is governed by three eigenvalues, 1 and $-\frac{1}{2} \pm \frac{\sqrt{3}}{2}i$. There is only monotonic blowup in the positive X direction due to a lone positive real eigenvalue. A numerical integration of Equation (13.4) yields the front cap asymptote:

$$H(X \to \infty) = \alpha^+ X^2 + \gamma^+ X + \beta^+ \tag{13.5}$$

The second coefficient can be changed due to an arbitrary shift of X but the quadratic coefficient is universal.

We then choose the origin of X until γ^+ vanishes. Equivalently, we can vary $H - 1$ with $H_X = H_{XX} = 0$ for the initial condition in our forward integration of Equation (13.4). This one-parameter iteration yields:

$$\alpha^+ = 0.32171 \qquad \beta^+ = 2.898 \tag{13.6}$$

Hence, the asymptotic curvature of the annular film toward the front cap is $H_{XX} = 2\alpha^+$ or, in the original dimensional coordinate:

$$h_{xx} = \frac{(3Ca)^{2/3}}{h_\infty} H_{XX} = \frac{2\alpha^+(3Ca)^{2/3}}{h_\infty} \tag{13.7}$$

This must match with the front spherical cap of radius R that makes the quadratic tangent with the capillary. Matching Equation (13.7) to this quadratic contact, we obtain the leading order estimate of the film thickness:

$$h_\infty/R = 0.6434(3Ca)^{2/3} \tag{13.8}$$

The back matching is more intricate. We note first that the complex eigenvalues suggest that the back film is undulating. A pronounced dimple due to this undulation is evident in the back profiles of Figure 13.1 computed by Lu and Chang (1988). This film oscillation is indeed confirmed by the photographs of Friz (1965). The arbitrary phase between these two complex modes must be specified. This extra degree of freedom is not present for the positive direction with only one real eigenvalue. Due to the quadratic contact of the back cap, we again iterate on the origin of X to obtain the back asymptote:

$$H(X \to -\infty) = \alpha^- X^2 + \beta^- \tag{13.9}$$

Because of the extra degree of freedom in the phase of the two complex conjugate modes, both α^- and β^- are functions of the phase, thus the pair (α^-, β^-) is a one-parameter family. To the leading order, this asymptote must also match a sphere of radius R that makes tangential contact with the capillary. Hence, $\alpha^- = \alpha^+ = 0.32171$. For this value of α^-, the corresponding value of β^- is $\beta^- = -0.8415$. (This is the most accurate estimate obtained by Chang and Demekhin, 1999. It is slightly different from many earlier values, including Bretherton's.)

The capillary pressure drops at the two caps arise from the $\beta^\pm$ terms. Consider the two spherical caps of radius $R' = R(1 + \varepsilon)$ different from the capillary radius R. Then, the expansion of the cap near the contact point, $\frac{dh(0)}{dx} = 0$, is:

$$h \sim \frac{x^2}{R} - R\varepsilon \tag{13.10}$$

Matching this expansion of the outer cap solution near the capillary to the two asymptotes of the inner film solutions, Equations (13.5) and (13.9), the front cap has a radius smaller than R, and the back cap has a larger radius. The difference is of order $Ca^{2/3}$, the scalings for H in both equations. Hence, the pressure drop across the entire bubble is the difference in the two cap capillary pressures σ/R':

$$\Delta p/(\sigma/R) = \frac{2}{1+\varepsilon^-} - \frac{2}{1+\varepsilon^+} \sim 2(\varepsilon^+ - \varepsilon^-)$$

$$= 2(\beta^+ - \beta^-)(h_\infty/R) = 10.0Ca^{2/3} \tag{13.11}$$

The scaling of this pressure drop is consistent with the order-of-magnitude arguments of the previous section. The unit-order coefficients are now specified by this classical matched asymptotic analysis. We note that an inner $X \ln X$ asymptotic behavior needs to be matched to similar expansions in the outer solution [Kalliadasis and Chang, 1996]. Such high-order matching becomes important only when contact lines appear.

13.3.1 Corrections to the Bretherton Results for Pressure-Driven Flow

At higher values of Ca, between 0.01 and 0.1, the film thickness and pressure drop across the bubble must be solved numerically instead of by matched asymptotics. This effort was carried out by Reinelt and Saffman (1985) and Lu and Chang (1988). The pressure drop can be correlated up to $Ca = 0.1$ as [Ratulowski and Chang, 1989]:

$$\Delta\rho/(\sigma/R) = 10.0Ca^{2/3} - 12.6Ca^{0.95} \tag{13.12}$$

However, the capillary number rarely exceeds 10^{-4} in microfluidics, and the Bretherton results of the previous section are usually adequate.

Bretherton finds his film thickness prediction to be smaller than the measured values at low *Ca*, exactly where the matched asymptotic analysis is most valid. This is confirmed by a series of experiments summarized by Ratulowski and Chang (1990), who attribute the deviation to Marangoni effects of surfactant contaminants that are most pronounced at the thin films of low *Ca*. The film thickness is determined only by how the front cap lays down a thin film by its capillary pressure. In this region, the film interface is stretched considerably, and the interfacial surfactant concentration decreases from the cap to the film. The film surface tension is then larger than the cap, and this Marangoni traction drags additional liquid into the film to thicken it.

For soluble surfactants, a complex model involving bulk-interface transport must be constructed to account for this new mechanism. For insoluble surfactants, a correction can be obtained almost trivially. In the limit of very small *Ca*, this traction approaches infinity, and the free surface in the transition region can be treated as a deformable but rigid interface that is laid onto the stagnant film. The velocity at the rigid interface vanishes, and the parabolic velocity of Equation (13.1) becomes:

$$u(y) = -\frac{\sigma}{\mu}\frac{\partial^3 h}{\partial x^3}\left(\frac{y^2}{2} - \frac{yh}{2}\right) \tag{13.13}$$

The flow rate q is then corrected by the factor of 4 due to the interface traction. The same correction yields a factor of 4 to the left-hand side of the Bretherton equation, Equation (13.3). Simply scale *Ca* by 4 and the same dimensionless Equation (13.4) results. Hence, in the limit of low *Ca*, soluble surfactants will correct the film thickness by a factor of $4^{2/3}$. This asymptote is approached by the experimental data in Figure 13.2 at low *Ca*. Ratulowski and Chang show that these asymptotic values at infinite traction are also the maximum values attainable for other more complex surfactant transport at low *Ca*. The correction to

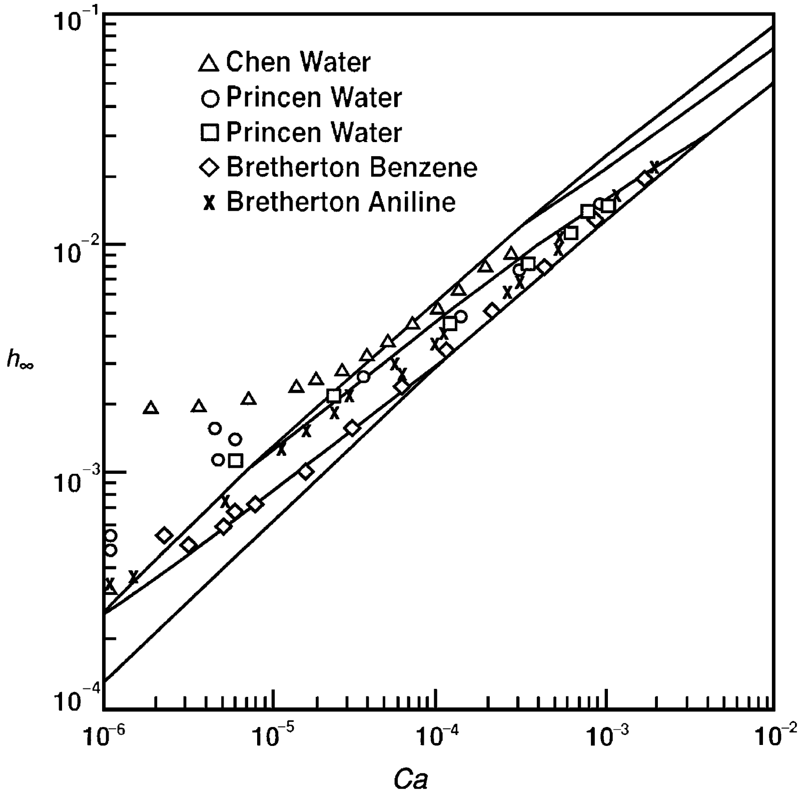


FIGURE 13.2 The film thickness of a bubble translating in various surfactant solutions. The capillary number *Ca* is a dimensionless speed, and the film thickness is scaled by the capillary radius. The theoretical curves correspond to different surfactant equilibrium constants between the interface and the bulk. At low *Ca*, they all approach the same asymptote derived in the text.

pressure drop is more intricate because it requires the resolution of the entire bubble. Because the surfactants accumulate at the back cap (or near a stagnation point near the back cap), correction requires a model for surfactant accumulation. Such a model was constructed by Park (1992) who then showed that the pressure drop across the bubble now has a $Ca^{1/3}$ scaling due to the accumulation. The pressure drop, increases by a factor of $Ca^{1/3}$ in the presence of surfactants.

One particularly interesting phenomenon concerning Marangoni effect is remobilization [Stebe et al., 1991] at high bulk surfactant concentrations when the entire interface can saturate even as it is being stretched. The Marangoni traction vanishes, and the mobile limit is again attained. This strategy reduces the pressure drop by only a factor of order unity and does not change the basic scalings.

For bubble trains whose bubbles are separated by thin lamellae instead of spherical caps (see Figure 13.3), Ratulowski and Chang (1989) show that the pressure drop remains constant to the leading order, while the film thickness decreases as adjacent bubbles are compressed (larger contact radius r_c in Figure 13.3. Geometric considerations clarify that a larger compression between adjacent bubbles will decrease the film thickness. An expansion of the Laplace–Young equation for the lamellae about zero contact radius shows that the film thickness is related to the free bubble thickness at $r_c = 0$ by:

$$h_\infty(r_c) = (1 - r_c)h_\infty(0) \tag{13.14}$$

Because the lamella is a constant curvature axisymmetric surface, its contribution to the curvatures of both asymptotes of the thin annular film is identical. The pressure drop across the bubble is independent of the contact radius.

Schwartz et al. (1986) examine drop transport and find that the thickness and pressure drop increase monotonically with respect to the viscosity ratio between the drop and the wetting fluid. The maximum occurs at infinitely large viscosity, corresponding to a solid drop, and the maximum is found to be larger than Bretherton's result by a factor of $2^{2/3}$. The difference between this correction factor and the Marangoni correction is a result of the differing films. The latter corresponds to a stationary rigid film while the former corresponds to a translating film. The Bretherton scaling results are robust estimates for circular capillaries. These estimates are only slightly corrected by Marangoni tractions due to surfactants, drop viscosity, and even bubble spacing.

The Bretherton scaling arguments break down for noncircular channels. Ratulowski and Chang (1989) examined the square channel numerically. Because the bubble caps of isolated bubbles are axisymmetric, contact must be made with the wall at low Ca, which is estimated to be at $Ca = 0.04$. Below this level, contact lines are expected, and the liquid does not wet the channel wall. Thus, favorable operating conditions only exist for Ca larger than 0.04, and the numerical results show that the film thickness and pressure drop show peculiar scaling:

$$h_\infty = 0.69 - 0.10 \ln Ca \qquad \Delta p/(\sigma/R) = 3.14Ca^{0.14} \tag{13.15}$$

The radius R corresponds to a cylindrical capillary with the same cross-section area.

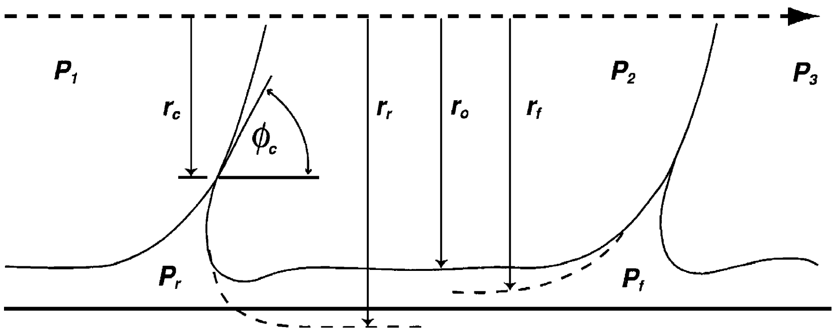


FIGURE 13.3 Schematic of a bubble train. The contact radius r_c represents the degree of compression.

Estimates for other channel geometries have not been computed, but the formulation by Ratulowski and Chang can be used. Solve the following two-dimensional unit cell equation for each dimensionless bubble radius *r*, scaled with respect to *R*:

$$\nabla^2 \psi = -1 \tag{13.16}$$

This fundamental solution is solved within a cross section of the straight capillary with a Dirichlet boundary condition at the capillary wall and Neumann condition at the circular interface with the dimensionless radius *r*. The flow rate–capillary pressure relation then becomes:

$$q = -K\frac{\partial p}{\partial x} = K(r)\frac{\partial^3 r}{\partial x^3} \tag{13.17}$$

The permeability constant $K(r)$ is the cross-section average of the previous fundamental solution multiplied by the factor $\sigma R^4/\mu$. A higher order version of the curvature can be used in place of the second derivative of *r*. To avoid contact between bubble and capillary, the capillary cross-section geometry must be nearly axisymmetric. As a result, one does not expect the pressure drop to be significantly different from Bretherton's estimate for circular capillaries, despite the difference in *Ca* scaling.

13.4 Bubble Transport by Electrokinetic Flow

The large pressure drop required to drive multiphase microchannel flow suggests the electrokinetic driving force is more desirable. If the electrokinetic flow behind the bubble is larger than that of the surrounding film, a high-pressure region can build up behind the bubble to drive it with the previously mentioned capillary pressure mechanism. The task is reducing the flow around the bubble without cutting the current required to drive the fluid. It is much easier to build the back pressure with electrokinetic flow than with pressure-driven flow behind the bubble because the required driving force is not as large. This is in direct contrast to single-phase channel flow where the hydrodynamic stress of the electrokinetic flow is confined to the thin double layer. As a result, the efficiency of single-phase electrokinetic flow is much lower than that of pressure-driven flow.

This design consideration requires some knowledge of electrokinetic flow [Russel et al., 1989; Probstein, 1994]. Electrokinetic flow occurs when the dielectric channel wall contains some surface charges that attract co-ions of opposite charge in the solution to a thin double layer of thickness λ. Also known as the Debye length, λ, ranges from 10 nm to microns depending on the bulk electrolyte concentration. The counter-ion concentration increases from the bulk value toward the wall within this double layer, while the co-ion decreases from its bulk value. Both bulk values are identical due to charge neutrality, therefore a net charge exists within the thin double layer. The total amount of this charge is determined through ionization equilibrium by the surface charge on the capillary.

Within the double layer, the potential ϕ is governed by the Poisson equation:

$$\frac{\partial^2 \phi}{\partial y^2} = \frac{F\rho}{\varepsilon} \tag{13.18}$$

The charge density is ρ, and the potential is set to zero at the bulk when *y* approaches infinity. The potential at the surface is called the zeta potential ζ. Due to the Boltzmann distributions of the co-ion and counter-ion, the counter-ion concentration increases much faster than the co-ion concentration decreases toward the wall. As a result, the total ion concentration in the double layer exceeds that in the bulk by a factor of $\exp(\zeta/kT)$, as seen in Figure 13.4. The charge density ρ also increases from zero at the bulk to a value at the wall equal to the bulk concentration multiplied by $\exp(\zeta/kT)$. Hence, at low ζ/kT, Equation (13.18) indicates that the scaling for λ is inversely proportional to the square root of the bulk ion concentration. By integrating Equation (13.18) over the double layer, its total charge scales linearly with respect to the

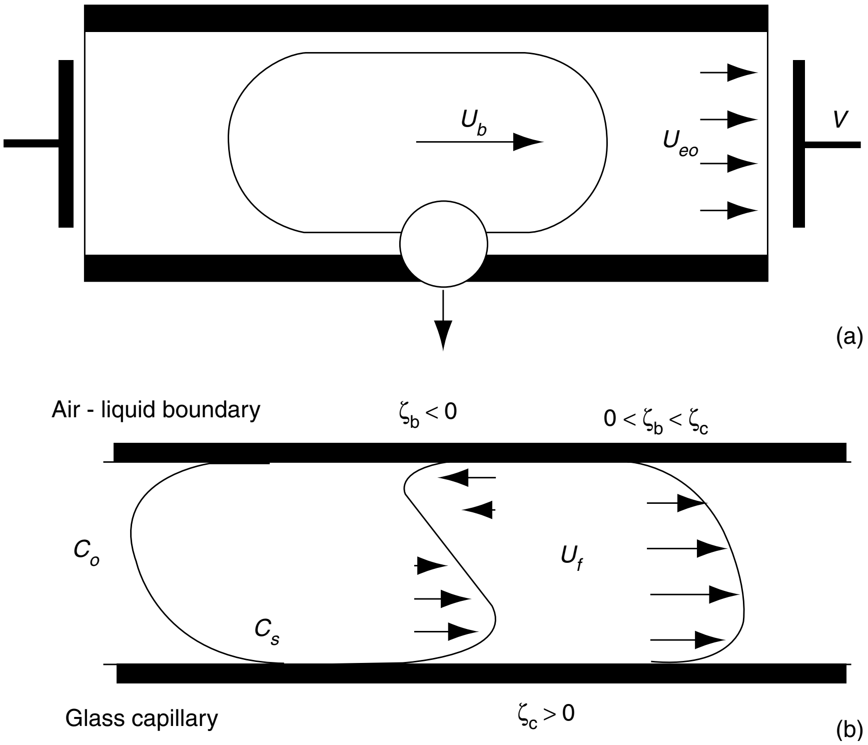


FIGURE 13.4 Electrokinetically driven bubble transport. The electrolyte ion concentration profile C_0 is shown with the velocity profiles U_f for both negative and positive zeta potentials at the bubble interface. The bubble translates with bubble speed U_b and the liquid with electrokinetic velocity U_{eo}.

potential gradient at the wall. The latter quantity scales as ζ/λ. One concludes that, for a given capillary–electrolyte pair, the zeta potential ζ scales as λ, inversely as the square root of the bulk electrolyte concentration.

In the presence of a tangential electric field E, there is a net body force on the electrolyte that scales as $E\rho$. This body force vanishes in the neutral bulk but accelerates the ions in the double layer to large speeds. These streaming ions drag the entire fluid body in the capillary along with them. The body force is concentrated in the thin double layer and acts like a surface force. The entire bulk liquid translates rigidly with a uniform tangential velocity, assuming there is no external pressure gradient. The momentum transfer in y for the tangential velocity field involves the viscous dissipation term ($\mu d^2u/d^2y$) balanced by the body force $E\rho$. Because this is in the same form as the Poisson equation, one sees that u scales linearly with respect to the electric potential ϕ but approaches a constant value away from the double layer. This asymptote is called the electrokinetic velocity:

$$u_c = -\frac{\varepsilon_0 \varepsilon \zeta_c E}{\mu} \tag{13.19}$$

The constants ε_0 and ε are the dielectric permittivities that we have omitted in the previous scaling arguments.

Because the electrokinetic velocity is flat away from the thin double layer (see Figure 13.4), the flow rate scales as the cross-section area, or R^2. For pressure-driven flow, the flow rate scales as the square of the area, or R^4. The electrokinetic velocity is independent of R, whereas the velocity of pressure-driven flow scales as the second power of R. Electrokinetic flow is much less efficient than pressure-driven flow, but electrokinetic flow is easier to scale up and down in microfluidic designs.

Unfortunately, the same flat electrokinetic velocity profile now serves to prevent cessation of film flow. By simple current–voltage calculation in the longitudinal direction, the local electric field E is shown to scale as the inverse of the cross-section area of the electrolyte across the capillary. By Equation (13.19),

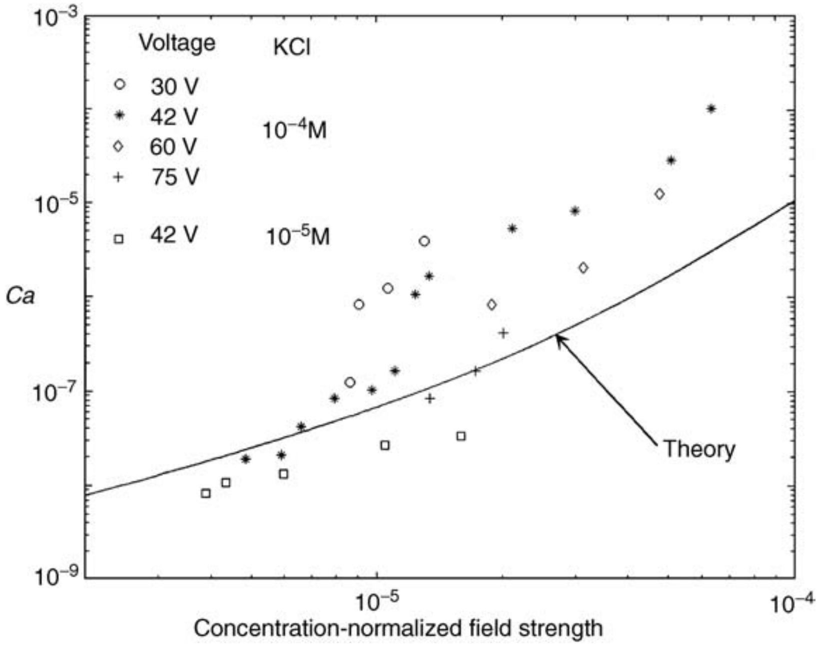


FIGURE 13.5 Electrokinetically driven bubble speed as a function of a concentration-normalized electric field for the KCl electrolyte of indicated concentrations. The unnormalized data scatter over 5 decades and are collapsed by the theory.

the electrokinetic velocity scales the same way. However, the flow rate scales as the electrokinetic velocity times the cross-section area and is independent of the cross-section area. The flow rate behind the bubble is the same as the flow rate in its surrounding film. As a result, there is no back pressure buildup, and the electrolyte simply flows around the bubble. This is observed when air bubbles are driven electrokinetically in a KCl/H_2SO_4 electrolyte (about 10^{-2} and 10^{-6} mol/L each) in a 5-cm-long glass capillary with a 1.0-mm inner diameter and with a voltage drop of 30 to 70 V [Takhistov et al., 2000]. At these conditions, the electrokinetic velocity of the electrolyte is 0.1 to 1.0 mm/sec, yet the bubble remains stationary as the electrolyte flows past it.

There are several possible means of breaking the flow rate invariance to cross-section area in order to reduce the film flow. One can endow the interface with traction by using finite viscosity drops or interfacial surfactants so that the film profile is no longer flat. The longitudinal electric field in the film can be reduced by lowering the electrolyte concentration such that the thickened double-layer thickness approaches that of the film. As a result, the higher ion concentration within the double layer can increase the film conductivity beyond the bulk value. More intriguingly, one can use an ionic surfactant to endow a double layer at the interface that has a different charge from the capillary double layer. The velocity at the interface is not zero in the moving frame but is negative (see Figure 13.4). This could effectively reduce the film flow to zero.

The surfactants act as a valve to film flow that requires no pressure expenditure and film flow leakage. The bubble front does not produce a pressure drop that counters the one in the back of the bubble. In contrast, the Bretherton problem in pressure-driven bubble flow requires a near-cancellation of these pressure drops, resulting in the small pressure buildup of Equation (13.12) and a similarly small bubble velocity. Takhistov et al. (2002) have experimentally established this major advantage of displacing air bubbles with electrokinetic flow.

Because the glass capillary surface of Takhistov et al.'s experiment has a positive charge such that its double layer contains a negative charge, an anionic surfactant, sodium dodecyl sulfate (SDS), tests the previous idea. Most glass surfaces are negatively charged, but the charge is reversed through chemical treatment to allow an interfacial double layer of the opposite charge. About 10^{-5} mol/L of the surfactant is added and, after some equilibration time, the bubbles begin to move. In Figure 13.5, the measured bubble

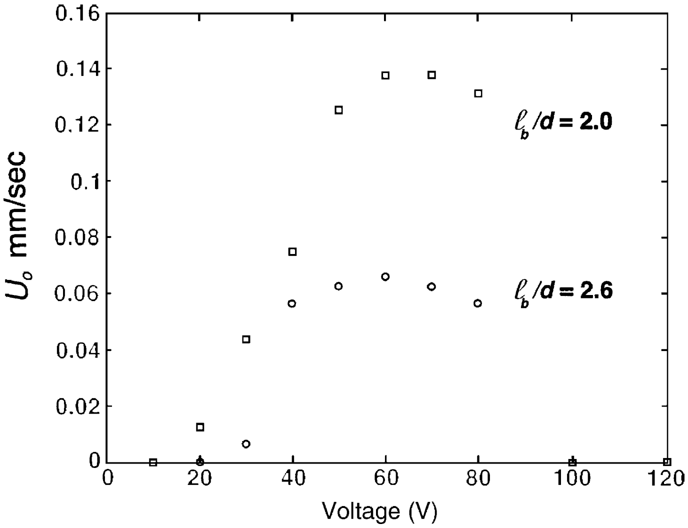


FIGURE 13.6 Raw bubble speed data U_o as a function of applied voltage and bubble length l_b normalized by the capillary diameter d. The bubble speed approaches that of the electrolyte electrokinetic velocity without bubble before dropping to zero abruptly at a critical voltage of 80 V.

speed Ca is recorded as a function of the electrolyte concentration, applied field, and the surfactant concentration. The last quantity is presented as a concentration-normalized field obtained from an electrokinetic theory [Takhistov et al., 2000]. Bubble speeds approaching the liquid electrokinetic velocity (without bubble) of $Ca = 10^{-4}$ are observed, indicating a complete cessation of film flow. Figure 13.6 shows a more specific set of data in dimensional quantities, showing a robust 0.14-mm/sec bubble speed.

There are limitations to such electrokinetically driven bubble flow in micochannels. The interfacial zeta potential endowed by the surfactants is a strong function of the electrolyte concentration due to the strong screening effects near the anions of the surfactants [Schultz, 1984]. At high concentrations, the interface double layer can become negligibly thin. As a result, the film velocity approaches a flat profile, and the bubble speed approaches zero. At very low bulk electrolyte concentrations, the Debye thickness approaches the film thickness. As a result, the ion concentration and conductivity in the film increase by a factor of $\exp(\zeta/kT)$. Because ζ is large at low concentrations, film conductivity increases significantly. This reduces the field strength E and further reduces the film flow. However, because the interfacial and capillary double layers are oppositely charged, their increased thickness and coulombic attraction to each other will eventually collapse the entire film. The experimental data for different KCl concentrations shown in Figure 13.7 demonstrates these limits. Bubble speeds approach zero at high concentrations, and this phenomenon suggests that interfacial traction provided by the surfactants is negligible. Therefore, a wide but finite window of electrolyte concentration exists where multiphase microfluidic flow can be achieved with an electrokinetic driving force.

The Bretherton analysis is extended to electrokinetic flow [Takhistov et al., 2000]. The theory now includes the electrolyte concentration dependence of the zeta potential and the important interfacial double layer. The resulting theoretical predictions collapse the data in Figure 13.5 and provide accurate estimates of the data in Figures 13.6 and 13.7. Also included in the theory are the transients necessary to establish a steadily translating bubble. Surfactant adsorption equilibration at the interface and capillary double-layer equilibration are just two of the important transients that must be considered for the design of microdevices.

13.5 Future Directions

Electrokinetic flow is the only means of overcoming the large capillary forces involved in transporting bubbles in microchannels. Electrokinetic displacement of bubbles in a circular capillary is possible only

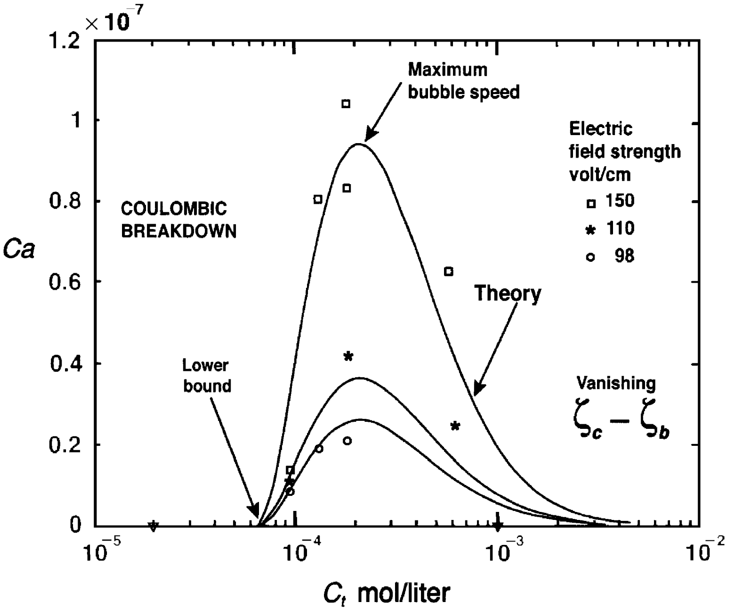


FIGURE 13.7 The window of total ion concentration C_t where bubble motion is possible.

within certain windows of operation. For noncircular channels, current and flow leakage at the corners are additional concerns. Such leaks must be prevented to have complete cessation of film flow. Further complications arise at channel junctions or constrictions where bubbles may break up or coalesce. Jet streaming of small drops from the front tip of a bubble being sheared in front of a constriction has also been observed. Most of these phenomena must be carefully avoided in the device design.

The rapid ion motion in the double layers of electrokinetic flow has certain desirable applications. We observed charge separation along the bubble that is sufficient to break up the bubble. This observation might suggest a means of electrophoretic bubble motion and bubble breakup in microreactors. The same polarization also induces bubble coalescence. Streaming potentials with opposite flows is another possibility but a rather difficult task because of the capillary pressure. Evaporation and condensation phenomena are also profoundly different in microchannels due to capillary forces. The DC field used to drive the electrokinetic flow can produce bubbles via electrode reactions that sustain the DC current. This mechanism could be used to generate the bubbles, as is done in some bubble pumps. In most cases, however, the electrodes should be separated from the microchannel by high-permittivity membranes to prevent bubble penetration. Biological cells with internal and external charges and of the same dimension as the microchannels exhibit a rich spectrum of electrophoretic, electrokinetic, and stress-induced adsorption dynamics in microfluidics. These and other known and new phenomena of ionic flow await future studies for applications in microdevice designs.

Acknowledgments

It has been my privilege to work with a remarkably talented group of Ph.D. and postdoctoral students in our studies of bubble transport. They are, in chronological order, J. Ratulowski (Shell), S. Kalliadasis (Leeds), A. Indeikina (Notre Dame), P. Takhistov (Rutgers), and L. Yeo (Notre Dame). Our research has been supported by grants from ACS-PRF and NSF (XYZ-on-a-Chip).

References

Bretherton, F.P. (1961) "The Motion of Long Drops and Bubbles in Tubes," *J. Fluid Mech.* **10**, pp. 166–88.

Chang, H.-C., and Demekhin, E.A. (1999) "Mechanism for Drop Formation on a Coated Vertical Fibre," *J. Fluid Mech.* **380**, pp. 233–55.

Friz, V.G. (1965) "Über den dynamischen Randwindel im Fall der vollstädigen Benetzung," *A. für angewandto Phys.* **19**, pp. 374–8.

Indeikina, A., and Chang, H.-C. (1999) "A Molecular Theory for Dynamic Contact Angles," *Proc. IUTAM Symp. on Nonlinear Singularities in Deformation and Flow*, D. Durban and J.R.A. Pearson, eds., Kluwer Academic, Dordrecht/Norwell, MA, pp. 321–68.

Kalliadasis, S., and Chang, H.-C. (1994) "Apparent Dynamic Contact Angle of an Advancing Gas–Liquid Meniscus," *Phys. Fluids.* **6**, pp. 12–23.

Kalliadasis, S., and Chang, H.-C. (1996) "Effects of Wettability on Spreading Dynamics," *Ind. Eng. Chem. Fluid* **35**, pp. 2860–74.

Lu, W.-Q., and Chang, H.-C. (1988) "A Boundary Integral Study of Bubble Formation and Transport in Channels Filled with a Viscous Fluid," *J. Comput. Phys.* **340**, pp. 77–89.

Park, C.-W. (1992) "Influence of Soluble Surfactants on the Motion of a Finite Bubble in a Capillary," *Phys. Fluids A* **4**, pp. 2335–47.

Probstein, R.F. (1994) *Physiochemical Hydrodynamics*, John Wiley & Sons, New York.

Ratulowski, J., and Chang, H.-C. (1989) "Transport of Gas Bubbles in Capillaries," *Phys. Fluids A* **1**, pp. 1642–55.

Ratulowski, J., and Chang, H.-C. (1990) "Maragoni Effects of Trace Impurities on the Motion of Long Gas Bubbles in Capillaries," *J. Fluid Mech.* **210**, pp. 303–28.

Reinelt, D.A., and Saffman, P.G. (1985) "The Penetration of a Finger into a Viscous Fluid in a Channel and Tube," *SIAM J. Stat. Comp.* **6**, pp. 542–61.

Russel, W.B., Saville, D.A., and Schowalter, W.R. (1989) *Colloidal Dispersion*, Cambridge University Press, Cambridge, U.K.

Schultz, H.J. (1984) *Physico-Chemical Elementary Processes in Flotation*, Elsevier, New York.

Schwartz, L.W., Princen, H.M., and Kiss, A.D. (1986) "On the Motion of Bubbles in Capillary Tubes," *J. Fluid Mech.* **172**, pp. 259–75.

Stebe, K.S., Lin, S.Y., and Maldarelli, C. (1991) "Remobilizing Surfactant Retarded Particle Interfaces," *Phys. Fluids A* **3**, pp. 3–20.

Takhistov, P., Indeikina, A., and Chang, H.-C. (2000) "Electrokinetically Driven Bubbles in Microchannels," *Phys. Fluids.* **14**, pp. 1–14.

Veretennikov, I., Indeikina, A., and Chang, H.-C. (1998) "Front Dynamics and Fingering of a Driven Contact Line," *J. Fluid Mech.* **373**, pp. 81–110.

14

Fundamentals of Control Theory

J. William Goodwine
University of Notre Dame

14.1 Introduction

This chapter reviews the fundamentals of linear and nonlinear control. This topic is particularly important in microelectromechanical systems (MEMS) applications for two reasons. First, as electromechanical systems, MEMS devices often must be controlled in order to be used in an effective manner. Second, important applications of MEMS technology are controls-related because of the utility of MEMS devices in sensor and actuator technologies. Because the area of control is far too vast to be entirely presented in one chapter, this chapter outlines a variety of techniques used for control system synthesis and analysis, provides at least a brief description of their mathematical foundation, discusses the advantages and disadvantages of the techniques, and provides references for the reader. The material varies from the basic (e.g., root locus design) to relatively advanced material (e.g., sliding mode control) to cutting-edge research (hybrid systems). Some examples are provided, and many references to the literature are provided to help the reader find additional examples of a particular analysis or synthesis technique.

This chapter is divided into three sections, all of which consider the stability and performance of a control system. The term performance includes: the qualitative nature of any transient response of the system, the reference signal tracking properties of the system, and the long-term or steady-state performance of the system. The first section considers classical control, which is the study of single-input, single-output (SISO) linear control systems. This section relies heavily upon mathematical techniques from complex variable theory, and outlines what is typically covered in an undergraduate controls course.

The second section considers so-called "modern control," which is the study of multi-input, multi-output (MIMO) control systems in state space. This section also includes what is sometimes called "post-modern control" [Zhou, 1996], which is a study of robust system performance and stability in the presence of unmodeled system dynamics. Finally, the third section considers nonlinear control techniques. Model-free control techniques based upon concepts from soft computing are outlined in Chapter 16. Nonlinear, open-loop control techniques are not covered in this chapter. (For recent advances in this area, refer to Lafferriere and Sussmann [1993], Bullo et al. [2000], and Goodwine and Burdick [2000].)

14.2 Classical Linear Control

Classical linear control relies heavily upon mathematical techniques from complex variable theory. This reliance is a historical consequence of the importance of frequency analyses of feedback amplifiers, which motivated much of the development of classical control theory. In addition, this reliance is a consequence of the fact that convolution in the time domain is simple multiplication in the frequency domain, which greatly simplifies the analysis of the natural input–output and "block diagram" structure of many control systems. Good references include Dorf (1992), Franklin et al. (1994), Gajec and Lelic (1996), Kuo (1995), Ogata (1997), Raven (1995), and Shinners (1992).

14.2.1 Mathematical Preliminaries

The main mathematical tool in classical linear control theory is the Laplace transform, which transforms the linear ordinary differential equation (ODE) into an algebraic equation, thus reducing the task of solving an ODE into simple algebra. The Laplace transform of a function $f(t)$ is defined as:

$$L[f(t)] = F(s) = \int_0^\infty e^{-st} f(t) dt \tag{14.1}$$

and the inverse Laplace transform of $F(s)$ as:

$$L^{-1}[F(s)] = f(t) = \frac{1}{2\pi j} \int_{c-j\omega}^{c+j\omega} F(s) e^{st} ds, \quad \text{for } t > 0 \tag{14.2}$$

A discussion of important mathematical details concerning convergence and the proper lower limit of integration is found in Ogata (1997). Evaluating the integrals in the definition of the Laplace transform and the inverse Laplace transform is rarely necessary because extensive tables of Laplace transform pairs are readily available. A few Laplace transform pairs for typical functions are listed in Table 14.1. More complete tables can be found in any undergraduate text on classical control theory such as the references listed previously.

Important properties of the Laplace transform are as follows:

1. Real differentiation: $L[\frac{d}{dt} f(t)] = sF(s) - f(0)$.
2. Linearity: $L[\alpha f_1(t) \pm \beta f_2(t)] = \alpha F_1(s) \pm \beta F_2(s)$.
3. Convolution: $L\left[\int_0^t f_1(t-\tau) f_2(\tau) d\tau\right] = F_1(s) F_2(s)$.
4. Final value theorem: If all the poles of $sF(s)$ are in the left half of the complex plane, then $\lim_{t\to\infty} f(t) = \lim_{s\to 0} sF(s)$.

A basic result from the first three properties is that to solve a linear ODE, one can take the Laplace transform of each side of the equation, which converts the differential equation into an algebraic equation in s. Then algebraically solve the expression for the Laplace transform of the dependent variable, and take the inverse Laplace transform of the resulting function.

TABLE 14.1 Laplace Transform Pairs for Basic Functions

	$F(t)$	$F(s)$
1	Unit impulse, $\delta(t)$	1
2	Unit step, $1(t)$	$\frac{1}{s}$
3	t	$\frac{1}{s^2}$
4	$t^n, n = 1, 2, 3, \ldots$	$\frac{n!}{s^{n+1}}$
5	e^{-at}	$\frac{1}{s+a}$
6	$t^n e^{-at}$	$\frac{n!}{(s+a)^{n+1}}$
7	$\sin \omega t$	$\frac{\omega}{s^2+\omega^2}$
8	$\cos \omega t$	$\frac{s}{s^2+\omega^2}$
9	$e^{-at}\cos bt$	$\frac{s+a}{(s+a)^2+b^2}$
10	$e^{-at}\sin bt$	$\frac{b}{(s+a)^2+b^2}$

Example

As a simple example, consider the differential equation:

$$\begin{aligned} \ddot{x} + x &= 0 \\ x(0) &= 0 \\ \dot{x}(0) &= 1 \end{aligned} \tag{14.3}$$

Taking the Laplace transform of the equation yields:

$$s^2X(s) - sx(0) - \dot{x}(0) + X(s) = 0 \tag{14.4}$$

Algebraic manipulation gives:

$$X(s) = \frac{1}{s^2 + 1} \tag{14.5}$$

Consequently, from the table of Laplace transform pairs:

$$x(t) = \sin(t) \tag{14.6}$$

For more examples, see Ogata (1997), Raven (1995), Kuo (1995), and Franklin et al. (1994).

Due to the convolution property of Laplace transforms, a convenient representation of a linear control system is the block diagram illustrated in Figure 14.1. In such a block diagram, each block contains the Laplace transform of the differential equation representing that component of the control system that relates the block's input to its output. Arrows between blocks indicate that the output from the preceding block is transferred to the input of the subsequent block. The output of the preceding block multiplies the contents of the block to which it is an input. Simple algebra will yield the overall transfer function of a block diagram representation for a system.

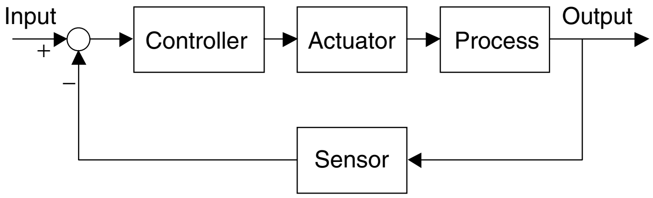


FIGURE 14.1 Typical block diagram representation of a control system.

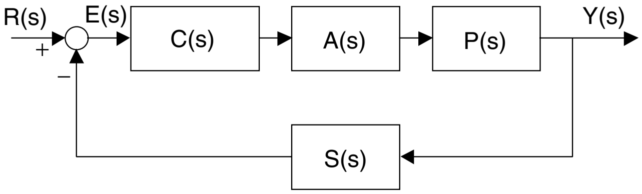


FIGURE 14.2 Generic block diagram including transfer functions.

Example

The transfer function for the system illustrated in Figure 14.2 can be computed by observing that:

$$E(s) = R(s) - Y(s)S(s) \tag{14.7}$$

and

$$Y(s) = E(s)C(s)A(s)P(s) \tag{14.8}$$

which can be combined to yield

$$\frac{Y(s)}{R(s)} = \frac{C(s)A(s)P(s)}{1 + C(s)A(s)P(s)S(s)} \tag{14.9}$$

A more complete exposition on block diagram algebra can be found in any of the previously cited undergraduate texts. Note that the numerator and denominator of the transfer function will typically be polynomials in s. The denominator is called the characteristic equation for the system. As entry 5 in Table 14.1 shows, if the characteristic polynomial has any roots with a positive real part, then the system will be unstable because it will correspond to an exponentially increasing solution. Given a reference input $R(s)$, determine the response of the system by multiplying the transfer function by the reference input, and perform a partial fraction expansion (i.e., expand):

$$Y(s) = \frac{R(s)C(s)A(s)P(s)}{1 + C(s)A(s)P(s)S(s)} = \frac{C_1}{s - p_1} + \frac{C_2}{s - p_2} + \cdots + \frac{C_n}{s - p_n} \tag{14.10}$$

where each term in the sum on the right-hand side of the equation is similar to one of the entries in Table 14.1. The contribution to the response of each individual term can be determined by referring to a Laplace transform table and can be superimposed to determine the overall solution:

$$y(t) = y_1(t) + y_2(t) + \cdots + y_n(t) \tag{14.11}$$

where each term in the sum is the inverse Laplace transform of the corresponding term in the partial fraction expansion.

Example

For the block diagram in Figure 14.2 if $C(s) = \dfrac{1}{s}$, $A(s) = 1$, $P(s) = \dfrac{\omega_n^2}{s + 2\zeta\omega_n}$, $S(s) = 1$ and $R(s) = \dfrac{1}{s}$ (a unit step input), then:

$$\begin{aligned}
Y(s) &= \frac{\omega_n^2}{s(s^2 + 2\zeta\omega_n s + \omega_n^2)} \\
&= \frac{1}{s} - \frac{s + 2\zeta\omega_n}{s^2 + 2\zeta\omega_n s + \omega_n^2} \\
&= \frac{1}{s} - \frac{s + \zeta\omega_n}{(s + \zeta\omega_n)^2 + \omega_d} - \frac{\zeta\omega_n}{(s + \zeta\omega_n)^2 + \omega_d} \\
&= \frac{1}{s} - \frac{s + \zeta\omega_n}{(s + \zeta\omega_n)^2 + \omega_d} - \frac{\zeta\omega_n}{\omega_d}\,\frac{\omega_d}{(s + \zeta\omega_n)^2 + \omega_d} \qquad (14.12)
\end{aligned}$$

where $\omega_d = \omega_n\sqrt{1-\zeta^2}$. Referring to Table 14.1 of Laplace transform pairs and assuming that $\zeta < 1$,

$$y(t) = 1 - e^{-\zeta\omega_n t}\left(\cos(\omega_d t) + \frac{\zeta}{\sqrt{1-\zeta^2}}\sin(\omega_d t)\right) \qquad (14.13)$$

14.2.2 Control System Analysis and Design

Control system analysis and design consider primarily stability and performance. The stability of a system with the closed-loop transfer function (note that in such a case a controller has already been specified):

$$T(s) = \frac{b_0 s^m + b_1 s^{m-1} + \cdots + b_m}{s^n + a_1 s^{n-1} + \cdots + a_n} \qquad (14.14)$$

is determined by the roots of the denominator, or characteristic equation. It is possible to determine whether the system is stable without actually computing the roots of the characteristic equation. A necessary condition for stability is that each of the coefficients a_i appearing in the characteristic equation be positive. Because this is a necessary condition, if any of the a_i are negative, then the system is unstable, but the converse is not necessarily true. Even if all the a_i are positive, the system may still be unstable. Routh (1975) devised a method to check necessary and sufficient conditions for stability.

The method is to construct the Routh array, defined as follows:

$$\begin{array}{llllll}
\text{Row } n & s^n: & 1 & a_2 & a_4 & \ldots \\
\text{Row } n-1 & s^{n-1}: & a_1 & a_3 & a_5 & \ldots \\
\text{Row } n-2 & s^{n-2}: & b_1 & b_2 & b_3 & \ldots \\
\text{Row } n-3 & s^{n-3}: & c_1 & c_2 & c_3 & \ldots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\text{Row } 2 & s^2: & * & * & & \\
\text{Row } 1 & s^1: & * & & & \\
\text{Row } 0 & s^0: & * & & &
\end{array}$$

in which the a_i are from the denominator of Equation (14.14). b_i and c_i are defined as:

$$b_1 = -\frac{\det\begin{bmatrix} 1 & a_2 \\ a_1 & a_3 \end{bmatrix}}{a_1} \quad b_2 = -\frac{\det\begin{bmatrix} 1 & a_4 \\ a_1 & a_5 \end{bmatrix}}{a_1} \quad b_3 = -\frac{\det\begin{bmatrix} 1 & a_6 \\ a_1 & a_7 \end{bmatrix}}{a_1}$$

$$c_1 = -\frac{\det\begin{bmatrix} a_1 & a_3 \\ b_1 & b_2 \end{bmatrix}}{b_1} \quad c_2 = -\frac{\det\begin{bmatrix} a_1 & a_5 \\ b_1 & b_3 \end{bmatrix}}{b_1} \quad c_3 = -\frac{\det\begin{bmatrix} a_1 & a_7 \\ b_1 & b_4 \end{bmatrix}}{b_1}$$

The basic result is that the number of poles in the right-half plane (i.e., unstable solutions) is equal to the number of sign changes among the elements in the first column of the Routh array. If they are all positive, the system is stable. When a zero is encountered, it should be replaced with a small positive constant ε which will then be propagated to lower rows in the array. The result can be obtained by taking the limit as $\varepsilon \rightarrow 0$.

Example

Construct the Routh array and determine the stability of the system described by the transfer function:

$$\frac{Y(s)}{R(s)} = \frac{1}{s^4 + 4s^3 + 9s^2 + 10s + 8} \tag{14.15}$$

The Routh array is

$$\begin{array}{llll} s^4: & 1 & 9 & 8 \\ s^3: & 4 & 10 & 0 \\ s^2: & \dfrac{-(10-36)}{1} = 26 & \dfrac{-(0-32)}{1} = 32 & 0 \\ s^1: & \dfrac{-(128-260)}{4} = 33 & 0 & 0 \\ s^0: & \dfrac{-(0-1056)}{26} = 40.6 & 0 & 0 \end{array} \tag{14.16}$$

The system is stable because there are no sign changes in the elements in the first column of the array.

One aspect of performance concerns the steady-state error exhibited by the system. For example, from the time-domain solution of the previous example, as $t \rightarrow \infty$, $y(t) \rightarrow 1$. However, the final value theorem can be used to determine this without actually solving for the time-domain solution.

Example

Determine the steady-state value for the time-domain function $y(t)$ if its Laplace transform is given by $Y(s) = \omega_n^2/s(s^2 + 2\zeta\omega_n s + \omega_n^2)$. Because all the solutions of $s^2 + 2\zeta\omega_n s + \omega_n^2 = 0$ have a negative real part, all the poles of $sY(s)$ lie in the left half of the complex plane. Therefore, the final value theorem can be applied to yield:

$$\lim_{t\rightarrow\infty} y(t) = \lim_{s\rightarrow 0} sY(s) = \lim_{s\rightarrow 0} s\frac{\omega_n^2}{s(s^2 + 2\zeta\omega_n s + \omega_n^2)} = 1 \tag{14.17}$$

which is identical to the limit of the time-domain solution as $t \rightarrow \infty$.

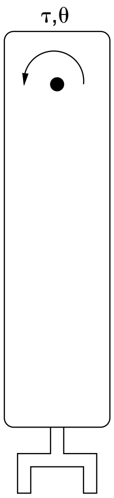


FIGURE 14.3 Robot arm model.

14.2.2.1 Proportional–Integral–Derivative (PID) Control

Perhaps the most common control implementation is so-called proportional–integral–derivative (PID) control, where the commanded control input (the output of the "controller" box in Figures 14.1 and 14.2) is equal to the sum of three terms: one term proportional to the error signal (the input to the "controller" box in Figures 14.1 and 14.2), the next term proportional to the derivative of the error signal, and the third term proportional to the time integral of the error signal. From Figure 14.2, $C(s) = K_P + (K_I/s) + K_d s$, where K_P is the proportional gain, K_I is the integral gain, and K_d is the derivative gain. A simple analysis of a second-order system shows that increasing K_P and K_I generally increases the speed of the response at the cost of reducing stability. Increasing K_d generally increases damping and stability of the response. With $K_I = 0$, there may be a nonzero steady-state error, but when K_I is nonzero, the effect of the integral control effort is to typically eliminate steady-state error.

Example — PID Control of a Robot Arm

Consider a robot arm illustrated in Figure 14.3. Linearizing the equations of motion about $\theta = 0$ (the configuration in Figure 14.3) gives:

$$I\ddot{\theta} + mg\theta = u \tag{14.18}$$

where I is the moment of inertia of the arm, m is the mass of the arm, θ is the angle of the arm, and u is a torque applied to the arm. For PID control,

$$u = K_p(\theta_{\text{desired}} - \theta_{\text{actual}}) + K_d(\dot{\theta}_{\text{desired}} - \dot{\theta}_{\text{actual}}) + K_I\int_0^t (\theta_{\text{desired}} - \theta_{\text{actual}})dt \tag{14.19}$$

If $I = 1$ and $m = 1/g$, the block diagram representation for the system is illustrated in Figure 14.4. Thus, the closed-loop transfer function is

$$T(s) = \frac{K_d s^2 + K_p s + K_I}{s^3 + K_d s^2 + (K_p + I)s + K_I} \tag{14.20}$$

Figure 14.5 illustrates the step response of the system for proportional control ($K_P = 1$, $K_I = 0$, $K_d = 0$), PD control ($K_P = 1$, $K_I = 0$, $K_d = 1$), and PID control ($K_P = 1$, $K_I = 1$, $K_d = 1$). Note that for proportional and PD controls, there is a final steady-state error that is eliminated with PI control. (Also note that both of these facts could be verified analytically using the final value theorem.) Finally, note that

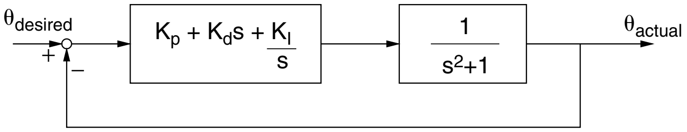


FIGURE 14.4 Robot arm block diagram.

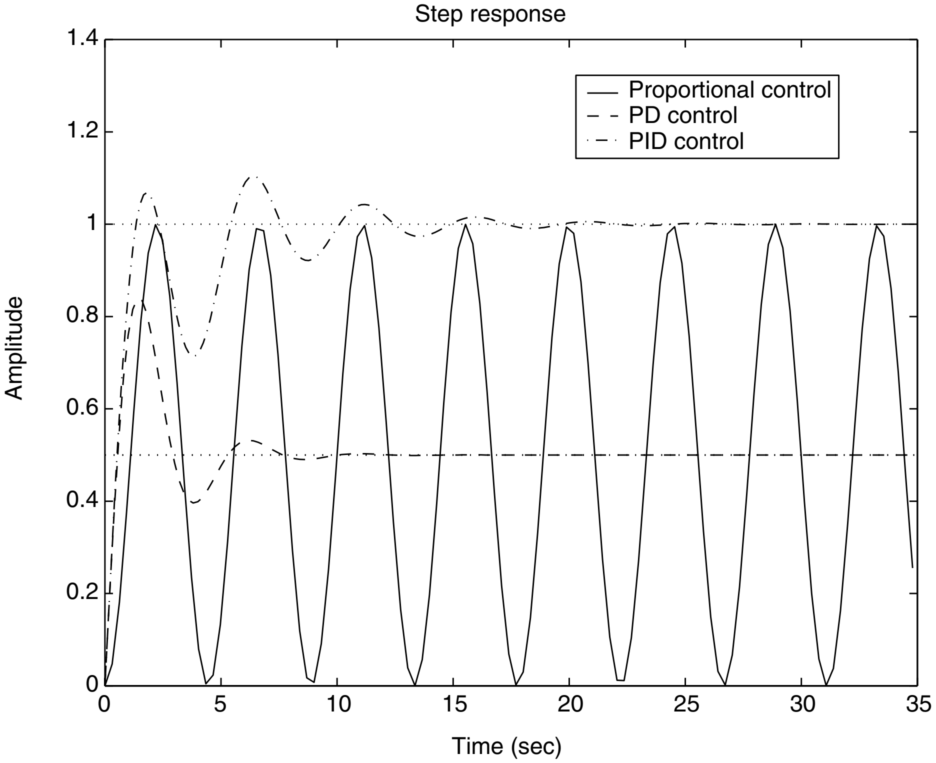


FIGURE 14.5 PID control response.

the system response for pure proportional control is oscillatory, whereas with derivative control the response is much more damped.

The subjects contained in the subsequent sections consider controller synthesis issues. For PID controllers, tuning methods exist. Refer to the undergraduate texts cited previously or to the papers by Ziegler and Nichols (1942, 1943).

14.2.2.2 The Root Locus Design Method

As mentioned previously in the discussion of PID control, various rules of thumb can be determined to relate system performance to changes in gains, however, a systematic approach is more desirable. Because pole locations determine the characteristics of the response of the system (recall the partial fraction expansion), one natural design technique is to plot how pole locations change as a system parameter or control gain is varied [Evans, 1948, 1950]. Because the real part of the pole corresponds to exponential solutions, if all the poles are in the left-half plane, the poles closest to the $j\omega$-axis will dominate the system response. If we focus a second-order system of the form:

$$H(s) = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2} \tag{14.21}$$

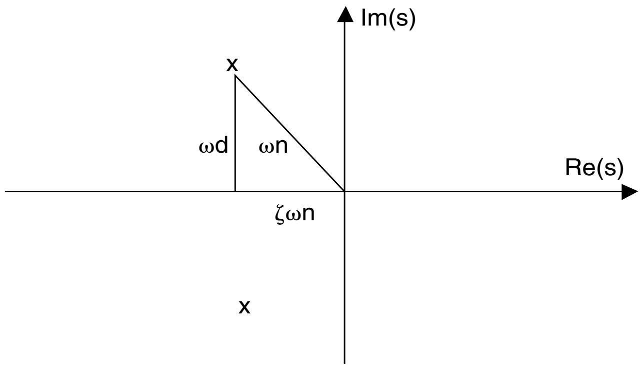


FIGURE 14.6 Complex conjugate poles, natural frequency, damped natural frequency, and damping ratio.

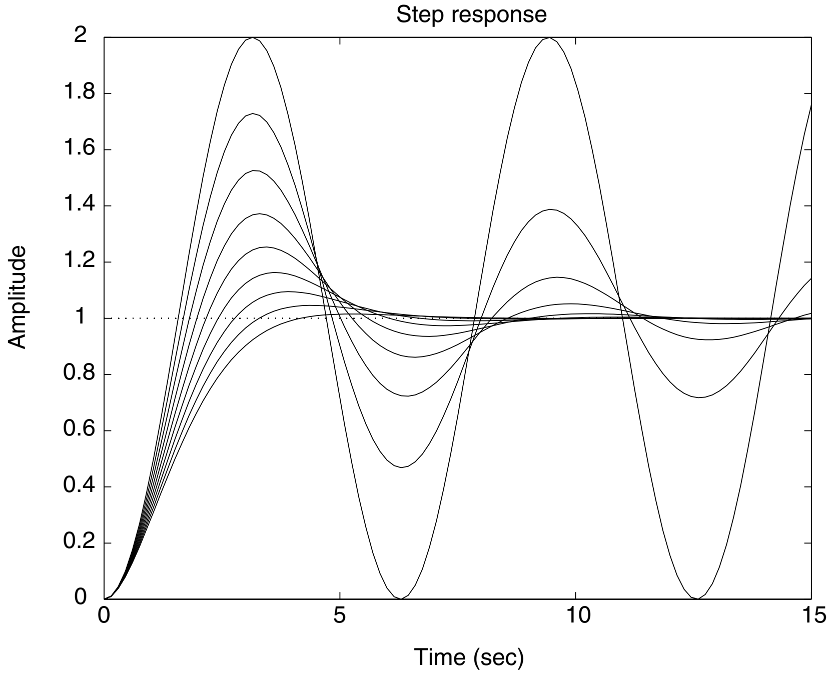


FIGURE 14.7 Step response for various damping factors.

the poles of the system are as illustrated in Figure 14.6. The terms ω_n, ω_d, and ζ are the natural frequency, the damped natural frequency, and the damping ratio, respectively. Multiplying $H(s)$ by $1/s$ (unit step), and performing a partial fraction expansion give:

$$Y(s) = \frac{1}{s} - \frac{s + \zeta\omega_n}{(s + \zeta\omega_n)^2 + \omega_n^2(1 - \zeta^2)} - \frac{\zeta\omega_n}{(s + \zeta\omega_n)^2 + \omega_n^2\,(1 - \zeta^2)} \tag{14.22}$$

so the time response for the system is

$$y(t) = 1 - e^{-\zeta\omega_n t}\left(\cos\omega_d t + \frac{\zeta}{\sqrt{1 - \zeta^2}}\sin\omega_d t\right) \tag{14.23}$$

where $\omega_d = \omega_n\sqrt{1-\zeta^2}$ and $0 \leq \zeta < 1$. Figure 14.7 illustrates plots of the response for various values of ζ. Referring to the previous equation and Figure 14.7, if the damping ratio is increased, the oscillatory nature of the response is increasingly damped.

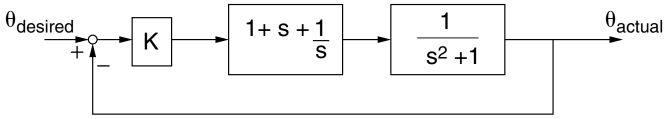


FIGURE 14.8 Robot arm block diagram.

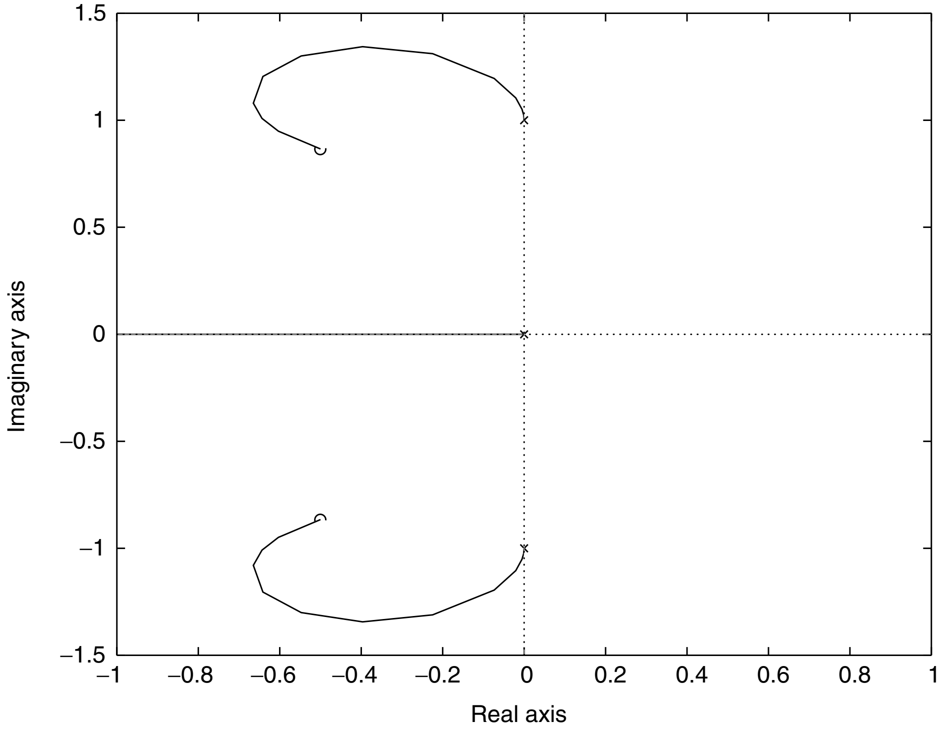


FIGURE 14.9 Root locus for robot arm PID controller.

Because the natural frequency and damping are directly related to the location of the poles, one effective approach to designing controllers is picking control gains based upon desired pole locations. A root locus plot is a plot of pole locations as a system parameter or controller gain is varied. Once the root locus has been plotted, pick the location on the root locus with the desired pole locations to give the desired system response. There is a systematic procedure to plot the root locus by hand (refer to the cited undergraduate texts), and computer packages such as Matlab (using the `rlocus()` and `rlocfind()` functions) make it even easier. Figure 14.9 illustrates a root locus plot for the previously noted robot arm with the block diagram as the single gain K is varied from 0 to ∞ as illustrated in Figure 14.8. Note that for the usual root locus plot, only one gain can be varied at a time. In the previous example, the ratio of the proportional, integral, and derivative gains was fixed, and a multiplicative scaling factor was varied in the root locus plot.

Because the roots of the characteristic equation start at each pole when $K = 0$ and approach each 0 of the characteristic equation as $K \to \infty$, the desired K can be determined from the root locus plot by finding the part of the locus that most closely matches the desired natural frequency ω_n and damping ratio ζ (recall Figure 14.7).

Typically, control system performance is specified in terms of time-domain conditions, such as rise time, maximum overshoot, peak time, and settling time, all of which are illustrated in Figure 14.10. Rough estimates of the relationship between the time-domain specifications and the natural frequency and damping ratio are given in Table 14.2 [Franklin et al., 1994].

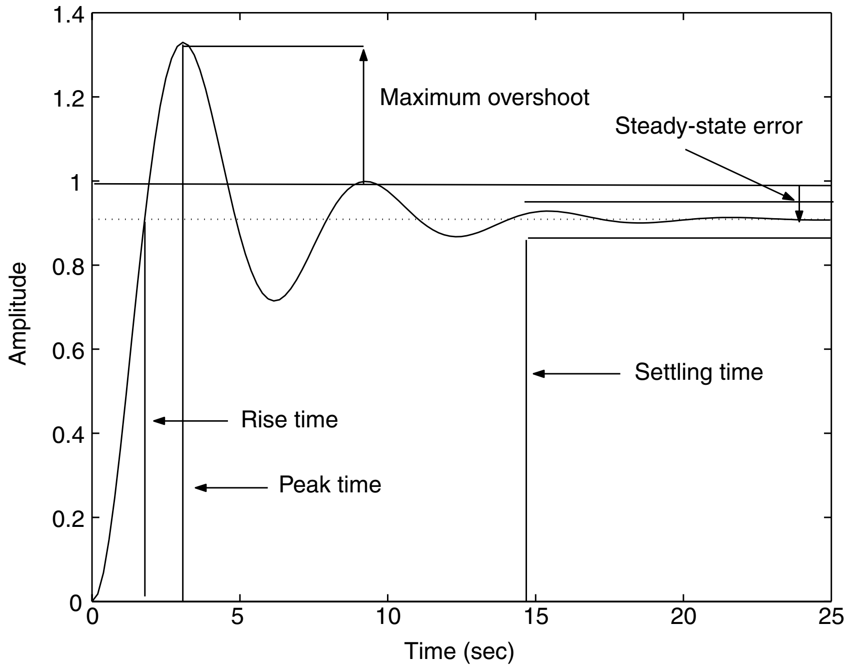


FIGURE 14.10 Time domain control specifications.

TABLE 14.2 Time-Domain Specifications as a Function of Natural Frequency, Damped Natural Frequency, and Damping Ratio

Rise time:	$t_r \cong \dfrac{1.8}{\omega_n}$
Peak time:	$t_p \cong \dfrac{\pi}{\omega_d}$
Overshoot:	$M_p = e^{-\pi\zeta/\sqrt{1-\zeta^2}}$
Settling time (1%):	$t_s = \dfrac{4.6}{\zeta\omega_n}$

Note: Results are from Franklin et al. (1994).

Example

Returning to the robot arm example, assume the desired system performance has a system rise time less than 1.4 sec, a maximum overshoot less than 30%, and a 1% settling time less than 10 sec. From the first row in Table 14.2, the natural frequency must be greater than 1.29. From the third and fourth rows, the damping ratio should be greater than approximately 0.4. Figure 14.11 illustrates the root locus plot, the pole locations and corresponding gain, and K (`rlocfind()` is the Matlab command for retrieving the gain value for a particular location on the root locus). These results provide a damping ratio of approximately .45 and a natural frequency of approximately 1.38. Figure 14.12 illustrates the step response of the system to a unit step input verifying these system parameters.

14.2.2.3 Frequency Response Design Methods

An alternative approach to controller design and analysis is the so-called frequency response method. Frequency response controller design techniques have two main advantages. They provide good controller

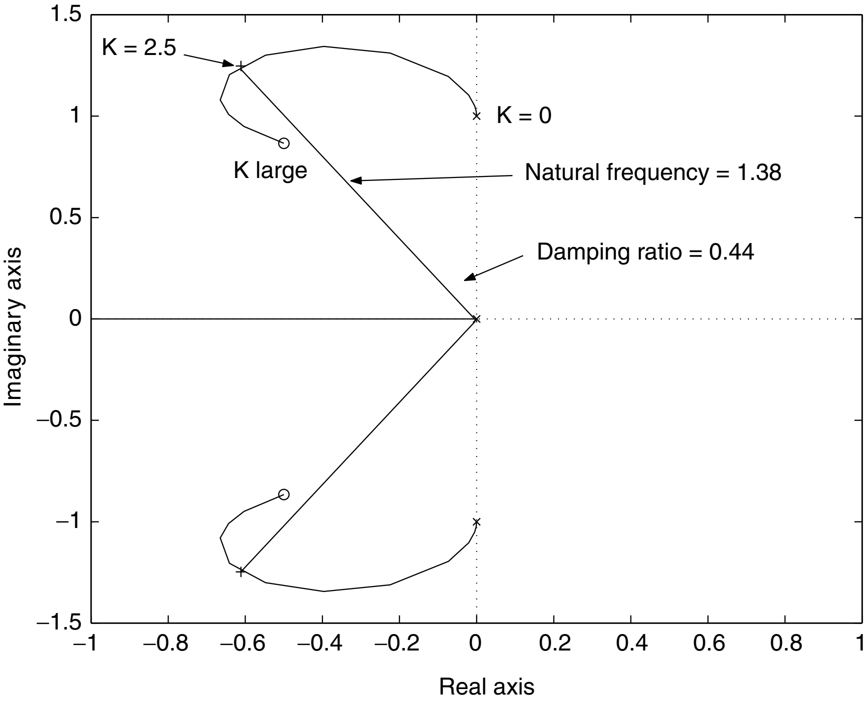


FIGURE 14.11 Selecting pole locations for a desired system response.

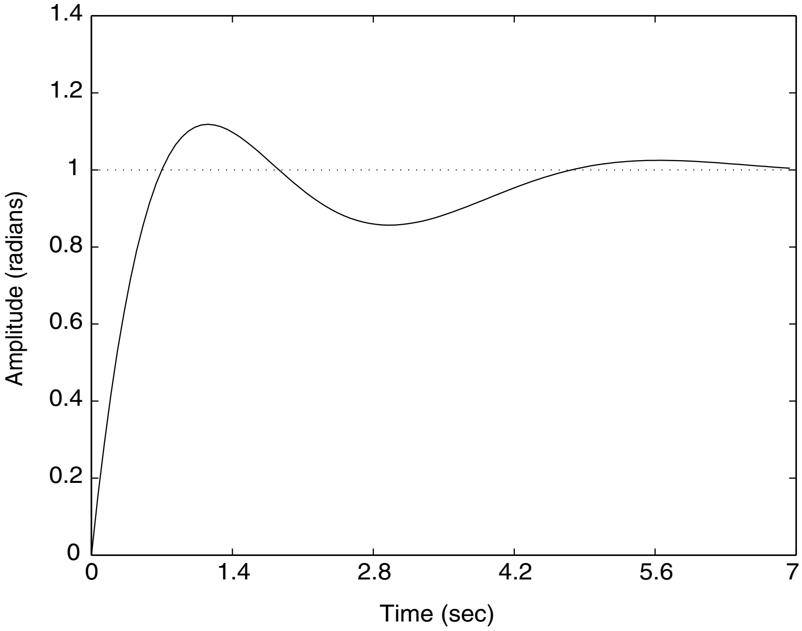


FIGURE 14.12 Robot arm step response.

design even with uncertainty with respect to high-frequency plant characteristics, and using experimental data for controller design purposes is straightforward. The two main tools are Bode and Nyquist plots (see [Bode, 1945] and [Nyquist, 1932] for first-source references), and stability analyses are considered first.

A Bode plot is a plot of two curves. The first curve is the logarithm of the magnitude of the response of the open-loop transfer function with respect to unit sinusoidal inputs of frequency ω. The second

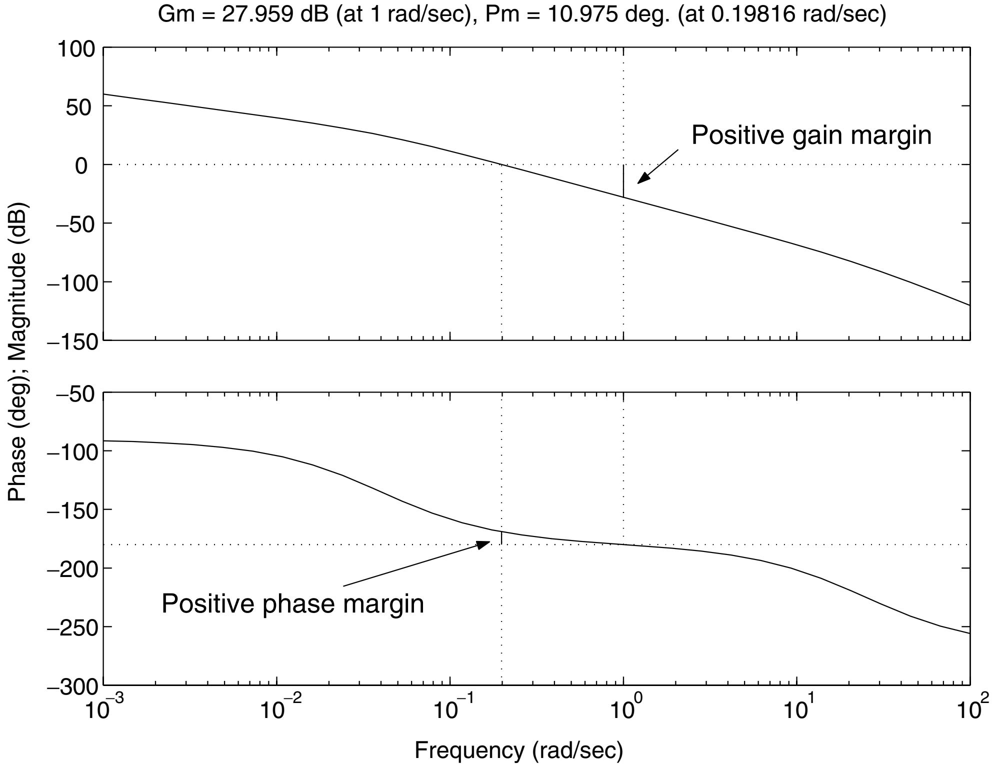


FIGURE 14.13 Bode plot.

curve is the phase of the open-loop transfer function response as a function of input frequency ω. Figure 14.13 illustrates the Bode plot for the transfer function:

$$G(s) = \frac{1}{s^3 + 25s^2 + s} \tag{14.24}$$

As the frequency of the sinusoidal input is increased, the magnitude of the system response decreases. The phase difference between the sinusoidal input and system response starts near −90° and approaches −270° as the input frequency becomes large.

An advantage of Bode plots is that they are easy to sketch by hand. Because the magnitude of the system response is plotted on a logarithmic scale, the contributions to the magnitude of the response due to individual factors in the transfer function add together. Due to basic facts related to the polar representation of complex numbers, the phase contributions of each factor add as well. Recipes for sketching Bode plots by hand can be found in any undergraduate controls text, such as Franklin et al. (1994), Raven (1995), Ogata (1997), and Kuo (1995).

For systems where the magnitude of the response passes through the value of 1 only one time and for systems where increasing the transfer function gain leads to instability (the most common, but not exclusive, scenario), the gain margin and phase margin can be determined directly from the Bode plot to provide a measure of system stability under unity feedback. Figure 14.13 also illustrates the definition of gain and phase margin. Positive gain and phase margins indicate stability under unity feedback. Conversely, negative gain and phase margins indicate instability under unity feedback. The class of systems for which Bode plots can be used to determine stability are called minimum phase systems. A system is minimum phase if all of its open-loop poles and zeros are in the left-half plane.

Bode plots also determine the steady-state error under unity feedback for various types of reference inputs (steps, ramps, etc.). In particular, if the low-frequency asymptote of the magnitude plot has a slope

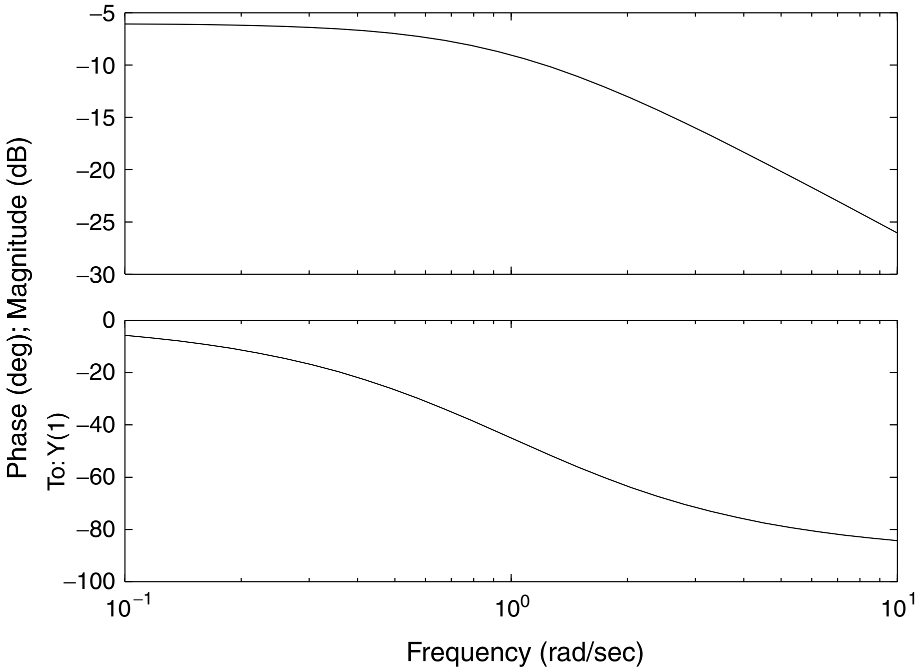


FIGURE 14.14 Bode plot for example problem.

of zero and if the value of this asymptote is denoted by K, then the steady-state error of the system under unity feedback to a step input is

$$\lim_{t\to\infty} e = \frac{1}{1+K} \tag{14.25}$$

If the slope of the magnitude plot at low frequencies is −20 dB/decade and if the value where the asymptote intersects the vertical line $\omega = 1$ is denoted by K, then the steady-state error to a ramp input is

$$\lim_{t\to\infty} e = \frac{1}{K} \tag{14.26}$$

Example

Consider the system illustrated in Figure 14.2 where $C(s) = A(s) = S(s) = 1$ and $P(s) = (1/2)/(s + 1)$. Figure 14.14 illustrates the Bode plot for the open-loop transfer function $P(s) = (1/2)/(s + 1)$. The low-frequency asymptote is approximately at −6, so $20 \log K = -6 \Rightarrow K \approx 0.5012 \Rightarrow y_{ss} \approx 0.6661$, where $y_{ss} = \lim_{t\to\infty} p(t)$. Figure 14.15 illustrates the unity feedback closed-loop step response of the system, verifying that the steady-state value for $y(t)$ is the same as computed from the Bode plot.

A Nyquist plot is a more sophisticated means to determine stability and is not limited to cases where only increasing gain leads to system instability. A Nyquist plot is based on the well-known result from complex variable theory called the principle of the argument. Consider the (factored) transfer function:

$$G(s) = \frac{\prod_i (s + z_i)}{\prod_j (s + p_j)} \tag{14.27}$$

By complex variable theory, $\angle G(s) = \Sigma_i \theta_i - \Sigma_i \varphi_j$, where θ_i are the angles between s and the zeros z_i, and ϕ_j are the angles between s and the poles p_j. Thus, a plot of $G(s)$ as s follows a closed contour (in the clockwise direction) in the complex plane will encircle the origin in the clockwise direction the same number of times that there are zeros of $G(s)$ within the contour minus the number of times that there are

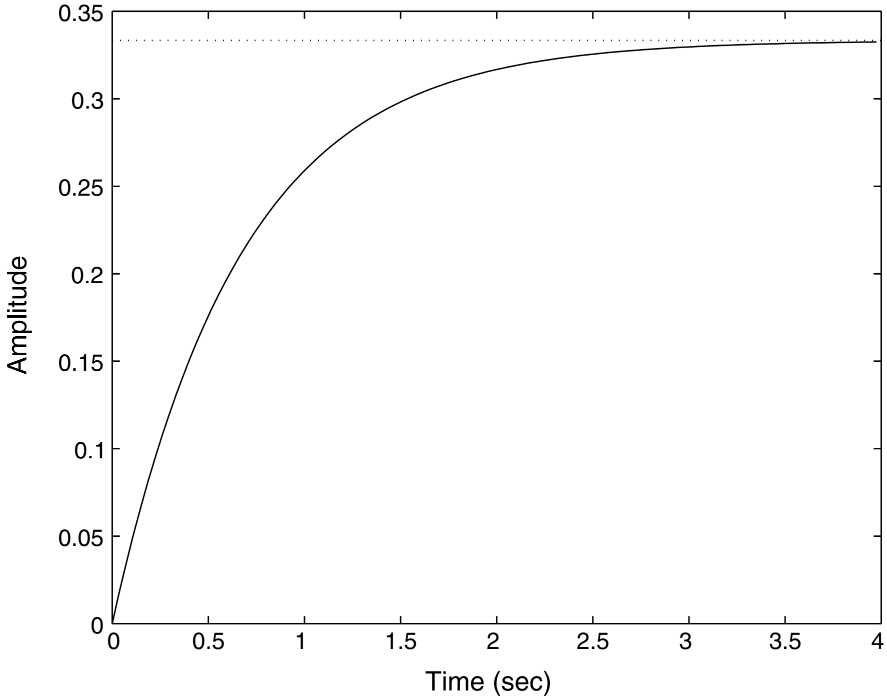


FIGURE 14.15 Step response for example problem.

poles of $G(s)$ within the contour. Therefore, an easy check for stability is to plot the open loop $G(s)$ on a contour that encircles the entire left-half complex plane. Assuming that G(s) has no right-half plane poles (poles of $G(s)$ itself, in contrast to poles of the closed-loop transfer function), an encirclement of -1 by the plot will indicate a right-half plane zero of $1 + G(s)$, which is an unstable right-half plane pole of the unity feedback closed-loop transfer function:

$$\frac{G(s)}{1 + G(s)} \tag{14.28}$$

Figure 14.16 illustrates the Nyquist plot for a unity feedback system with open-loop transfer function given by:

$$G(s) = \frac{1}{(s + 1)(s + 1)} \tag{14.29}$$

which is stable under unity feedback. Figure 14.17 illustrates a Nyquist plot for a system that is unstable under unity feedback.

14.2.2.4 Lead–Lag Compensation

Lead–lag controller design is another popular compensation technique. In this case, the compensator (the $C(s)$ block in Figure 14.2) is of the form:

$$C(s) = K\beta \frac{As + 1}{\alpha As + 1} \frac{Bs + 1}{\beta Bs + 1} \tag{14.30}$$

where $\alpha < 1$ and $\beta > 1$. The first fraction is the lead portion of the compensator and can provide increased stability with an appropriate choice for A. The second term is the lag compensator and provides decreased steady-state error. Figure 14.18 plots the Bode plot for a lead compensator for various values of the parameter A. Because the lead compensator shifts the phase plot up, by an appropriate choice of the parameter A, the crossover point where the magnitude plot crosses through the value of 0 dB can be shifted to the right, increasing the gain margin.

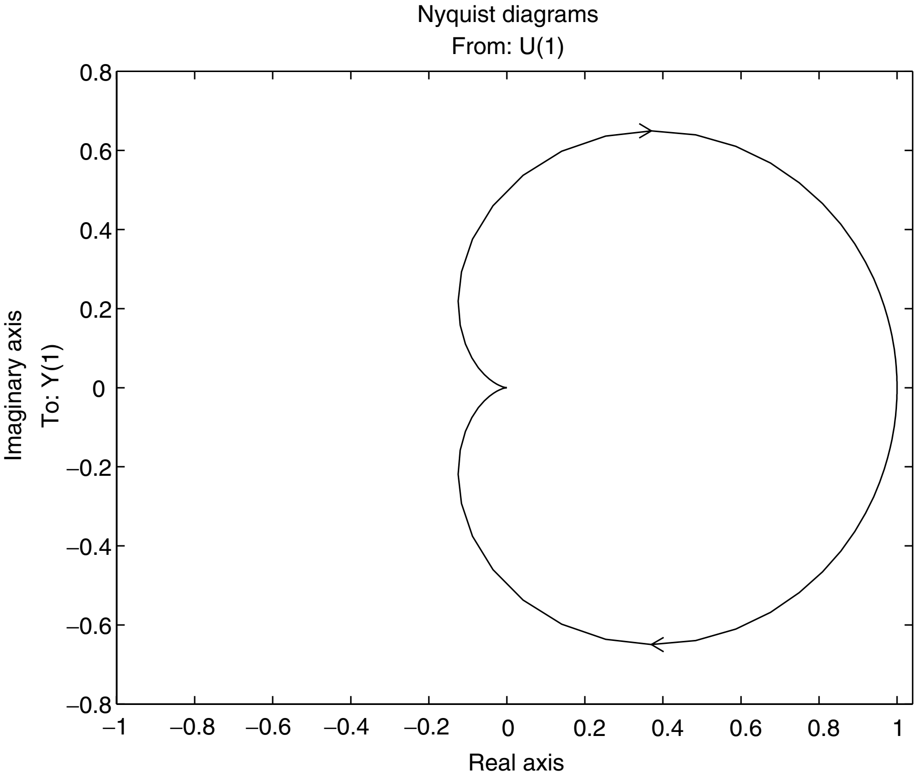


FIGURE 14.16 Nyquist plot for a stable system.

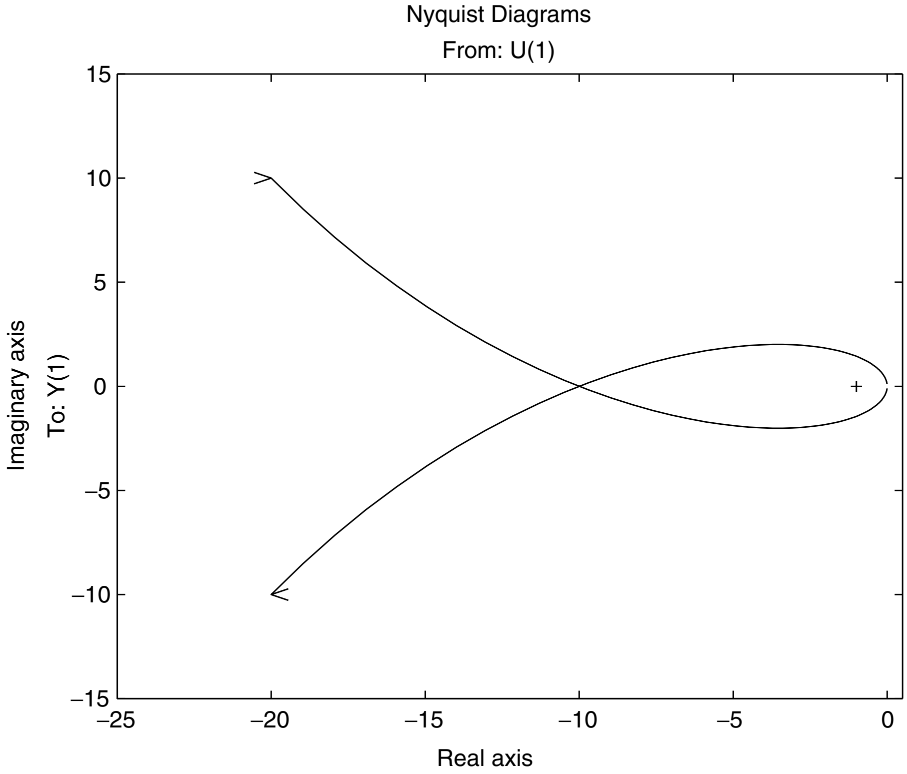


FIGURE 14.17 Nyquist plot of an unstable system.

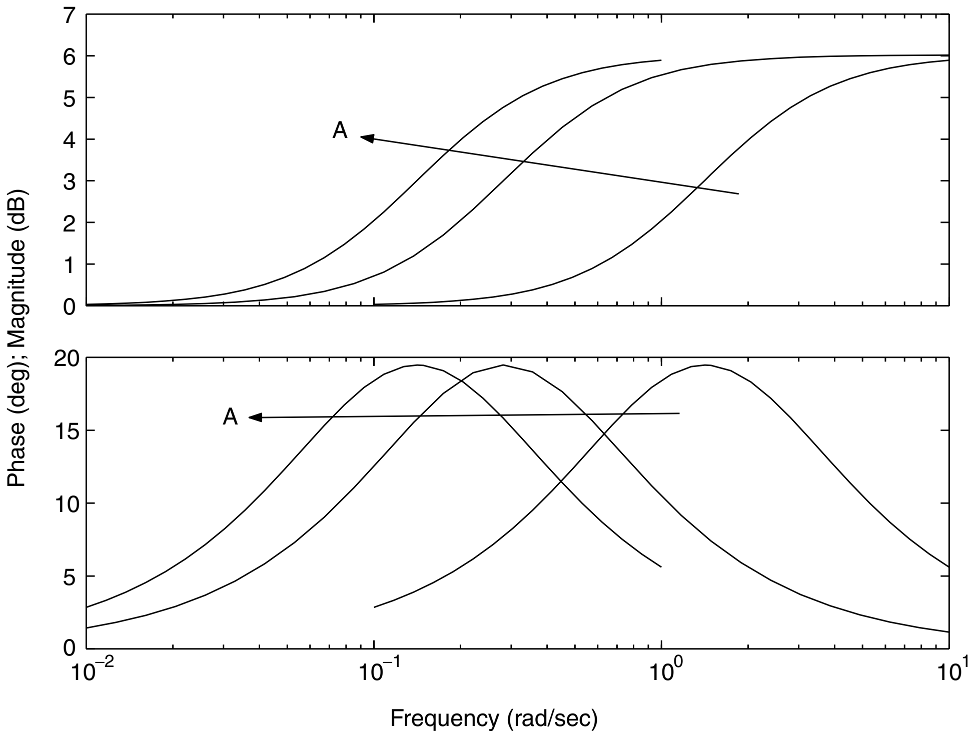


FIGURE 14.18 Bode plots of various lead compensators.

Example

Figure 14.19 plots the Bode plot for the compensated system:

$$G(s) = \frac{As + 1}{\alpha As + 1}\,\frac{1}{s^3 + 25s^2 + s} \tag{14.31}$$

where $A = 0$, 1 and $\alpha = 0.5$. The magnitude crossover point has been shifted to the left, increasing the gain margin. In a similar manner, unstable systems (which would originally have negative gain and phase margins) can possibly be stabilized.

Lag compensation works in a similar manner to increase the magnitude plot for low frequencies, which decreases the steady-state error for the system. Lead and lag controllers can be used in series to increase stability and decrease steady-state error. Systematic approaches for determining the parameters α, β, A, and B can be found in the references, particularly Franklin et al. (1994).

14.2.3 Other Topics

Various other topics are typically considered in classical control but will not be outlined here. Such topics include, but are not limited to, systematic methods for tuning PID regulators, lead–lag compensation, and techniques for considering and modeling time delay. Interested readers should consult the references, particularly Franklin et al. (1994), Ogata (1997), Kuo (1995), and Raven (1995).

14.3 "Modern" Control

In contrast to classical control, which is essentially a complex-variable, frequency-based approach for SISO systems, modern control is a time-domain approach that is amenable to MIMO systems. The basic

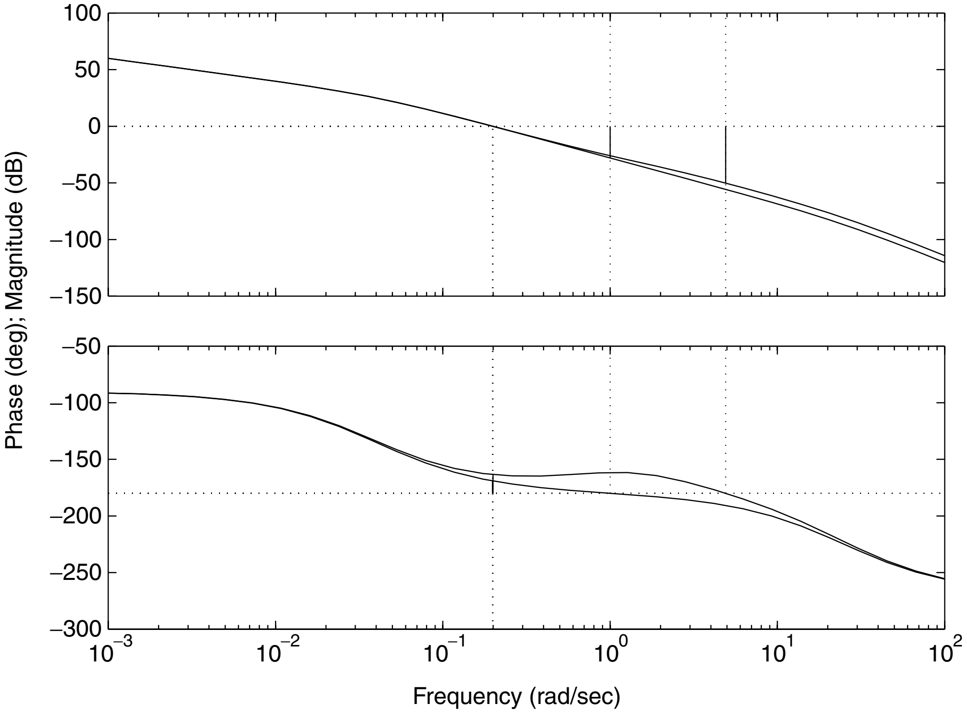


FIGURE 14.19 Lead compensated example system.

tools are from the theory of ODEs and matrix algebra. The topics outlined in this section are the pole placement, linear quadratic regulator (LQR) problems, and the basics of robust control.

14.3.1 Pole Placement

First, a multistate but single-input control system will be examined. Consider a control system written in state space:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u \tag{14.32}$$

where **x** is the $1 \times n$ state vector, u is the scalar input, A is an $n \times n$ constant matrix, and **B** is an $n \times 1$ constant matrix. If we assume that the control input u can be expressed as a combination of the current state variables (called full state feedback), we can write:

$$u = -k_1x_1 - k_2x_2 - \cdots - k_nx_n = -\mathbf{K}\mathbf{x} \tag{14.33}$$

where **K** is a row vector comprised of each of the gains k_i. Then, the state-space description of the system becomes:

$$\dot{\mathbf{x}} = (\mathbf{A} - \mathbf{B}\mathbf{K})\mathbf{x} \tag{14.34}$$

so that the solution of this equation is

$$\mathbf{x}(t) = e^{(\mathbf{A}-\mathbf{B}\mathbf{K})t}\mathbf{x}(0) \tag{14.35}$$

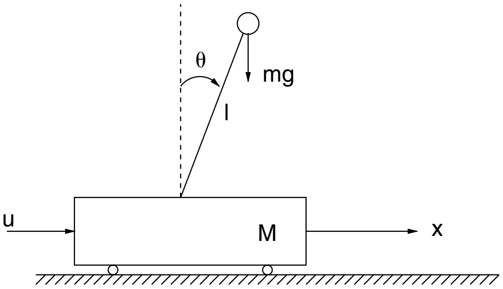


FIGURE 14.20 Cart and pendulum system.

where $e^{(\mathbf{A}-\mathbf{BK})t}$ is the matrix exponential of the matrix $\mathbf{A} - \mathbf{BK}$ defined by:

$$e^{(\mathbf{A}-\mathbf{BK})t} = \mathrm{I} + (\mathbf{A} - \mathbf{BK})\mathrm{t} + \frac{(\mathbf{A} - \mathbf{BK})^2 t^2}{2!} + \frac{(\mathbf{A} - \mathbf{BK})^3 t^3}{3!} + \cdots \tag{14.36}$$

Basic theory from linear algebra and ODEs [Hirsch and Smale, 1974] indicates that the stability and characteristics of the transient response will be determined by the eigenvalues of the matrix $\mathbf{A} - \mathbf{BK}$. If:

$$\mathrm{rank}[\mathbf{B}|\mathbf{AB}|\mathbf{A}^2\mathbf{B}|\mathbf{A}^3\mathbf{B}| \ldots |\mathbf{A}^{n-1}\mathbf{B}] = n \tag{14.37}$$

then it can be shown that the eigenvalues of $\mathbf{A} - \mathbf{BK}$ can be placed arbitrarily as a function of the elements of $\mathbf{K}$. Techniques to solve the problem by hand by way of a similarity transformation exist (see the standard undergraduate controls books), and Matlab has functions for the computations as well.

Example — Pole Placement for Inverted Pendulum System

Consider the cart and pendulum system illustrated in Figure 14.20. In state-space form the equations of motion are:

$$\frac{d}{dt}\begin{bmatrix} x \\ \dot{x} \\ \theta \\ \dot{\theta} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \dfrac{-gm}{M} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \dfrac{-(m+M)g}{lM} & 0 \end{bmatrix} \begin{bmatrix} x \\ \dot{x} \\ \theta \\ \dot{\theta} \end{bmatrix} + \begin{bmatrix} 0 \\ \dfrac{1}{M} \\ 0 \\ \dfrac{-1}{lM} \end{bmatrix} u \tag{14.38}$$

Setting $M = 10$, $m = 1$, $g = 9.81$, and $l = 1$ and letting $u = -k_1 x + k_2 \dot{x} + k_3 \theta + k_4 \dot{\theta}$ if the desired pole locations for the system are at:

$$\begin{aligned} \lambda_1 &= -1 - i \\ \lambda_2 &= -1 + i \\ \lambda_3 &= -8 \\ \lambda_4 &= -9 \end{aligned} \tag{14.39}$$

the Matlab function `place()` can be used to compute the values for the corresponding k_i. For this problem, the gain values are:

$$\begin{aligned} k_1 &= 122.32 \\ k_2 &= 151.21 \\ k_3 &= -849.77 \\ k_4 &= -38.79 \end{aligned} \tag{14.40}$$

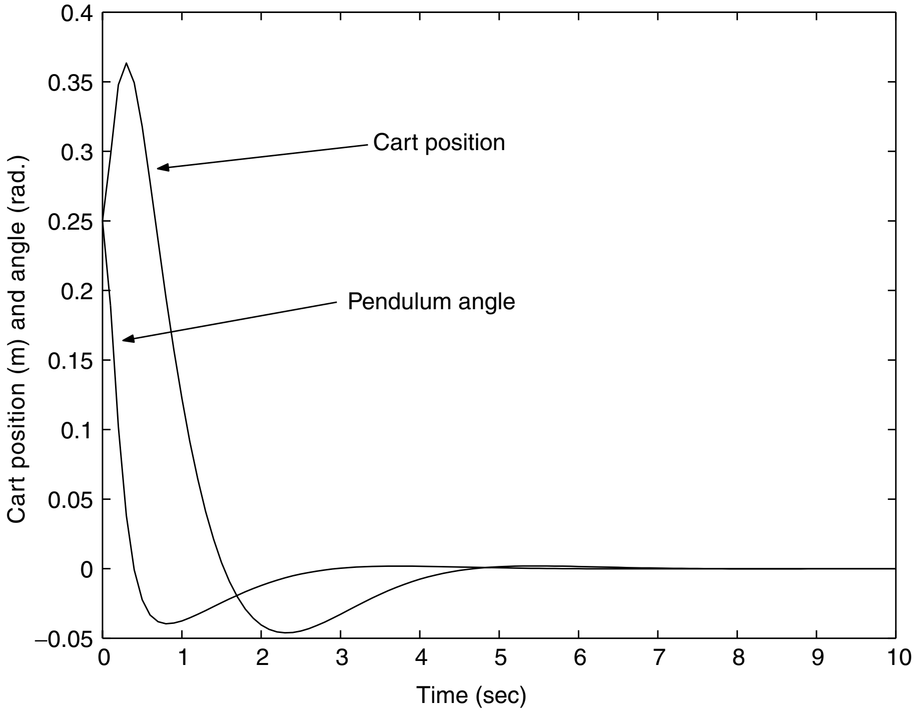


FIGURE 14.21 Cart and pendulum system pole placement response.

With initial conditions $x(0) = 0.25$, $\dot{x}(0) = 0$, $\theta(0) = 0.25$ and $\dot{\theta}(0) = 0$, Figure 14.21 illustrates the response of the system. Note that the cart position x initially moves in the "wrong" direction in order to compensate for the pendulum position.

14.3.2 The Linear Quadratic Regulator (LQR)

The LQR problem is not limited to scalar input problems and seeks to find a control input:

$$\mathbf{u} = -\mathbf{K}\mathbf{x}(t) \tag{14.41}$$

for the system:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \tag{14.42}$$

that minimizes the performance index:

$$J = \int_0^\infty (\mathrm{x}^T\mathbf{Q}\mathbf{x} + \boldsymbol{u}^T\mathbf{R}\mathbf{u})dt \tag{14.43}$$

where **Q** and **R** are positive definite, real symmetric matrices. By the second method of Lyapunov [Khalil, 1996; Sastry, 2000], the control input that minimizes the performance index is:

$$u = -\mathrm{R}^{-1}\mathbf{B}^T\mathbf{P}\mathbf{x}(t) \tag{14.44}$$

where **R** and **B** are from the performance index and equations of motion, respectively, and **P** satisfies the reduced matrix Riccati equation:

$$\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A} - \mathbf{P}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P} + \mathbf{Q} = \mathbf{0} \tag{14.45}$$

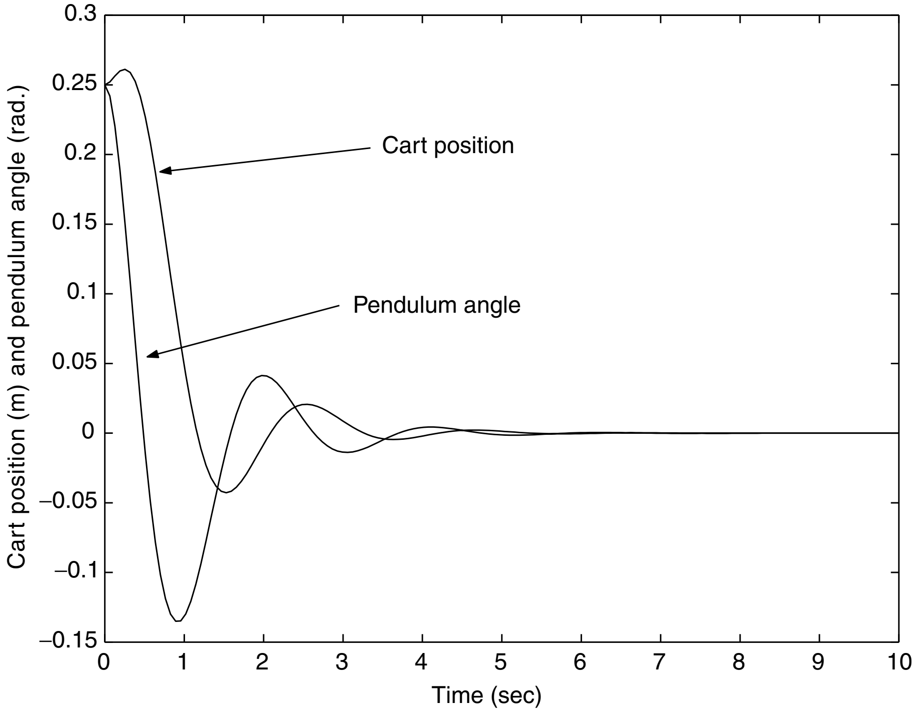


FIGURE 14.22 Cart and pendulum LQR response.

Example — LQR for Inverted Pendulum System

For the same cart and pole system as in the previous example with:

$$Q = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{14.46}$$

(which weights all the states equally) and $R = 0.001$, the optimal gains (computed via the Matlab `lqr()` function) are:

$$\begin{aligned} k_1 &= 31.62 \\ k_2 &= 145.75 \\ k_3 &= -95.53 \\ k_4 &= -21.65 \end{aligned} \tag{14.47}$$

and the response of the system with initial conditions $x(0) = 0.25$, $\dot{x}(0) = 0$, $\theta(0) = 0.25$ and $\dot{\theta}(0) = 0$ is illustrated in Figure 14.22. If the Q matrix is modified to provide a heavy weighting for the θ state:

$$Q = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 100 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{14.48}$$

Figure 14.23 illustrates the system response. Note that the pendulum angle goes to zero very rapidly but at the "expense" of a slower response and greater deviation for the cart position.

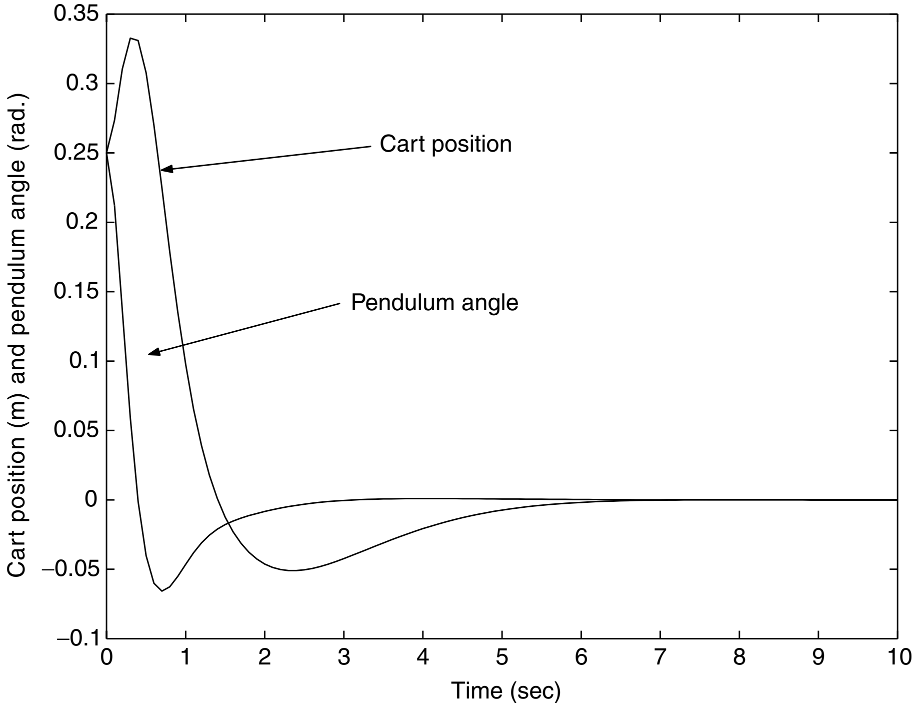


FIGURE 14.23 Cart and pendulum LQR response with large pendulum angle weighting.

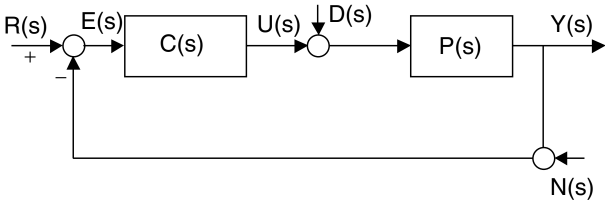


FIGURE 14.24 Robust control feedback block diagram.

14.3.3 Basic Robust Control

The main idea motivating modern robust control techniques is explicitly incorporating plant uncertainty representations into system modeling and control synthesis methods. The material here outlines the presentation in Doyle et al. (1992), and the more advanced material is from Zhou (1996). Modern robust control is a very involved subject and only the briefest outline is provided here.

Consider the unity feedback SISO system illustrated in Figure 14.24, where P and C are the plant and controller transfer functions; $R(s)$ is the reference signal; $Y(s)$ is the output; $D(s)$ and $N(s)$ are external disturbances and sensor noise, respectively; $E(s)$ is the error signal; and $U(s)$ is the control input.

Define the loop transfer function $L = CP$ and the sensitivity function:

$$S = \frac{1}{1 + L} \tag{14.49}$$

which is the transfer function from the reference input $R(s)$ to the error $E(s)$ which provides a measure of the sensitivity of the closed loop (or complementary sensitivity) transfer function:

$$T = \frac{PC}{1 + PC} \tag{14.50}$$

to infinitesimal variations in the plant P. Given a (frequency-dependent) weighting function $W_1(s)$, a natural performance specification (relating tracking error to classes of reference signals) is

TABLE 14.3 Internal Stability Conditions

Perturbation	Condition
$(1 + \Delta W_2)P$	$\|\|W_2T\|\|_\infty < 1$
$P + \Delta W_2$	$\|\|W_2CS\|\|_\infty < 1$
$\dfrac{P}{1 + \Delta W_2 P}$	$\|\|W_2PS\|\|_\infty < 1$
$\dfrac{P}{1 + \Delta W_2}$	$\|\|W_2S\|\|_\infty < 1$

$$||W_1S||_\infty < 1 \tag{14.51}$$

where $||\cdot||_\infty$ denotes the infinity norm. An easy graphical test for the performance specification is that the Nyquist plot of L must always lie outside a disk of radius $|W_1|$ centered at -1.

To incorporate plant uncertainty into the model, consider a nominal plant P and perturbed plant $\tilde{P}$ where P and $\tilde{P}$ differ by some multiplicative or other type of uncertainty. Let W_2 be a stable transfer function and Δ be a variable stable transfer function satisfying $||\Delta||_\infty \leq 1$. Common uncertainty models are constructed by appropriate combinations of P, Δ, and W. It is shown that the system is internally stable (this is a stronger definition than simple input–output stability; see [Doyle et al., 1992]) for the conditions shown in Table 14.3.

Recall that the nominal performance condition was $||W_1S||_\infty < 1$. The robust performance condition is a combination of the two (for the $(1 + \Delta W_2)P$ perturbation):

$$|||W_1S| + |W_2T|||_\infty < 1 \tag{14.52}$$

Other robust performance measures for various types of uncertainty are found in Doyle et al. (1992) and Zhou (1996).

Recall that W_1 is the performance specification weighting function and W_2 is the plant uncertainty transfer function. Consider the following facts:

1. Plant uncertainty is greatest for high frequencies.
2. It is only reasonable to demand high performance for low frequencies.

Typically,

$$|W_1| > 1 > |W_2| \tag{14.53}$$

for low frequencies, and

$$|W_1| < 1 < |W_2| \tag{14.54}$$

for high frequencies (it can be shown that the magnitude of either W_1 or W_2 must be less than 1). By considering the relationship between L, S, and T, the following is derived:

$$|W_1| \gg 1 > |W_2| \;\Rightarrow\; |L| > \frac{|W_1|}{1 - |W_2|} \tag{14.55}$$

and

$$|W_1| < 1 \ll |W_2| \;\Rightarrow\; |L| < \frac{1 - |W_1|}{|W_2|} \tag{14.56}$$

Loopshaping [Bower and Schultheiss, 1961; Horowitz, 1963] controller design is the task of determining an L (and hence C) that satisfies the low-frequency performance criterion as well as the high-frequency

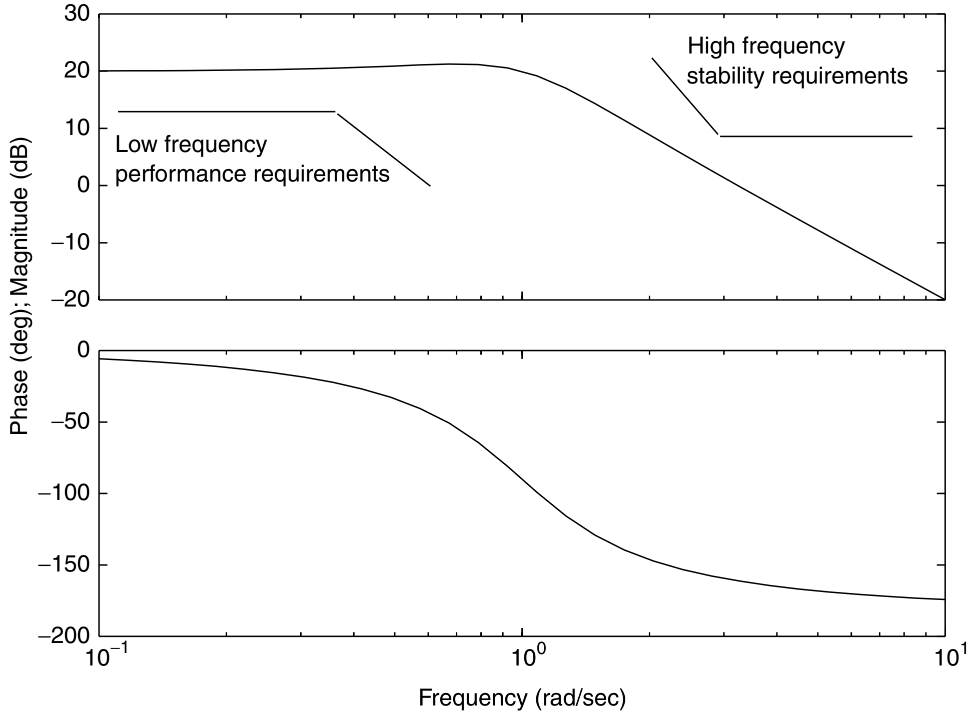


FIGURE 14.25 Loopshaping concepts.

robustness criterion. The task is to design C so that the magnitude versus frequency plot of L appears as in Figure 14.25. In the figure, the indicated low-frequency performance bound is a plot of:

$$\frac{|W_1|}{1-|W_2|} \tag{14.57}$$

for low frequencies, and the high-frequency stability bound is a plot of:

$$\frac{1-|W_1|}{|W_2|} \tag{14.58}$$

for high frequencies.

Two more aspects of this problem have been developed in recent years. The first concerns optimality, and the second concerns multivariable systems. For both aspects of these recent developments, refer to the comprehensive book by Zhou (1996).

14.4 Nonlinear Control

Aside from the developments of robust optimal control briefly outlined in the previous section, the area of most recent development in control theory has been nonlinear control. Nonlinear control does not ignore nonlinear effects via linearization, the nonlinearities in the control system are either expressly recognized or are even exploited for control purposes. Much, but not all, development in nonlinear control uses tools from differential geometry. While the control techniques will be outlined here, the basics of differential geometry will not, and the interested reader is referred to Abraham et al. (1988), Boothby (1986), Isidori (1996), and Nijmeijer and van der Schaft (1990) for details.

The general nonlinear model considered here is of the form:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \sum_{i=1}^{n} \mathbf{g}_{\mathrm{i}}(\mathbf{x})u_i \tag{14.59}$$

where **x** is a $1 \times n$ vector, the $\mathbf{f}(\mathbf{x})$ and $\mathbf{g}_\mathrm{i}(\mathbf{x})$ are smooth vector fields, and the u_i are scalar control inputs. Note that this is not the most general form for nonlinear systems, as the u_i are assumed to enter the equations in an affine manner (i.e., they simply multiply the $\mathbf{g}_\mathrm{i}(\mathbf{x})$ vector fields). For some aerodynamic problems, this assumption may not be true.

14.4.1 SISO Feedback Linearization

In contrast to the standard Jacobian linearization of a nonlinear control system, feedback linearization is a technique to construct a nonlinear change of coordinates which converts a nonlinear system in the original coordinates to a linear system in the new coordinates. Whereas the Jacobian linearization is an approximation of the original system, a feedback linearized system is exactly the original system. SISO systems will be considered first, followed by MIMO systems. References for feedback linearization are Isidori (1996), Nijmeijer and van der Schaft (1990), Krener (1987), Khalil (1996), and Sastry (2000). Developmental papers or current research in this area are considered in Slotine and Hedrick (1993), Brockett (1978), Dayawansa et al. (1985), Isidori et al. (1981a; 1981b), and Krener (1987).

Consider the nonlinear system:

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x})\boldsymbol{u} \\ y &= \boldsymbol{h}(\mathbf{x}) \end{aligned} \tag{14.60}$$

where the function $h(\mathbf{x})$ is called the output function. Let $L_\mathrm{f}h$ denote the Lie derivative of the function h with respect to the vector field f, which is defined in coordinates as:

$$L_\mathrm{f}h(x) = \sum_{i=1}^{n} \frac{\partial h}{\partial x_i}(\mathbf{x})f_i(\mathbf{x}) \tag{14.61}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} \quad \text{and} \quad \mathbf{f} = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ \mathrm{f}_{n-1}(\mathbf{x}) \\ f_n(\mathbf{x}) \end{bmatrix} \tag{14.62}$$

so it is simply the directional derivative of h along **f**. Because the system evolves according to the state equations, the time derivative of the output function $\dot{y}$ is simply the directional derivative of the output function along the control system:

$$\dot{y} = \dot{h} = \frac{\partial h}{\partial \mathbf{x}}\dot{\mathbf{x}} = \frac{\partial h}{\partial \mathbf{x}}(\mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x})u) = L_{\mathrm{f}+gu}h = L_\mathrm{f}h + L_\mathrm{g}hu \tag{14.63}$$

The relative degree of a system is defined as follows: a SISO nonlinear system is said to have strict relative degree γ at the point x if:

1. $L_gL_\mathrm{f}^ih(x) \equiv 0 \quad i = 0, 1, 2, \ldots, \gamma - 2$ (14.64)
2. $L_gL_\mathrm{f}^{\gamma-1}h(x) \neq 0$ (14.65)

In the case where $\gamma = n$, the system is full state feedback linearizable, and it is possible to construct the following change of coordinates where the original coordinates x_i are mapped to a new set of coordinates

ξ_i as follows:

$$\begin{aligned}
\xi_1 &= h(\mathrm{x}) \\
\xi_2 &= \dot{\xi}_1 = \dot{h} = L_{\mathrm{f}}h \\
\xi_3 &= \dot{\xi}_2 = \ddot{h} = L_{\mathrm{f}}^2 h \\
&\vdots \\
\xi_n &= \dot{\xi}_{n-1} = L_{\mathrm{f}}^{\gamma-1} h
\end{aligned} \tag{14.66}$$

Computing derivatives, the control system becomes:

$$\begin{aligned}
\dot{\xi}_1 &= \xi_2 \\
\dot{\xi}_2 &= \xi_3 \\
&\vdots \\
\dot{\xi}_{n-1} &= \xi_n \\
\dot{\xi}_n &= L_{\mathrm{f}}^{\gamma} h + L_{\mathrm{g}} L_{\mathrm{f}}^{\gamma-1} hu
\end{aligned} \tag{14.67}$$

or, setting

$$u = \frac{1}{L_{\mathrm{g}}\mathrm{L}_{\mathrm{f}}^{\gamma-1}\, h}(-L_{\mathrm{f}}^{\gamma} h + v) \tag{14.68}$$

the system is

$$\begin{aligned}
\dot{\xi}_1 &= \xi_2 \\
\dot{\xi}_2 &= \xi_3 \\
&\vdots \\
\dot{\xi}_{n-1} &= \xi_n \\
\dot{\xi}_n &= v
\end{aligned} \tag{14.69}$$

which is both linear and in controllable canonical form. One approach to determine an appropriate v to stabilize the system or track desired values of $h(x)$ is pole placement (i.e., $v = -\mathrm{K}\xi$). Note that the overall approach is to determine an output function h that could be differentiated n times before the control input appeared. This approach essentially constructs a system known as a chain of integrators, as the derivative of the ith state in the ξ variables is equal to the $(i + 1)$th state variable.

There are two main limitations to feedback linearization approaches. The first is that not all systems are feedback linearizable, although analytical tests exist to determine whether a particular system is linearizable. Second, determining the output function $h(x)$ involves solving a system of partial differential equations.

Example — SISO Full State Feedback Linearization

Consider the following system as a mathematical example of the computations involved in feedback linearization:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} x_3 \\ x_4 \\ x_1 + x_2 + x_3 \\ x_1 - x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} u = f(x) + g(x)u \tag{14.70}$$

with output function $y = h(x) = x_1$. The system has a relative degree equal to 4, so the system is full state feedback linearizable and the coordinate transformation is given by:

$$\begin{aligned}
\xi_1 &= h(x) = x_1 \\
\xi_2 &= L_f h(x) = x_3 \\
\xi_3 &= L_f^2 \mathrm{h}(x) = x_1 + x_2 + x_3 \\
\xi_4 &= L_f^3 h(x) = x_1 + x_2 + 2x_3 + x_4
\end{aligned} \tag{14.71}$$

The above equations and the fact that the system has a relative degree of 4 is verified by the following detailed calculations:

$$\begin{aligned}
L_g h(x) &= L_g x_1 = [1 \quad 0 \quad 0 \quad 0]\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = 0, \\
L_g L_f h(x) &= L_g l_f x_1 = L_g[1 \quad 0 \quad 0 \quad 0]\begin{bmatrix} x_3 \\ x_4 \\ x_1 + x_2 + x_3 \\ x_1 - x_3 \end{bmatrix} = L_g x_3 = [0 \quad 0 \quad 1 \quad 0]\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = 0, \\
L_g L_f^2 h(x) &= L_g L_f^2 x_1 = L_g L_f x_3 = L_g[0 \quad 0 \quad 1 \quad 0]\begin{bmatrix} x_3 \\ x_4 \\ x_1 + x_2 + x_3 \\ x_1 - x_3 \end{bmatrix} \\
&= L_g(x_1 + x_2 + x_3) = [1 \quad 1 \quad 1 \quad 0]\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = 0, \\
L_g L_f^3 h(x) &= L_g L_f^3 x_1 = L_g L_f(x_1 + x_2 + x_3) = L_g[1 \quad 1 \quad 1 \quad 0]\begin{bmatrix} x_3 \\ x_4 \\ x_1 + x_2 + x_3 \\ x_1 - x_3 \end{bmatrix} \\
&= L_g(x_1 + x_2 + 2x_3 + x_4) = [1 \quad 1 \quad 2 \quad 1]\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = 1
\end{aligned} \tag{14.72}$$

Therefore, a controller of the form:

$$\begin{aligned}
u &= \frac{1}{L_g L_f^3 h}(-L_f^4 h + v) \\
&= (-(3x_1 + 2x_2 + 2x_3 + x_4) + k_1 x_1 + k_2 x_3 + k_3(x_1 + x_2 + x_3) + k_4(x_1 + x_2 + 2x_3 + x_{er4}))
\end{aligned} \tag{14.73}$$

with the gains k_i picked via pole placement, for example, will allow the system to track trajectories of the output function $h(x) = x_1$.

So far, this section has considered full state feedback linearization where the relative degree of a system is equal to the dimension of its state space. Partial feedback linearization is also possible where the relative degree is less than the dimension of the state space. However for such systems, an analysis of the stability of the zero dynamics is necessary. In particular, if the relative degree $\gamma < n$, then the change of coordinates is typically expressed in the form:

$$\begin{aligned}
\xi_1 &= h(x) \\
\xi_2 &= \dot{\xi}_1 = \dot{h} = L_{\mathrm{f}} h \\
\xi_3 &= \dot{\xi}_2 = \ddot{h} = L_{\mathrm{f}}^2 h \\
&\vdots \\
\xi_\gamma &= \dot{\xi}_{\gamma-1} = L_f^{\gamma-1} h \\
\eta_1 &= \eta_1(x) \\
\eta_2 &= \eta_2(x) \\
&\vdots \\
\eta_{n-\gamma} &= \eta_{n-\gamma}(x)
\end{aligned} \tag{14.74}$$

where the η_i are chosen so that the matrix:

$$\begin{bmatrix} dh(x) \\ dL_f h(x) \\ \vdots \\ dL_f^{\gamma-1} h(x) \\ d\eta_1(x) \\ \vdots \\ d\eta_{n-\gamma}(x) \end{bmatrix} \tag{14.75}$$

is full rank. The dynamics of the system in the new coordinates will be of the form:

$$\begin{aligned}
\dot{\xi}_1 &= \xi_2 \\
\dot{\xi}_2 &= \xi_3 \\
&\vdots \\
\dot{\xi}_\gamma &= b(\xi, \eta) + a(\xi, \eta) u \\
\dot{\eta}_1 &= q_1(\xi, \eta) \\
&\vdots \\
\dot{\eta}_{n-\gamma} &= q_{n-\gamma}(\xi, \eta)
\end{aligned} \tag{14.76}$$

The zero dynamics are the dynamics expressed by the η equations, the stability of which must be considered independently of the linearized ξ equations. Refer to texts by Isidori (1996), Khalil (1996), Nijmeijer and van der Schaft (1990), and Sastry (2000) for the relevant details.

14.4.2 MIMO Full-State Feedback Linearization

The MIMO feedback linearization is a slight extension of the SISO feedback linearization by which the SISO linearization construction is repeated for m output functions for a system with m control inputs:

$$\begin{aligned}
\dot{x} &= \mathbf{f}(\mathbf{x}) + \mathbf{g}_1(\mathbf{x}) u_1 + \mathbf{g}_2(\mathbf{x}) u_2 + \cdots + \mathbf{g}(\mathbf{x})_m u_m \\
y_1 &= h_1(\mathbf{x}) \\
y_2 &= h_2(\mathbf{x}) \\
&\vdots \\
y_m &= h_m(\mathbf{x})
\end{aligned} \tag{14.77}$$

The vector relative degree is defined as a combination of relative degrees for each of the output functions. Considering the *j*th output y_j,

$$\dot{y}_{\mathrm{j}} = L_{\mathrm{f}} h_j + L_{\mathrm{g}_1} h_j u_1 + L_{\mathrm{g}_2} h_j u_2 + \cdots + L_{\mathrm{g}_m} h_j u_m \tag{14.78}$$

If $L_{g_i} h_j \equiv 0$ for each i, then the inputs do not appear in the derivative. Now let γ_j be the smallest integer such that $L_g L_f^{\gamma_j - 1} h_j \neq 0$ for at least one i. Define the matrix:

$$A(x) = \begin{bmatrix} L_{\mathrm{g}_3} L_{\mathrm{f}}^{\gamma_1-1} h_1 & \cdots & L_{\mathrm{g}_m} L_{\mathrm{f}}^{\gamma_1-1} h_1 \\ \vdots & \ddots & \vdots \\ L_{\mathrm{g}_1} L_{\mathrm{f}}^{\gamma_m-1} h_m & \cdots & L_{\mathrm{g}_m} L_{\mathrm{f}}^{\gamma_m-1} h_m \end{bmatrix} \tag{14.79}$$

The system has vector relative degree $\gamma_1, \gamma_2, \ldots, \gamma_m$ at x if $L_{g_i} L_f^k h_1 \equiv 0\ 0 \leq k \leq \gamma_i - 2$ for $i = 1, \ldots, m$ and the matrix $A(x)$ is nonsingular.

14.4.3 Control Applications of Lyapunov Stability Theory

Lyapunov theory for autonomous differential equations states that if $x = 0$ is an equilibrium point for a differential equation $\dot{x} = f(x)$, and there exists a continuously differentiable function, $V(x) > 0$ except for $V(0) = 0$ and $\dot{V}(x) < 0$ and in some domain containing zero where:

$$\dot{V}(x) = \sum_{i=1}^{n} \frac{\partial V}{\partial x_i} \dot{x}_i = \sum_{i=1}^{n} \frac{\partial V}{\partial x_i} f_i(x) \tag{14.80}$$

then the point $x = 0$ is an asymptotically stable equilibrium point for the differential equation. The utility of Lyapunov theory in control is that controller synthesis techniques can be designed to ensure the negative definiteness of a Lyapunov function to ensure stability or boundedness of the system trajectories.

As fully described in Khalil (1996), the main applications of Lyapunov stability theory to control system design are Lyapunov redesign, backstepping, sliding mode control, and adaptive control. The basic concepts of all of these will be briefly outlined here.

Lyapunov redesign is an instance of nonlinear robust control design. However, there is a severe restriction upon how the uncertainties are expressed in the equations of motion with a corresponding restriction on the types of systems that are amenable to this technique. In particular, consider the system:

$$\dot{x} = f(t, x) + G(t, x)u + G(t, x)\delta(t, x, u) \tag{14.81}$$

where f and G are known; δ is unknown but is bounded by a known, but not necessarily small, function. The main restriction here is that the uncertainty enters the system in exactly the same manner as the control input. In order to use Lyapunov redesign, a stabilizing control law exists for the nominal system (ignoring δ), and a Lyapunov function for the nominal system must be known. (Note that one nice aspect of the feedback linearization discussed previously is that if a controller is designed using that technique, a Lyapunov function is straightforward to determine because of the simple form of the equations of motion after the nonlinear coordinate transformation.) Because of the way that the uncertainty enters the system, it is easy to modify the nominal control law to compensate for the uncertainty. References concerning Lyapunov redesign include Corless (1993), Corless and Heitmann (1981), Barmish et al. (1983), and Spong and Vidyasager (1989).

Backstepping is a recursive controller design procedure where the entire control system is decomposed into smaller, simpler subsystems for which it may be easier to design a stabilizing controller. By considering the appropriate way to modify a Lyapunov function after each smaller subsystem is designed, a stabilizing controller for the full system may be obtained. The main restriction for this technique is a limitation on the structure of the equations of motion (a type of hierarchical structure is required). Extensions of this procedure to account for certain system uncertainties also have been developed. For references, see Krstic et al. (1995), Qu (1993), and Slotine and Hedrick (1993).

The basic idea in sliding mode control is to drive the system in finite time to a certain submanifold of the configuration space, called the sliding manifold, upon which the system should indefinitely evolve. Because the sliding manifold has a lower dimension than the full state space for the system, a lower order model describes the evolution of the system on the sliding manifold. If a stabilizing controller is designed for the sliding manifold, the problem reduces to designing a controller to drive the system to the sliding manifold. The advantage of sliding mode control is that it is very robust with respect to system uncertainties. One disadvantage is that it is a bit mathematically quirky as there are discontinuities in the control law when switching from the full state of the system to the sliding manifold. Additionally, "chattering," wherein the system constantly alternates between the two sides of the submanifold, is a common problem. See Utkin (1992) and DeCarlo et al. (1988) for overviews of the approach.

Finally, there is vast literature in the area of adaptive control. In adaptive control, some system performance index is measured, and the adaptive controller modifies adjustable parameters in the controller in order to maintain the performance index of the control system close to a desired value (or set of desired values). This is desirable in cases where system parameters are unknown or change with time. Representative references concerning adaptive control include Anderson et al. (1986), Ioannou and Sun (1995), Krstic et al. (1995), Landau et al. (1998), Narendra and Annaswamy (1989), and Sastry and Bodson (1989).

14.4.4 Hybrid Systems

Hybrid systems are systems characterized by both continuous and discrete dynamics. Examples of hybrid systems include, but are not limited to, digital computer-controlled systems, distributed control systems governed by a hierarchical logical interaction structure, multi-agent systems (such as the air traffic management system [Tomlin, 1998]), and systems characterized by intermittent physical contact [Goodwine and Burdick, 2000]. Recent papers considering modeling and control synthesis methods for such complicated systems include Alur and Henzinger (1996), Antsaklis et al. (1995, 1997), Branicky et al. (1998), Henzinger and Sastry (1998), and Lygeros et al. (1999).

14.5 Parting Remarks

This chapter provides a brief overview of the fundamental concepts in analysis and design of control systems. This chapter includes an outline of classical linear control including stability concepts (the Routh array) and controller design techniques (root locus and lead-lag synthesis). Additionally, more recent advances in control including pole placement, the LQR, and the basic concepts from robust control are outlined and examples are provided. Finally, recent developments in nonlinear control, including feedback linearization (for both single-input, single-output and multi-input, multi-output systems), are outlined along with basic approaches using Lyapunov stability theory.

References

Alur, R., and Henzinger, T., eds. (1996) *Hybrid Systems III: Verification and Control*, Springer-Verlag, New York.

Anderson, B.D.O., Bitmead, R.R., Johnson, C.R., Kokotovic, P.V., Kosut, R.L., Mareels, I.M.Y., Praly, L., and Riedle, B.D. (1986) *Stability of Adaptive Systems*, MIT Press, Cambridge, MA.

Antsaklis, P., Kohn, W., Nerode, A., and Sastry, S., eds. (1995) *Hybrid Systems II*, Springer-Verlag, New York.

Antsaklis, P., Kohn, W., Nerode, A., and Sastry, S., eds. (1997) *Hybrid Systems IV*, Springer-Verlag, New York.

Barmish, B.R., Corless, M., and Leitmann, G. (1983) "A New Class of Stabilizing Controllers for Uncertain Dynamical Systems," *SIAM J. Control Optimization* **21**, pp. 246–355.

Bode, H.W. (1945) *Network Analysis and Feedback Amplifier Design*, D. Van Nostrand, Princeton, NJ.

Boothby, W.M. (1986) *An Introduction to Differentiable Manifolds and Reimannian Geometry*, Academic Press, Boston.

Bower, J.L., and Schultheiss, P. (1961) *Introduction to the Design of Servomechanisms*, Wiley, New York.

Branicky, M., Borkar, V., and Mitter, S.K. (1998) "A Unified Framework for Hybrid Control: Model and Optimal Control Theory," *IEEE Trans. Autom. Control* **AC-43**, pp. 31–45.

Brockett, R.W. (1978) "Feedback Invariants for Nonlinear Systems," in *Proceedings of the 1978 IFAC Congress*, Helsinki, Finland, Pergamon Press, Oxford.

Bullo, F., Leonard, N.E., and Lewis, A. (2000) "Controllability and Motion Algorithms for Underactuated Lagrangian Systems on Lie Groups," *IEEE Trans. Autom. Control* **45**, pp. 1437–54.

Corless, M. (1993) "Control of Uncertain Nonlinear Systems," *J. Dyn. Syst. Meas. Control* **115**, pp. 362–72.

Corless, M., and Leitmann, G. (1981) "Continuous State Feedback Guaranteeing Uniform Ultimate Boundedness for Uncertain Dynamic Systems," *IEEE Trans. Autom. Control* **AC-26**, pp. 1139–44.

Dayawansa, W.P., Boothby, W.M., and Elliott, D. (1985) "Global State and Feedback Equivalence of Nonlinear Systems," *Syst. Control Lett.* **6**, pp. 517–35.

DeCarlo, R.A., Zak, S.H., and Matthews, G.P. (1988) "Variable Structure Control of Nonlinear Multivariable Systems: A Tutorial," *Proc. IEEE* **76**, pp. 212–32.

Dorf, R.C. (1992) *Modern Control Systems*, Addison-Wesley, Reading, MA.

Doyle, J.C., Francis, B.A., and Tannenbaum, A.R. (1992) *Feedback Control Theory*, Macmillan, New York.

Evans, W.R. (1948) "Graphical Analysis of Control Systems," *AIEE Trans. Part II*, pp. 547–51.

Evans, W.R. (1950) "Control System Synthesis by Root Locus Method," *AIEE Trans. Part II*, pp. 66–9.

Franklin, G.F., Powell, D.J., and Emami-Naeini, A. (1994) *Feedback Control of Dynamic Systems*, Addison-Wesley, Reading, MA.

Gajec, Z., and Lelic, M.M. (1996) *Modern Control Systems Engineering*, Prentice-Hall, London.

Goodwine, B., and Burdick, J. (2000) "Motion Planning for Kinematic Stratified Systems with Application to Quasistatic Legged Locomotion and Finger Gaiting," *IEEE J. Robotics Autom.*, accepted for publication.

Henzinger, T., and Sastry, S., eds. (1998) *Hybrid Systems: Computation and Control*, HSCC 2000, Pittsburgh, PA.

Hirsch, M., and Smale, S. (1974) *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, Boston.

Horowitz, I.M. (1963) *Synthesis of Feedback Mechanisms*, Academic Press, New York.

Ioannou, P.A., and Sun, J. (1995) *Robust Adaptive Control*, Prentice-Hall, Englewood Cliffs, NJ.

Isidori, A. (1996) *Nonlinear Control Systems*, Springer-Verlag, Berlin.

Isidori, A., Krener, A.J., Gori-Giorgi, C., and Monaco, S. (1981a) "Locally (f, g)-Invariant Distributions," *Syst. Control Lett.* **1**, pp. 12–5.

Isidori, A., Krener, A.J., Gori-Giorgi, C., and Monaco, S. (1981b) "Nonlinear Decoupling Via Feedback: A Differential Geometric Approach," *IEEE Trans. Autom. Control* **AC-26**, pp. 331–45.

Khalil, H.K. (1996) *Nonlinear Systems*, Prentice-Hall, Englewood Cliffs, NJ.

Krener, A.J. (1987) "Normal Forms for Linear and Nonlinear Systems," *Contempor. Math* **68**, pp. 157–89.

Krstic, M., Kanellakopoulos, I., and Kokotovic, P. (1995) *Nonlinear and Adaptive Control Systems Design*, John Wiley & Sons, New York.

Kuo, B.C. (1995) *Automatic Control Systems*, Prentice-Hall, Englewood Cliffs, NJ.

Lafferriere, G., and Sussmann, H. (1993) "A Differential Geometric Approach to Motion Planning," in *Nonholonomic Motion Planning*, Z. Li and J. F. Canny, eds., Academic Press, New York, pp. 235–70.

Landau, Y.D., Lozano, R., and M'Saad, M. (1998) *Adaptive Control*, Springer-Verlag, Berlin.

Lygeros, J., Godbole, D., and Sastry, S. (1999) "Controllers for Reachability Specifications for Hybrid Systems," *Automatica* **35**, pp. 349–70.

Narendra, K.S., and Annaswamy, A.M. (1989) *Stable Adaptive Systems*, Prentice-Hall, Englewood Cliffs, NJ.

Nijmeijer, H., and van der Schaft, A.J. (1990) *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York.

Nyquist, H. (1932) "Regeneration Theory," *Bell Syst. Technol. J.* **11**, pp. 126–47.

Ogata, K. (1997) *Modern Control Engineering*, Prentice-Hall, Englewood Cliffs, NJ.

Qu, Z. (1993) "Robust Control of Nonlinear Uncertain Systems under Generalized Matching Conditions," *Automatica* **29**, pp. 985–98.

Raven, F.H. (1995) *Automatic Control Engineering*, McGraw-Hill, New York.

Routh, E.J. (1975) *Stability of Motion*, Taylor & Francis, London.

Sastry, S. (2000) *Nonlinear Systems: Analysis, Stability and Control*, Springer-Verlag, New York.

Sastry, S., and Bodson, M. (1989) *Adaptive Systems: Stability, Convergence and Robustness*, Prentice-Hall, Englewood Cliffs, NJ.

Shinners, S.M. (1992) *Modern Control System Theory and Design*, John Wiley & Sons, New York.

Slotine, J.-J.E., and Hedrick, J.K. (1993) "Robust Input—Output Feedback Linearization," *Int. J. Control* **57**, pp. 1133–9.

Spong, M.W., and Vidyasager, M. (1989) *Robot Dynamics and Control*, Wiley, New York.

Tomlin, C., Pappas, G., and Sastry, S. (1998) "Conflict Resolution in Air Traffic Management: A Study in Multi-Agent Hybrid Systems," *IEEE Trans. Autom. Control* **43**(4), pp. 509–21.

Utkin, V.I. (1992) *Sliding Modes in Optimization and Control*, Springer-Verlag, New York.

Zhou, K. (1996) *Robust and Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ.

Ziegler, J.G., and Nichols, N.B. (1942) "Optimum Settings for Automatic Controllers," *ASME Trans.* **64**, pp. 759–68.

Ziegler, J.G., and Nichols, N.B. (1943) "Process Lags in Automatic Control Circuits," *ASME Trans.* **65**, pp. 433–44.

15

Model-Based Flow Control for Distributed Architectures

Thomas R. Bewley
University of California

As traditional scientific disciplines individually grow toward their maturity, many new opportunities for significant advances lie at their intersection. For example, remarkable developments in control theory in the last few decades have considerably expanded the selection of available tools which may be applied to regulate physical and electrical systems. When combined with microelectromechanical systems (MEMS) techniques for distributed sensing and actuation, as highlighted elsewhere in this handbook, these techniques hold great promise for several applications in fluid mechanics, including the delay of transition and the regulation of turbulence. Such applications of control theory require a very balanced perspective in which one considers the relevant flow physics when designing the control algorithms and, conversely, takes into account the requirements and limitations of control algorithms when designing both reduced-order flow models and the fluid-mechanical systems to be controlled. Such a balanced perspective is elusive, however, as both the research establishment in general and universities in particular are accustomed only to the dissemination and teaching of component technologies in isolated fields. To advance, we must not toss substantial new interdisciplinary questions over the fence for fear of them being "outside our area;" rather, we must break down these very fences that limit us and attack these challenging new questions with a Renaissance approach. In this spirit, this chapter surveys a few recent attempts at bridging the gaps between the several scientific disciplines comprising the field of flow control, in an attempt to clarify the author's perspective on how recent advances in these constituent disciplines fit together in a manner that opens up significant new research opportunities.

15.1 Introduction

Flow control is perhaps the most difficult grand challenge application area for MEMS technology. Potentially, it is one of the most rewarding because a common feature in many fluid systems is the existence of natural instability mechanisms by which a small input, when coordinated correctly, can lead to a large response in the overall system. As one of the key driving application areas for MEMS, it is appropriate to survey recent developments in the fundamental framework for flow control in this handbook.

The area of flow control plainly resides at the intersection of disciplines, incorporating essential and nontrivial elements from control theory, fluid mechanics, Navier–Stokes mathematics, numerical methods, and fabrication technology for "small" (millimeter-scale), self-contained, durable devices which can integrate the functions of sensing, actuation, and control logic. Recent developments in the integration of these disciplines, while grounding us with appropriate techniques to address some fundamental open questions, hint at the solution of several new questions. To follow up on these new directions, it is essential to have a clear vision of how recent advances in these fields fit together and to know where the significant unresolved issues at their intersection lie.

This chapter attempts to elucidate the utility of an interdisciplinary perspective to this type of problem by focusing on the control of a prototypical and fundamental fluid system: plane channel flow. The control of the flow in this simple geometry embodies a myriad of complex issues and interrelationships. These issues and relationships require us to draw from a variety of traditional disciplines. Only when these issues and perspectives are combined is a complete understanding of the state of the art achieved and a vision of where to proceed identified.

Though plane channel flow will be the focus problem we discuss here, the purpose of this work goes well beyond simply controlling this particular flow with a particular actuator/sensor configuration. At its core, the research effort we describe is devoted to the development of an integrated, interdisciplinary understanding that allows us to synthesize the necessary tools to attack a variety of flow control problems in the future. The focus problem of control of channel flow is chosen not simply because of its technological relevance or fundamental character, but because it embodies many of the important unsolved issues encountered in the assortment of new flow control problems that will inevitably follow. The primary objective of this work is to lay a solid, integrated footing upon which these future efforts may be based.

To this end, this chapter will describe mostly the efforts with which the author has been directly involved, in an attempt to weave the story that threads these projects together as part of the fabric of a substantial new area of interdisciplinary research. Space does not permit complete development of these projects; rather, the chapter will survey a selection of recent results that bring the relevant issues to light. Refer to the appropriate full journal articles for all of the relevant details and careful placement of these projects in context with the works of others. Space limitations also do not allow this brief chapter to adequately review the various directions all my friends and colleagues are taking in this field. Rather than attempt such a review and fail, refer to a host of other recent reviews which span only a fraction of the current work being done in this active area of research. For an experimental perspective, refer to several other chapters in this handbook and to the recent reviews of Ho and Tai (1996, 1998), McMichael (1996), Gad-el-Hak (1996), and Löfdahl and Gad-el-Hak (1999). For a mathematical perspective, refer to the recent dedicated volumes compiled by Banks (1992), Banks et al. (1993), Gunzburger (1995), Lagnese et al. (1995), and Sritharan (1998) for a sampling of recent results in this area.

15.2 Linearization: Life in a Small Neighborhood

As a starting point for the introduction of control theory into the fluid-mechanical setting, we first consider the linearized system arising from the equation governing small perturbations to a laminar flow. From a physical point of view, such perturbations are quite significant because they represent the initial stages of the complex process of transition to turbulence. Therefore, their mitigation or enhancement has a substantial effect on the evolution of the flow.

An enlightening problem that captures the essential physics of many important features of both transition and turbulence in wall-bounded flows is that of plane channel flow, as illustrated in Figure 15.1. Assume the walls are located at $y = \pm 1$. We begin our study by analyzing small perturbations $\{u, v, w, p\}$ to the (parabolic) laminar flow profile $U(y)$ in this geometry, which are governed by the linearized incompressible Navier–Stokes equation:

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0, \tag{15.1a}$$

$$\dot{u} + U\frac{\partial}{\partial x}u + U'v = -\frac{\partial p}{\partial x} + \frac{1}{\mathrm{Re}}\Delta u, \tag{15.1b}$$

$$\dot{v} + U\frac{\partial}{\partial x}v = -\frac{\partial p}{\partial y} + \frac{1}{\mathrm{Re}}\Delta v, \tag{15.1c}$$

$$\dot{w} + U\frac{\partial}{\partial x}w = -\frac{\partial p}{\partial z} + \frac{1}{\mathrm{Re}}\Delta w. \tag{15.1d}$$

Equation (15.1a), the continuity equation, constrains the solution of Equations (15.1b) to (15.1d), the momentum equations, to be divergence free. This constraint is imposed through the ∇p terms in the momentum equations, which act as Lagrange multipliers to maintain the velocity field on a divergence-free submanifold of the space of square-integrable vector fields. In the discretized setting, such systems are

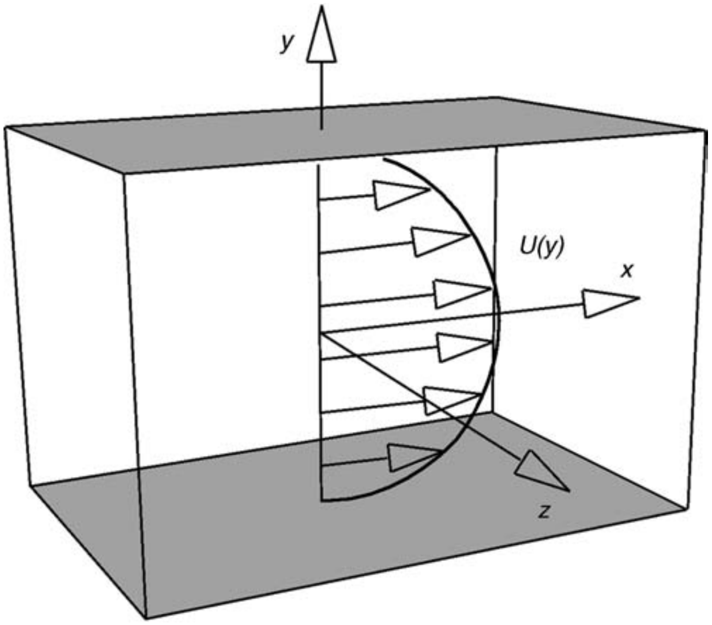


FIGURE 15.1 Geometry of plane channel flow. The flow is sustained by an externally applied pressure gradient in the *x* direction. This canonical problem provides an excellent testbed for the study of both transition and turbulence in wall-bounded flows. Many of the important flow phenomena in this geometry, in both the linear and nonlinear setting, are fundamentally three dimensional. A nonphysical assumption of periodicity of the flow perturbations in the *x* and *z* directions is often assumed for numerical convenience, with the box size chosen to be large enough that this nonphysical assumption has minimal effect on the observed flow statistics. It is important to evaluate critically the implications of such assumptions during the process of control design, as discussed in detail in Sections 15.4 and 15.5.

called descriptor systems or differential-algebraic equations and, defining a state vector **x** and a control vector **u**, may be written in the generalized state-space form:

$$E\dot{\mathbf{x}} = A\mathbf{x} + B\mathbf{u}. \tag{15.2}$$

If the Navier–Stokes Equation (15.1) is put directly into this form, *E* is singular. This is an essential feature of the Navier–Stokes equation that necessitates careful treatment in both simulation and control design to avoid spurious numerical artifacts. A variety of techniques exist to express the system of Equations (15.1) with a reduced set of variables or spatially distributed functions with only two degrees of freedom per spatial location, referred to as a divergence-free basis. In such a basis, the continuity equation is applied implicitly, and the pressure is eliminated from the set of governing equations. All three velocity components and the pressure (up to an arbitrary constant) may be determined from solutions represented in such a basis. When discretized and represented in the form of Equation (15.2), the Navier–Stokes equation written in such a basis leads to an expression for E that is nonsingular.

For the geometry indicated in Figure 15.1, a suitable choice for this reduced set of variables, which is convenient in terms of the implementation of boundary conditions, is the wall-normal velocity *v* and the wall-normal vorticity, $\omega \triangleq \partial u/\partial z - \partial w/\partial x$. Taking the Fourier transform of Equation (15.1) in the streamwise and spanwise directions and manipulating these equations and their derivatives leads to the classical Orr–Sommerfeld/Squire formulation of the Navier–Stokes equation at each wavenumber pair $\{k_x, k_z\}$:

$$\hat{\Delta}\dot{\hat{v}} = \{-ik_x U\hat{\Delta} + ik_x U'' + \hat{\Delta}(\hat{\Delta}/\mathrm{Re})\}\hat{v}, \tag{15.3a}$$

$$\dot{\hat{\omega}} = \{-ik_z U'\}\hat{v} + \{-ik_x U + \hat{\Delta}/\mathrm{Re}\}\hat{\omega}, \tag{15.3b}$$

where the hats (ˆ) indicate Fourier coefficients and the Laplacian now takes the form $\hat{\Delta} \triangleq \partial^2/\partial y^2 - k_x^2 - k_z^2$. Particular care is needed when solving this system; to invert the Laplacian on the LHS of Equation

(15.3a), the boundary conditions on v must be accounted for properly. By manipulating the governing equations and casting them in a derivative form, we effectively trade one numerical difficulty (singularity of E) for another (a tricky boundary condition inclusion to make the Laplacian on the LHS of Equation (15.3a) invertible).

Note the spatially invariant structure of the present geometry: every point on each wall is, statistically speaking, identical to every other point on that wall. Canonical problems with this sort of spatially invariant structure in one or more directions form the backbone of much of the literature on flow transition and turbulence. It is this structure that facilitates the use of Fourier transforms to completely decouple the system state $\{\hat{v}, \hat{\omega}\}$ at each wavenumber pair $\{k_x, k_z\}$ from the system state at every other wavenumber pair, as indicated in Equation (15.3). Such decoupling of the Fourier modes of the unforced linear system in the directions of spatial invariance is a classical result upon which much of the available linear theory for the stability of Navier–Stokes systems is based. As noted by Bewley and Agarwal (1996), taking the Fourier transform of both the control variables and the measurement variables maintains this system decoupling in the control formulation, greatly reducing the complexity of the control design problem to several smaller, completely decoupled control design problems at each wavenumber pair $\{k_x, k_z\}$, each of which requires spatial discretization in the y direction only.

Once a tractable form of the governing equation has been selected, to pose the flow control problem completely, several steps remain:

- the state equation must be spatially discretized,
- boundary conditions must be chosen and enforced,
- the variables representing the controls and the available measurements must be identified and extracted,
- the disturbances must be modeled, and
- the "control objective" must be precisely defined.

To identify a fundamental yet physically relevant flow control problem, the decisions made at each of these steps require engineering judgment. Such judgment is based on physical insight concerning the flow system to be controlled and how the essential features of such a system may be accurately modeled. An example of how to accomplish these steps is described in some detail by Bewley and Liu (1998). In short, we may choose:

- a Chebyshev spatial discretization in y,
- no-slip boundary conditions ($u = w = 0$ on the walls) with the distribution of v on the walls (the blowing/suction profile) prescribed as the control,
- skin friction measurements distributed on the walls,
- idealized disturbances exciting the system, and
- an objective of minimizing flow perturbation energy.

As we learn more about the physics of the system to be controlled, there is significant room for improvement in this problem formulation, particularly in modeling the structure of relevant system disturbances and in the precise statement of the control objective.

Once the previously mentioned steps are complete, the present decoupled system at each wavenumber pair $\{k_x, k_z\}$ may finally be manipulated into the standard state-space form:

$$\begin{aligned} \dot{\mathbf{x}} &= A\mathbf{x} + B_1\mathbf{w} + B_2\mathbf{u}, \\ \mathbf{y} &= C_2\mathbf{x} + D_{21}\mathbf{w}, \end{aligned} \tag{15.4}$$

with

$$B_1 \triangleq (G_1\ 0), \quad C_2 \triangleq G_2^{-1}C, \quad D_{21} \triangleq (0\ \alpha I), \quad \mathbf{w} \triangleq \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{pmatrix},$$

where $\mathbf{x}$ denotes the state, $\mathbf{u}$ denotes the control, $\mathbf{y}$ denotes the available measurements (scaled as discussed below), and $\mathbf{w}$ accounts for the external disturbances (including the state disturbances $\mathbf{w}_1$ and the measurement noise $\mathbf{w}_2$, scaled as discussed below). Note that $C\mathbf{x}$ denotes the raw vector of measured variables, and G_1 and αG_2 represent the square root of any known or expected covariance structure of the state disturbances and measurement noise, respectively. The scalar α^2 is identified as an adjustable parameter that defines the ratio of the maximum singular value of the covariance of the measurement noise divided by the maximum singular value of the covariance of the state disturbances; without loss of generality, we take $\bar{\sigma}(G_1) = \bar{\sigma}(G_2) = 1$. Effectively, the matrix G_1 reflects which state disturbances are strongest, and the matrix G_2 reflects which measurements are most corrupted by noise. Small a implies relatively high overall confidence in the measurements, whereas large α implies relatively low overall confidence in the measurements.

Not surprisingly, there is a wide body of theory surrounding how to control a linear system in the standard form of Equation (15.4). The application of one popular technique (to a related two-dimensional problem), called proportional–integral (PI) control and generally referred to as "classical" control design, is presented in Joshi et al. (1997). The application of another technique, called $\mathcal{H}_\infty$ control and generally referred to as "modern" control design, is laid out in Bewley and Liu (1998). The application of a related modern control strategy (to the two-dimensional problem), called *loop transfer recovery* (LTR), is presented in Cortelezzi and Speyer (1998). More recent publications by these groups further extend these seminal efforts.

It is useful to understand the various theoretical implications of the control design technique chosen. Ultimately, however, flow control is the design of a control that achieves the desired engineering objective (transition delay, drag reduction, mixing enhancement, etc.) to the maximum extent possible. The theoretical implications of the particular control technique chosen are useful only to the degree to which they help attain this objective. Engineering judgment, based on an understanding of the merits of the various control theories and based on the suitability of such theories to the structure of the fluid-mechanical problem of interest, guides the selection of an appropriate control design strategy. In the following section, we summarize the $\mathcal{H}_\infty$ control design approach, illustrate why this approach is appropriate for the structure of the problem at hand, and highlight an important distinguishing characteristic of the present system when controls computed via this approach are applied.

15.3 Linear Stabilization: Leveraging Modern Linear Control Theory

As only a limited number of noisy measurements $\mathbf{y}$ of the state $\mathbf{x}$ are available in any practical control implementation, it is beneficial to develop a filter that extracts as much useful information as possible from the available flow measurements before using this filtered information to compute a suitable control. In modern control theory, a model of the system itself is used as this filter, and the filtered information extracted from the measurements is simply an estimate of the state of the physical system. This intuitive framework is illustrated schematically in Figure 15.2. By modeling (or neglecting) the influence of the unknown disturbances in Equation (15.4), the system model takes the form:

$$\dot{\hat{\mathbf{x}}} = A\hat{\mathbf{x}} + B_1\hat{\mathbf{w}} + B_2\mathbf{u} - \mathbf{v}, \tag{15.5a}$$

$$\hat{\mathbf{y}} = C_2\hat{\mathbf{x}} + D_{21}\hat{\mathbf{w}}, \tag{15.5b}$$

where $\hat{\mathbf{x}}$ is the state estimate, $\hat{\mathbf{w}}$ is a disturbance estimate, and $\mathbf{v}$ is a feedback term based on the difference between the measurement of the state $\mathbf{y}$ and the corresponding quantity in the model, $\hat{\mathbf{y}}$, such that:

$$\mathbf{v} = L(\mathbf{y} - \hat{\mathbf{y}}). \tag{15.5c}$$

The control $\mathbf{u}$, in turn, is based on the state estimate $\hat{\mathbf{x}}$ such that:

$$\mathbf{u} = K\hat{\mathbf{x}}. \tag{15.6}$$

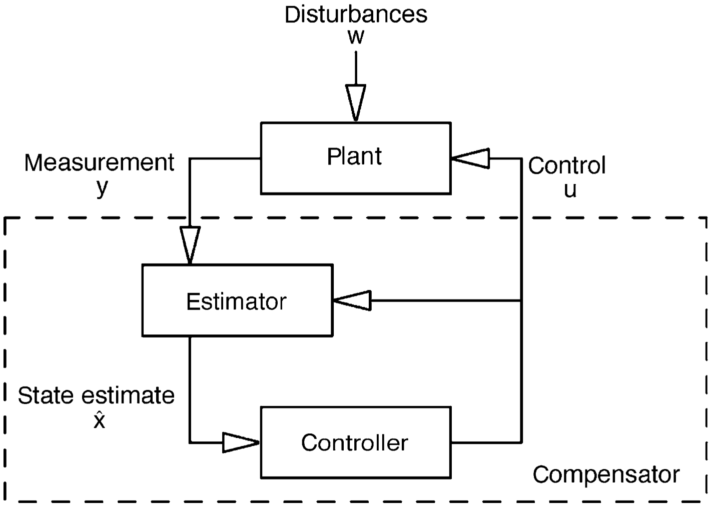


FIGURE 15.2 Flow of information in a modern control realization. The plant, forced by external disturbances, has an internal state **x** which cannot be observed. Instead, a noisy measurement **y** is made, with which a state estimate $\hat{\mathbf{x}}$ is determined. This state estimate is then used to determine the control **u** to be applied to the plant to regulate **x** to zero. Essentially, the full equation for the plant (or a reduced model thereof) is used in the estimator as a filter to extract useful information about the state from the available measurements.

Equation (15.4) is referred to as the "plant," Equation (15.5) is referred to as the "estimator," and Equation (15.6) is referred to as the "controller." The estimator and the controller, taken together, will be referred to as the "compensator." The problem at hand is to compute linear time-invariant (LTI) matrices *K* and *L* and some estimate of the disturbance, $\hat{\mathbf{w}}$, such that:

1. the estimator feedback **v** forces $\hat{\mathbf{x}}$ toward **x**, and
2. the controller feedback **u** forces **x** toward zero,

even as unknown disturbances **w** both disrupt the system evolution and corrupt the available measurements of the system state.

15.3.1 The $\mathcal{H}_\infty$ Approach to Control Design

Several textbooks describe in detail how the $\mathcal{H}_\infty$ technique determines *K*, *L*, and $\hat{\mathbf{w}}$ for systems of the form Equations (15.4) to (15.6) in the presence of structured and unstructured disturbances **w**. Refer to the seminal paper by Doyle et al. (1989), the more accessible textbook by Green and Limebeer (1995), and the more advanced texts by Zhou et al. (1996) and Zhou and Doyle (1998) for derivation and further discussion of these control theories. Refer to Bewley and Liu (1998) for an extended discussion in the context of the present problem. To summarize this approach briefly, a cost function $\mathcal{J}$ describing the control problem at hand is defined that weighs together the state **x**, the control **u**, and the disturbance **w** such that:

$$\mathcal{J} \triangleq E[\mathbf{x}^* Q \mathbf{x} + \ell^2 \mathbf{u}^* \mathbf{u} - \gamma^2 \mathbf{w}^* \mathbf{w}] \triangleq E[\mathbf{z}^* \mathbf{z} - \gamma^2 \mathbf{w}^* \mathbf{w}], \tag{15.7a}$$

where

$$\mathbf{z} \triangleq C_1 \mathbf{x} + D_{12} \mathbf{u}, \quad C_1 \triangleq \begin{pmatrix} Q^{1/2} \\ 0 \end{pmatrix}, \quad D_{12} \triangleq \begin{pmatrix} 0 \\ \ell I \end{pmatrix}. \tag{15.7b}$$

The matrix *Q*, shaping the dependence on the state in the cost function $\mathbf{x}^* Q \mathbf{x}$, may be selected to numerically approximate any of a variety of physical properties of the flow, such as the flow perturbation energy, its enstrophy, the mean square of the drag measurements, etc. The matrix Q may also be biased to place extra penalty on flow perturbations in a specific region in space of particular physical significance. The choice of Q has a profound effect on the final closed-loop behavior, and it must be selected with care. Based on our numerical tests to date, cost functions related to the energy of the flow perturbations have been the most successful for the purpose of transition delay. To simplify the algebra that follows, we have set the matrices R and S shaping the $\mathbf{u}^*R\mathbf{u}$ and $\mathbf{w}^*S\mathbf{w}$ terms in the cost function equal to I. As shown in Lauga and Bewley (2000), it is straightforward to generalize this result to other positive-definite choices for R and S. Such a generalization is particularly useful when designing controls for a discretization of a partial differential equation (PDE) in a consistent manner such that the feedback kernels converge to continuous functions as the computational grid is refined.

Given the structure of the system defined in Equations (15.4) to (15.6) and the control objective defined in Equation (15.7), the $\mathcal{H}_\infty$ compensator is determined by simultaneously minimizing the cost function $\mathcal{J}$ with respect to the control $\mathbf{u}$ and maximizing $\mathcal{J}$ with respect to the disturbance $\mathbf{w}$. In such a way, a control $\mathbf{u}$ is found that maximally attains the control objective even in the presence of a disturbance $\mathbf{w}$ that maximally disrupts this objective. For sufficiently large γ and a system that is both stabilizable and *detectable* via the controls and measurements chosen, this results in finite values for $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$, the magnitudes of which may be adjusted by variation of the three scalar parameters ℓ, α, and γ, respectively. Reducing ℓ, modeling the "price of the control" in the engineering design, generally results in increased levels of control feedback $\mathbf{u}$. Reducing α, modeling the "relative level of corruption" of the measurements by noise, generally results in increased levels of estimator feedback $\mathbf{v}$. Reducing γ, modeling the "price" of the disturbance to nature (in the spirit of a noncooperative game), generally results in increased levels of disturbances $\mathbf{w}$ of maximally disruptive structure to be accounted for during the design of the compensator.

The $\mathcal{H}_\infty$ control solution [Doyle et al., 1989] may be described as follows: a compensator that minimizes $\mathcal{J}$ in the presence of that disturbance which simultaneously maximizes $\mathcal{J}$ is given by:

$$K = -\frac{1}{\ell^2}B_2^*X, \quad L = -\frac{1}{\alpha^2}ZYC_2^*, \quad \hat{\mathbf{w}} = \frac{1}{\gamma^2}B_1^*X\hat{\mathbf{x}}, \tag{15.8}$$

where

$$X = \mathrm{Ric}\begin{pmatrix} A & \dfrac{1}{\gamma^2}B_1B_1^* - \dfrac{1}{\ell^2}B_2B_2^* \\ -C_1^*C_1 & -A^* \end{pmatrix},$$

$$Y = \mathrm{Ric}\begin{pmatrix} A^* & \dfrac{1}{\gamma^2}C_1^*C_1 - \dfrac{1}{\alpha^2}C_2^*C_2 \\ -B_1B_1^* & -A \end{pmatrix},$$

$$Z = \left(1 - \frac{YX}{\gamma^2}\right)^{-1},$$

where Ric $(\cdot)$ denotes the positive-definite solution of the associated Riccati equation [Laub, 1991]. The simple structure of the above solution, and its profound implications in terms of the performance and robustness of the resulting closed-loop system, is one of the most elegant results of linear control theory. The following comments touch on a few of the more salient features of this result.

Algebraic manipulation of Equations (15.4) to (15.8) leads to the closed-loop form:

$$\begin{aligned} \dot{\tilde{\mathbf{x}}} &= \tilde{A}\mathbf{x} + \tilde{B}\mathbf{w}, \\ \mathbf{z} &= \tilde{C}\tilde{\mathbf{x}}, \end{aligned} \tag{15.9}$$

where

$$\tilde{\mathbf{x}} = \begin{pmatrix} \mathbf{x} \\ \mathbf{x} - \hat{\mathbf{x}} \end{pmatrix},$$

$$\tilde{A} = \begin{pmatrix} A + B_2K & -B_2K \\ -\gamma^{-2}B_1B_1^* & A + LC_2 + \gamma^{-2}B_1B_1^* \end{pmatrix},$$

$$\tilde{B} = \begin{pmatrix} B_1 \\ B_1 + LD_{21} \end{pmatrix},$$

$$\tilde{C} = (C_1 + D_{12}K - D_{12}K).$$

Taking the Laplace transform of Equation (15.9), it is easy to define the transfer function $T_{\mathrm{zw}}(s)$ from $\mathbf{w}(s)$ to $\mathbf{z}(s)$ (the Laplace transforms of $\mathbf{w}$ and $\mathbf{z}$) such that:

$$\mathbf{z}(s) = \tilde{C}(sI - \tilde{A})^{-1}\tilde{B}\mathbf{w}(s) \triangleq T_{\mathrm{zw}}(s)\mathbf{w}(s).$$

Norms of the system transfer function $T_{\mathrm{zw}}(s)$ quantify how the system output of interest $\mathbf{z}$ responds to disturbances $\mathbf{w}$ exciting the closed-loop system.

The expected value of the root mean square (rms) of the output $\mathbf{z}$ over the rms of the input $\mathbf{w}$ for disturbances $\mathbf{w}$ of maximally disruptive structure is denoted by the ∞–norm of the system transfer function,

$$\|T_{\mathrm{zw}}\|_\infty \triangleq \sup_{\omega} \overline{\sigma}\, [T_{\mathrm{zw}}(j\omega)].$$

$\mathcal{H}_\infty$ control is often referred to as "robust" control, as $\|T_{\mathrm{zw}}\|_\infty$, reflecting the worst-case amplification of disturbances by the system from the input $\mathbf{w}$ to the output $\mathbf{z}$, is in fact bounded from above by the value of γ used in the problem formulation. Subject to this ∞–norm bound, $\mathcal{H}_\infty$ control minimizes the expected value of the rms of the output $\mathbf{z}$ over the rms of the input $\mathbf{w}$ for white Gaussian disturbances $\mathbf{w}$ with identity covariance, denoted by the 2–norm of the system transfer function:

$$\|T_{\mathrm{zw}}\|_2 \triangleq \left(\frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{trace}[T_{\mathrm{zw}}(j\omega)^* T_{\mathrm{zw}}(j\omega)]\,d\omega\right)^{1/2}.$$

Note that $\|T_{\mathrm{zw}}\|_2$ is often cited as a measure of performance of the closed-loop system, whereas $\|T_{\mathrm{zw}}\|_\infty$ is often cited as a measure of its robustness. Further motivation for consideration of control theories related to these particular norms is elucidated by Skogestad and Postlethwaite (1996). Efficient numerical algorithms to solve the Riccati equations for X and Y in the compensator design and to compute the transfer function norms $\|T_{\mathrm{zw}}\|_2$ and $\|T_{\mathrm{zw}}\|_\infty$ quantifying the closed-loop system behavior are well developed and are discussed further in the standard texts.

For high-dimensional discretizations of infinite dimensional systems, it is not feasible to perform a parametric variation on the individual elements of the matrices defining the control problem. The control design approach taken here represents a balance of engineering judgment in the construction of the matrices defining the structure of the control problem $\{B_1, B_2, C_1, C_2\}$ and parametric variation of the three scalar parameters involved $\{\ell, \alpha, \gamma\}$ to achieve the desired trade-offs between performance, robustness, and the control effort required. This approach retains a sufficient but not excessive degree of flexibility in the control design process. In general, intermediate values of the three parameters $\{\ell, \alpha, \gamma\}$ lead to the most suitable control designs.

$\mathcal{H}_2$ control (also known as linear quadratic Gaussian control, or LQG) is an important limiting case of $\mathcal{H}_\infty$ control. It is obtained in the present formulation by relaxing the bound γon the infinity norm of the closed-loop system, taking the limit as $\gamma \to \infty$ in the controller formulation. Such a control formulation focuses solely on performance (i.e., minimizing $\|T_{\mathrm{zw}}\|_2$). As LQG does not provide any guarantees about system behavior for disturbances of particularly disruptive structure ($\|T_{\mathrm{zw}}\|_\infty$), it is often referred to as "optimal" control. Though one might confirm *a* posteriori that a particular LQG design has favorable robustness properties, such properties are not guaranteed by the LQG control design process. When designing a large number of compensators for an entire array of wavenumber pairs $\{k_x, k_z\}$ via an automated algorithm, as is necessary in the current problem, it is useful to have a control design tool that inherently builds in system robustness, such as $\mathcal{H}_\infty$. For isolated low-dimensional systems, as often encountered in many industrial processes, a posteriori robustness checks on hand-tuned LQG designs are often sufficient.

It is also interesting that certain favorable robustness properties may be assured by the LQG approach by strategies involving either:

1. setting $B_1 = (B_2\ 0)$ and taking $\alpha \to 0$, or
2. setting $C_1 = \begin{pmatrix} C_2 \\ 0 \end{pmatrix}$ and taking $\ell \to 0$.

These two approaches are referred to as loop transfer recovery (LQG/LTR), and are further explained in Stein and Athans (1987). Such a strategy is explored by Cortelezzi and Speyer (1998) in the two-dimensional setting of the current problem. In the present system, both B_2 and C_2 are very low rank because there is only a single control variable and a single measurement variable at each wall in the Fourier-space representation of the physical system at each wavenumber pair $\{k_x, k_z\}$. However, the state itself is a high-dimensional approximation of an infinite-dimensional system. It is beneficial in such a problem to allow the modeled state disturbances $\mathbf{w}_1$ to input the system, via the matrix B_1, at more than just the actuator inputs, and to allow the response of the system $\mathbf{x}$ to be weighted in the cost function, via the matrix C_1, at more than just the sensor outputs. The LQG/LTR approach of assuring closed-loop system robustness, however, requires us to sacrifice one of these features in the control formulation, in addition to taking $\alpha \to 0$ or $\ell \to 0$, to apply one of the two strategies listed above. It is noted here that the $\mathcal{H}_\infty$ approach, when soluble, allows for the design of compensators with inherent robustness guarantees without such sacrifices of flexibility in the definition of the control problem of interest, thereby giving significantly more latitude in the design of a "robust" compensator.

The names $\mathcal{H}_2$ and $\mathcal{H}_\infty$ are derived from the system norms $\|T_{zw}\|_2$ and $\|T_{zw}\|_\infty$ that these control theories address, with the symbol $\mathcal{H}$ denoting the particular "Hardy space" in which these transfer function norms are well defined. It deserves mention that the difference between $\|T_{zw}\|_2$ and $\|T_{zw}\|_\infty$ might be expected to be increasingly significant as the dimension of the system is increased. Neglecting, for the moment, the dependence on ω in the definition of the system norms, the matrix Frobenius norm (trace$[T^*T]^{1/2}$) and the matrix 2–norm $\bar{\sigma}[T]$ are "equivalent" up to a constant. Indeed, for scalar systems these two matrix norms are identical, and for low-dimensional systems their ratio is bounded by a constant related to the dimension of the system. For high-dimensional discretizations of infinite-dimensional systems, however, this norm equivalence is relaxed, and the differences between these two matrix norms may be substantial. The temporal dependence of the two system norms $\|T_{zw}\|_2$ and $\|T_{zw}\|_\infty$ distinguishes them even for low-dimensional systems. For high-dimensional systems, the important differences between these two system norms are even more pronounced, and control techniques such as $\mathcal{H}_\infty$ that account for both such norms might prove to be beneficial. Techniques (such as $\mathcal{H}_\infty$) that bound $\|T_{zw}\|_\infty$ are especially appropriate for the present problem, as transition is often associated with the triggering of a "worst-case" phenomenon, which is well characterized by this measure.

15.3.2 Advantages of Modern Control Design for Non-Normal Systems

Matrices A arising from the discretization of systems in fluid mechanics are often highly "non-normal," which means that the eigenvectors of A are highly nonorthogonal. This is especially true for transition in a plane channel. Important characteristics of this system, such as $O(1000)$ transient energy growth and large amplification of external disturbance energy in stable flows at subcritical Reynolds numbers, cannot be explained by examination of its eigenvalues alone. Discretizations of Equation (15.3), when put into the state-space form of Equation (15.4), lead to system matrices of the form:

$$A = \begin{pmatrix} L & 0 \\ C & S \end{pmatrix}. \tag{15.10}$$

For certain wavenumber pairs (specifically, those with $k_x \approx 0$ and $k_z = O(1)$), the eigenvalues of A are real and stable, the matrices L and S are quite similar in structure, and $\bar{\sigma}(C)$ is disproportionately large.

To illustrate the behavior of a system matrix with such structure, consider a reduced system matrix of the previous form but where L, C, and S are scalars. Specifically, compare the two stable closed-loop system matrices:

$$A_1 = \begin{pmatrix} -0.01 & 0 \\ 0 & -0.011 \end{pmatrix},$$

$$A_2 = \begin{pmatrix} -0.01 & 0 \\ 1 & -0.011 \end{pmatrix}.$$

Both matrices have the same eigenvalues. However, the eigenvectors of A_1 are orthogonal, whereas the eigenvectors of A_2 are

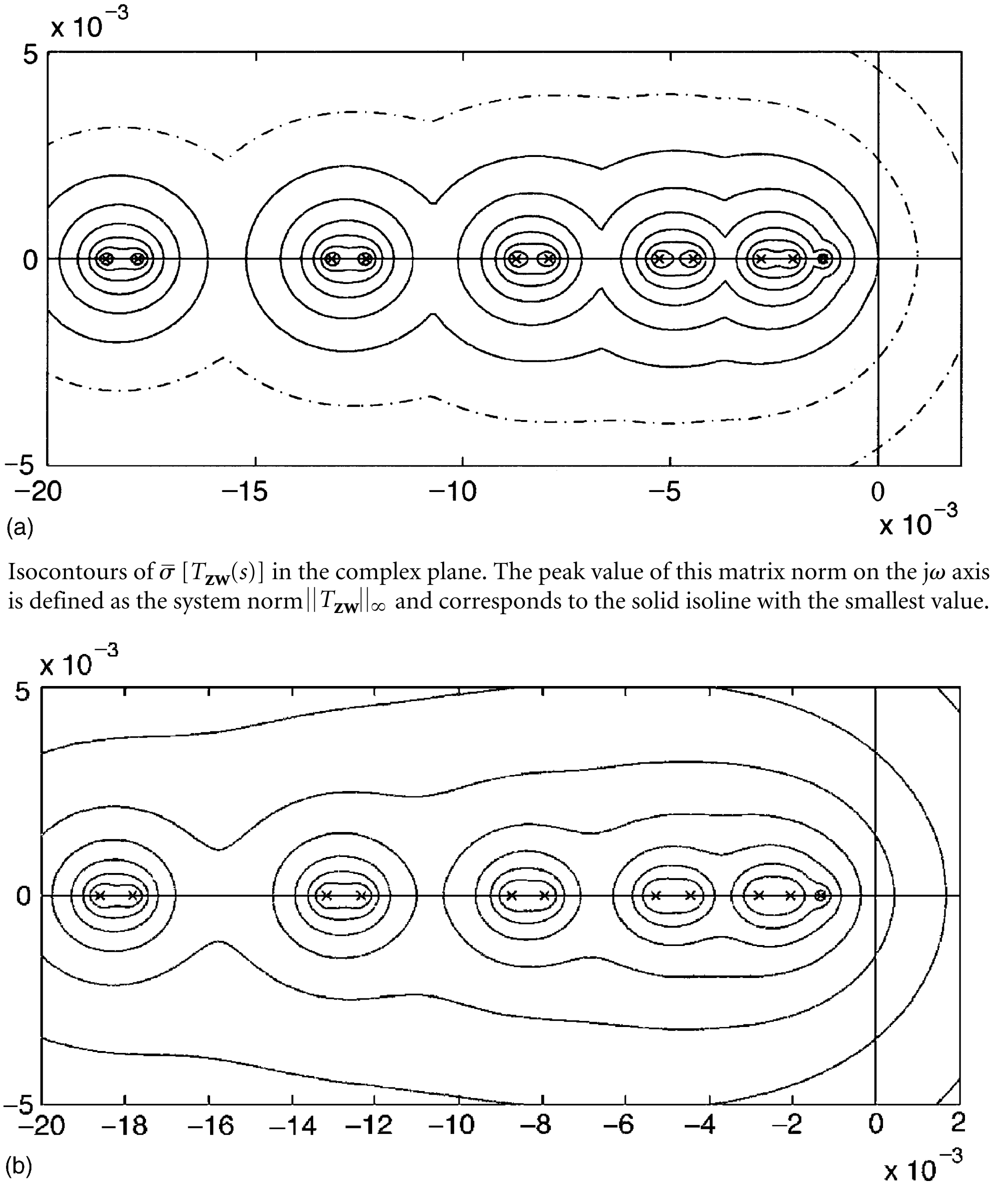
$$\xi_1 = \begin{pmatrix} 0.001 \\ 1.000 \end{pmatrix} \quad \text{and} \quad \xi_2 = \begin{pmatrix} 0 \\ 1.000 \end{pmatrix}.$$

Even though its eigenvalues differ by 10%, the eigenvectors of A_2 are less than 0.06° from being exactly parallel. It is in this sense that we define this system as being "non-normal" or "nearly defective." This severe nonorthogonality of the system eigenvectors is a direct result of the disproportionately large coupling term C. Compensators that reduce C will make the eigenvectors of A_2 closer to orthogonal without necessarily changing the system eigenvalues.

The consequences of nonorthogonality of the system eigenvectors are significant. Though the "energy" (the Euclidean norm) of the state of the system $\dot{\mathbf{x}} = A_1\mathbf{x}$ uniformly decreases in time from all initial conditions, the "energy" of the state of the system $\dot{\mathbf{x}} = A_2\mathbf{x}$ from the initial condition $\mathbf{x}(0) = \xi_1 - \xi_2$ grows by a factor of over a thousand before eventually decaying due to the stability of the system. This is referred to as the transient energy growth of the stable non-normal system and is a result of the reduced destructive interference exhibited by the two modes of the solution as they decay at different rates. In fluid mechanics, transient energy growth is thought to be an important linear mechanism leading to transition in subcritical flows, which are linearly stable but nonlinearly unstable [Butler and Farrell, 1992].

The excitation of such systems by external disturbances is well described in terms of the system norms $\|T_{\mathrm{zw}}\|_2$ and $\|T_{\mathrm{zw}}\|_\infty$, which (as described previously) quantify the rms amplification of Gaussian and worst-case disturbances by the system. For example, consider a closed-loop system of the form of Equation (15.9) with $\tilde{B} = \tilde{C} = I$. Taking the system matrix $\tilde{A} = A_1$, the norms of the system transfer function are $\|T_{\mathrm{zw}}\|_2 = 9.8$ and $\|T_{\mathrm{zw}}\|_\infty = 100$. Alternatively, taking the system matrix $\tilde{A} = A_2$, the 2–norm of the system transfer function is 48 times larger and the ∞–norm is 91 times larger, though the two systems have identical closed-loop eigenvalues. Large system-transfer-function norms and large values of maximum transient energy growth are often highly correlated because they both are a result of nonnormality in a stable system.

Graphical interpretations of $\|T_{\mathrm{zw}}\|_2$ and $\|T_{\mathrm{zw}}\|_\infty$ for the present channel flow system are given in Figures 15.3 and 15.4 by examining contour plots of the appropriate matrix norms of $T_{\mathrm{zw}}(s)$ in the complex plane s. Recall that $T_{\mathrm{zw}}(s) \triangleq \tilde{C}(sI - \tilde{A})^{-1}\tilde{B}$, therefore these contours approach infinity in the neighborhood of each eigenvalue of $\tilde{A}$. Contour plots of this type have recently become known as the pseudospectra of an input/output system and have become a popular generalization of plots of the eigenvalues of $\tilde{A}$ in recent efforts to study nonnormality in uncontrolled fluid systems [Trefethen et al., 1993]. For the open-loop systems depicted in these figures, we define $\tilde{A} = A$, $\tilde{B} = B_1$, and $\tilde{C} = C_1$. The severe non-normality of the present fluid system for Fourier modes with $k_x \approx 0$ is reflected by the elliptical isolines surrounding each pair of eigenvalues with nearly parallel eigenvectors in these pseudospectra, a feature that is much more pronounced in the system depicted in Figure 15.3 than in that depicted in Figure 15.4. The severe non-normality of the system depicted in Figure 15.3 is also reflected by its much larger value of $\|T_{\mathrm{zw}}\|_\infty$. As $\{\tilde{A}, \tilde{B}, \tilde{C}\}$ may be defined for either the open-loop or the closed-loop case, this technique for analysis of non-normality extends directly to the characterization of controlled fluid systems.

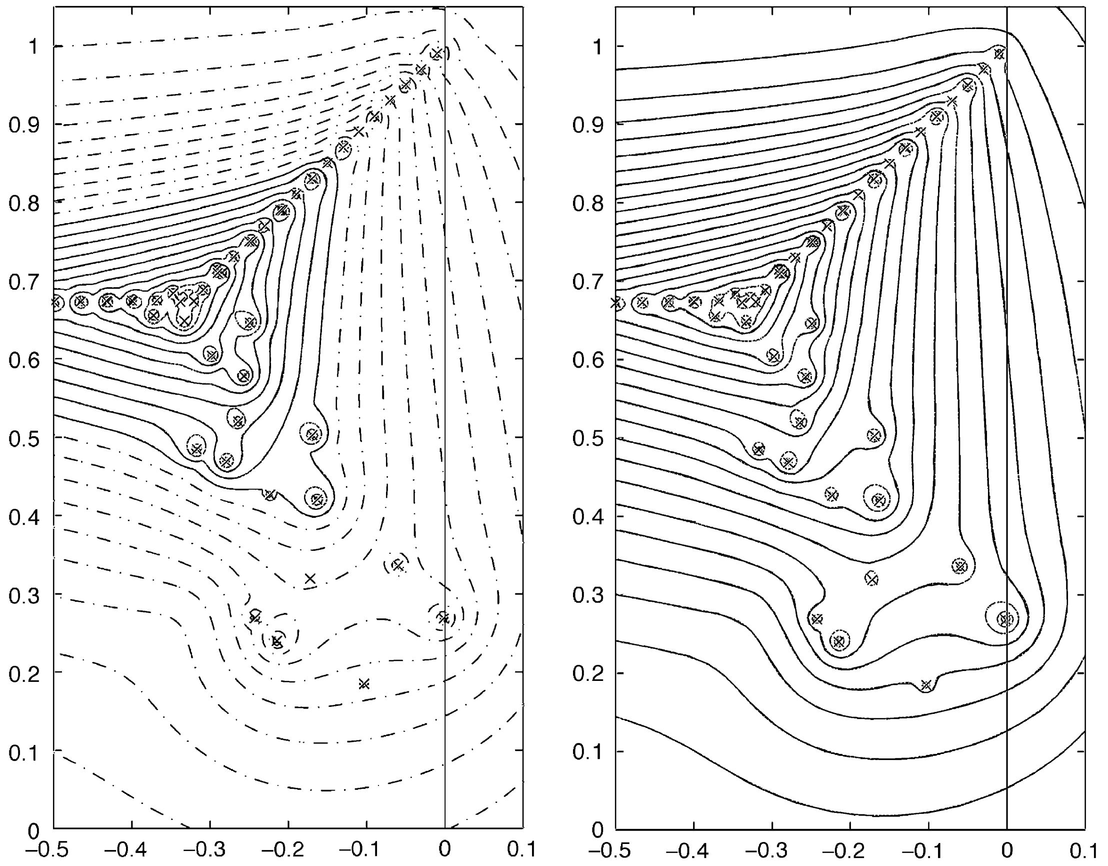


Isocontours of $\bar{\sigma}$ [$T_{\mathbf{zw}}(s)$] in the complex plane. The peak value of this matrix norm on the $j\omega$ axis is defined as the system norm$||T_{\mathbf{zw}}||_\infty$ and corresponds to the solid isoline with the smallest value.

Isocontours of (trace[T^*T])$^{1/2}$ in the complex plane s. The system norm$||T_{\mathbf{zw}}||_2$is related to the inetegral of the square of this matrix norm over the $j\omega$ axis.

FIGURE 15.3 Graphical interpretations (a.k.a. "pseudospectra") of the transfer function norms $\|T_{zw}\|_\infty$ (a) and $\|T_{zw}\|_2$ (b) for the present system in open loop, obtained at $k_x = 0$, $k_z = 2$, and $Re = 5000$. The eigenvalues of the system matrix A are marked with an $\times$. All isoline values are separated by a factor of 2, and the isolines with the largest value are those nearest to the eigenvalues. For this system, $\|T_{zw}\|_\infty = 2.6 \times 10^5$.

The $\mathcal{H}_\infty$ control technique is based on minimizing the 2–norm of the system transfer function while simultaneously bounding the ∞–norm of the system-transfer function. In the current transition problem, our control objective is to inhibit the (linear) formation of energetic flow perturbations that can lead to nonlinear instability and transition to turbulence. It is natural that control techniques such as $\mathcal{H}_\infty$, which are designed upon the very transfer function norms that quantify the excitation of such flow perturbations by external disturbances, will have a distinct advantage for achieving this objective over control techniques that account for the eigenvalues only, such as those based on the analysis of root-locus plots.



(a) Isocontours of $\bar{\sigma}\ [T_{\mathbf{zw}}(s)]$.

(b) Isocontours of (trace$[T^*T\ ]$)$^{1/2}$

FIGURE 15.4 Pseudospectra interpretations of $\|T_{zw}\|_\infty$(a) and $\|T_{zw}\|_2$ (b) for the open loop system at $k_x = -1$, $k_z = 0$, and $Re = 5000$. For plotting details, see Figure 15.3. For this system, $\|T_{zw}\|_\infty = 1.9 \times 10^4$.

15.3.3 Effectiveness of Control Feedback at Particular Wavenumber Pairs

The application of the modern control design approach described in Section 15.3.1 to the Orr-Sommerfeld/Squire problem laid out in Section 15.2 was explored extensively in Bewley and Liu (1998) for two particular wavenumber pairs and Reynolds numbers. The control effectiveness was quantified using several different techniques, including eigenmode analysis, transient energy growth, and transfer function norms. The control was remarkably effective and the trends with $\{\ell, \alpha, \gamma\}$ were all as expected. Refer to the journal article for complete tabulation of the results. One of the most notable features of this paper is that the application of the control resulted in the closed-loop eigenvectors becoming significantly closer to orthogonal, as illustrated in Figure 15.5. Note especially the high degree of correlation between the second and third eigenvectors of Figure 15.5a, and how this correlation is disrupted in Figure 15.5b. This was accompanied by concomitant reductions in both transient energy growth and the system transfer function norms in the controlled system. The nearly parallel nature of the pairs of eigenvectors $\{\xi_2, \xi_3\}$, $\{\xi_4, \xi_5\}$, $\{\xi_6, \xi_7\}$, and $\{\xi_8, \xi_9\}$ in the uncontrolled case (Figure 15.5a) is also reflected by the elliptical isolines surrounding the corresponding eigenvalues illustrated by the pseudospectra of Figure 15.3.

Note the nonzero value of $\hat{v}$ at the walls in Figure 15.5b; this reflects the wall blowing/suction applied as the control. Note also that half of the eigenvectors in Figure 15.5a have zero $\hat{v}$ components. These are commonly referred to as the Squire modes of the system and are decoupled from the perturbations in $\hat{v}$ because of the block of zeros in the upper-right corner of A. Such decoupling is not seen in Figure 15.5b because the closed-loop system matrix $A + B_2K$ is full.

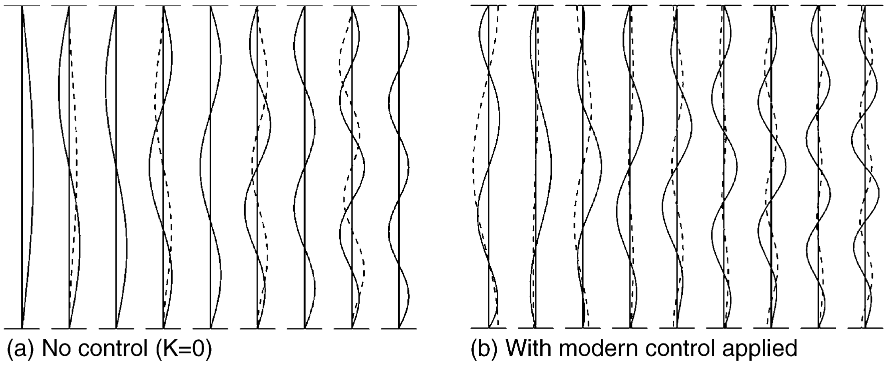


FIGURE 15.5 The nine least stable eigenmodes of the closed-loop system matrix $A + B_2K$ for $k_x = 0$, $k_z = 2$, and $Re = 5000$. Plotted are the nonzero part of the $\hat{\omega}$ component of the eigenvectors (solid) and the nonzero part of the $\hat{v}$ component of the eigenvectors (dashed) as a function of y from the lower wall (bottom) to the upper wall (top). In (a), the dashed line is magnified by a factor of 1000 with respect to the solid line; in (b), the dashed line is magnified by a factor of 300. The eigenvectors become significantly closer to orthogonal by the application of the control. (From Bewley, T.R., and Liu, S. (1998) *J. Fluid Mech.* **365**, 305–49. Reprinted with permission of Cambridge University Press.)

15.4 Decentralization: Designing for Massive Arrays

As illustrated in Figures 15.6 and 15.7, there are two possible approaches for experimental implementation of linear compensators for this problem:

1. a centralized approach, applied in Fourier space, or
2. a decentralized approach, applied in physical space.

Both of these approaches may be used to apply boundary control (such as distributions of blowing/suction) based on wall information (such as distributions of skin friction measurements). Both approaches may be used to implement the $\mathcal{H}_\infty$ compensators developed in Section 15.3, LQG/LTR compensators, PID feedback, or a host of other types of control designs. However, there are important differences in terms of the applicability of these two approaches to physical systems. The pros and cons of these approaches are now presented.

15.4.1 Centralized Approach

The centralized approach is simplest in terms of its derivation, as most linear compensators in this geometry are designed in Fourier space, leveraging the spatially invariant structure of this system mentioned previously and the complete decoupling into Fourier modes which this structure provides [Bewley and Agarwal, 1996]. As indicated in Figure 15.6, implementation of this approach is straightforward. This type of experimental realization was recommended by Cortelezzi and Speyer (1998) in related work. There are two major shortcomings of this approach:

1. The approach requires an online two-dimensional fast Fourier transform (FFT) of the entire measurement vector and an online two-dimensional inverse FFT (iFFT) of the entire control vector.
2. The approach assumes spatial periodicity of the flow perturbations.

With regard to point 1, it is important to note that the expense of centralized computations of two-dimensional FFTs and iFFTs will grow rapidly with the size of the array of sensors and actuators. Specificly, the computational expense is proportional to $N_x N_z \log(N_x N_z)$. This will rapidly decrease the bandwidth possible as the array size (and the number of Fourier modes) is increased for a fixed speed of the central processing unit (CPU). Communication of signals to and from the CPU is also an important limiting factor as the array size grows. Thus, this approach does not extend well to massive arrays of sensors and actuators.

With regard to point 2, it is important to note that transition phenomena in physical systems, such as boundary layers and plane channels, are not spatially periodic, though it is often useful to characterize the

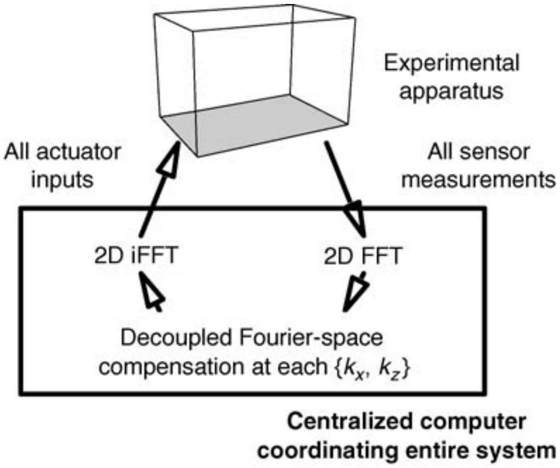

Experimental apparatus
All actuator inputs
All sensor measurements
2D iFFT
2D FFT
Decoupled Fourier-space compensation at each $\{k_x, k_z\}$
Centralized computer coordinating entire system

FIGURE 15.6 Centralized approach to the control of plane channel flow in Fourier space.

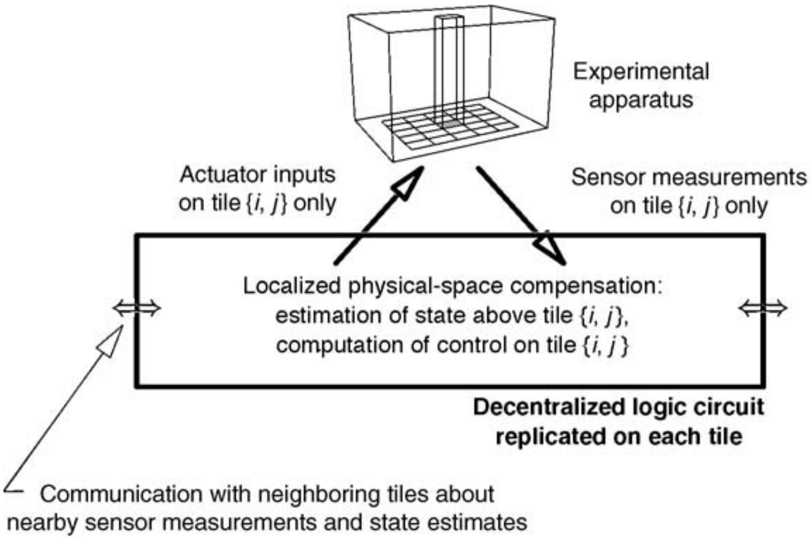

Experimental apparatus
Actuator inputs on tile $\{i, j\}$ only
Sensor measurements on tile $\{i, j\}$ only
Localized physical-space compensation: estimation of state above tile $\{i, j\}$, computation of control on tile $\{i, j\}$
Decentralized logic circuit replicated on each tile
Communication with neighboring tiles about nearby sensor measurements and state estimates

FIGURE 15.7 Decentralized approach to the control of plane channel flow in physical space.

solutions of such systems with Fourier modes. The application of Fourier-space controllers that assume spatial periodicity in their formulation to physical systems that are not spatially periodic will be corrupted by Gibb's phenomenon, the well-known effect in which a Fourier transform is spoiled across all frequencies when the data one is transforming are not themselves spatially periodic. To correct for this phenomenon in formulations based on Fourier-space computations of the control, windowing functions such as the Hanning window are appropriate. Windowing functions filter the signals coming into the compensator such that they are driven to zero near the edges of the physical domain under consideration, thus artificially imposing spatial periodicity on the non-spatially-periodic measurement vector.

15.4.2 Decentralized Approach

The decentralized approach, applied in physical space, is not as convenient to derive. Riccati equations of the size of the entire discretized three-dimensional system pictured in Figure 15.1 and governed by Equation (15.1), represented in physical space appear numerically intractable.

However, if such a problem could be solved, one would expect that the controller feedback kernels relating the state estimate $\hat{\mathbf{x}}$ inside the domain to the control forcing $\mathbf{u}$ at some point on the wall should decay quickly as a function of distance from the control point, as the control authority of any blowing/suction hole drilled into the wall on the surrounding flow decays rapidly with distance in a distributed viscous system.

Similarly, the estimator feedback kernels relating measurement errors ($\mathbf{y} - \hat{\mathbf{y}}$) at some point on the wall to the estimator forcing terms $\mathbf{v}$ on the system model inside the domain should decay quickly as a function of distance from the measurement point, as the correlation of any two flow-perturbation variables is known to decay rapidly with distance in a distributed viscous system.

Finally, due to the spatially invariant structure of the problem at hand, the control and estimation kernels for each sensor and actuator on the wall should be identical, though spatially shifted.

In other words, the physical-space kernels sought to determine the control and estimator feedback are spatially localized convolution kernels. If their spatial decay rate is rapid enough (e.g., exponential), then we will be able to truncate them at a finite distance from each actuator and sensor while maintaining a prescribed degree of accuracy in the feedback computation, resulting in spatially compact convolution kernels with finite support.

With such spatially compact convolution kernels, decentralized control of the present system becomes possible, as illustrated in Figure 15.7. In such an approach, several tiles are fabricated, each with sensors, actuators, and an identical logic circuit. The computations on each tile are limited in spatial extent, with the individual logic circuit on each tile responsible for the (physical-space) computation of the state estimate only in the volume immediately above that tile. Each tile communicates its local measurements and state estimates with its immediate neighbors, with the number of tiles over which such information propagates in each direction depending on the tile size and spatial extent of the truncated convolution kernels. By replication, we can extend such an approach to arbitrarily large arrays of sensors and actuators. Though additional truncation of the kernels will disrupt the effectiveness of this control strategy near the edges of the array, such edge effects are limited to the edges in this case (unlike Gibbs' phenomenon) and should become insignificant as the array size is increased.

15.5 Localization: Relaxing Nonphysical Assumptions

As discussed previously, though the physical-space representation of the three-dimensional linear system is intractable in the controls setting, the (completely decoupled) one-dimensional systems at each wavenumber pair $\{k_x, k_z\}$ in the Fourier-space representation of this problem are easily managed. Remarkably, these two representations are completely equivalent. Performing a Fourier transform (which is simply a linear change of variables) of the entire three-dimensional system (including the state, the controls, the measurements, and the disturbances) block diagonalizes all of the matrices involved in the three-dimensional physical-space control problem. With such block-diagonal structure, the constituent $\mathcal{H}_\infty$ control problems at each wavenumber pair $\{k_x, k_z\}$ may be solved independently and, once solved, reassembled in physical space with an inverse Fourier transform. If the numerics are handled properly, this approach is equivalent to solving the three-dimensional physical-space control problem directly.

Recent theoretical work on this problem by Bamieh et al. (2000), and related work by D'Andrea and Dullerud (2000), further support the notion that an array of $\mathcal{H}_\infty$ compensators developed at each wavenumber pair, when inverse-transformed back to the physical domain, should in fact result in spatially localized convolution kernels with exponential decay. This exponential decay, in turn, allows truncation of the kernels to any prescribed degree of accuracy. Thus, if the truncated kernels are allowed to be sufficiently large in streamwise and spanwise extent, favorable closed-loop system properties, such as robust stability and reduced system transfer function norms, may be retained. Until very recently, however, it has not been possible to obtain such kernels for Navier–Stokes systems, due to an assortment of numerical challenges.

In Högberg and Bewley (2000), spatially localized convolution kernels for both the control and estimation of plane channel flow have finally been obtained. The technique used was based on that described previously, deriving (in our initial efforts) $\mathcal{H}_2$ compensation at an array of wavenumber pairs $\{k_x, k_z\}$ and then inverse-transforming the lot, with special attention paid to the details of the control formulation and the numerical method. In particular, a numerical discretization technique not plagued by spurious eigenvalues was chosen, and the control formulation was slightly modified such that the time derivative of the

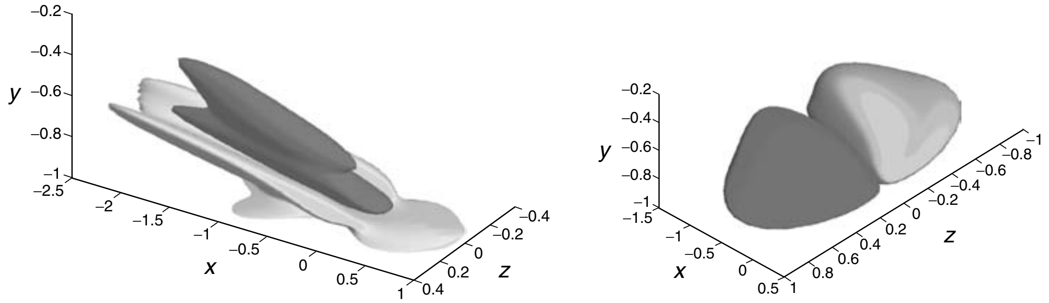


FIGURE 15.8 (**See color insert following** page 10-34.) Localized controller gains relating the state estimate $\hat{\mathbf{x}}$ inside the domain to the control forcing **u** at the point $\{x = 0, y = -1, z = 0\}$ on the wall. Visualized are a positive and negative isosurface of the convolution kernels for (left) the wall-normal component of velocity and (right) the wall-normal component of vorticity. (Högberg, M., Bewley, T.R., and Henningson, D.S. (2003) "Linear Feedback Control and Estimation of Transition in Plane Channel Flow," *J. Fluid Mech.* **481**, pp. 149–75. Reprinted with permission from Elsevier Science.)

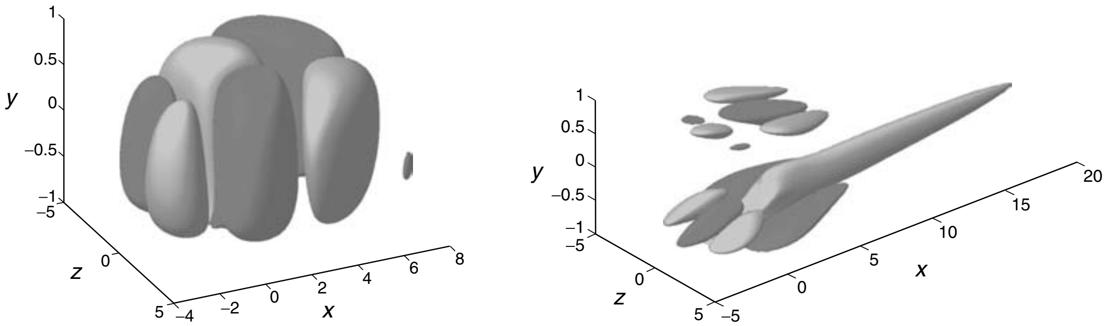


FIGURE 15.9 (**See color insert following** page 10-34.) Localized estimator gains relating the measurement error $(\mathbf{y} - \hat{\mathbf{y}})$ at the point $\{x = 0, y = -1, z = 0\}$ on the wall to the estimator forcing terms **v** inside the domain. Visualized are a positive and negative isosurface of the convolution kernels for (left) the wall-normal component of velocity and (right) the wall-normal component of vorticity. (Högberg, M., Bewley, T.R., and Henningson, D.S. (2003) "Linear Feedback Control and Estimation of Transition in Plane Channel Flow," *J. Fluid Mech.* **481**, pp. 149–75. Reprinted with permission from Elsevier Science.)

blowing/suction velocities is penalized in the cost function. The resulting localized kernels are illustrated in Figures 15.8 and 15.9. Such kernels facilitate the decentralized control implementation discussed in Section 15.4.2 and depicted in Figure 15.7, paving the way for experimental implementation with massive arrays of tiles integrating sensing, actuating, and the control logic.

The control convolution kernels shown in Figure 15.8 angle away from the wall in the upstream direction. Coupled with the mean flow profile indicated in Figure 15.1, this accounts for the convective delay which requires us to anticipate flow perturbations on the interior of the domain with actuation on the wall somewhere downstream. The estimation convolution kernels shown in Figure 15.9, on the other hand, extend well downstream of the measurement point. This accounts for the delay between the motions of the convecting flow structures on the interior of the domain and the eventual influence of these motions on the local drag profile on the wall; during this time delay, the flow structures responsible for these motions convect downstream. The upstream bias of the control kernels and the downstream bias of the estimation kernels, though physically tenable, were not prescribed in the problem formulation. A posteriori study of the streamwise, spanwise, and wall-normal extent, the symmetry, and the shape of such control and estimation kernels provides us with a powerful new tool with which the fundamental physics of this distributed fluid-mechanical system may be characterized.

The localized convolution kernels illustrated in Figures 15.8 and 15.9 are approximately independent of the size of the computational box in which they were computed, so long as this box is sufficiently large. Thus, when implementing these kernels, we may effectively assume that they were derived in an infinite-sized box, relaxing the nonphysical assumption of spatial periodicity used in the problem formulation and modeling the physical situation of spatially evolving flow perturbations in a spatially invariant geometry and mean flow.

The localized convolution kernels illustrated in Figures 15.8 and 15.9 are also approximately independent of the computational mesh resolution with which they were computed, when this computational mesh is sufficiently fine. A computational mesh sufficient to resolve the flow under consideration also adequately resolves these convolution kernels.

15.5.1 Open Questions

As we have shown, the framework for decentralized $\mathcal{H}_\infty$ control of the fully resolved transition problem in the geometry depicted in Figure 15.1 is now established. Obtaining spatial localization of the convolution kernels in physical space was the final remaining conceptual and numerical hurdle to be overcome. This work paves the way for decentralized application of such compensation with massive arrays of identical control tiles integrating sensing, actuation, and the control logic (Figure 15.7). Though in some sense "complete," this effort has also exposed several fundamental open questions, which are now briefly discussed.

For a given choice of the matrices $\{B_1, B_2, C_1, C_2\}$ and design parameters $\{\ell, \alpha, \gamma > \gamma_0\}$ selected, decentralized $\mathcal{H}_\infty$ compensators may be determined using the procedure previously described, and performance and robustness benchmarks may be obtained via simulation. As a final step in the control design process, explore how much the computational effort required by the logic on each tile may be reduced without significant degradation in the closed-loop system behavior. This can lead to a significant reduction in the number of floating point operations per second required by the logic circuit on each tile. However, as is discussed in Section 15.6, compensator reduction in the decentralized setting remains a significant unsolved problem; standard reduction strategies developed for finite, closed systems are not applicable and new research is motivated.

With the decentralized linear control framework established and prototypical numerical examples solved, we are now in a position to explore the effectiveness of compensators computed via this framework to the finite-amplitude perturbations that actually lead to transition and to the "large" amplitude perturbations of fully developed turbulence, in the nonlinear equations of fluid motion. An extensive analytical and numerical study within this framework is underway. Issues regarding our preliminary efforts in this direction are briefly reviewed in Section 15.7. As emphasized in the introduction, such a study should be guided by an interdisciplinary perspective to be maximally successful. Specifically, such a study should fully incorporate the known or postulated linear mechanisms leading to transition or, in the case of turbulence, the linear mechanisms thought to be at least partially responsible for sustaining the turbulent cascade of energy. In addition, this effort motivates the development of new analytical tools that might help clarify the types of state disturbances and flow perturbations that are particularly important in such phenomena. Armed with such an understanding, large benefits might be realized in the compensator design because the modeling of the structure of the state disturbances exciting the system G_1 and the weighting on the flow perturbations of interest in the cost function Q are important design criteria. In fact, we fully expect that the transfer of information between our physical understanding of fundamental flow phenomena and our knowledge of how to control such phenomena will be a two-way transfer. Such a strategy promises to provide powerful new tools for obtaining fundamental physical understanding of classical problems in fluid mechanics while we gain new insight in how to modify these phenomena by the action of control feedback.

A host of other canonical flow control problems, including the control of spatially developing boundary layers, bluff-body flows, and free shear layers, should also be amenable to linear control application using the framework outlined here. A few such extensions are discussed briefly in Section 15.8.

15.6 Compensator Reduction: Eliminating Unnecessary Complexity

Strategies for the development of reduced-order decentralized compensators of the present form remain a key unsolved issue. With the $\mathcal{H}_2/\mathcal{H}_\infty$ approach, as described previously, a physical-space state estimate in the volume immediately above each tile must be updated online by the logic circuit on each tile as the flow evolves. However, it is not necessary for the compensator to compute an accurate state estimate as an intermediate variable; indeed, our only requirement is that, based on whatever filtered information the dynamic compensator does extract from the noisy system measurements, suitable controls may be determined to achieve the desired closed-loop system behavior. It should be possible to reduce substantially the complexity of the dynamic compensator and still achieve this more modest objective.

There are two possible representations in which the complexity of the compensator can be reduced: in Fourier space (where the compensator is designed) or in physical space (where the decentralized compensation is applied).

15.6.1 Fourier-Space Compensator Reduction

At any particular wavenumber pair $\{k_x, k_z\}$, there is one actuator variable at each wall, one sensor variable at each wall, and a spatial discretization in y of the state variables across the domain stretching between these walls. Because of the complete decoupling of the control problem into separate Fourier modes, the system model used in the estimator at each particular wavenumber pair is not referenced by the compensator at any other wavenumber pair. Thus, the compensators at each wavenumber pair are completely decoupled and may be reduced independently. At certain wavenumber pairs, it might be important to retain several degrees of freedom in the dynamic compensator, while at other wavenumber pairs, it might be possible to retain significantly fewer degrees of freedom without significant degradation in the closed-loop system behavior. Several existing compensator reduction strategies are well suited to this problem, and their application in this setting is straightforward. Cortelezzi and Speyer (1998) successfully applied the balanced truncation technique of open-loop model reduction in this Fourier-space framework to facilitate the design of a reduced-complexity dynamic compensator.

As mentioned earlier, it is the nonorthogonality of the entire set of system eigenvectors that leads to the peculiar (and important) possibilities for energy amplification in these systems, so compensator reduction techniques mindful of the relevant transfer function norms are necessary. In addition, as eloquently described by Obinata and Anderson (2000), it is most appropriate when designing low-order compensators for high-order plants to reduce the compensator while accounting for how it performs in the closed loop. An assortment of closed-loop compensator reduction techniques are now available and should be tested in future work.

In the setting of designing a decentralized compensator, there is an important shortcoming to performing standard compensator reductions in Fourier space. As the compensator reduction problem is independent at each wavenumber pair, we might be left with a different number of degrees of freedom in the reduced-order compensator at each wavenumber pair, leaving us with a dynamical system model that is impossible to inverse transform back into the physical domain. Even if we restrict the compensator reduction algorithm to reduce to the same number of degrees of freedom at each wavenumber pair (a restrictive assumption that should be unnecessary), there appears to be no appropriate strategy currently available to coordinate this reduction process across all wavenumbers in a consistent manner such that the inverse transform of the reduced dynamic model is spatially localized. Without such coordination, it seems inevitable that the ordering and representation of the various modes of this dynamic model will be scrambled during the process of compensator reduction at each wavenumber pair, resulting in an inverse-transform back in physical space that does not exhibit the spatial localization which is essential to facilitate decentralized control.

15.6.2 Physical-Space Compensator Reduction

As an alternative to Fourier-space compensator reduction, one might consider instead the reduction of the physical-space model and its associated localized convolution kernels. This has several advantages linked to the fact that this is the actual compensation to be computed on each tile. The first advantage is that spatial localization will be retained, as compensator reduction is applied after the localized kernels are obtained. Another important advantage is that this setting allows us to keep more degrees of freedom in the dynamical system model to represent streamwise and spanwise fluctuations of the state near the wall than we retain to represent the behavior of the state on the interior of the domain. This effectively relaxes the restrictive assumption referred to in the previous paragraph. Such an emphasis on resolving the state near the wall is motivated by inspection of the convolution kernels plotted in Figures 15.8 and 15.9, in which it is clear that the details of the flow near the wall are of increased importance when computing the feedback.

However, the system model simulated on each individual tile is not self-contained, due to the interconnections with neighboring tiles indicated in Figure 15.7. Thus, if one reduces the system model above a single tile, all neighboring tiles that reference this state estimate will be affected. As the system model is not self-contained, as it was in the Fourier-space case, existing compensator reduction approaches are not applicable.

An important observation, however, is that the structure of the system model carried by each tile is identical. Due to the repeated structure of the model represented on the array, it is sufficient to optimize the system model carried by a single tile. The repeated structure of the distributed physical-space model should make the compensator reduction problem tractable. This fundamental problem of reducing distributed, interconnected dynamic compensators in the decentralized closed-loop setting remains, as yet, unsolved.

15.6.3 Nonspatially Invariant Systems

Finally, it should be stated that the Fourier-space decoupling leveraged at the outset of this problem formulation has been one of the key ingredients that have permitted accurate solution of well-resolved canonical flow control problems to date. The linear control technique we have used to solve these control problems involves the solution of matrix Riccati equations, which are accurately soluble for state dimensions only up to $O(10^3)$. As we move to more applied flow control problems in which such Fourier-space decoupling is either more restrictive or not available, if we continue to use Riccati-based control approaches, creative new compensator reduction strategies will be required. We might need to apply "open-loop" model reduction strategies (in advance of computing the control feedback and closing the loop) to make manageable the dimension of the Riccati equations to be solved in the compensator formulation. As mentioned earlier, it is most appropriate when designing low-order compensators for high-order plants to reduce the compensator while accounting for how it performs in the closed loop. Unfortunately, extremely high-order discretizations of nonspatially invariant PDE systems will not likely afford us this luxury, as such systems do not decouple (via Fourier transforms) into constituent lower-order control problems amenable to matrix-based compensator design strategies.

15.7 Extrapolation: Linear Control of Nonlinear Systems

Once a decentralized linear compensator of the present form is developed, a verification of its utility for the transition problem may be obtained by applying it to the laminar flow depicted in Figure 15.1 with either finite-amplitude (but sufficiently small) initial flow perturbations or finite-amplitude (but sufficiently small) applied external disturbances. The resulting finite-amplitude flow perturbations are governed by the fully nonlinear Navier–Stokes equation and have been simulated in well resolved direct numerical simulations (DNS) with the code benchmarked in Bewley et al. (2001). Representative simulations are shown in Figure 15.10, indicating that linear compensators can indeed relaminarize perturbed flows that would otherwise proceed rapidly towards transition to turbulence. With the framework presented here, extensive numerical studies promise to significantly extend our fundamental understanding of the process of transition and how this process may be inhibited by control feedback.

It is also of interest to consider the application of decentralized linear compensation to the fully nonlinear problem of a turbulent flow, such as that shown in Figure 15.11. The first reason to try such an

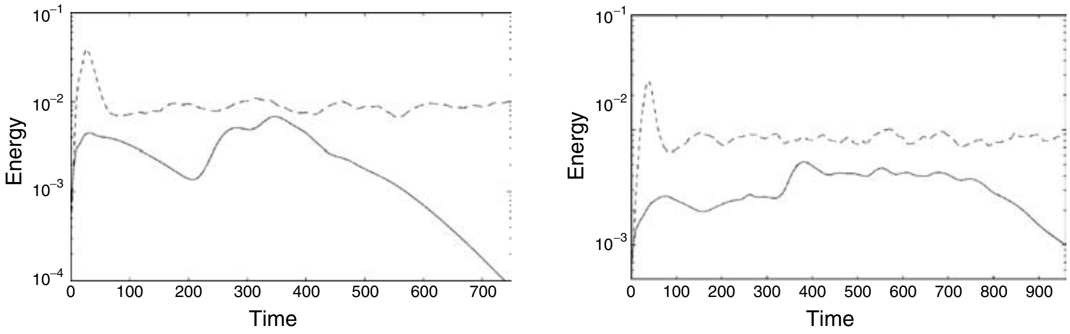


FIGURE 15.10 Evolution of oblique waves (left) and an initially random flow perturbation (right) added to a laminar flow at $Re = 2000$, with and without decentralized linear control feedback. The magnitude of the initial flow perturbations in these simulations greatly exceed the thresholds reported by Reddy et al. (1998) that lead to transition to turbulence in an uncontrolled flow (by a factor of 225 for the oblique waves and by a factor of 15 for the random initial perturbation). Solid lines indicate the energy evolution in the controlled case, dashed lines indicate the energy evolution in the uncontrolled case. Both of the uncontrolled systems lead quickly to transition to turbulence, whereas both of the controlled systems relaminarize. For the controlled cases, initial perturbations with greater energy fail to relaminarize, whereas initial perturbations with less energy relaminarize earlier. (Högberg, M., Bewley, T.R., and Henningson, D.S. (2003) "Linear Feedback Control and Estimation of Transition in Plane Channel Flow," *J. Fluid Mech.* **481**, pp. 149–75. Reprinted with permission from Elsevier Science.)

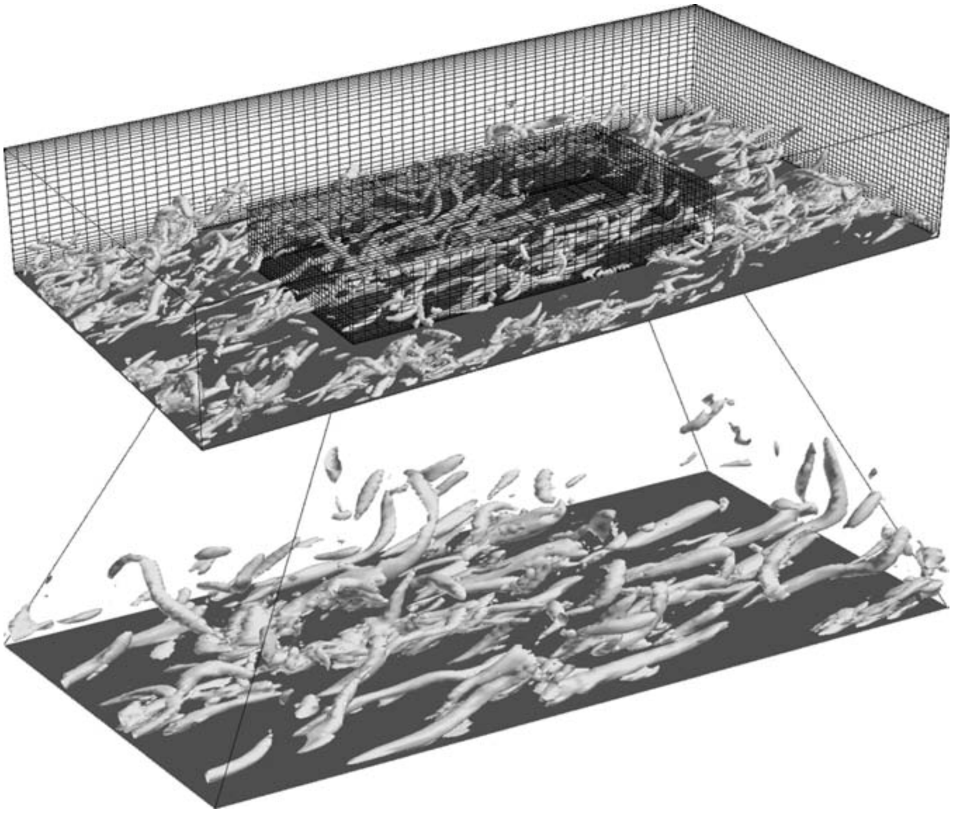


FIGURE 15.11 (**See color insert following page 10-34.**) Visualization of the coherent structures of uncontrolled near-wall turbulence at $Re_\tau = 180$. Despite the geometric simplicity of this flow (see Figure 15.1), it is phenomenologically rich and is characterized by a large range of length scales and time scales over which energy transport and scalar mixing occur. The relevant spectra characterizing these complex nonlinear phenomena are continuous over this large range of scales, thus such flows have largely eluded accurate description via dynamic models of low state dimension. The nonlinearity, the distributed nature, and the inherent complexity of its dynamics make turbulent flow systems particularly challenging for successful application of control theory. (Simulation by Bewley, T.R., Moin, P., and Temam, R. (2001) *J. Fluid Mech.* Reprinted with permission of Cambridge University Press.)

approach is simply because we can: linear control theory leads to implementable control algorithms and grants a lot of flexibility in the compensator design. Nonlinear turbulence control strategies, though currently under active development (see Sections 15.9 to 15.13), are much more difficult to design and implement and require substantial further research before they will provide implementable control strategies as flexible and powerful as those which we currently have at our disposal in the linear setting.

There is some evidence in the fluids literature that applying linear control feedback to turbulence might be at least partially effective. Though the significance of this result has been debated in the fluid mechanics community, Farrell and Ioannou (1993) have clearly shown that linearized Navier–Stokes systems in plane channel flows, when excited with the appropriate stochastic forcing, exhibit behavior reminiscent of the streamwise vortices and streamwise streaks that characterize actual near-wall turbulence. The present linear control framework (perhaps restricted to a finite horizon) should be able to exploit whatever information the linearized Navier–Stokes equation actually contains about the mechanisms sustaining these turbulence structures. Though the life cycle of the near-wall coherent structures of turbulence appears to involve important nonlinear phenomena [see, e.g., Hamilton et al., 1995], that in itself does not disqualify the utility of linear control strategies to effectively disrupt critical linear terms of this nonlinear process. Recent numerical experiments by Kim and Lim (2000) support this idea by conclusively demonstrating the importance of the coupling term C in the linearized system matrix A (see Equation (15.10)) for maintaining near-wall turbulence in nonlinear simulations.

To understand the possible pitfalls of applying linear feedback to nonlinear systems, a low-order nonlinear convection problem governed by the Lorenz equation was studied by Bewley (1999). As with the problem of turbulent channel flow, but in a low-order system easily amenable to analysis, control feedback was determined with linear control theory by linearizing the governing equation about a desired fixed point. Once a linear controller was determined by such an approach, it was then applied directly to the fully nonlinear system. The result is depicted in Figure 15.12.

For control feedback determined by linear control theory with a large weighting ℓ, on the control effort, direct application of linear feedback to the full nonlinear system stabilizes both the desired state and an undesired state, indicated by the two trajectories marked in Figure 15.12a. An unstable manifold exists between these two states, indicated by the contorted surface shown. Any initial state on one side of this manifold will converge to the desired state, and any initial state on the other side of this manifold will converge to the undesired state.

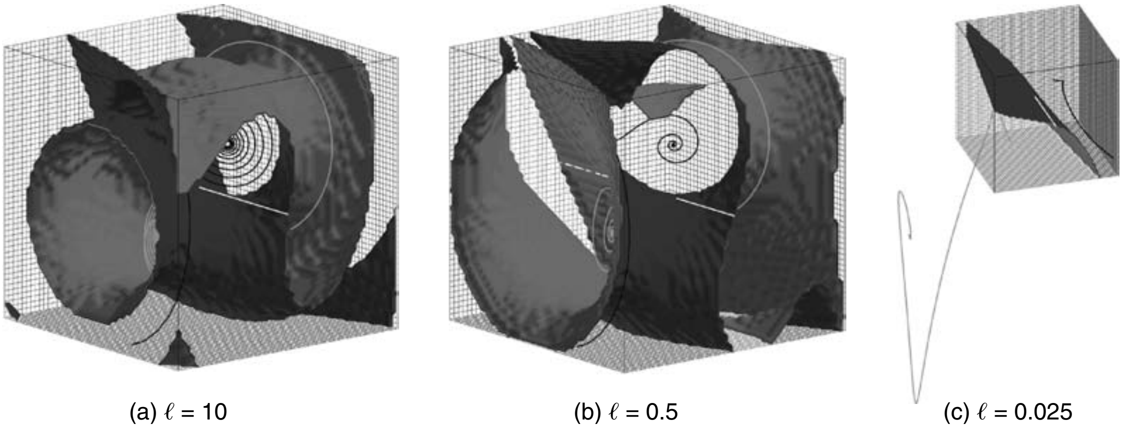

(a) $\ell = 10$ (b) $\ell = 0.5$ (c) $\ell = 0.025$

FIGURE 15.12 (**See color insert following page 10-34.**) Example of the spectacular failure of linear control theory to stabilize a simple nonlinear chaotic convection system governed by the Lorenz equation. Plotted are the regions of attraction to the desired stationary point (blue) and to an undesired stationary point (red) in the linearly controlled nonlinear system, and typical trajectories in each region (black and green, respectively). The cubical domain illustrated is $\Omega = (-25, 25)^3$ in all subfigures. For clarity, different viewpoints are used in each subfigure. (Reprinted with permission from Bewley, T.R. (1999) *Phys. Fluids* **11**, 1169–86. Copyright 1999, American Institute of Physics.)

As seen in Figures 15.12b and c, as the weighting on the control effort ℓ, is turned down and the desired stationary state is stabilized more aggressively, the domain of convergence to the undesired stabilized state remains large. This undesired state is "aggravated" by the enhanced control feedback, moving farther from the origin. The undesired state eventually escapes to infinity for sufficiently small ℓ, indicating instability of the nonlinear system from a wide range of initial conditions even though the desired stationary point is endowed with a high degree of linear stability. Implication: strong linear stabilization of a desired system state (such as laminar flow) will not necessarily eliminate undesired nonlinear system behavior (such as turbulence) in a chaotic system.

Some form of nonlinearity in the feedback rule was required to eliminate this undesired behavior. One effective technique is to apply a switch such that the linear control feedback is turned on only when the state $\mathbf{x}(t)$ is within some sufficiently small neighborhood of the desired stabilized state $\overline{\mathbf{x}}$ in the linearly controlled system. The chaotic dynamics of the uncontrolled Lorenz system will bring the system into this neighborhood in finite time, after which control may be applied to "catch" the system at the desired equilibrium state.

Thus, even in this simple model problem, linear feedback can have a destabilizing influence if applied outside the neighborhood for which it was designed. For the full Navier–Stokes problem, though a certain set of linear feedback gains might stabilize the laminar state, on the "other side of the manifold" might lie a turbulent state aggravated by the same linear controls. Application of linear control to nonlinear chaotic systems must therefore be done with vigilance, lest nonlinearities destabilize the closed-loop system, as shown here. The easy fix for this low-order model problem (that is, simply turn off the control until the chaotic dynamics bring the state into a neighborhood of the desired state) might not be available for the (high-dimensional) problem of turbulence because fully turbulent flows appear to remain at all times far from the laminar state.

In our preliminary attempts at applying the decentralized compensators previously developed to turbulence, we have succeeded in reducing the drag of a fully developed turbulent flow by 25% with state-feedback controllers, as shown in Figure 15.13. Interestingly, for the choice of control parameters selected here, there is no evidence of an aggravated turbulent state. A 25% drag reduction, though significant, is comparable to the drag reductions obtained with a variety of other ad hoc control approaches in this flow. We are actively pursuing modification of this linear control feedback to improve upon this result. Interdisciplinary considerations, such as those involved in the design of linear compensation for the problem of transition, are essential in this effort. Specifically, the (unmodeled) nonlinear terms in the Navier–Stokes equation provide insight as to the structure of the disturbances, G_1, to be accounted for in the linear control formulation to best compensate for their unmodeled effects. Additionally, the coherent structures of fully developed near-wall turbulence, believed to be a major player in the self-sustaining nonlinear process of turbulence generation near the wall, provide a phenomenological target that may be exploited in the selection of the weighting on the flow perturbations Q in the cost function.

15.8 Generalization: Extending to Spatially Developing Flows

Extension of the decentralized linear control framework developed here to a large class of slightly nonparallel flows is heuristic but straightforward. To accomplish this, the parabolic mean flow profile $U(y)$ indicated in Figure 15.1 is replaced with an appropriate "quasi-one-dimensional" profile, such as the Blasius boundary layer profile. As long as the mean flow profile evolves slowly enough in space (as compared to the wavelengths of the significant instabilities in the problem), it may be assumed to be constant in space for the purpose of developing the linear control feedback. Such an assumption of slow spatial divergence forms the foundation of the study of local and global modes used in the characterization of absolute and convective instabilities [Huerre and Monkewitz, 1990] and has proven to be a powerful concept. For the appropriate flows, we believe this concept is also appropriate in the context of the development of control feedback.

Implementation of the decentralized control concept in this setting is a heuristic extension of the approach presented in Figure 15.7. Gradual variations in the mean flow are accounted for by local extension of the

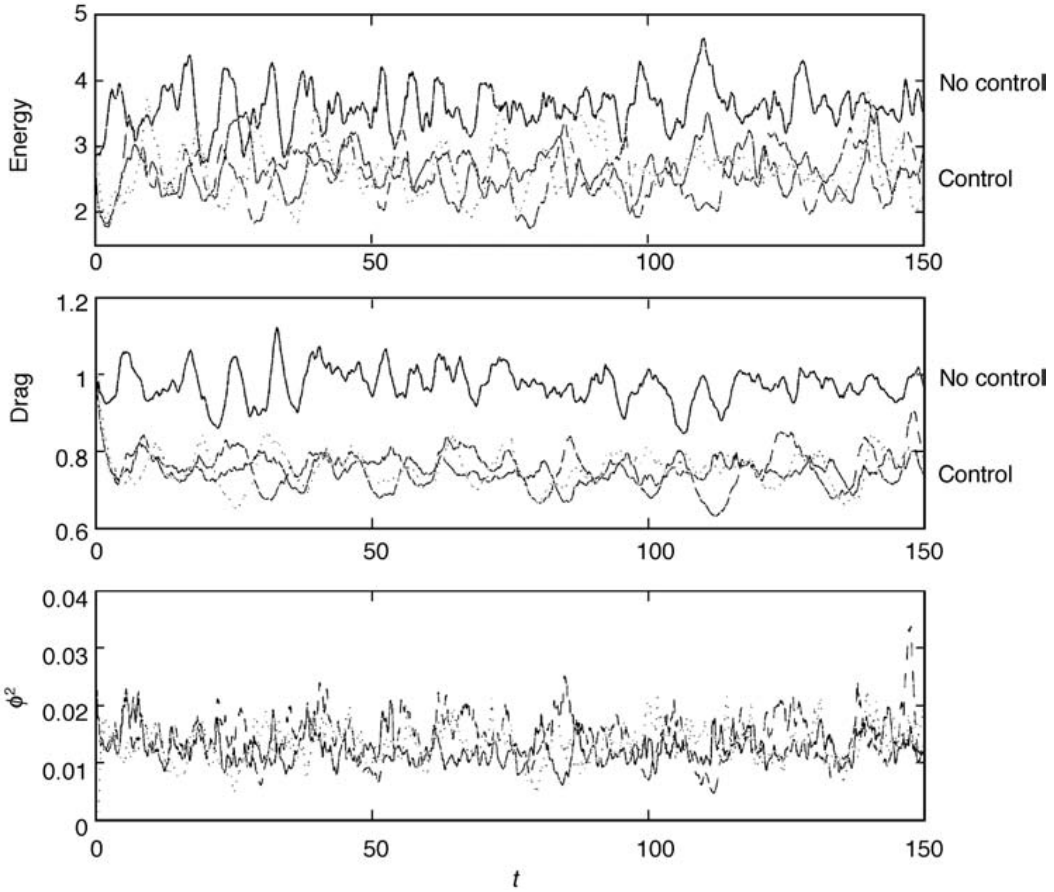


FIGURE 15.13 Evolution of fully developed turbulence at $RE_\tau = 100$ with and without decentralized linear control feedback. This flow has approximately the same mass flux as the laminar flow at $Re = 2000$. (Top) Energy of flow perturbation. (Middle) Drag (note approximately 25% reduction in the controlled cases). (Bottom) Control effort used. The uncontrolled energy and drag are the (upper) solid lines in the top and middle figures. A gain scheduling approach is used to tune the control feedback gains to the instantaneous mean flow profile. (Högberg, M., Bewley, T.R., and Henningson, D.S. (2003) "Linear Feedback Control and Estimation of Transition in Plane Channel Flow," *J. Fluid Mech.* **481**, pp. 149–75. Reprinted with permission from Elsevier Science.)

mean flow profile in the compensator derivation for each tile, gradually scaling the compensation rules from one tile to the next as the flow develops downstream. For example, we may consider developing this strategy for the laminar boundary layer (LBL) solutions of the Falkner–Skan–Cooke family, found by solving the ordinary differential equation (ODE)

$$f''' + ff'' + \beta(1 - f'^{\,2}) = 0$$

with $f(0) = f'(0) = 0$ and $f'(\infty) \to 1$ and defining

$$U = U_0 f'(\eta) \quad \text{and} \quad V = \sqrt{\frac{\nu U_0}{2x}}\,[\eta f'(\eta) - f(\eta)].$$

Cases of interest include the Blasius profile, modeling a zero-pressure-gradient, flat-plate LBL with

$$U_0 = U_\infty, \quad \beta = 0, \quad \eta = y\sqrt{\frac{U_0}{2\nu x}},$$

the Falkner–Skan profile, modeling a nonzero-pressure-gradient LBL or wedge flow by taking

$$U_0 = Kx^m, \quad \beta = \frac{2m}{1+m}, \quad \eta = y\sqrt{\frac{(m+1)U_0}{2\nu x}},$$

and the Falkner–Skan–Cooke profile, which models the addition of sweep to the leading edge by solving the supplemental ODE

$$g'' + fg' = 0$$

with $g(0) = 0$ and $g(\infty) \to 1$ and defining $W = W_\infty g(\eta)$. The self-similarity of the LBL profiles might lead to simplified parameterizations of the convolution kernels for the control and estimation problems. Extension of this approach to a variety of other spatially developing flows (self-similar or otherwise) should also be straightforward.

15.9 Nonlinear Optimization: Local Solutions for Full Navier–Stokes

Given an idealized setting of full state information, no disturbances, and extensive computational resources, significant finite-horizon optimization problems may be formulated and (locally) solved for complex nonlinear systems using iterative, adjoint-based, gradient optimization strategies. Such optimization problems can now be solved for high-dimensional discretizations of turbulent flow systems, incorporating the full nonlinear Navier–Stokes equation, locally minimizing cost functionals representing a variety of control problems of physical interest within a given space of feasible control variables. The mathematical framework for such optimizations will be reviewed briefly in Section 15.9.1 and is described in greater detail by Bewley et al. (2001).

The optimizations obtained via this approach are only "local" over the domain of feasible controls (that is, unless restrictive assumptions are made in the formulation of the control problem). Thus, the performance obtained via this approach usually cannot be guaranteed to be "globally optimal." However, the performance obtained with such nonlinear optimizations often far exceeds that possible with other control design approaches (see, e.g., Figure 15.14). In addition, this approach is quite flexible because it can iteratively improve high-dimensional control distributions directly, as is illustrated below. Alternatively, this approach can optimize open-loop forcing schedules, shape functions, or the coefficients of practical, implementable, and possibly nonlinear feedback control rules. Thus, interest in adjoint-based optimization strategies for turbulent flow systems goes far beyond that of establishing performance benchmarks via predictive optimizations of the control distribution itself. Establishing such benchmarks is only a first step toward a much wider range of applications for adjoint-based tools in turbulent flow systems.

The general idea of this approach, often referred to as model predictive control, is well motivated by comparing and contrasting it to massively parallel brute-force algorithms recently developed to play the game of chess. The goal when playing chess is to capture the other player's king through an alternating series of discrete moves with the opponent; at any particular turn, a player has to select one move out of at most 20 or 30 legal alternatives.

To accomplish its optimization, a computer program designed to play the comparatively "simple" game of chess, such as Deep Blue [Newborn, 1997], must, in the worst case, plan ahead by iteratively examining a tree of possible evolutions of the game several moves into the future [Atkinson, 1993], a strategy based on "function evaluations" alone. At each step, the program selects the move that leads to its best expected outcome, given that the opponent is doing the same in a truly noncooperative competition. The version of Deep Blue that defeated Garry Kasparov in 1997 was able to calculate up to 200 billion moves in the three minutes it was allowed to conduct each turn. Even with this extreme number of function evaluations at its disposal on this relatively simple problem, the algorithm was only about an even match with Kasparov's human intuition.

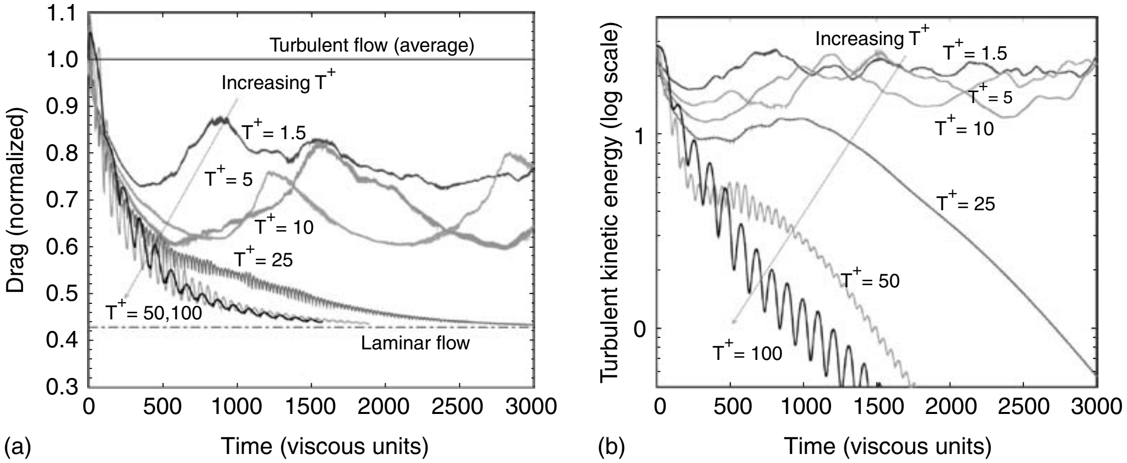


FIGURE 15.14 (**See color insert following page 10-34.**) Performance of optimized blowing/suction controls for formulations based on minimizing $J_o(\phi)$, case c (see Section 15.9.1.2), as a function of the optimization horizon T^+. The direct numerical simulations of turbulent channel flow reported here were conducted at $Re_\tau = 100$. For small optimization horizons ($T^+ = O(1)$, sometimes called the "suboptimal approximation"), approximately 20% drag reduction is obtained, a result that can be obtained with a variety of other approaches. For sufficiently large optimization horizons ($T^+ \geq 25$), the flow is returned to the region of stability of the laminar flow, and the flow relaminarizes with no further control effort required. No other control algorithm tested in this flow to date has achieved this result with this type of flow actuation. (From Bewley, T.R., Moin, P., and Temam, R. (2001) *J. Fluid Mech.*, to appear. Reprinted with permission of Cambridge University Press.)

An improved algorithm compared to those based on function evaluations alone, suitable for optimizing the present problem in a reasonable amount of time, is available because we know the equation governing the evolution of the present system, and we can state the problem of interest as a functional to be minimized. Taking these two facts together, we may devise an iterative procedure based on gradient information, derived from an adjoint field, to optimize the controls for the desired purpose on the prediction horizon of interest in an efficient manner. Only by exploiting such gradient information can the high-dimensional optimization problem at hand (up to $O(10^7)$ control variables per optimization horizon in some of our simulations) be made tractable.

15.9.1 Adjoint-Based Optimization Approach

15.9.1.1 Governing Equation

The problem we consider here is the control of a fully developed turbulent channel flow with full flowfield information and copious computational resources available to the control algorithm. The flow is governed by the incompressible Navier–Stokes equation inside a three-dimensional rectangular domain (Figure 15.15) with unsteady wall-normal velocity boundary conditions ϕ applied on the walls as the control. Three vector fields are first defined: the flow state $\mathbf{q}$, the flow perturbation state $\mathbf{q}'$, and the adjoint state $\mathbf{q}^\star$:

$$\mathbf{q}(\mathbf{x},t) = \begin{pmatrix} p(\mathbf{x},t) \\ \mathbf{u}(\mathbf{x},t) \end{pmatrix}, \quad \mathbf{q}'(\mathbf{x},t) = \begin{pmatrix} p'(\mathbf{x},t) \\ \mathbf{u}'(\mathbf{x},t) \end{pmatrix}, \quad \mathbf{q}^\star(\mathbf{x},t) = \begin{pmatrix} p^\star(\mathbf{x},t) \\ \mathbf{u}^\star(\mathbf{x},t) \end{pmatrix}.$$

Each of these vector fields is composed of a pressure component and a velocity component, all of which are continuous functions of space $\mathbf{x}$ and time t. The velocity components themselves are also vectors, with components in the streamwise direction x_1, the wall-normal direction x_2, and the spanwise direction x_3. Partial differential equations governing all three of these fields will be derived in due course, and the motivation for introducing $\mathbf{q}'$ and $\mathbf{q}^\star$ will be given as the need for these fields arises in the control derivation. Only after the optimization approach has been derived completely in differential form is it discretized in space

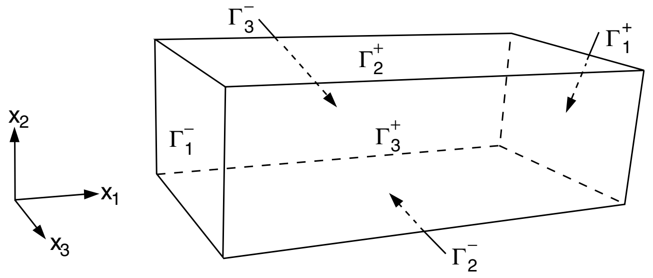


FIGURE 15.15 Channel flow geometry. The interior of the domain is denoted Ω and the boundaries of the domain in the x_i direction are denoted $\Gamma_i^{\pm}$. Unsteady wall-normal velocity boundary conditions are applied on the walls $\Gamma_2^{\pm}$ as the control, with periodic boundary conditions applied in the streamwise direction x_1 and spanwise direction x_3. An external pressure gradient is applied to induce a mean flow in the x_1 direction.

and time. An alternative strategy, discretizing the state equation in space before determining the adjoint operator, is discussed in Section 15.9.2.

The governing equation is written as

$$\begin{aligned} \mathcal{N}(\mathbf{q}) &= \mathbf{F} && \text{in } \Omega, \\ \mathbf{u} &= -\phi\mathbf{n} && \text{on } \Gamma_2^{\pm}, \\ \mathbf{u} &= \mathbf{u}_0 && \text{at } t = 0, \end{aligned} \tag{15.11}$$

where $\mathcal{N}(\mathbf{q})$ is the (nonlinear) Navier–Stokes operator

$$\mathcal{N}(\mathbf{q}) = \begin{pmatrix} \dfrac{\partial u_j}{\partial x_j} \\ \dfrac{\partial u_i}{\partial t} + \dfrac{\partial u_j u_i}{\partial x_j} - \nu \dfrac{\partial^2 u_i}{\partial x_j^2} + \dfrac{\partial p}{\partial x_i} \end{pmatrix},$$

$\mathbf{F}$ is a forcing vector accounting for an externally applied mean pressure gradient driving the flow in the streamwise direction, and $\mathbf{n}$ is the unit outward normal to the boundary $\partial\Omega$. The boundary conditions on the state $\mathbf{q}$ are periodic in the streamwise and spanwise directions. A wall-normal control velocity ϕ is distributed over the walls as indicated, and is constrained to inject zero net mass such that $\forall t$, $\int_{\Gamma_2^+} \phi d\mathbf{x} = \int_{\Gamma_2^-} \phi d\mathbf{x} = 0$. Initial conditions on the velocity $\mathbf{u}_0$ of fully developed turbulent channel flow are prescribed.

15.9.1.2 Cost Functional

As in the linear setting, an essential step in the framing of the nonlinear optimization problem is the representation of the control objective as a cost functional to be minimized. Several cases of physical interest may be represented by a cost functional of the generic form

$$\mathcal{J}_0(\phi) = \frac{1}{2}\int_0^T\int_\Omega |C_1\mathbf{u}|^2 d\mathbf{x}\, dt + \frac{1}{2}\int_\Omega |C_2\mathbf{u}(\mathbf{x}, T)|^2 d\mathbf{x} - \int_0^T\int_{\Gamma_2^{\pm}} C_3 \nu \frac{\partial \mathbf{u}}{\partial n} \cdot \mathbf{r}\, d\mathbf{x}\, dt + \frac{\ell^2}{2}\int_0^{\mathrm{T}}\int_{\Gamma_2^{\pm}} |\phi|^2\, d\mathbf{x}\, dt.$$

Four cases of particular interest are:

a. $C_1 = d_1 I$ and $C_2 = C_3 = 0 \Rightarrow$ regulation of turbulent kinetic energy;
b. $C_1 = d_2 \nabla \times$ and $C_2 = C_3 = 0 \Rightarrow$ regulation of the square of the vorticity;
c. $C_2 = d_3 I$ and $C_1 = C_3 = 0 \Rightarrow$ terminal control of turbulent kinetic energy;
d. $C_3 = d_4 I$ and $C_1 = C_2 = 0 \Rightarrow$ minimization of the time-averaged skin friction in the direction $\mathbf{r}$ integrated over the boundary of the domain, where $\mathbf{r}$ is a unit vector in the streamwise direction.

All four of these cases, and many others, may be considered in the current framework, and the extension to other cost functionals is straightforward. The dimensional constants d_i (which are the appropriate functions of the kinematic viscosity, the channel width and the bulk velocity), as well as ℓ, are included to make the cost functional dimensionally consistent and to account for the relative weight of each individual term.

In both the chess problem and the turbulence problem, the further into the future one can optimize the problem the better (Figure 15.14). However, both problems get exponentially more difficult to optimize as the prediction horizon is increased. Because only intermediate-term optimization is tractable, representing the final objective in the cost functional is not always the best approach. In the chess problem, though the final aim is to capture the other player's king, it is most effective to adopt a mid-game strategy of establishing good board position and achieving material advantage. Similarly, if the turbulence control objective is reducing drag, Bewley et al. (2001) found that it is most effective along the way to minimize a finite-horizon cost functional related to the turbulent kinetic energy of the flow because the turbulent transport of momentum is responsible for inducing a substantial portion of the drag in a turbulent flow. In a sense, turbulence is the "cause" and high drag is the "effect," and it is most effective to target the "cause" in the cost functional when optimizations on only intermediate prediction horizons are possible.

In addition, a smart optimization algorithm allows for excursions in the short term if it leads to a long-term advantage. For example, in chess, a good player is willing to sacrifice a lesser piece if, by so doing, a commanding board position is attained or a restoring exchange is forced a few moves later. Similarly, by allowing a turbulence control scheme to increase (temporarily) the turbulent kinetic energy of a flow, a transient may ensue which, eventually, effectively diminishes the strength of the near-wall coherent structures. Bewley et al. (2001) found that terminal control strategies, aimed at minimizing the turbulence only at the end of each optimization period, have a decided advantage over regulation strategies, which penalize excursions of, for example, the turbulent kinetic energy over the entire prediction horizon.

15.9.1.3 Gradient of Cost Functional

As suggested by Abergel and Temam (1990), a rigorous procedure may be developed to determine the sensitivity of a cost functional $\mathcal{J}$ to small modifications of the control ϕ for nonlinear problems of this sort. To do this, consider the perturbation to the cost functional resulting from a small perturbation to the control ϕ in the direction ϕ'. (This control perturbation direction ϕ' is arbitrary and scaled to have unit norm.) Define $\mathcal{J}'$ as the Fréchet differential [Vainberg, 1964] of a cost functional $\mathcal{J}$ such that

$$\mathcal{J}' \triangleq \lim_{\varepsilon \to 0} \frac{\mathcal{J}(\phi + \varepsilon\phi') - \mathcal{J}(\phi)}{\varepsilon} \triangleq \int_0^T \int_{\Gamma_2^{\pm}} \frac{\mathcal{D}\mathcal{J}(\phi)}{\mathcal{D}(\phi)} \phi' \, dt \, d\mathbf{x}.$$

The quantity $\mathcal{J}'$ is the cost functional perturbation due to a control perturbation $\varepsilon\phi'$ scaled by the inverse of the control perturbation magnitude ε in the limit that $\varepsilon \to 0$. The above relation, considered for arbitrary ϕ', also defines the gradient of the cost functional $\mathcal{J}$ with respect to the control ϕ, which is written $\mathcal{D}\mathcal{J}(\phi)/\mathcal{D}\phi$.

In the current approach, the cost functional perturbation $\mathcal{J}'$ defined previously will be expressed as a simple linear function of the direction of the control perturbation ϕ' through the solution of an adjoint problem. By the above formula, such a representation then reveals the gradient direction $\mathcal{D}\mathcal{J}(\phi)/\mathcal{D}\phi$ directly. With this gradient information, the control ϕ is updated on $(0, T]$ in the direction that, at least locally (i.e., for infinitesimal control updates), most effectively reduces the cost functional. The finite distance the control is updated in this direction is then found by a line search routine, which makes this iteration procedure stable even when controlling nonlinear phenomena. The flow resulting from this modified control is then computed according to the (nonlinear) Navier–Stokes Equation (15.11). The sensitivity of this new flow to further control modification is computed, and the process is repeated. Upon convergence of this iteration, the flow is advanced over the interval $(0, T_1]$, where $T_1 \leqslant T$, and an iteration for the optimal control over a new time interval $(T_1, T_1 + T]$ begins anew.

The cost functional perturbation $\mathcal{J}'$ resulting from a control perturbation in the direction ϕ' is given by

$$\mathcal{J}_0'(\phi) = \frac{1}{2}\int_0^T\int_\Omega C_1^\star C_1\mathbf{u}\cdot\mathbf{u}'d\mathbf{x}\,dt + \frac{1}{2}\int_\Omega (C_2^\star C_2\mathbf{u}\cdot\mathbf{u}')_{t=T}d\mathbf{x} - \int_0^T\int_{\Gamma_2^\pm} \nu C_3^\star\mathbf{r}\cdot\frac{\partial\mathbf{u}'}{\partial n}d\mathbf{x}\,dt$$

$$+\ \ell^2\int_0^T\int_{\Gamma_2^\pm}\phi\phi' d\mathbf{x}\,dt \triangleq \int_0^T\int_{\Gamma_2^\pm}\frac{\mathcal{D}\mathcal{J}_0(\phi)}{\mathcal{D}(\phi)}\phi' d\mathbf{x}\,dt,$$

where $\mathbf{u}'$ is the Fréchet differential of $\mathbf{u}$, as defined in the following subsection. Adjoint calculus is used simply to re-express the integrals involving $\mathbf{u}'$ as a linear function of ϕ'. Once this is accomplished, ϕ' is factored out of the integrands and, as the equation holds for arbitrary ϕ', an expression for the gradient $\mathcal{D}\mathcal{J}_0(\phi)/\mathcal{D}\phi$ is identified.

15.9.1.4 Linearized Perturbation Field

Now consider the linearized perturbation $\mathbf{q}'$ to the flow $\mathbf{q}$ resulting from a perturbation ϕ' to the control ϕ. Again, the quantity $\mathbf{q}'$ may be defined by the limiting process of a Fréchet differential such that

$$\mathbf{q}' \triangleq \lim_{\varepsilon\to 0}\frac{\mathbf{q}'(\phi+\varepsilon\phi') - \mathbf{q}(\phi)}{\varepsilon}.$$

For the purpose of gaining physical intuition, the quantity $\mathbf{q}'$, previously described as a differential quantity, may instead be defined as the small perturbation to the state $\mathbf{q}$ arising from a small control perturbation ϕ' to the control ϕ. In such derivations, the notations $\delta\phi$ and $\delta\mathbf{q}$, denoting small perturbations to ϕ and $\mathbf{q}$, are used instead of the differential quantities ϕ' and $\mathbf{q}'$. The two derivations are roughly equivalent, though the present derivation does not assume that primed quantities are small.

The equation governing the dependence of the linearized flow perturbation $\mathbf{q}'$ on the control perturbation ϕ' may be found by taking the Fréchet differential of the state Equation (15.11). The result is

$$\begin{aligned} \mathcal{N}'(\mathbf{q})\mathbf{q}' &= 0 && \text{in } \Omega, \\ \mathbf{u}' &= -\phi'\mathbf{n} && \text{on } \Gamma_2^\pm, \\ \mathbf{u}' &= 0 && \text{at } t = 0, \end{aligned} \tag{15.12}$$

where the linearized Navier–Stokes operation $\mathcal{N}'(\mathbf{q})\mathbf{q}'$ is given by

$$\mathcal{N}'(\mathbf{q})\mathbf{q}' = \begin{bmatrix} \dfrac{\partial u_j'}{\partial x_j} \\ \dfrac{\partial u_i'}{\partial t} + \dfrac{\partial}{\partial x_j}(u_j u_i' + u_j' u_i) - \nu\dfrac{\partial^2 u_i'}{\partial x_j^2} + \dfrac{\partial p'}{\partial x_i} \end{bmatrix}.$$

The operation $\mathcal{N}'(\mathbf{q})\mathbf{q}'$ is a linear operation on the perturbation field $\mathbf{q}'$, though the operator $\mathcal{N}'(\mathbf{q})\mathbf{q}'$ is itself a function of the solution $\mathbf{q}$ of the Navier–Stokes problem. Equation (15.12) thus reflects the linear dependence of the perturbation field $\mathbf{q}'$ in the interior of the domain on the control perturbation ϕ' at the boundary. However, the implicit linear relationship $\mathbf{q}' = \mathbf{q}'(\phi')$ given by this equation is not yet tractable for expressing $\mathcal{J}_0'$ in a simple form from which $\mathcal{D}\mathcal{J}_0(\phi)/\mathcal{D}\phi$ may be deduced. For the purpose of determining a more useful relationship with which we may determine $\mathcal{D}\mathcal{J}_0(\phi)/\mathcal{D}\phi$, we now appeal to an adjoint identity.

15.9.1.5 Statement of Adjoint Identity

This subsection derives the adjoint of the linear partial differential operator $\mathcal{N}'(\mathbf{q})\mathbf{q}'$. For readers not familiar with this approach, a review of the derivation of an adjoint operator for a very simple case in the present notation is given in Appendix A of Bewley et al. (2001). The adjoint derivation presented below extends in a straightforward manner to more complex equations, such as the compressible Euler equation, as shown in Appendix B of Bewley et al. (2001) (again, using the same notation). Such generality highlights the versatility of the present approach.

Define an inner product over the domain in space-time under consideration such that

$$\langle \mathbf{q}^\star, \mathbf{q}' \rangle = \int_0^T \int_\Omega \mathbf{q}^\star \cdot \mathbf{q}' d\mathbf{x}\, dt$$

and consider the identity

$$\langle \mathbf{q}^\star,\ \mathcal{N}'(\mathbf{q})\mathbf{q}' \rangle = \langle \mathcal{N}'(\mathbf{q})^\star \mathbf{q}^\star, \mathbf{q}' \rangle + b. \tag{15.13}$$

Integration by parts may be used to move all differential operations from $\mathbf{q}'$ on the left-hand side of Equation (15.13) to $\mathbf{q}^*$ on the right-hand side, resulting in the derivation of the adjoint operator

$$\mathcal{N}'(\mathbf{q})^\star \mathbf{q}^\star = \begin{pmatrix} \dfrac{\partial u_j^\star}{\partial x_j} \\ -\dfrac{\partial u_i^\star}{\partial t} - u_j\left(\dfrac{\partial u_i^\star}{\partial x_j} + \dfrac{\partial u_j^\star}{\partial x_i}\right) - \nu \dfrac{\partial^2 u_i^\star}{\partial x_i^2} - \dfrac{\partial p^\star}{\partial x_i} \end{pmatrix},$$

where, again, the operation $\mathcal{N}'(\mathbf{q})^\star \mathbf{q}^\star$ is a linear operation on the adjoint field $\mathbf{q}^\star$, and the operator $\mathcal{N}'(\mathbf{q})^\star$ is itself a function of the solution $\mathbf{q}$ of the Navier–Stokes problem. From the integrations by parts, we also get several boundary terms:

$$b = \int_\Omega (u_j^\star u_i')_{t=0}^{t=T}\, d\mathbf{x} + \int_0^T \int_\Omega n_j \left[u_i^\star (u_j u_i' + u_j' u_i) + p^\star u_j' - \nu \left(u_i^\star \frac{\partial u_i'}{\partial x_j} - u_i' \frac{\partial u_i^\star}{\partial x_j} \right) + u_j^\star p' \right] d\mathbf{x}\, dt.$$

The identity Equation (15.13) is the key to expressing $\mathcal{J}'$ in the desired form. An adjoint field $\mathbf{q}^*$ is first defined using the operator $\mathcal{N}'(\mathbf{q})^\star$ together with appropriate forcing on an interior equation with appropriate boundary conditions and initial conditions. There is some flexibility which we exploit to obtain a simple expression of $\mathcal{J}'$. Combining this definition of $\mathbf{q}^\star$ with the definitions of $\mathbf{q}$ in Equation (15.11) and $\mathbf{q}'$ in Equation (15.12), the identity Equation (15.13) reveals the desired expression, as is now shown.

15.9.1.6 Definition of Adjoint Field

Consider an adjoint state defined (as yet, arbitrarily) by

$$\begin{aligned} \mathcal{N}'(\mathbf{q})^\star \mathbf{q}^\star &= \begin{pmatrix} 0 \\ C_1^\star C_1 \mathbf{u} \end{pmatrix} && \text{in } \Omega, \\ \mathbf{u}^\star &= C_3^\star \mathbf{r} && \text{on } \Gamma_2^\pm, \\ \mathbf{u}^\star &= C_2^\star C_2 \mathbf{u} && \text{at } t = T, \end{aligned} \tag{15.14}$$

where the adjoint operation $\mathcal{N}'(\mathbf{q})^\star$ is derived in the previous subsection. Note by Equation (15.14) that, depending on where the cost functional weighs the flow perturbations (see Section 15.9.1.2), the adjoint problem may be driven by the initial conditions, by the boundary conditions, or by the RHS of the adjoint PDE itself. Note also that the adjoint "initial" conditions are defined at $t = T$ and are thus best referred to as "terminal" conditions. With this definition, the adjoint field must be marched backward in time over the optimization horizon. Because of the sign of the time derivative and viscous terms in the adjoint operator $\mathcal{N}'(\mathbf{q})^\star$, this is the natural direction for this time march. However, as both the adjoint operator $\mathcal{N}'(\mathbf{q})^\star$ and the RHS forcing on Equation (15.14) are functions of $\mathbf{q}$, computation of the adjoint field $\mathbf{q}^\star$ requires storage of the flow field $\mathbf{q}$ on $t \in [0, T]$, which itself must be computed with a forward march. This storage issue presents one of the numerical complications that preclude solution of the present optimization problem for large optimization intervals T. However, this storage issue is not insurmountable for intermediate values of $T^+ < O(100)$. The adjoint problem Equation (15.14), though linear, has complexity similar to that of the Navier–Stokes problem, Equation (15.11), and may be solved with similar numerical methods.

15.9.1.7 Identification of Gradient

The identity Equation (15.13) is now simplified using the equations defining the state field Equation (15.11), the perturbation field Equation (15.12), and the adjoint field Equation (15.14). Due to the judicious choice of the forcing terms driving the adjoint problem, the identity Equation (15.13) reduces (after some manipulation) to

$$\int_0^T \int_\Omega C_1^\star C_1 \mathbf{u} \cdot \mathbf{u}' d\mathbf{x}\, dt + \int_\Omega (C_2^\star C_2 \mathbf{u} \cdot \mathbf{u}')_{t=T} d\mathbf{x} - \int_0^T \int_{\Gamma_2^\pm} \nu C_3^\star \mathbf{r} \cdot \frac{\partial \mathbf{u}'}{\partial n} d\mathbf{x}\, dt = \int_0^T \int_{\Gamma_2^\pm} P^\star \phi' d\mathbf{x}\, dt.$$

Using this equation, the cost functional perturbation $\mathcal{J}_0'$ may be rewritten as

$$\mathcal{J}_0'(\phi; \phi') = \int_0^T \int_{\Gamma_2^\pm} (p^\star + \ell^2 \phi)\phi' \, d\mathbf{x}\, dt \triangleq \int_0^T \int_{\Gamma_2^\pm} \frac{\mathcal{D}\mathcal{J}_0(\phi)}{\mathcal{D}(\phi)} \phi' d\mathbf{x}\, dt.$$

Because ϕ' is arbitrary, we may identify (weakly) the desired gradient as

$$\frac{\mathcal{D}\mathcal{J}_0(\phi)}{\mathcal{D}(\phi)} = p^\star + \ell^2 \phi.$$

The desired gradient $\mathcal{D}\mathcal{J}_0(\phi)/\mathcal{D}\phi$ is a simple function of the solution of the adjoint problem proposed in Equation (15.14). Specifically, in the present case of boundary forcing by wall-normal blowing and suction, the gradient is a simple function of the adjoint pressure on the walls.

In fact, this simple result hints at the more fundamental physical interpretation of what the adjoint field actually represents: *The adjoint field $\mathbf{q}^\star$, when properly defined, is a measure of the sensitivity of the terms of the cost functional that appraise the state $\mathbf{q}$ to additional forcing of the state equation.*

There are exactly as many components of the adjoint field $\mathbf{q}^\star$ as there are components of the state PDE on the interior of the domain. Also note that the adjoint field may take nontrivial values at the initial time $t = 0$ and on the boundaries $\Gamma_2^\pm$. Depending upon where the control is applied to the state Equation (15.11), (i.e., on the RHS of the mass or momentum equations on the interior of the domain, on the boundary conditions, or on the initial conditions), the adjoint field will appear in the resulting expression for the gradient accordingly.

To summarize, the forcing on the adjoint problem is a function of where the flow perturbations are weighed in the cost functional. The dependence of the gradient $\mathcal{D}\mathcal{J}(\phi)/\mathcal{D}\phi$ on the resulting adjoint field, however, is a function of where the control enters the state equation.

15.9.1.8 Gradient Update to Control

A control optimization strategy using a steepest descent algorithm may now be proposed such that

$$\phi^k = \phi^{k-1} - \alpha^k \frac{\mathcal{D}\mathcal{J}_0(\phi^{k-1})}{\mathcal{D}\phi}$$

over the entire time interval $t \in (0, T]$, where k indicates the iteration number and α^k is a parameter of descent that governs how large an update is made, which is adjusted at each iteration step to be the value that minimizes $\mathcal{J}$. This algorithm updates ϕ at each iteration in the direction of maximum decrease of $\mathcal{J}$. As $k \to \infty$, the algorithm should converge to some local minimum of $\mathcal{J}$ over the domain of the control ϕ on the time interval $t \in (0, T]$. Convergence to a global minimum will not in general be attained by such a scheme and that, as time proceeds, $\mathcal{J}$ will not necessarily decrease.

The steepest descent algorithm previously described illustrates the essence of the approach, but is usually not very efficient. Even in linear low-dimensional problems, for cases in which the cost functional has a long, narrow "valley," the lack of a momentum term from one iteration to the next tends to cause the steepest descent algorithm to bounce from one side of the valley to the other without turning to proceed along the valley floor. Standard nonlinear conjugate gradient algorithms [e.g., Press et al., 1986] improve this behavior considerably with relatively little added computational cost or algorithmic complexity, as discussed further in Bewley et al. (2001).

As mentioned previously, the dimension of the control in the present problem (once discretized) is quite large, which precludes the use of second-order techniques based on the computation or approximation of the Hessian matrix $\partial^2\mathcal{J}/\partial\phi_i\partial\phi_j$ or its inverse during the control optimization. The number of elements in such a matrix scales with the square of the number of control variables and is unmanageable in the present case. However, reduced-storage variants of variable metric methods [Vanderplaats, 1984], such as the Davidon–Fletcher–Powell (DFP) method, the Broydon–Fletcher–Goldfarb–Shanno (BFGS) method, and the sequential quadratic programming (SQP) method, approximate the inverse Hessian information by outer products of stored gradient vectors and thus achieve nearly second-order convergence without storage of the Hessian matrix itself. Such techniques should be explored further for very large-scale optimization problems.

15.9.2 Continuous Adjoint vs. Discrete Adjoint

Direct numerical simulations (DNS) of the current three-dimensional nonlinear system necessitate carefully chosen numerical techniques involving a stretched, staggered grid, an energy-conserving spatial discretization, and a mixture of implicit and multistep explicit schemes for accurate time advancement, with incompressibility enforced by an involved fractional step algorithm. The optimization approach previously described, which will be referred to as "optimize then discretize" (OTD), avoids all of these cumbersome numerical details by deriving the gradient of the cost functional in the continuous setting, discretizing in time and space as the final step before implementation in numerical code. The remarkable similarity of the flow and adjoint systems allows both to be coded with similar numerical techniques. For systems which are well resolved in the numerical discretization, this approach is entirely justifiable and yields adjoint systems which are easy to derive and implement in numerical code.

Unfortunately, many PDE systems, such as high-Reynolds-number turbulent flows, are difficult or impossible to simulate with sufficient resolution to capture accurately all of the important dynamic phenomena of the continuous system. Such systems are often simulated on coarse grids, usually with some "subgrid-scale model" to account for the unresolved dynamics. This setting is referred to as large eddy simulation (LES), and a variety of techniques are currently under development to model the significant subgrid-scale effects.

There are important unresolved issues concerning how to approach large eddy simulations in the optimization framework. If we continue with the OTD approach, in which the optimization equations are determined before the numerical discretization is applied, it is not yet clear at what point the LES model should be introduced. Professor Scott Collis' group (Rice University) has modified the numerical code of Bewley et al. (2001) to study this issue; Chang and Collis (1999) report on their preliminary findings.

An alternative approach to the OTD setting, in which one spatially discretizes the governing equation before determining the optimization equations, may also be considered. After spatially discretizing the governing equation, this approach, which will be referred to as "discretize then optimize" (DTO), follows an analogous sequence of steps as the OTD approach presented previously, with these steps now applied in the discrete setting. Derivation of the adjoint operator is significantly more cumbersome in this discrete setting. In general, the processes of optimization and discretization do not commute, and thus the OTD and DTO approaches are not necessarily equivalent even upon refinement of the space/time grid [Vogel and Wade, 1995]. However, by carefully framing the discrete identity defining the DTO adjoint operator as a discrete approximation of the identity given in Equation (15.13), these two approaches can be posed in an equivalent fashion for Navier–Stokes systems.

It remains the topic of some debate whether or not the DTO approach is better than the OTD approach for marginally resolved PDE systems. The argument for DTO is that it clearly is the most direct way to optimize the discrete problem actually being solved by the computer. The argument against DTO is that one really wants to optimize the continuous problem, so gradient information that identifies and exploits deficiencies in the numerical discretization that can lead to performance improvements in the discrete problem might be misleading when interpreting the numerical results in terms of the physical system.

15.10 Robustification: Appealing to Murphy's Law

Though optimal control approaches possess an attractive mathematical elegance and are now proven to provide excellent results in terms of drag and turbulent kinetic energy reduction in fully developed turbulent flows, they are often impractical. One of the most significant drawbacks of this nonlinear optimization approach is that it tends to "over-optimize" the system, leaving a high degree of design-point sensitivity. This phenomenon has been encountered frequently in, for example, the adjoint-based optimization of the shape of aircraft wings. Overly optimized wing shapes might work quite well at exactly the flow conditions for which they were designed, but their performance is often abysmal at off-design conditions. To abate such system sensitivity, the noncooperative framework of robust control provides a natural means to "detune" the optimized results. This concept can be applied easily to a broad range of related applications. *The noncooperative approach to robust control, one might say, amounts to Murphy's law taken seriously: If a worst-case disturbance can disrupt a controlled closed-loop system, it will.*

When designing a robust controller, therefore, one might plan on a finite component of the worst-case disturbance aggravating the system, and design a controller suited to handle this extreme situation. A controller designed to work in the presence of a finite component of the worst-case disturbance will also be robust to a wide class of other possible disturbances which, by definition, are not as detrimental to the control objective as the worst-case disturbance. This concept leads to the $\mathcal{H}_\infty$ control formulation discussed previously in the linear setting, and can easily be extended to the optimization of nonlinear systems.

Based on the ideas of $\mathcal{H}_\infty$ control theory presented in Section 15.3, the extension of the nonlinear optimization approach presented in Section 15.9 to the noncooperative setting is straightforward. A disturbance is first introduced to the governing Equation (15.11). As an example, consider disturbances that perturb the state PDE itself such that

$$\mathcal{N}(\mathbf{q}) = \mathbf{F} + \mathbf{B}_1(\psi) \quad \text{in } \Omega.$$

(Accounting for disturbances to the boundary conditions and initial conditions of the governing equation is also straightforward.) The cost functional is then extended to penalize these disturbances in the noncooperative framework, as was also done in the linear setting

$$\mathcal{J}_r(\psi, \phi) = \mathcal{J}_0 - \frac{\gamma^2}{2} \int_0^T \int_\Omega |\psi|^2 \, d\mathbf{x} \, dt.$$

This cost functional is simultaneously minimized with respect to the controls ϕ and maximized with respect to the disturbances ψ (Figure 15.16). The parameter γ is used to scale the magnitude of the disturbances accounted for in this noncooperative competition, with the limit of large γ recovering the optimal approach discussed in Section 15.9 (i.e., $\psi \to 0$). A gradient-based algorithm may then be devised to march to the saddle point, such as the simple algorithm given by:

$$\phi^k = \phi^{k-1} - \alpha^k \frac{\mathcal{D}\mathcal{J}_r(\psi^{k-1}; \phi^{k-1})}{\mathcal{D}\phi},$$

$$\psi^k = \psi^{k-1} + \beta^k \frac{\mathcal{D}\mathcal{J}_r(\psi^{k-1}; \phi^{k-1})}{\mathcal{D}\psi}.$$

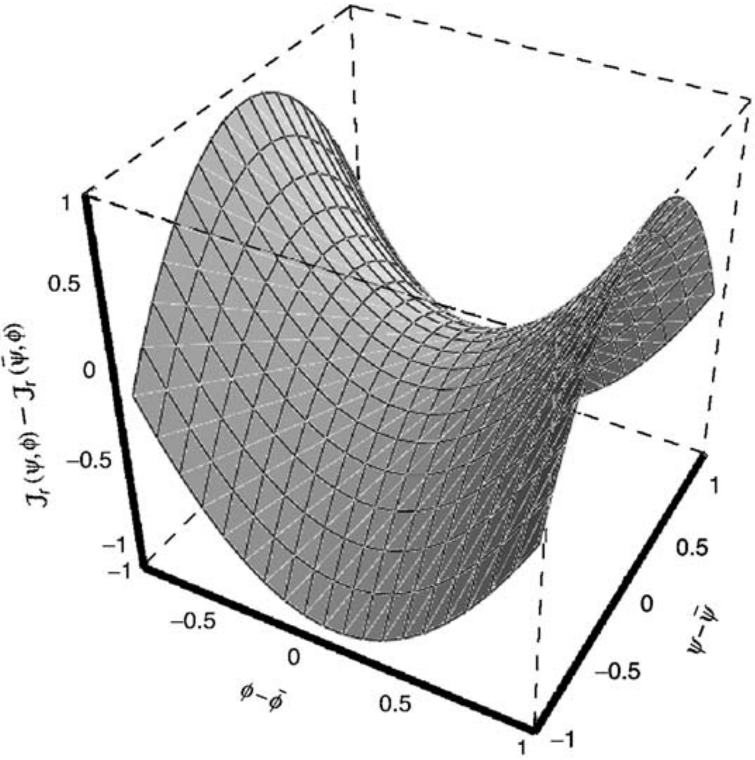


FIGURE 15.16 Schematic of a saddle point representing the neighborhood of a solution to a robust control problem with one scalar disturbance variable ψ and one scalar control variable ϕ. When the robust control problem is solved, the cost function $\mathcal{J}_r$ is simultaneously maximized with respect to ψ and minimized with respect to ϕ, and a saddle point such as $(\bar{\psi}, \bar{\phi})$ is reached. An essentially infinite- dimensional extension of this concept might be formulated to achieve robustness to disturbances and insensitivity to design point in fluid-mechanical systems. In such approaches, the cost $\mathcal{J}_r$ is related to a distributed disturbance ψ and a distributed control ϕ through the solution of the Navier–Stokes equation.

The robust control problem is considered to be solved when a saddle point $(\bar{\psi}, \bar{\phi})$ is reached; such a solution, if it exists, is not necessarily unique.

The gradients $\mathcal{D}\mathcal{J}_r(\psi; \phi)/\mathcal{D}\phi$ and $\mathcal{D}\mathcal{J}_r(\psi; \phi)/\mathcal{D}\psi$ may be found in a manner analogous to that leading to $\mathcal{D}\mathcal{J}_0(\phi)/\mathcal{D}\phi$ discussed in Section 15.9. In fact, both gradients may be extracted from the single adjoint field defined by Equation (15.14). Thus, the additional computational complexity introduced by the noncooperative component of the robust control problem is simply a matter of updating and storing the appropriate disturbance variables.

15.10.1 Well-Posedness

Based on the extensive mathematical literature on the Navier–Stokes equation, Abergel and Temam (1990) established the well-posedness of the mathematical framework for the optimization problem presented in Section 15.9. This characterization was generalized and extended to the noncooperative framework of Section 15.10 in Bewley et al. (2000).

Because the inequalities currently available for estimating the magnitude of the various terms of the Navier–Stokes equation are limited, the mathematical characterizations in both of these articles are quite conservative. In our numerical simulations, we regularly apply numerical optimization techniques to control problems that are well outside the range over which we can mathematically establish well-posedness. However, such mathematical characterizations are still quite important because they give us confidence that, for example, if ℓ, and γ are at least taken to be large enough, a saddle point of the noncooperative optimization problem will exist. Once such mathematical characterizations are derived, numerically

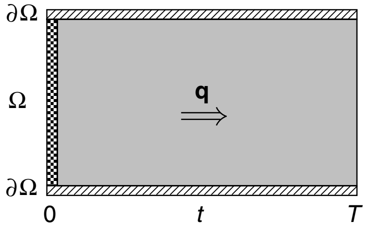


FIGURE 15.17 Schematic of the space–time domain over which the flow field **q** is defined. The possible regions of forcing in the system defining **q** are: (1) the right-hand side of the PDE, indicated with shading, representing flow control by interior volume forcing (e.g., externally applied electromagnetic forcing by wall-mounted magnets and electrodes); (2) the boundary conditions, indicated with diagonal stripes, representing flow control by boundary forcing (e.g., wall transpiration); and (3) the initial conditions, indicated with checkerboard, representing optimization of the initial state in a data assimilation framework (e.g., the weather forecasting problem).

determining the values of ℓ, and γ for which solutions of the control problem may still be obtained is reduced to a simple matter of implementation.

15.10.2 Convergence of Numerical Algorithms

Saddle points are typically more difficult to find than minimum points, and particular care needs to be taken to craft efficient but stable numerical algorithms for finding them. In the approach described previously, sufficiently small values of α^k and β^k must be selected to ensure convergence. Fortunately, the same mathematical inequalities used to characterize well-posedness of the control problem can also be used to characterize convergence of proposed numerical algorithms. Such characterizations lend valuable insight when designing practical numerical algorithms. Preliminary work in the development of such saddle point algorithms is reported by Tachim Medjo (2000).

15.11 Unification: Synthesizing a General Framework

The various cost functionals considered previously led to three possible sources of forcing for the adjoint problem: the right-hand side of the PDE, the boundary conditions, and the initial conditions. Similarly, three different locations of forcing may be identified for the flow problem. As illustrated in Figures 15.17 and 15.18 and discussed further in Bewley et al. (2000), the various regions of forcing of the flow and adjoint problems together form a general framework that can be applied to a wide variety of problems in fluid mechanics including both flow control (e.g., drag reduction, mixing enhancement, and noise control) and flow forecasting (e.g., weather prediction and storm forecasting). Related techniques, but applied to the time-averaged Navier–Stokes equation, have also been used extensively to optimize the shapes of airfoils [see, e.g., Reuther et al., 1996].

By identifying a range of problems that all fit into the same general framework, we can better understand how to extend, for example, the idea of noncooperative optimizations to a full suite of related problems in fluid mechanics. Though advanced research projects must often be highly focused and specialized to obtain solid results, the importance of making connections of such research to a large scope of related problems must be recognized to realize fully the potential impact of the techniques developed.

15.12 Decomposition: Simulation-Based System Modeling

For the purpose of developing model-based feedback control strategies for turbulent flows, reduced-order nonlinear models of turbulence that are effective in the closed-loop setting are highly desired. Recent

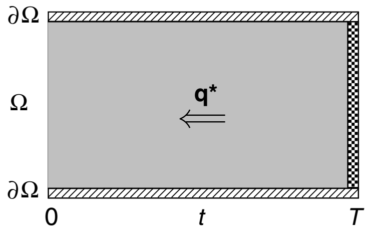


FIGURE 15.18 Schematic of the space–time domain over which the adjoint field **q*** is defined. The possible regions of forcing in the system defining **q***, corresponding exactly to the possible domains in which the cost functional can depend on **q**, are: (1) the right-hand side of the PDE, indicated with shading, representing regulation of an interior quantity (e.g., turbulent kinetic energy); (2) the boundary conditions, indicated with diagonal stripes, representing regulation of a boundary quantity (e.g., wall skin friction); and (3) the terminal conditions, indicated with checkerboard, representing terminal control of an interior quantity (e.g., turbulent kinetic energy).

work in this direction, using proper orthogonal decompositions (POD) to obtain these reduced-order representations, is reviewed by Lumley and Blossey (1998).

The POD technique uses analysis of a simulation database to develop an efficient reduced-order basis for the system dynamics represented within the database [Holmes et al., 1996]. One of the primary challenges of this approach is that the dynamics of the system in closed loop (after the control is turned on) is often quite different than the dynamics of the open-loop (uncontrolled) system. Thus, development of simulation-based reduced-order models for turbulent flows should probably be coordinated with the design of the control algorithm itself to determine system models that are maximally effective in the closed-loop setting. Such coordination of simulation-based modeling and control design is largely an unsolved problem. A particularly sticky issue is that, as the controls are turned on, the dynamics of the turbulent flow system are nonstationary (they evolve in time). The system eventually relaminarizes if the control is sufficiently effective. In such nonstationary problems, it is not clear which dynamics the POD should represent (of the flow shortly after the control is turned on, of the nearly relaminarized flow, or of something in between), or if in fact several PODs should be created and used in a scheduled approach in an attempt to capture several different stages of the nonstationary relaminarization process.

Reduced-order models that are effective in the closed-loop setting need not capture the majority of the energetics of the unsteady flow. Rather, the essential feature of a system model for the purpose of control design is that the model capture the important effects of the control on the system dynamics. Future control-oriented modeling efforts might benefit by deviating from the standard POD mindset of simply attempting to capture the energetics of the system dynamics, instead focusing on capturing the significant effects of the control on the system in a reduced-order fashion.

15.13 Global Stabilization: Conservatively Enhancing Stability

Global stabilization approaches based on Lyapunov analysis of the system energetics have been explored recently for two-dimensional channel-flow systems (in the continuous setting) by Balogh et al. (2001). In the setting considered there, localized tangential wall motions are coordinated with local measurements of skin friction via simple proportional feedback strategies. Analysis of the flow at $Re \leq 0.125$ motivates such feedback rules, indicating appropriate values of proportional feedback coefficients that enhance the L^2 stability of the flow. Though such an approach is very conservative, rigorously guaranteeing enhanced stability of the channel-flow system only at extremely low Reynolds numbers, extrapolation of the feedback strategies so determined to much higher Reynolds numbers also indicates effective enhancements of system stability, even for three-dimensional systems up to $Re = 2000$ (A. Balogh, pers. comm.).

An alternative approach for achieving global stabilization of a nonlinear PDE is the application of nonlinear backstepping to the discretized system equation. Boškovic and Krstic (2001) report on recent efforts in this direction (applied to a thermal convection loop). Backstepping is typically an aggressive approach to stabilization. One of the primary difficulties with this approach is that proofs of convergence to a continuous, bounded function upon refinement of the grid are difficult to attain due to increasing controller complexity as the grid is refined. Significant advancements are necessary before this approach will be practical for turbulent flow systems.

15.14 Adaptation: Accounting for a Changing Environment

Adaptive control algorithms, such as least mean squares (LMS), neural networks (NN), genetic algorithms (GA), simulated annealing, extremum seeking, and the like, play an important role in the control of fluid-mechanical systems when the number of undetermined parameters in the control problem is fairly small (O(10)) and individual "function evaluations" (i.e., quantitative characterizations of the effectiveness of the control) can be performed relatively quickly. Many control problems in fluid mechanics are of this type, and are readily approachable by a wide variety of well-established adaptive control strategies. A significant advantage of such approaches over those discussed previously is that they do not require extensive analysis or coding of localized convolution kernels, adjoint fields, etc., but may instead be applied directly "out of the box" to optimize the parameters of interest in a given fluid-mechanical problem. This also poses a bit of a disadvantage, however, because the analysis required during the development of model-based control strategies can sometimes yield significant physical insight that black-box optimizations fail to provide.

To apply the adaptive approach, one needs an inexpensive simulation code or an experimental apparatus in which the control parameters of interest can be altered by an automated algorithm. Any of a number of established methodological strategies can then be used to search the parameter space for favorable closed-loop system behavior. Given enough function evaluations and a small enough number of control parameters, such strategies usually converge to effective control solutions. Koumoutsakos et al. (1998) demonstrate this approach (computationally) to determine effective control parameters for exciting instabilities in a round jet. Rathnasingham and Breuer (1998) demonstrate this approach (experimentally) for the feed-forward reduction of turbulence intensities in a boundary layer.

Unfortunately, due to an effect known as "the curse of dimensionality," as the number of control parameters to be optimized is increased, the ability of adaptive strategies to converge to effective control solutions based on function evaluations alone is diminished. For example, in a system with 1000 control parameters, it takes 1000 function evaluations to determine the gradient information available in a single adjoint computation. Thus, for problems in which the number of control variables to be optimized is large, the convergence of adaptive strategies based on function evaluations alone is generally quite poor. In such high-dimensional problems, for cases in which the control problem of interest is plagued by multiple minima, a blend of an efficient adjoint-based gradient optimization approach with GA-type management of parameter "mutations" or the simulated annealing approach of varying levels of "noise" added to the optimization process might prove to be beneficial.

Adaptive strategies are also quite valuable for recognizing and responding to changing conditions in the flow system. In the low-dimensional setting, they can be used online to update controller gains directly as the system evolves in time (for instance, as the mean speed or direction of the flow changes or as the sensitivity of a sensor degrades). In the high-dimensional setting, adaptive strategies can be used to identify certain critical aspects of the flow (such as the flow speed), and based on this identification, an appropriate control strategy may be selected from a look-up table of previously computed controller gains.

The selection of what level of adaptation is appropriate for a particular flow control problem of interest is a consideration that must be guided by physical insight of the particular problem at hand.

15.15 Performance Limitation: Identifying Ideal Control Targets

Another important, but as yet largely unrealized, role for mathematical analysis in the field of flow control is in the identification of fundamental limitations on the performance that can be achieved in certain

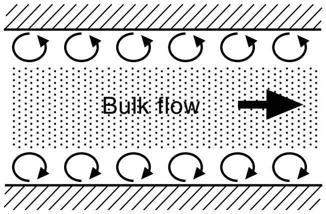


FIGURE 15.19 An enticing picture: fundamental restructuring of the near-wall unsteadiness to insulate the wall from the viscous effects of the bulk flow. It has been argued [Nosenchuck, 1994; Koumoutsakos, 1999] that it might be possible to maintain a series of so-called "fluid rollers" to effectively reduce the drag of a near-wall flow. Such rollers are depicted in the figure above by indicating total velocity vectors in a reference frame convecting with the vortices themselves; in this frame, the generic picture of fluid rollers is similar to a series of stationary Kelvin–Stuart cat's eye vortices. A possible mechanism for drag reduction might be akin to a series of solid cylinders serving as an effective conveyor belt, with the bulk flow moving to the right above the vortices and the wall moving to the left below the vortices. It is still the topic of some debate whether or not a continuous flow can be maintained in such a configuration by an unsteady control in such a way as to sustain the mean skin friction below laminar levels. Such a control might be implemented either by interior electromagnetic forcing (applied with wall-mounted magnets and electrodes) or by boundary controls such as zero-net mass-flux blowing/suction.

flow control problems. For example, motivated by the active debate surrounding the proposed physical mechanism for channel-flow drag reduction illustrated in Figure 15.19, we formally state the following, as yet unproven, conjecture:

Conjecture: The lowest sustainable drag of an incompressible constant mass-flux channel flow, in either two or three dimensions, when controlled via a distribution of zero-net mass-flux blowing/suction over the channel walls, is exactly that of the laminar flow.

By "sustainable drag" we mean the long-time average of the instantaneous drag, given by:

$$D_\infty = \lim_{(T\to\infty)} \frac{-1}{T} \int_0^T \int_{\Gamma_2^{\pm}} \nu \frac{\partial u_1}{\partial n}\, d\mathbf{x}\, dt$$

Proof (by mathematical analysis) or disproof (by counterexample) of this conjecture would be quite significant and lead to greatly improved physical understanding of the channel flow problem. If proven to be correct, it would provide rigorous motivation for targeting flow relaminarization when the problem one actually seeks to solve is minimization of drag. If shown to be incorrect, our target trajectories for future flow control strategies might be substantially altered.

Similar fundamental performance limitations may also be sought for exterior flow problems, such as the minimum drag of a circular cylinder subject to a class of zero-net control actions, such as rotation or transverse oscillation (B. Protas, pers. comm.).

15.16 Implementation: Evaluating Engineering Trade-Offs

We are still some years away from applying the distributed control techniques discussed herein to microelectromechanical systems (MEMS) arrays of sensors and actuators, such as that depicted in Figure 15.20. One of the primary hurdles to bringing us closer to actual implementation is that of accounting for practical designs of sensors and actuators in the control formulations, rather than the idealized distributions of blowing/suction and skin-friction measurements that we have assumed here. Detailed simulations, such as that shown in Figure 15.21, of proposed actuator designs are essential for developing reduced-order models of the effects of the actuators on the system of interest to make control design for realistic arrays of sensors and actuators tractable.

By performing analysis and control design in a high-dimensional, unconstrained setting, as discussed in this chapter, it is believed that we can obtain substantial insight into the physical characteristics of

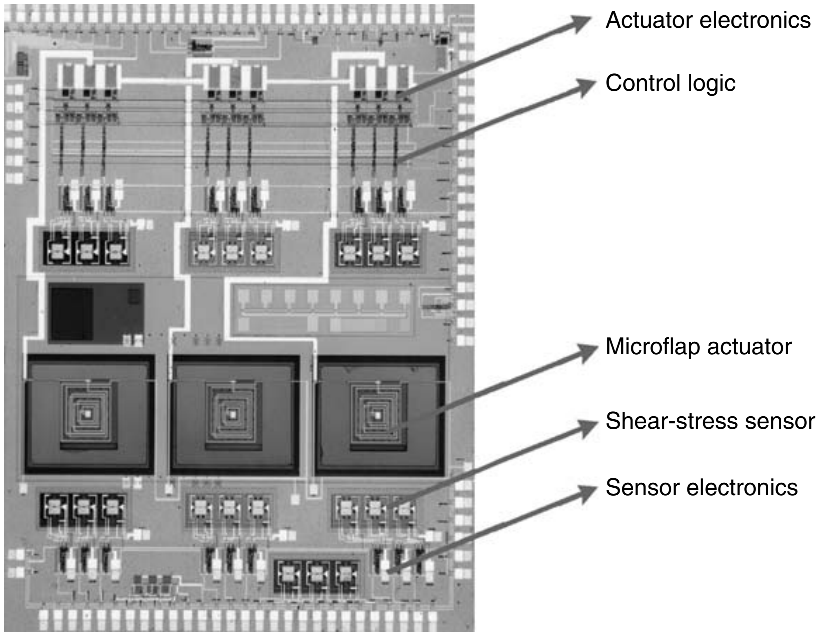


FIGURE 15.20 (**See color insert following page 10-34.**) A MEMS tile integrating sensors, actuators and control logic for distributed flow control applications. (Developed by Professors Chih-Ming Ho, UCLA, and Yu-Chong Tai, Caltech.)

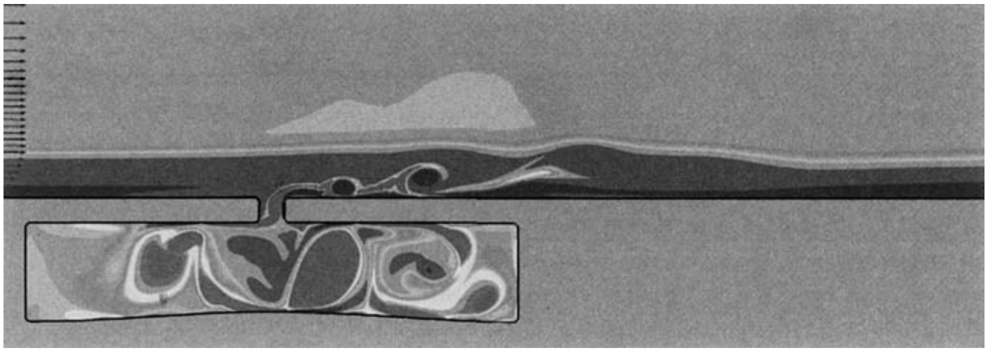


FIGURE 15.21 Simulation of a proposed driven-cavity actuator design (Professor Rajat Mittal, University of Florida). The fluid-filled cavity is driven by vertical motions of the membrane along its lower wall. Numerical simulation and reduced-order modeling of the influence of such flow-control actuators on the system of interest will be essential for the development of feedback control algorithms to coordinate arrays of realistic sensor/actuator configurations.

highly effective control strategies. Such insight naturally guides the engineering trade-offs that follow to make the design of the turbulence control system practical. Particular traits of the present control solutions in which we are especially interested include the times scales and the streamwise and spanwise length scales that are dominant in the optimized control computations (which shed insight on suitable actuator bandwidth, dimensions, and spacing) and the extent and structure of the convolution kernels (which indicate the distance and direction over which sensor measurements and state estimates should propagate when designing the communication architecture of the tiled array).

It is recognized that the control algorithm finally to be implemented must be kept fairly simple for its realization in the on-board electronics to be feasible. We believe that an appropriate strategy for determining implementable feedback algorithms that are both effective and simple is to learn how to solve the high-dimensional, fully resolved control problem first, as discussed herein. This results in high-dimensional compensator designs that are highly effective in the closed-loop setting. Compensator reduction strategies combined with engineering judgment may then be used to distill the essential features of such well-resolved control solutions to implementable feedback designs with minimal degradation of the closed-loop system behavior.

15.17 Discussion: A Common Language for Dialog

It is imperative that an accessible language be developed that provides a common ground upon which people from the fields of fluid mechanics, mathematics, and controls can meet, communicate, and develop new theories and techniques for flow control. Pierre-Simon de Laplace (quoted by Rose, 1998) once said

> Such is the advantage of a well-constructed language that its simplified notation often becomes the source of profound theories.

Similarly, it was recognized by Gottfried Wilhelm Leibniz (quoted by Simmons, 1992) that

> In symbols one observes an advantage in discovery which is greatest when they express the exact nature of a thing briefly … then indeed the labor of thought is wonderfully diminished.

Profound new theories are still possible in this young field. We have not yet homed in on a common language in which such profound theories can be framed. Such a language needs to be actively pursued. Time spent on identifying, implementing, and explaining a clear "compromise" language that is approachable by those from the related "traditional" disciplines is time well spent.

In particular, care should be taken to respect the meaning of certain "loaded" words which imply specific techniques, qualities, or phenomena in some disciplines but only general notions in others. When both writing and reading papers on flow control, one must be especially alert, as these words are sometimes used outside of their more narrow, specialized definitions, creating undue confusion. With time, a common language will develop. In the meantime, avoiding the use of such words outside of their specialized definitions, precisely defining such words when they are used, and identifying and using the existing names for specialized techniques already well established in some disciplines when introducing such techniques into other disciplines, will go a long way toward keeping us focused and in sync as an extended research community.

There are, of course, some significant obstacles to the implementation of a common language. For example, fluid mechanicians have historically used **u** to denote flow velocities and **x** to denote spatial coordinates, whereas the controls community overwhelmingly adopts **x** as the state vector and **u** as the control. The simplified two-dimensional system that fluid mechanicians often study examines the flow in a vertical plane, whereas the simplified two-dimensional system that meteorologists often study examines the flow in a horizontal plane. Thus, when studying three-dimensional problems such as turbulence, those with a background in fluid mechanics usually introduce their third coordinate z in a horizontal direction, whereas those with a background in meteorology normally have "their zed in the clouds." Writing papers in a manner conscious to such different backgrounds and notations, elucidating, motivating, and distilling the suitable control strategies, the relevant flow physics, the useful mathematical inequalities, and the appropriate numerical methods to a general audience of specialists from other fields is certainly extra work. However, such efforts are necessary to make flow control research accessible to the broad audience of scientists, mathematicians, and engineers whose talents will be instrumental in advancing this field in the years to come.

15.18 The Future: A Renaissance

The field of flow control is now poised for explosive growth and exciting new discoveries. The relative maturity of the constituent traditional scientific disciplines contributing to this field provides us with key

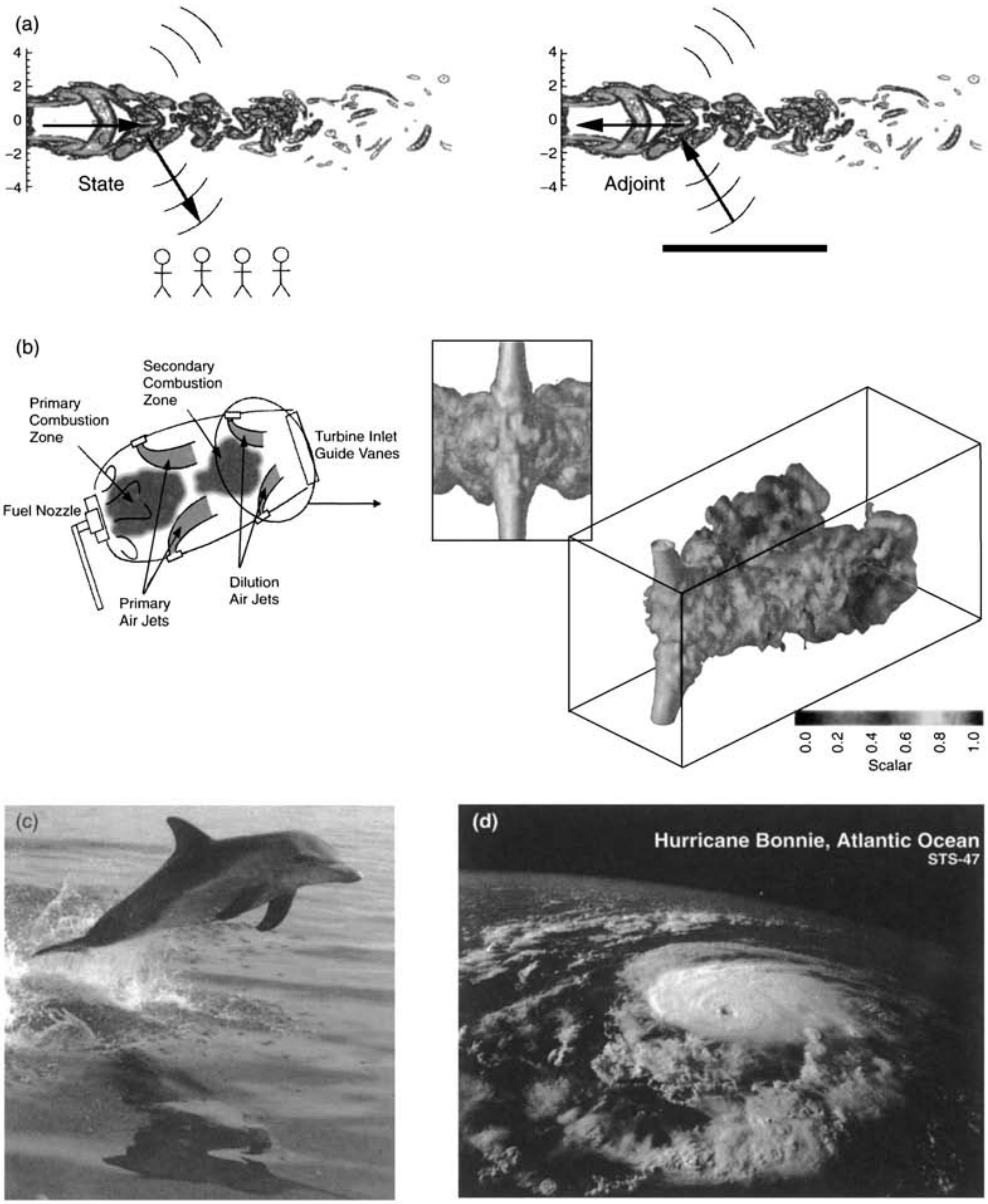


FIGURE 15.22 (**See color insert following page 10-34.**) Future interdisciplinary problems in flow control amenable to adjoint-based analysis: (a) minimization of sound radiating from a turbulent jet (simulation by Prof. Jon Freund, UCLA), (b) maximization of mixing in interacting cross-flow jets (simulation by Dr. Peter Blossey, UCSD) [Schematic of jet engine combustor is shown at left. Simulation of interacting cross-flow dilution jets, designed to keep the turbine inlet vanes cool, are visualized at right.], (c) optimization of surface compliance properties to minimize turbulent skin friction, and (d) accurate forecasting of inclement weather systems.

elements that future efforts in this field may leverage. The work described herein represents only our first, preliminary steps towards laying an integrated, interdisciplinary footing upon which future efforts in this field may be based. Many technologically significant and fundamentally important problems lie before us, awaiting analysis and new understanding in this setting. With each of these new applications come significant new questions about how best to integrate the constituent disciplines. The answers to these difficult questions will only come about through a broad knowledge of what these disciplines have to offer and how they can best be used in concert. A few problems that might be studied in the near future in the present interdisciplinary framework are highlighted in Figure 15.22.

Unfortunately, there are particular difficulties in pursuing truly interdisciplinary investigations of fundamental problems in flow control in our current society because it is impossible to conduct such investigations from the perspective of any particular traditional discipline alone. Though the language of interdisciplinary research is in vogue, many university departments, funding agencies, technical journals, and college professors fall back on the pervasive tendency of the twentieth-century scientist to categorize and isolate difficult scientific questions, often to the exclusion of addressing the fundamentally interdisciplinary issues. The proliferation and advancement of science in the twentieth century was, in fact, largely due to such an approach; by isolating specific and difficult problems with single-minded focus into narrowly defined scientific disciplines, great advances could once be achieved. To a large extent, however, the opportunities once possible with such a narrow focus have stagnated in many fields, though we are left with the scientific infrastructure in which that approach once flourished. To advance, we must courageously lead our research groups outside of the various neatly defined scientific domains into which this infrastructure injects us, and pursue the significant new opportunities appearing at their intersection. University departments and technical journals can and will follow suit as increasingly successful interdisciplinary efforts, such as those in the field of flow control, gain momentum. The endorsement that professional societies, technical journals, and funding agencies might bring to such interdisciplinary efforts holds the potential to significantly accelerate this reformation of the scientific infrastructure.

To promote interdisciplinary work in the scientific community at large, describing oneself as working at the intersection of disciplines X and Y (or, where they are still disjoint, the bridge between such disciplines) needs to become more commonplace. People often resort to the philosophy "I do X … oh, and I also sometimes dabble a bit with Y," but the philosophy "I do $X * Y$," where * denotes something of the nature of an integral convolution, has not been in favor since the Renaissance. Perhaps the primary reason for this is that X and Y (and Z, W, …) have gotten progressively more and more difficult. By specialization (though often to the point of isolation), we are able to "master" our more and more narrowly defined disciplines. In the experience of the author, not only is it often the case that X and Y are not immiscible, but the solution sought may often not be formulated with the ingredients of X or Y alone. To advance, the essential ingredients of X and Y must be crystallized and communicated across the artificial disciplinary boundaries. New research must then be conducted at the intersection of X and Y. To be successful in the years to come, we must prepare ourselves and our students with the training, perspective, and resolve to seize the new opportunities appearing at such intersections with a Renaissance approach.

Acknowledgments

This chapter is adapted from the article by the same author, "Flow Control: New Challenges for a New Renaissance," *Progress in Aerospace Sciences*, 37, pp. 21–58, 2001, with permission from Elsevier Science.

The author is thankful to Professor Robert Skelton, for his vision to promote integrated controls research at the intersection of disciplines, to Professors Roger Temam and Mohammed Ziane, for their patient attention to the new mathematical challenges laying the foundation for this development, and to Professors Dan Henningson, Patrick Huerre, John Kim, and Parviz Moin, for their physical insight and continued support of this sometimes unconventional effort within the fluid mechanics community. The author has also benefited from numerous technical discussions with Professors Jeff Baggett, Bassam Bamieh, Bob Bitmead, Scott Collis, Brian Farrell, Jon Freund, Petros Ioannou, and Miroslav Krstic and remains indebted to the several graduate students and postdoctoral students who are making it all happen, including Dr. Peter Blossey, Markus Högberg, Eric Lauga, and Scott Miller. We also thank the AFOSR, NSF, and DARPA for the foresight to sponsor several of these investigations. This article is dedicated to the memory of Professor Mohammed Dahleh, whose charisma and intellect endure as a continual inspiration to all who knew him.

References

Abergel, F., and Teman, R. (1990) "On Some Control Problems in Fluid Mechanics," *Theor. Comput. Fluid Dyn.* **1**, pp. 303–25.

Atkinson, G.W. (1993) *Chess and Machine Intuition*, Ablex, Norwood, NJ.

Balogh, A., Liu, W.-J., and Krstic, M. (2001) "Stability Enhancement by Boundary Control in 2D Channel Flow," *IEEE Trans. Autom. Control* (submitted).

Bamieh, B., Paganini, F., and Dahleh, M. (2002) "Distributed Control of Spatially Invariant Systems," *IEEE Trans. Autom. Control* **47**(7), pp. 1091–107.

Banks, H.T., ed. (1992) "Control and Estimation in Distributed Parameter System," in *Frontiers in Applied Mathematics*, vol. 11, SIAM.

Banks, H.T., Fabiano, R.H., and Ito, K., eds. (1993) "Identification and Control in Systems Governed by Partial Differential Equations," in *Proceedings in Applied Mathematics*, vol. 68, SIAM.

Bewley, T.R. (1999) "Linear Control and Estimation of Nonlinear Chaotic Convection: Harnessing the Butterfly Effect," *Phys. Fluids* **11**, pp. 1169–86.

Bewley, T.R., and Agarwal, R. (1996) "Optimal and Robust Control of Transition," *Proc. of the Summer Program 1996*, pp. 405–32, Center for Turbulence Research, Stanford University/NASA Ames.

Bewley, T.R., and Liu, S. (1998) "Optimal and Robust control and Estimation of Linear Paths to Transition," *J. Fluid Mech.* **365**, pp. 305–49.

Bewley, T.R., Moin, P., and Temam, R. (2001) "DNS-Based Predictive Control of Turbulence: An Optimal Benchmark for Feedback Algorithms," *J. Fluid Mech.* **447**, pp. 179–225.

Bewley, T.R., Temam, R., and Ziane, M. (2000) "A General Framework for Robust Control in Fluid Mechanics," *Physica D* **138**, pp. 360–92.

Boškovic, D.M., and Krstic, M. (2001) "Global Stabilization of a Thermal Convection Loop," *Automatica* (submitted).

Butler, K.M., and Farrell, B.F. (1992) "Three-Dimensional Optimal Perturbations in Viscous Shear Flows," *Phys. Fluids A* **4**(8), pp. 1637–50.

Chang, Y., and Collis, S.S. (1999) "Active Control of Turbulent Channel Flows Based on Large-Eddy Simulation," *Proc. of the 3rd ASME/JSME Joint Fluids Engineering Conf.*, FEDSM 99-6929, July 18–23, San Francisco, CA.

Cortelezzi, L., and Speyer, J.L. (1998) "Robust Reduced-Order Controller of Laminar Boundary Layer Transitions," *Phys. Rev. E* **58**, pp. 1906–10.

D'Andrea, R., and Dullerud, G.E. (2000) "Distributed Control of Spatially Interconnected Systems," *IEEE Trans. Autom. Control* (submitted).

Doyle, J.C., Glover, K., Khargonekar, P.P., and Francis, B.A. (1989) "State-Space Solutions to Standard and Control Problems," *IEEE Trans. Autom. Control* **34**, pp. 831–47.

Farrell, B.F., and Ioannou, P.J. (1993) "Stochastic Forcing of the Linearized Navier–Stokes Equation," *Phys. Fluids A* **5**, pp. 2600–9.

Gad-el-Hak, M. (1996) "Modern Developments in Flow Control," *Appl. Mech. Rev.* **49**, pp. 365–79.

Green, M., and Limebeer, D.J.N. (1995) *Linear Robust Control*, Prentice-Hall, Englewood Cliffs, NJ.

Gunzburger, M.D., ed. (1995) *Flow Control*, Springer-Verlag, Berlin.

Hamilton, J.M., Kim, J., and Waleffe, F. (1995) "Regeneration Mechanisms of Near-Wall Turbulence Structures," *J. Fluid Mech.* **287**, pp. 317–48.

Ho, C.-M., and Tai, Y.-C. (1996) "Review: MEMS and Its Applications for Flow Control," *ASME J. Fluid Eng.* **118**, 437–47.

Ho, C.-M., and Tai, Y.-C. (1998) "Micro-Electro-Mechanical Systems (MEMS) and Fluid Flows," *Annu. Rev. Fluid Mech.* **30**, pp. 579–612.

Högberg, M., Bewley, T.R., and Henningson, D.S. (2003) "Linear Feedback Control and Estimation of Transition in Plane Channel Flow," *J. Fluid Mech.* **481**, pp. 149–75.

Holmes, P., Lumley, J.L., and Berkooz, G. (1996) *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, Cambridge University Press, Cambridge, U.K.

Huerre, P., and Monkewitz, P.A. (1990) "Local and Global Instabilities in Spatially Developing Flows," *Annu. Rev. Fluid Mech.* **22**, pp. 473–537.

Joshi, S.S., Speyer, J.L., and Kim, J. (1997) "A Systems Theory Approach to the Feedback Stabilization of Infinitesimal and Finite-Amplitude Disturbances in Plane Poiseuille Flow," *J. Fluid Mech.* **332**, pp. 157–84.

Kim, J., and Lim, J. (2000) “A Linear Process in Wall-Bounded Turbulent Shear Flows,” *Phys. Fluids* **12**(8), pp. 1885–8.

Koumoutsakos, P. (1999) “Vorticity Flux Control for a Turbulent Channel Flow,” *Phys. Fluids* **11**(2), pp. 248–50.

Koumoutsakos, P., Freund, J., and Parekh, D. (1998) “Evolution Strategies for Parameter Optimization in Controlled Jet Flows,” *Proc. of the 1998 CTR Summer Program*, Center for Turbulence Research, Stanford University/NASA Ames.

Lagnese, J.E., Russell, D.L., and White, L.W., eds. (1995) *Control and Optimal Design of Distributed Parameter Systems*, Springer-Verlag, Berlin.

Laub, A.J. (1991) “Invariant Subspace Methods for the Numerical Solution of Riccati Equations,” in *The Riccati Equation*, Bittaini, Laub, and Willems, eds., Springer-Verlag, Berlin, pp. 163–96.

Lauga, E., and Bewley, T.R. (2000) “Robust Control of Linear Global Instability in Models of Non-Parallel Shear Flows” (under preparation).

Löfdahl, L., and Gad-el-Hak, M. (1999) “MEMS Applications in Turbulence and Flow Control,” *Prog. Aerosp. Sci.* **35**, pp. 101–203.

Lumley, J., and Blossey, P. (1998) “Control of Turbulence,” *Annu. Rev. Fluid Mech.* **30**, pp. 311–27.

McMichael, J.M. (1996) “Progress and Prospects for Active Flow Control Using Microfabricated Electro-Mechanical Systems (MEMS),” AIAA Paper 96-0306.

Newborn, M. (1997) *Kasparov Versus Deep Blue: Computer Chess Comes of Age*, Springer-Verlag, Berlin.

Nosenchuck, D.M. (1994) “Electromagnetic Turbulent Boundary-Layer Control,” *Bull. Am. Phys. Soc.* **39**, p. 1938.

Obinata, G., and Anderson, B.D.O. (2000) *Model Reduction for Control System Design* (in press).

Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. (1986) *Numerical Recipes*, Cambridge University Press, Cambridge, U.K.

Rathnasingham, R., and Breuer, K.S. (1997) “System Identification and Control of a Turbulent Boundary Layer,” *Phys. Fluids* **9**, pp. 1867–9.

Reddy, S.C., Schmid, P.J., Baggett, J.S., and Henningson, D.S. (1998) “On Stability of Streamwise Streaks and Transition Thresholds in Plane Channel Flows,” *J. Fluid Mech.* **365**, pp. 269–303.

Reuther, J., Jameson, A., Farmer, J., Martinelli, L., and Saunders, D. (1996) “Aerodynamic Shape Optimization of Complex Aircraft Configurations via an Adjoint Formulation,” AIAA Paper 96-0094.

Rose, N., ed. (1998) *Mathematical Maxims and Minims*, Rome Press, Raleigh, NC.

Simmons, G. (1992) *Calculus Gems*, McGraw-Hill, New York.

Skogestad, S., and Postlethwaite, I. (1996) *Multivariable Feedback Control*, John Wiley & Sons, New York.

Sritharan, S.S. (1998) *Optimal Control of Viscous Flows*, SIAM.

Stein, G., and Athans, M. (1987) “The LQG/LTR Procedure for Multivariable Feedback Control Design,” *IEEE Trans. Autom. Control* **32**, pp. 105–14.

Tachim Medjo, T. (2000) “Iterative Methods for Robust Control Problems” (in preparation).

Trefethen, L.N., Trefethen, A.E., Reddy, S.C., and Driscoll, T.A. (1993) “Hydrodynamic Stability without Eigenvalues,” *Science* **261**, pp. 578–84.

Vainberg, M. (1964) *Variational Methods for the Study of Nonlinear Operators*, Holden-Day, Oakland, CA.

Vanderplaats, G. (1984) *Numerical Optimization Techniques for Engineering Design*, McGraw-Hill, New York.

Vogel and Wade (1995) “Analysis of Costate Discretizations in Parameter Estimation for Linear Evolution Equations,” *SIAM J. Control Optimization* **33**, pp. 227–54.

Zhou, K., and Doyle, J.C. (1998) *Essentials of Robust Control*, Prentice-Hall, Englewood Cliffs, NJ.

Zhou, K., Doyle, J.C., and Glover, K. (1996) *Robust and Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ.

16

Soft Computing in Control

Mihir Sen and
J. William Goodwine
University of Notre Dame

16.1 Introduction

Many important applications of micro-electro-mechanical systems (MEMS) devices involve the control of complex systems, be they fluid, solid, or thermal. For example, MEMS were used for microsensors and microactuators [Subramanian et al., 1997; Nagaoka et al., 1997]. They were also used in conjunction with optimal closed-loop control to increase the critical buckling load of a pinned-end column and for structural stability [Berlin et al., 1998]. Another example entails using MEMS sensors placed on an enclosure wall to actively control noise inside the enclosure from exterior noise sources. Varadan et al. (1995) used them for the active vibration and noise control of thin plates. Vandelli et al. (1998) developed a MEMS microvalve array for fluid flow control. Nelson et al. (1998) applied control theory to the microassembly of MEMS devices.

Solving the problem of control of other complex systems, such as fluid flows and structures, using these techniques appears to be promising. Ho and Tai (1996, 1998) reviewed the applications of MEMS to flow control. Gad-el-Hak (1999) discussed the fluid mechanics of microdevices, and Löfdahl and Gad-el-Hak (1999) provided an overview of the applications of MEMS technology to turbulence and flow control. Sen and Yang (2000) reviewed applications of artificial neural networks and genetic algorithms to thermal systems.

A previous chapter outlined the basics of control theory and some of its applications. Apart from the traditional approach, another perspective can be taken towards control, that of artificial intelligence (AI). This is a body of diverse techniques that were recently developed in the computer science community to solve problems that could not be solved, or were difficult to solve, by other means. AI is often defined as using a computer to mimic how a human being would solve a given problem. The objective here is not to discuss AI but to point out that some of the techniques developed in the context of AI can be transported to applications that involve MEMS. If the latter is the hardware of the future, the former might be the software. AI encompasses a broad spectrum of computational techniques and methods which are based on heuristics rather than algorithms. Thus, they are not guaranteed to work, but have a high probability of success. AI imitates nature as a general characteristic, though the difference between computers and nature is so vast that the analogy is far from perfect.

There are some specific techniques within AI, collectively known as soft computing (SC), that have matured to the point of being computationally useful for complex engineering problems. SC includes artificial neural networks, genetic algorithms, fuzzy logic, and related techniques in probabilistic reasoning. SC techniques are especially useful for complex systems. For purposes of this discussion, complex systems are defined as those that can be broken into a number of subsystems. These subsystems are individually simple and may be analytically or numerically computed but together cannot be analyzed in real time for control purposes. The physical phenomena behind these complex systems might not be known, or the system might not be possible to model mathematically. Often the model equations, as in the case of turbulent fluid flow, are too many or too difficult to permit analytical solutions or rapid numerical computations.

The objective of this chapter is to describe the basic SC techniques that can be applied to control problems relevant to MEMS. The following sections describe the artificial neural network, genetic algorithms, and fuzzy logic methodologies. The descriptions are introductory so that readers can decide whether the technique is useful for their own application. SC techniques are tools and, as such, work much better in some circumstances than in others. Caution must always be used. Some of the applications that are reported in the literature give an idea of the kind of problems that can be approached.

Several excellent books and texts include information on the general subject of SC. Aminzadeh and Jamshidi (1994), Yager and Zadeh (1994), Bouchon-Meunier et al. (1995), and Jang et al. (1997) cover broad aspects of SC. Schalkoff (1997) and Haykin (1999) deal with artificial neural networks. Fogel (1999) provides an outline of evolutionary programming. Goldberg (1989) presents an exposition on genetic algorithms, and Mordeson and Nair (1988) introduce the topic of fuzzy logic. Books covering more specific areas include those on the application of SC to robotic systems by Jain and Fukuda (1998) and those on neuro-fuzzy systems by Buckley and Feuring (1999) and Pal and Mitra (1999).

16.2 Artificial Neural Networks

One of the most common SC-based techniques is an artificial neural networks (ANN). Excellent introductory texts including those by Schalkoff (1997) and Haykin (1999), entail the history and mathematical background of ANN. The technique has been applied to diverse fields such as philosophy, psychology, business and economics, and science and engineering. What all these applications have in common is complexity, for which the ANN is particularly suitable.

16.2.1 Background

Inspiration for the ANN comes from the study of biological neurons in humans and other animals. These neurons learn from experience and are also able to handle and store information that is not precise [Eeckman, 1992]. Each neuron in a biological network of interconnecting neurons receives input signals from other neurons and, if the accumulation of inputs exceeds a certain threshold, puts out a signal that is sent to other neurons to which it is connected. The decision to fire or not represents the ability of the ANN to learn and store information. In spite of the analogy between the biological and computational neurons, there are significant differences that must be remembered. Though the biological processes in a neuron are slower, the connections are massively parallel as compared to its computational analogue, which is limited by the speed of the currently available hardware.

Artificial neural networks are designed to mimic the biological behavior of natural neurons. Each artificial neuron (or node) in this network has connections (or synapses) with other neurons and has an input and output characteristic function. An ANN is composed of a number of artificial neurons. Each interneural connection is associated with a certain weight, and each neuron is associated with a certain bias. Training, which is also called the learning process, is central to the use of an ANN. Training uses existing data to find a suitable set of weights and biases.

Many different ANN structures and configurations have been proposed as well as various training methodologies [Warwick et al., 1992]. The configuration we discuss in some detail is the multilayer ANN operating in the feedforward mode, and we will use the backpropagation algorithm for training. This combination has been useful for many engineering and control purposes [Zeng, 1998]. The ANN, once trained, works as an input–output system with multiple inputs and outputs. Learning is accomplished by adjusting the weights and biases so that the training data are reproduced.

16.2.2 Feedforward ANN

Figure 16.1 shows a feedforward ANN consisting of a series of layers, each with a number of neurons. The first and last layers are for input and output, respectively, while the ones in between are the hidden layers. The ANN is said to be fully connected when any neuron in a given layer is connected to all the neurons in the adjacent layers.

Though notation in this subject is not standard, we will use the following. The *j*th neuron in the *i*th layer will be written (i, j). The input of the neuron (i, j), is x_{ij}, its output is y_{ij}, its bias is θ_{ij}, and $w^{i,j}_{i-1,k}$ is the synaptic weight between neurons $(i-1, k)$ and (i, j). A number of parameters determine the configuration: I is the total number of layers, and J_i is the number of neurons in the *i*th layer. There are J_1 input values and J_1 output values to the ANN.

Each neuron processes the information between its input and output. The input of a neuron is the sum of all the outputs from the previous neurons modified by the respective internodal synaptic weights and a bias at the neuron. Thus, the relation between the output of the neurons $(i-1, k)$ for $k = 1, \ldots, J_{i-1}$ in one layer and the input of a neuron (i, j) in the following layer is:

$$x_{i,j} = \theta_{i,j} + \sum_{k=1}^{J_{i-1}} w^{i,j}_{i-1,k}\, y_{i-1,k} \tag{16.1}$$

The input and output of the neuron (i, j) are related by:

$$y_{i,j} = \phi_{i,j}(x_{i,j}) \quad \text{for } i > 1 \tag{16.2}$$

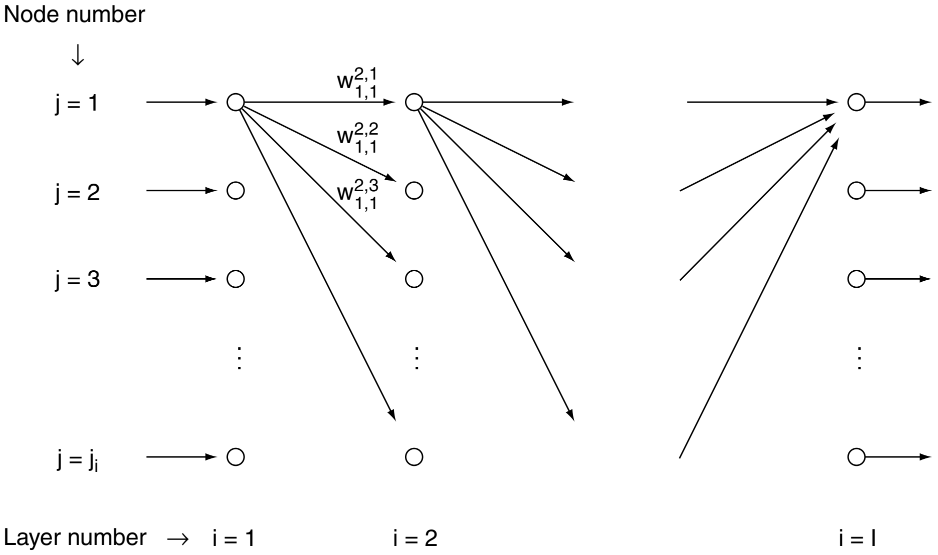


FIGURE 16.1 Schematic of feedforward neural network.

and

$$y_{i,j} = x_{i,j} \quad \text{for } i = 1 \tag{16.3}$$

The function $\phi_{i,j}(x)$, called the activation function, plays a central role in the processing of information by the ANN. When the input signal is small, the neuron suppresses the signal altogether, resulting in a small output. When the input exceeds a certain threshold, the neuron fires and sends a signal to all the neurons in the next layer. Several appropriate activation functions have been used, the most popular being the logistic sigmoid function:

$$\phi_{i,j}(x) = \frac{1}{1 + \exp(-x/c)} \tag{16.4}$$

where c is a parameter that determines the steepness of the function. This function has several useful characteristics because it is an approximation to the step function that simulates the operation of firing and not firing but with continuous derivatives, and because its output always lies between 0 and 1.

16.2.3 Training

Once the configuration of an ANN is fixed, the weights between the neurons and the bias at each neuron define its input–output characteristics. The weights determine the relative importance of each one of the signals received by a neuron from those of the previous layer, and the bias is the propensity for the combined input to trigger a response from the neuron. The training process is the adjustment of the weights and biases to reproduce a known set of provided input–output values.

Though there are many methods in use, the backpropagation technique is a widely used deterministic training algorithm for this type of ANN [Rumelhart et al., 1986]. This method is based on the minimization of an error function by the method of steepest descent. Descriptions of this algorithm exist in many recent texts on ANN (for instance, Rzempoluck [1998]), and only a brief outline is given here. Initial values are assigned to the weights and biases and, for a given input to the ANN, the output is determined. The synaptic weights and biases are iteratively modified until the output values differ little from the target outputs.

In the backpropagation method, an error $\delta_{I,j}$ is quantified for the last layer by:

$$\delta_{I,j} = (y^T_{I,j} - y_{I,j})y_{I,j}(1 - y_{I,j}) \tag{16.5}$$

where $y^T_{I,j}$ is the target output for the jth neuron of the last layer. This equation comes from a finite-difference approximation of the derivative of the sigmoid function. After all the $\delta_{I,j}$ have been calculated, the computation moves back to the layer $I - 1$. There are no target outputs for this layer, so the value,

$$\delta_{I-1,k} = y_{I-1,k}(1 - y_{I-1,k}) \sum_{j=1}^{J_I} \delta_{I,j}\, w^{I,j}_{I-1,k} \tag{16.6}$$

is used instead. A similar procedure is used for all the inner layers until layer 2 is reached. After all these errors have been determined, changes in the weights and biases are calculated from:

$$\Delta w^{I,j}_{i-1,k} = \lambda \delta_{i,j} y_{i-1,k} \tag{16.7}$$

$$\Delta \theta_{i,j} = \lambda \delta_{i,j}$$

for $i < I$, where λ is the learning rate. From this the new weights and biases are determined.

In one cycle of training, a new set of synaptic weights and biases is determined for all training data after which the error defined by:

$$E = \frac{1}{2} \sum_{j=1}^{J_I} (y^T_{I,j} - y_{I,j})^2 \tag{16.8}$$

is calculated. The error of the ANN at the end of each cycle can be based on a maximum or averaged value for the output errors. The process is repeated over many cycles with the weights and biases being continuously updated throughout the training runs and cycles. The training is terminated when the error is low enough to be satisfactory in some pre-determined sense.

16.2.4 Implementation Issues

Several choices must be made to construct a suitable ANN for a given problem, and these choices are fairly important to achieve good results. There is no general theoretical basis for these choices, and experience combined with trial and error are the best guides.

16.2.4.1 Configuration

The first choice that must be made in using an ANN is its configuration (i.e., the number of layers and the number of neurons in each layer). Though the accuracy of prediction sometimes becomes better (at other times, it picks up additional noise) as the number of layers and neurons becomes larger, the number of cycles to achieve this accuracy also increases. It is possible to do some optimization by beginning with one hidden layer as a starting point and then adding more neurons and layers while checking the prediction error [Flood and Kartam, 1994]. Practical considerations dictate a compromise between accuracy and computational speed. Many users prefer only one hidden layer, and it is unusual to go beyond two or three hidden layers.

There are other suggestions for choosing the parameters of the ANN. Karmin (1990) used a relatively large ANN reduced in size by removing neurons that do not significantly affect the results. In the so-called radial-Gaussian system, hidden neurons are added to the ANN in a systematic way during the training process [Gagarin et al., 1994]. It is also possible to use evolutionary programming to optimize the ANN configuration [Angeline et al., 1994]. Some authors, for example, Thibault and Grandjean (1991), present studies of the effect of varying these parameters.

16.2.4.2 Normalization

The data that the ANN handles are usually dimensional and thus have to be normalized. Furthermore, the slope of the sigmoid function $\phi_{i,j}(x)$ used as the activation function becomes smaller as $x \to \pm\infty$. To use the central part of the function, it is desirable to normalize all physical variables. In other words, the range between the minimum and maximum values in the training data is linearly mapped into a restricted range such as [0.15, 0.85] or [0.1, 0.9]. The exact choice is somewhat arbitrary, and the operation of the ANN is not very sensitive to these values.

16.2.4.3 Learning Rate

The learning rate λ is another parameter that must be arbitrarily assumed. If the learning rate is large, the changes in the weights and biases in each step will be large. The ANN will learn quickly, but the training process could be oscillatory or unstable. Small learning rates lead to a longer training period to achieve the same accuracy. The learning rate is usually around 0.4 and is determined by trial and error. Other possibilities also exist [Kamarthi et al., 1992].

16.2.4.4 Initial Values

Initial values of weights and biases are assigned at the beginning of the training process. Both the final values reached after training and the number of cycles needed to reach a reasonable convergence depend on these initial values. One method is assigning the values in a random fashion, though Wessels and Barnard (1992), Drago and Ridella (1992), and Lehtokangas et al. (1995) have suggested other methods for determining the initial assignment. Sometimes, the ANN is trained or upgraded on new data for which the old values can be used as the initial weights and biases.

16.2.4.5 Training Cutoff

Training is repeated until a certain criterion is reached. A simple criterion is a fixed number of cycles. It is also common to specify the minimum in the error-number of cycles curve as the end of training. This has a possible pitfall in that there may be a local minimum, beyond which the error may decrease some more.

16.2.5 Neurocontrol

The ANN described up to now is a static input–output system; that is, given an input vector, the ANN is able to predict an output. For purposes of real-time control, the procedure must be extended to variables that are changing in time. This means that time t must also be a variable for both training and predictions. There are two ways in which this can be done: either the time t or a time step Δt between predictions can be additional inputs to the ANN. The latter procedure, which is convenient for microprocessor applications, has an advantage because the initial values that are not really relevant to a control system after a long time quickly become irrelevant. The time step Δt in this procedure may be constant or may vary according to the needs of the prediction as time goes on.

The ANN can be trained as before. The trained ANN predicts values of variables at an instant $t + \Delta t$ if the values at t are given. The dynamic simulation can be introduced into a prediction-based controller to control the behavior of a system.

16.2.6 Heat Exchanger Application

The previous sections discussed the general procedures and methodology of ANNs. In this section, we apply the method to the specific problem of heat exchangers. A heat exchanger is an example of a system that is complex. Though the physical phenomena are well known, in the face of turbulence, secondary flows, developing flows, complicated geometry, property variations, conduction along walls, etc., it becomes impossible to compute the desired operating variables for prediction (Pacheco-Vega, 2001). Computing in real time for control purposes is even more difficult. ANNs can be used for this purpose. Much of the work reported here is in Díaz (2000).

The heat exchanger used in tests is schematically shown in Figure 16.2. The experiments were carried out in a variable-speed, open wind-tunnel facility [Zhao, 1995]. Hot water flows inside the tubes of the heat exchanger, and room air is drawn over the outside of the tubes. The flow rates of air and water and the temperatures of the two fluids going in and out are measured.

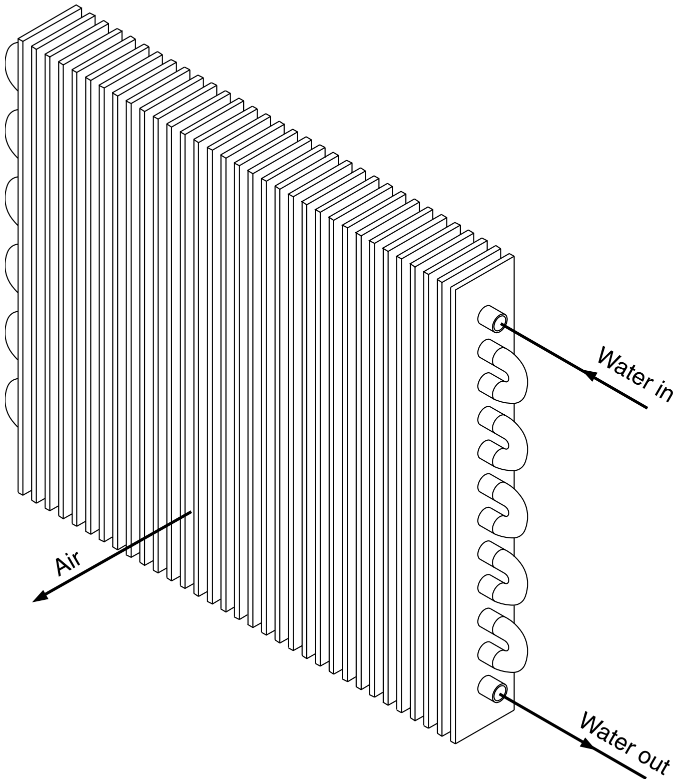


FIGURE 16.2 Schematic of compact heat exchanger.

16.2.6.1 Steady State

For a given heat exchanger, the heat transfer rate $\dot{Q}$ under steady-state conditions depends on the flow rates of air and water, $\dot{m}_a$ and $\dot{m}_w$, respectively, and their inlet temperatures, T_a^{in} and T_w^{in}, respectively. From the heat transfer rate, secondary quantities such as the fluid outlet temperatures, T_a^{out} and T_w^{out}, respectively, are determined. For the present experiments, a total of 259 runs were made, of which data for only 197 runs were used for training, while the rest were used for testing the predictions. For each test run the six quantities, $\dot{Q}$, $\dot{m}_a$, $\dot{m}_w$, T_a^{in}, T_w^{in}, T_a^{out} and T_w^{out} were measured or determined from measurements. The work here is described in detail in Díaz et al. (1999).

From the data an ANN was trained. This network had four inputs, $\dot{m}_a$, $\dot{m}_w$, T_a^{in} and T_w^{in}, and a single output. Many different configurations and numbers of training cycles were tried. Good results were found for a 4–5–5–1 (i.e., four layers with 4, 5, 5, and 1 neurons in each layer, respectively) configuration trained for 200,000 cycles. The trained ANN was tested on the dataset provided for the purpose. Figure 16.3 shows the results of the ANN prediction,$\dot{Q}^p_{\text{ANN}}$, plotted against the actual measurement, $\dot{Q}^e$. The 45° line that is shown is an exact prediction; the dotted lines represent errors of ±10%. Figure 16.3 also shows the prediction of a power-law correlation for the same data, $\dot{Q}^p_{\text{cor}}$ [Zhao, 1995]. The ANN does a better job than the correlation.

16.2.6.2 Thermal Neurocontrol

The same heat exchanger was used to develop the neurocontrol methodology. Dynamic data were obtained by varying the water inlet temperatures by changing the heater settings while keeping the other variables constant. Training data were obtained from experiments in which the water inlet temperature was varied in small increments of 5.56°C from 32.2°C up to 65.6°C. Díaz et al. (2001a) provides further details. A nonlinear system may be controlled in many different ways. The method chosen for neurocontrol testing is shown in Figure 16.4. An inverse-ANN controller, *C*, controls the heat exchanger, while a forward ANN, *M*, models the plant, *P*. The controller is trained as a dynamic ANN. The desired control objective was to keep the outlet air temperature T_a^{out} constant.

In the first test, the system was subjected to a step change in the set temperature. The system was stabilized around $T_a^{\text{out}} = 32°\text{C}$, after which the set temperature was suddenly changed to 36°C. Figure 16.5 shows how the neurocontroller behaved as compared to conventional PID (proportional–integral–derivative) and PI (proportional–integral) controllers. All the controllers work properly, but the neurocontroller has

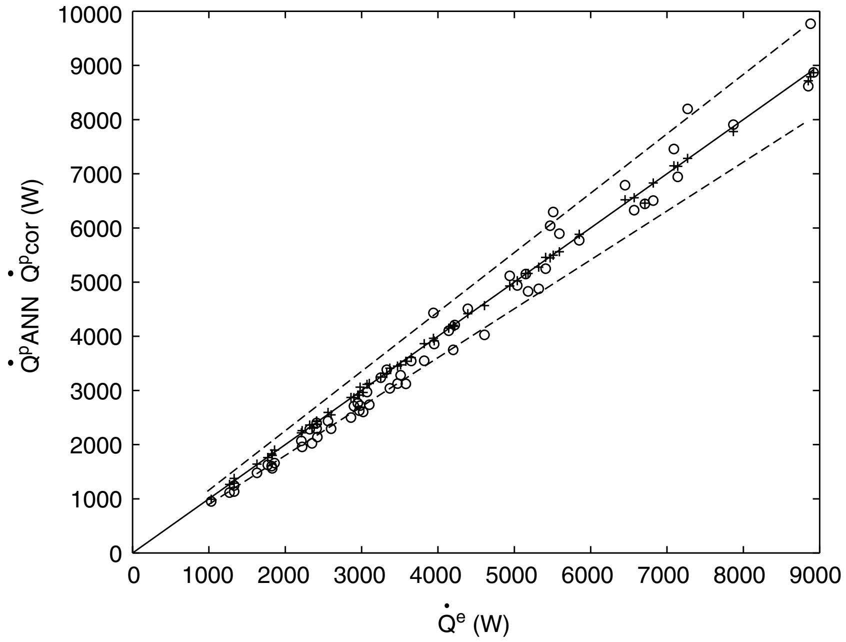


FIGURE 16.3 Predictions of 4–5–5–1 neural network (+) and correlation (○).

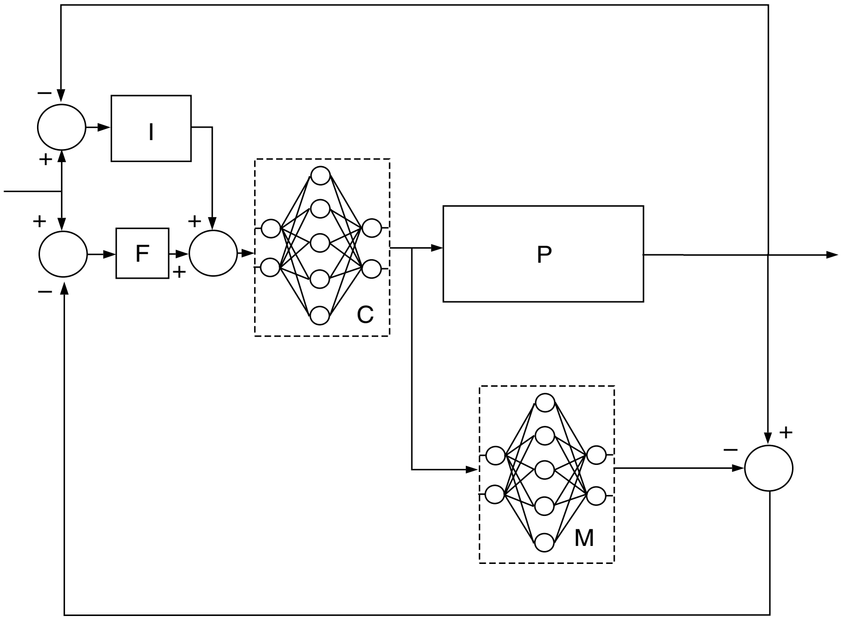


FIGURE 16.4 Closed-loop neurocontrol (P = plant, M = ANN model of plant, C = neurocontroller, F = filter, I = integral controller).

fewer oscillations. In a second test, we looked at the disturbance rejection ability of the controller. At steady operation, the water flow is completely shut down for a short interval between $t = 40$ s and $t = 70$ s. Figure 16.6 shows how the system variables respond to PID and neurocontrol during and after the disturbance pulse. In the neurocontroller, oscillations are quickly damped out.

The procedure previously outlined in which the ANN simulates the plant to be controlled is fairly straightforward. Other special aspects should be examined further.

16.2.6.2.1 Stabilization of Feedback Loop

(See [Díaz et al., 2001b].) The static ANN, once incorporated into the feedback loop, may lead to a dynamical system that is unstable. In order to avoid this possibility, the ANN has to be trained not only to make accurate predictions but also to give dynamical stability to the loop. Because the weights are not unique, it is possible to come up with an algorithm that does both. The stabilization algorithm was designed and tested on the heat exchanger facility. Figure 16.7 shows the behavior of two controllers C_1 and C_2, where the former is a stable controller and the latter is unstable. In each case, the air flow rate is being controlled, and the air outlet temperature is shown as a function of time. The stable controller works well, while in the unstable case the air flow rate is increased as far as possible without achieving the desired result.

16.2.6.2.2 Adaptive Neurocontrol

(See [Díaz et al., 2001c].) The major advantage of neurocontrollers is that they can be adaptive. The ANN can go through a process of retraining if its predictions are not accurate due to change in system characteristics. The adaptive controller is tested in two different ways. The first is a disturbance on the water side, shown in Figure 16.8; v_a in the figure is the air velocity. Initially, the neurocontroller keeps the system close to $T_a^{\text{out}} = 34$°C. From $t = 100$ s to $t = 130$ s, the water is shut off. The neurocontroller tries to control until $t = 110$ s, at which point it hands off to a backup PID controller while it adapts to the new circumstances. Then, once it is able to make reasonable predictions, it resumes control at $t = 170$ s. The second test is a sudden reduction of inlet area of the wind tunnel test facility, shown in Figure 16.9. The controller keeps T_a^{out} at 34°C, and then half of the inlet is suddenly blocked at $t = 150$ s. The neurocontroller adapts until it learns the behavior of the new system, and finally at $t = 240$ s it resumes control of the system.

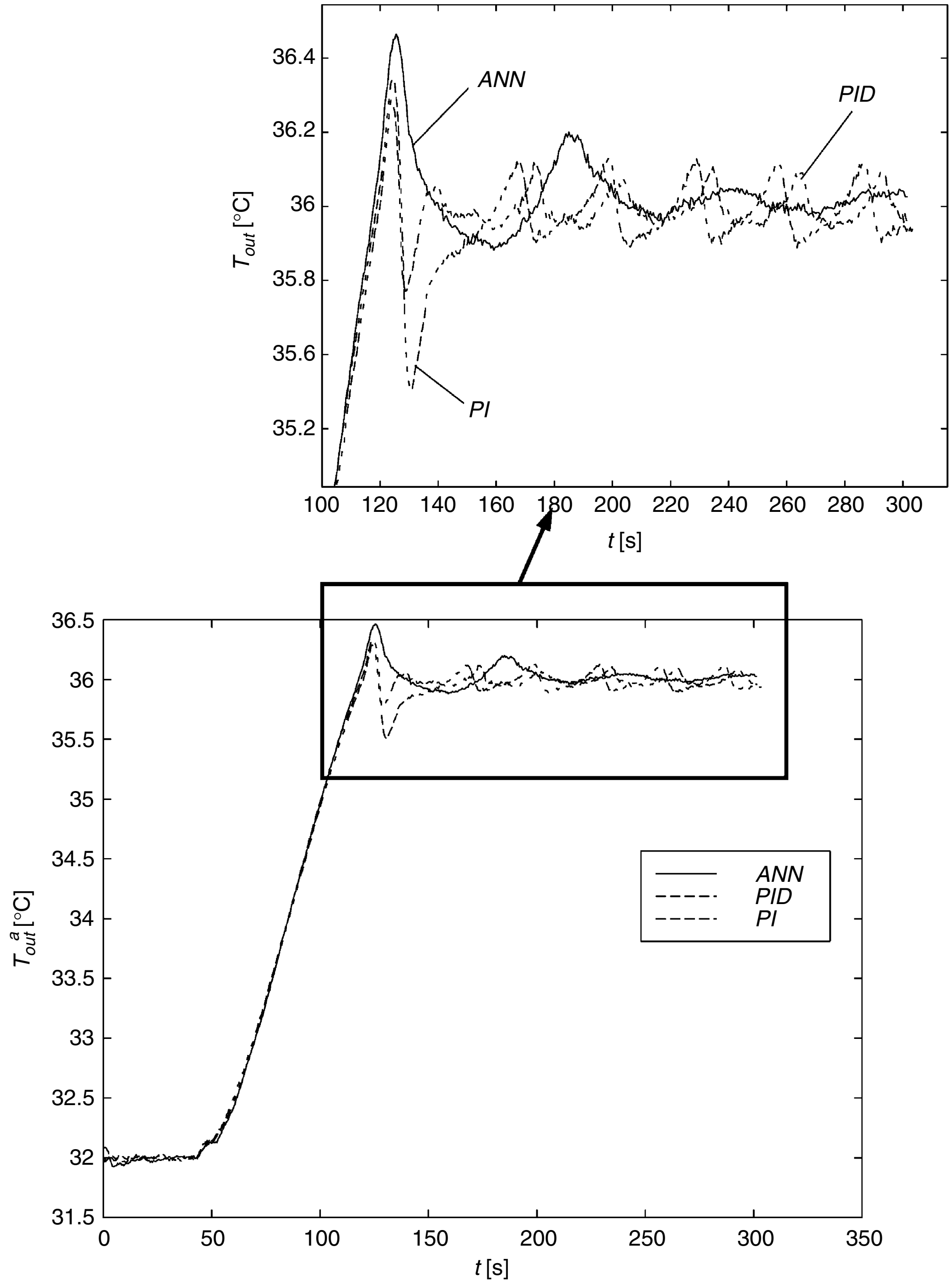


FIGURE 16.5 Response to change in set point.

16.2.6.2.3 Optimal Control

For systems where there is more than one variable to be controlled, it is often required that they be controlled under an optimizing constraint. For thermal systems, for example, it may be possible to require that the system use the least energy at the new setpoint. Figure 16.10 shows the neurocontroller being used in this mode with the water and air flow rates being the control variables. In this case, the ANN has been provided information about the energy consumption of the following components of the facility: the hydraulic pump, the fan, and the electric heater. We first let the controller stabilize T_a^{out} at 34°C. Around $t = 130$ s, the energy minimization routine turns on, and the controller adjusts both the water and air flow rates to minimize energy usage while keeping T_a^{out} roughly at its set value.

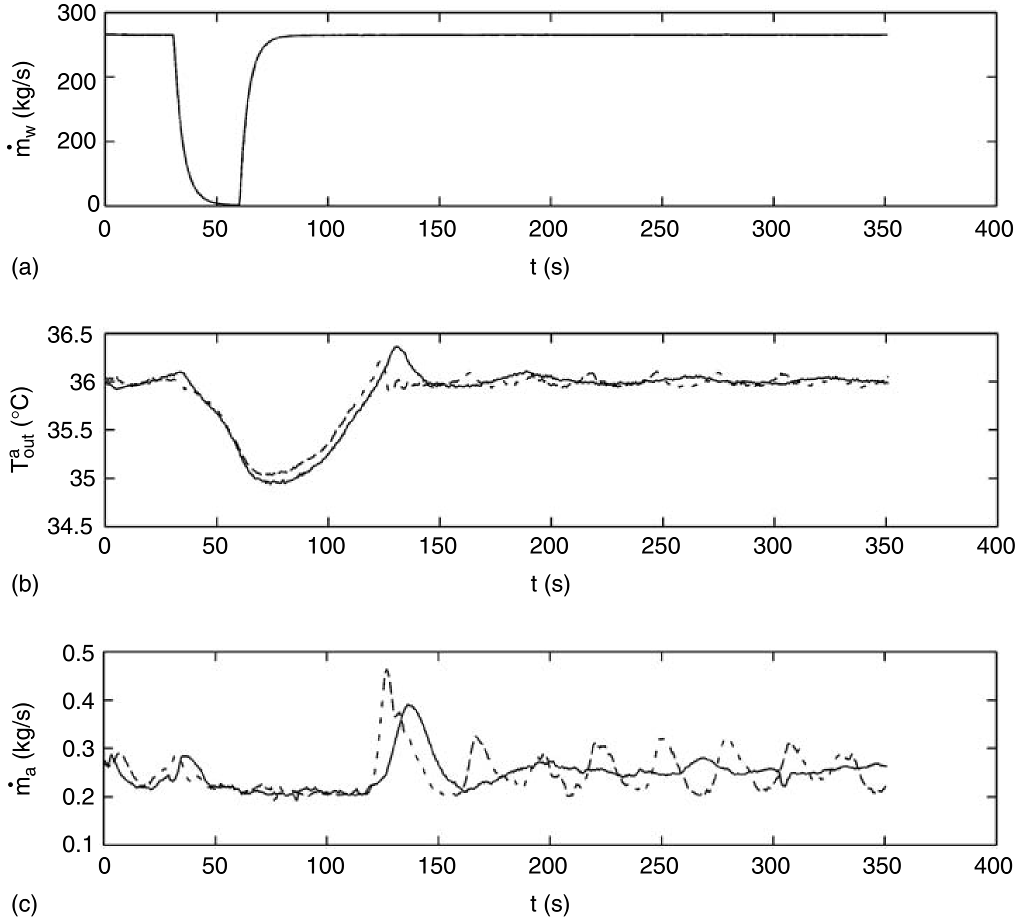


FIGURE 16.6 Disturbance rejection (continuous line is neural network, broken line is PID).

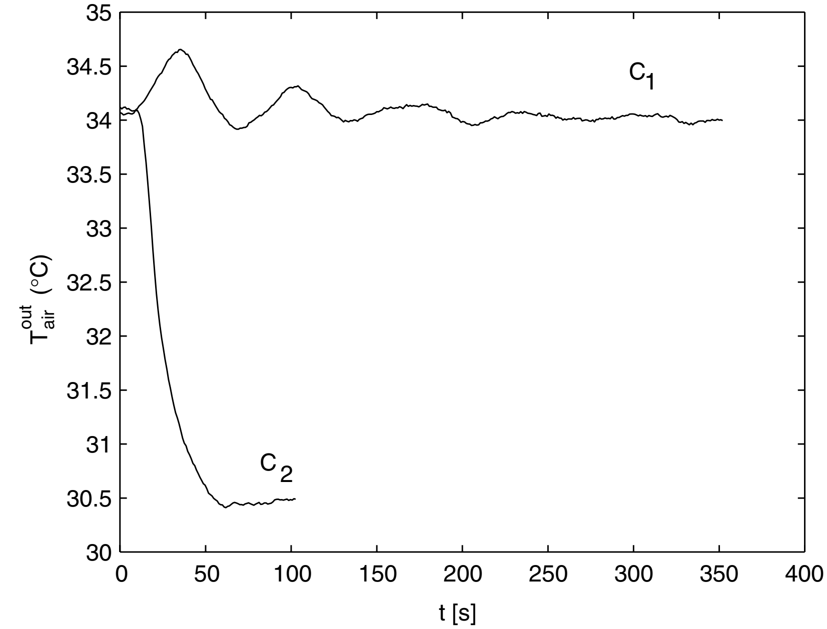


FIGURE 16.7 Performance of stable and unstable neurocontrollers.

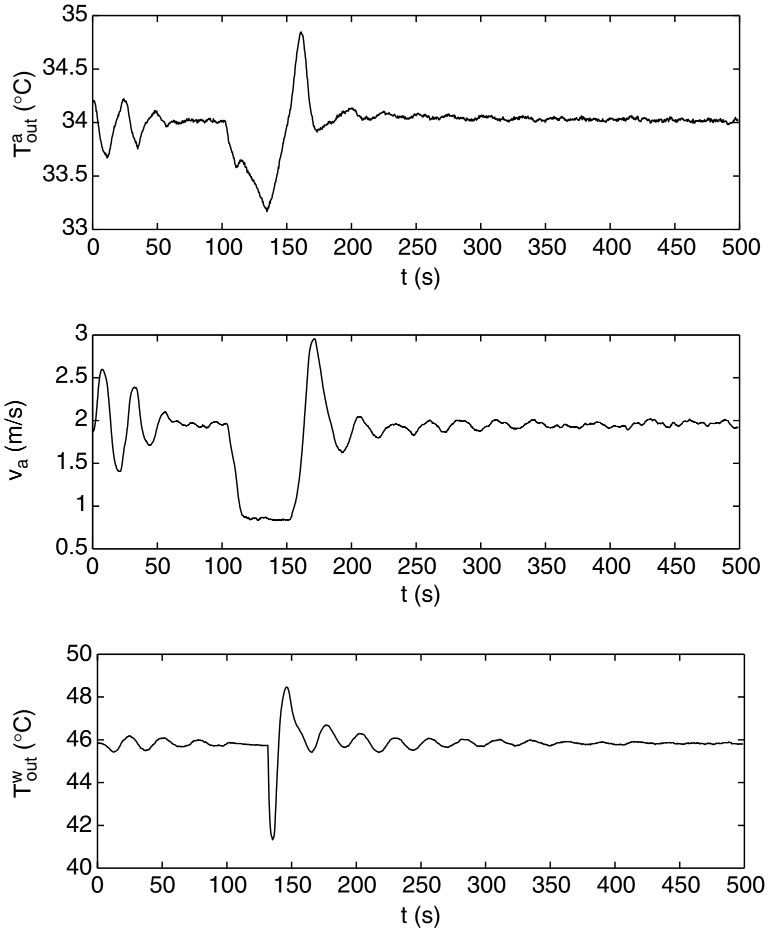


FIGURE 16.8 Response of adaptive neurocontroller to pulsed stoppage in water flow.

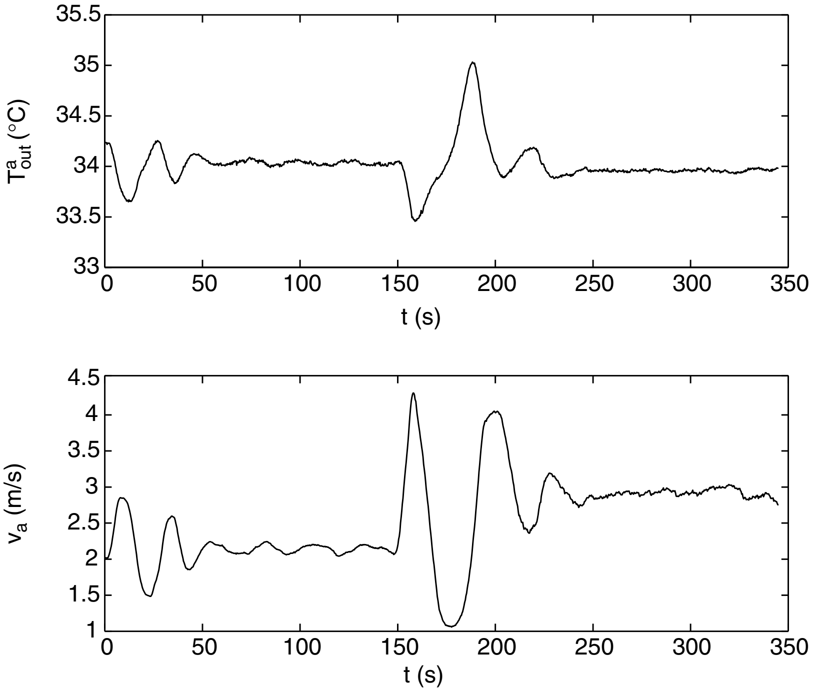


FIGURE 16.9 Response of adaptive neurocontroller to sudden change in wind-tunnel inlet area.

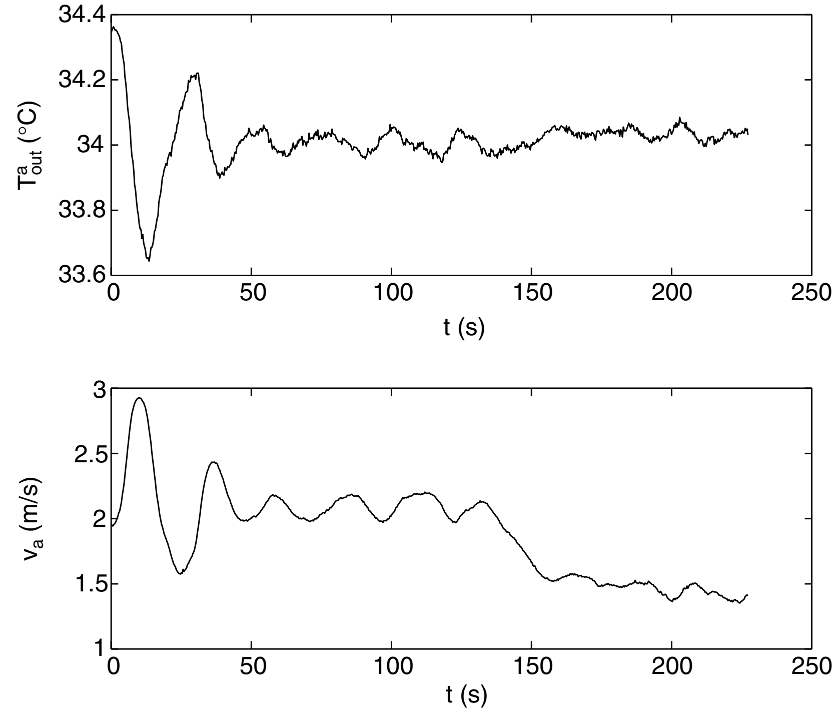


FIGURE 16.10 Application of energy minimization procedure.

16.2.7 Other Applications

Many other applications of neurocontrol have been reported in the literature, including inflow-related problems. Gad-el-Hak (1994) gave an overview of the problem of flow control in turbulent boundary layers. Control was attempted by Jacobson and Reynolds (1993). Lee et al. (1997) used ANNs for turbulence control for drag reduction purposes. Suzuki and Kasagi (1997) were able to use ANNs for the optimal control of vortex shedding behind a square cylinder, minimizing the angular motion of the cylinder at the same time. Chan and Rad (2000) used ANNs for the purpose of real-time flow control.

16.2.8 Concluding Remarks

The implementation of ANN procedures in a complex problem is straightforward. It relieves the necessity of having a first-principles model that may not be available for relatively new devices, such as those based on MEMS. Predictions can be made and systems can be controlled on the basis of available information without the need for models. The ANN is also extremely adaptable to changing circumstances.

16.3 Genetic Algorithms

Evolutionary algorithms change, or evolve, as they do their work. The genetic algorithm (GA) is a specific type of search technique based on the Darwinian evolutionary principles of natural selection to attain its objective of optimization. Optimization is fundamental to many applications in engineering, including the design of systems. Although analysis permits the prediction of the behavior of a given system, optimization is the technique that searches among all possible designs of the system to find the one that is the best for the application. Optimization is also intimately related to the control problem where parameters have to be chosen. The configuration of neural networks has to be selected, and the constants of PID control have to be set. The importance of this problem has given rise to a wide variety of techniques that help search for the optimum. In this context, local and global optima must be distinguished. If one visualizes a multivariable function with many peaks, any one of these is a local maximum, while only one is the highest.

Genetic algorithms are described in monographs by Goldberg (1989), Michalewicz (1992), Mitchell (1997), and Man et al. (1999). GAs are generally used for the purpose of global optimization. They are not gradient-based and are an alternative to other global optimization techniques such as simulated annealing. Because local gradient information is not used, a GA usually finds the global optimum as opposed to a local one, a characteristic that is often useful. The fact that gradients are not used may be significant in problems in which the variables to be optimized are functions of discrete quantities and their derivatives are not possible.

Robustness, a central characteristic for the survival of natural species, is also a feature of genetic algorithms. Goldberg (1989) compared the genetic algorithm from this perspective with other search techniques. Methods based on the calculus, such as equating to zero the derivatives of a function to obtain the extrema, are indirect. In a direct method, the iteration moves in a direction determined by the local gradient of the function. Both these methods are local and depend on the local existence of derivatives. Another class of methods is based on evaluation of a function at every point on a fine but finite grid. For most practical applications this approach is too time consuming. In yet another class of techniques, randomness is used. Simulated annealing and genetic algorithms are examples of these techniques. They search from a population of points rather than from a single point. They use only the function rather than its derivative, and they use probabilistic rather than deterministic rules. The search using many points makes the method global rather than local.

16.3.1 Procedure

There are many variants of the genetic algorithm procedure. A simple approach is described here, but it is not the only one. We can illustrate the GA procedure by finding the maximum of a function $f(x)$ within a given domain $x \in U \subset \Re^n$. In a gradient-based method, the slope of the tangent plane at a given value of x indicates which way to "go up" within an iterative procedure. The GA does not work this way. For simplicity, in the following we will assume that $U = [a, b] \subset \Re$, though the method can be easily generalized to higher dimensions. Furthermore, we will map the interval $[a, b]$ to $[0, 2^c - 1]$. This way, the independent variable x can be represented by a binary string of length c running from 000…000 up to 111…111. The example function chosen for maximization is $f(x) = x(1 - x)$.

Step 1: We begin by randomly selecting r candidate numbers within the desired domain (i.e., $x_1, \dots, x_r$). This is the first generation. An example is shown in Table 16.1, where we have taken $r = 10$ and $c = 6$. The second column shows the numbers in decimal form, the third column the same numbers in binary form, and the fourth the normalized binary version of the same. The function $f(x_i)$ for each number which, in the context of this algorithm, is called the fitness is indicated in the fifth column. The reason for this name is that, the higher the fitness is, the closer x_i is to its value where $f(x_i)$ is a maximum. In other words, we seek the value of x for which $f(x)$ is the fittest. The maximum fitness of the members of this generation is 0.2469 for $x_i = 0.4444$. The last column is the normalized fitness $s(x_i)$, the values of the previous column divided by the sum of the fitnesses; thus, $s(x_i) = f(x)/\Sigma_f(x_i)$.

Step 2: From the first generation of numbers, pairs of parents are chosen which then give rise to offspring that form the next generation. To visualize the process we draw a pie chart, shown in Figure 16.11, with slices that have angles proportional to the normalized fitness $s(x_i)$. To form parents, pairs of numbers are randomly selected from the chart as if it were a roulette wheel. Of course, the numbers with larger normalized fitnesses have the higher probability of being selected. The result of such a selection process is shown in the second column of Table 16.2 as generation $G = 1/4$. These are then shuffled to produce column $G = 1/2$. The first two entries in this column are a set of parents, the next two another set, and so on.

Step 3: Each pair of parents produces a pair of offspring by crossover, which can be done in different ways. We have randomly chosen a point along the two binary strings that form the parents and have

TABLE 16.1 Binary, Decimal, and Normalized Forms of the First Generation of Candidate Numbers and Their Absolute and Normalized Fitnesses.

i	$x_i(d)$	$x_i(b)$	$x_i(d, n)$	$f(x_i)(d)$	$s(x_i)(d)$
1	18	010010	0.2857	0.2041	0.1096
2	53	110101	0.8413	0.1335	0.0717
3	43	101011	0.6825	0.2167	0.1164
4	11	001011	0.1746	0.1441	0.0774
5	22	010110	0.3492	0.2273	0.1221
6	46	101110	0.7302	0.1970	0.1058
7	28	011100	0.4444	0.2469	0.1326
8	42	101010	0.6667	0.2222	0.1194
9	25	011001	0.3968	0.2394	0.1286
10	61	111101	0.9683	0.0307	0.0165

Note: b = binary, d = decimal, n = normalized.

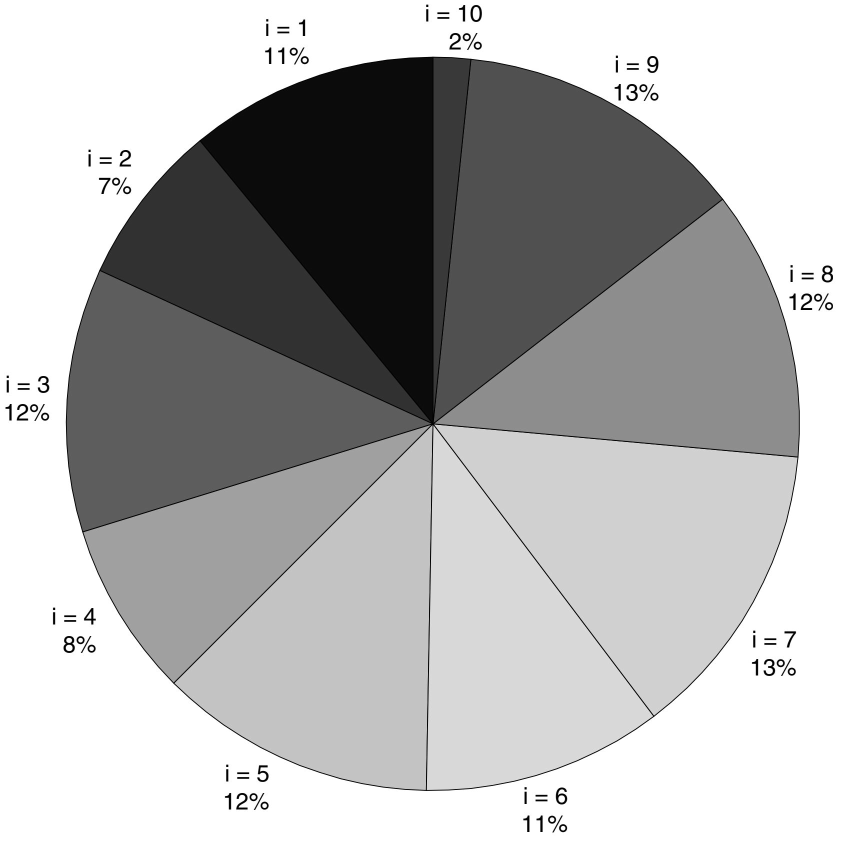


FIGURE 16.11 "Old" fitness.

interchanged the part of the string beyond this point. This is a single-point crossover. An example is shown in Figure 16.12, where the crossover point is in the middle of the string and the numbers 011100 and 011001 produce the offspring 011001 and 011100. The crossover point in the other pairs might be different. The final result of crossover between the parents is column $G = 3/4$.

Step 4: The column $G = 1$ is obtained from $G = 3/4$ by mutation. This is obtained by selecting a randomly chosen bit and changing it from 0 to 1, or vice versa. Figure 16.12 shows that one bit in the fourth number changed. This procedure gives a new generation with members that are generally fitter (i.e., give

TABLE 16.2 From One Generation to the Next.

$G = 0$	$G = 1/4$	$G = 1/2$	$G = 3/4$	$G = 1$
010010	011100	011100	011001	011001
110101	010110	011001	011100	011100
101011	011001	010110	010110	010110
001011	101010	010110	010110	010010
010110	010010	101010	101001	101001
101110	101011	011001	011010	011010
011100	011100	010010	010010	010010
101010	010110	101010	101010	101010
011001	011001	101011	101010	101010
111101	010110	011100	011101	011101

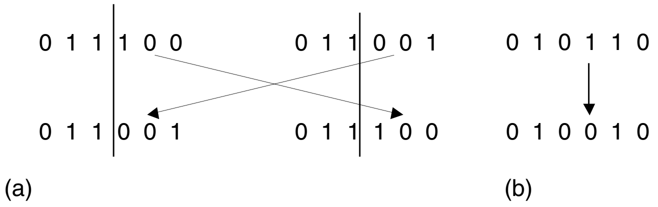


FIGURE 16.12 (a) Crossover and (b) mutation in a genetic algorithm.

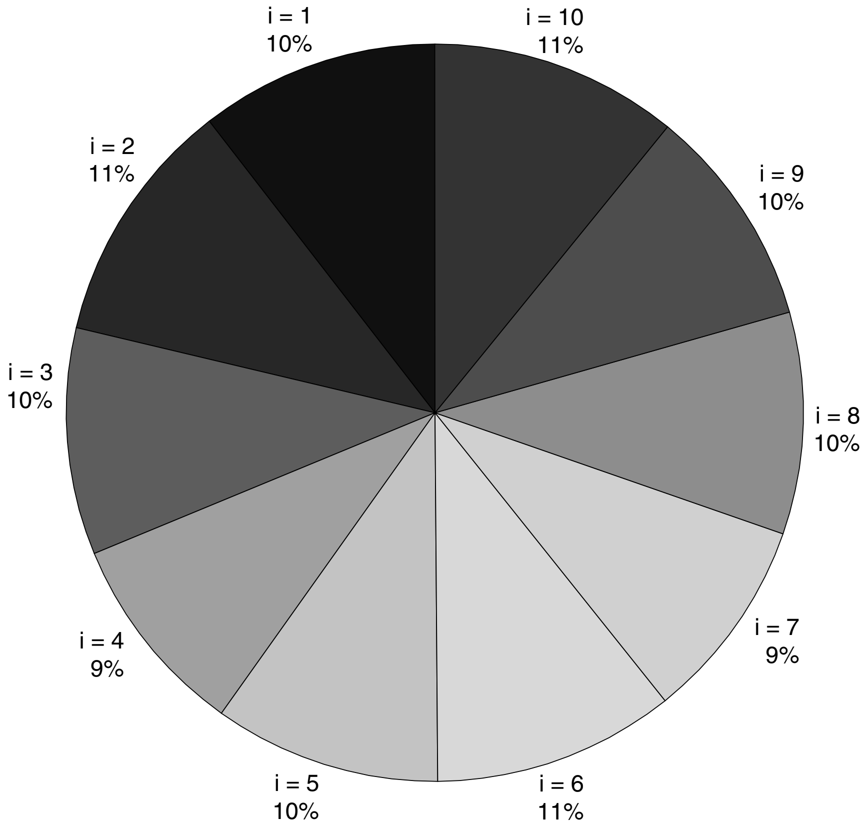


FIGURE 16.13 "New" fitness.

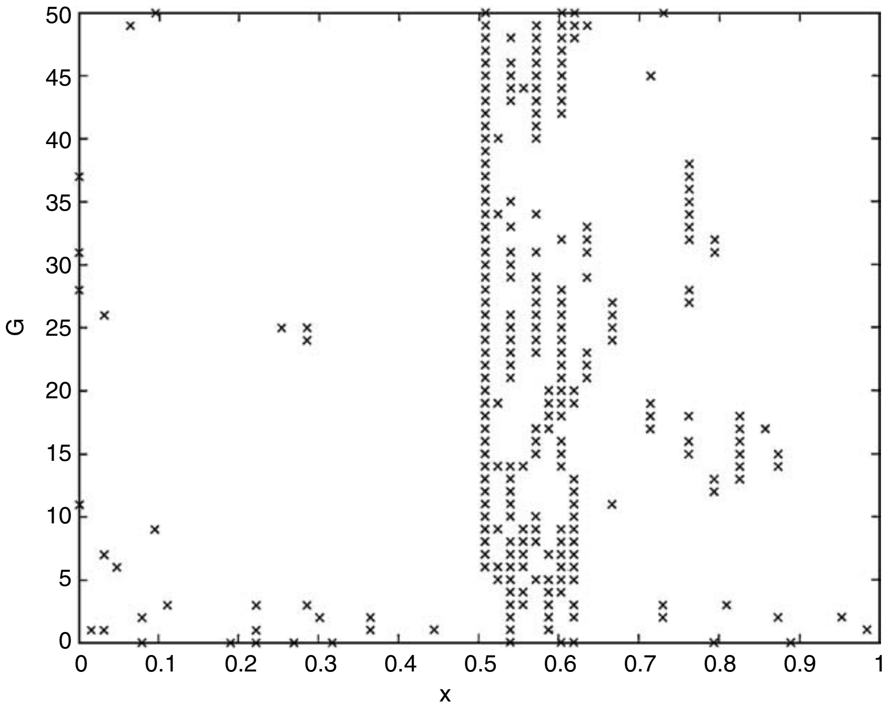


FIGURE 16.14 Change in population as a function of generation number.

a value of the function that is closer to the maximum). The new pie-chart of the normalized fitness, Figure 16.13, shows much more uniformity in fitnesses since the "unfit" have disappeared. The maximum fitness of this generation turns out to be $f(x_i) = 0.2484$ at $x_i = 0.4603$ compared to the value of $f(x_i) = 0.2469$ at $x_i = 0.4444$ for the previous generation. The process of finding a new generation is repeated several times until some criterion is satisfied. This criterion could be one of the following: a desired number of cycles have been completed, the maximum fitness of the generation does not change much, or the value of x at which this maximum fitness is obtained does not change significantly.

Some programming details need to be taken care of in actual implementation. Parameters such as the number of members of a generation and the length of a binary string must be decided. In addition, a finite probability for crossover and mutation must be prescribed. The crossover does the bulk of the work in selecting the new generation. Although the probability of mutation is generally kept small, it is vital because mutation enables a possible solution to break out of the neighborhood of a local optimum and go somewhere else that may turn out to have a better local optimum. It is also common to keep the best of each generation in the next to make sure that the fitness of the generation is nondecreasing. The algorithm itself is probabilistic so that it cannot be guaranteed to find the global optimum. In fact, every time it is run, the exact results obtained will be different.

Figure 16.14 shows the result of running a GA code (written by A. Pacheco-Vega) to continue the process indicated in Tables 16.1 and 16.2. The probability of crossover is taken to be unity, and that of mutation is 0.03. The abscissa shows the distribution of the population $x_1, \ldots, x_r$ at a generation number G indicated in the ordinate. The initial population at $G = 0$ is indicated by 10 different crosses (the exact values are different from those in Tables 16.1 and 16.2 because of the generation of different random numbers). In the following generations, some of the crosses overlap as numbers may repeat themselves within a population. The code has been terminated after 50 generations. A certain crowding of the population around the correct value $x = 0.5$ is observed as well as the presence of values relatively far from it. This is a consequence of the global nature of the search and is a characteristic of the GA. In addition, even at $G = 50$ the value of x_i that gives the highest value of the function is close to the correct value but is not exact.

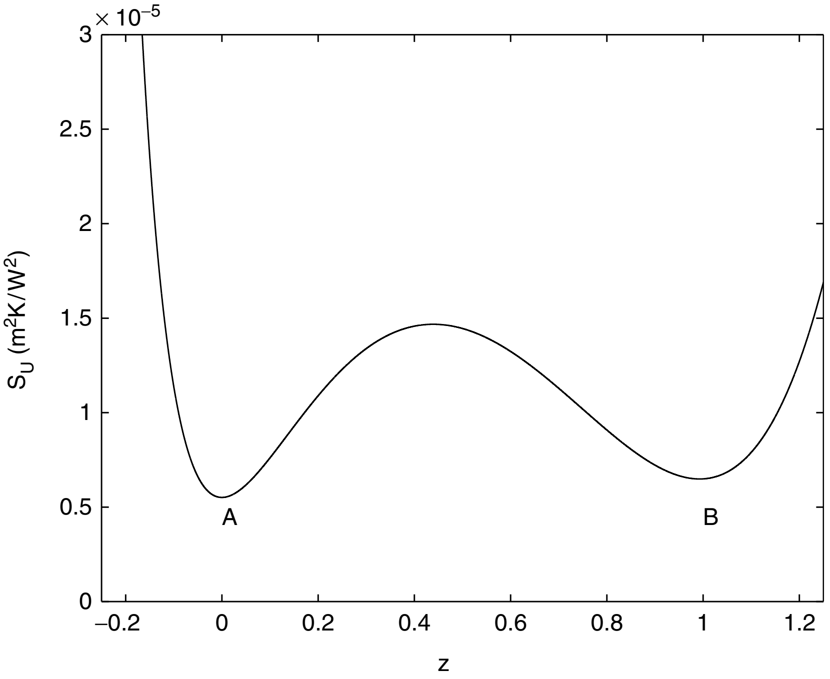


FIGURE 16.15 Global vs. local minima in optimization problem.

16.3.2 Heat Exchanger Application

This SC technique is applied to the heat exchanger described before. The optimization problem here is to find the best correlation that fits experimental data. A set of $N = 214$ experimental runs provided the database. In each case, the heat rate $\dot{Q}$ is found as a function of the two flow rates m_w and m_a as well as the two inlet fluid temperatures I_a^{in} and I_w^{in}. Details are in Pacheco-Vega et al. (1998).

There are two resistances to the flow of heat by convection: on the inside with water and on the outside with air. The conventional way of handling data is determining correlations for the inner and outer heat transfer coefficients. For example, power-law relations of the form $Nu = aRe^n$ between the Nusselt and Reynolds numbers, *Nu* and *Re*, respectively, on both sides of the tube wall are often assumed. There are then four constants to determine: a_1, a_2, n_1, and n_2. One possible procedure is to minimize the root mean square (rms) error S_U (a_1, a_2, n_1, n_2) in total thermal resistance to heat transfer between prediction and data in the least-square sense. The total resistance is the sum of the air-side and water-side resistances.

This procedure leads to a large number of local minima due to the nonlinearity of the function to be minimized. Figure 16.15 shows a pair of such minima. In the figure, a section of the error surface S_U (a_1, a_2, n_1, n_2) that passes through two local minima *A* and *B* is shown. The coordinate *z* is a linear combination of a_1, a_2, n_1, and n_2 such that it is zero at *A* and unity at *B*, and the ordinate is the rms error. The values S_U of the two correlations obtained at *A* and at *B* are very similar, and the heat rate predictions for the resulting correlations are also almost equally accurate. However, a_1, a_2, n_1, n_2, and the predictions of the thermal resistances on either side are very different. This shows the importance of using global minimization techniques for nonlinear regression analysis. If the *GA* is used to find the global minimum, the point *A* is the global minimum. The correlation (not shown) found as a result of the global search is the best that fits the assumed power laws and is closest to the experimental data.

16.3.3 Other Applications

Many other applications of GAs to optimization and control problems include optimization of a control scheme by Seywald et al. (1995), Michalewicz et al. (1992), Perhinschi (1998), and Tang et al. (1996b). Reis et al. (1997) and Kao (1999) have used the GA to find the optimal location of control valves in a piping network. Gaudenzi et al. (1998) optimized the control of a beam using the technique. Several workers have applied the method to the motion of robots [Nakashima et al., 1998; Nordin et al., 1998]. Katisikas et al. (1995) and Tang et al. (1996a) used the genetic algorithm for active noise control. Nagaya and Ryu (1996) controlled the shape of a flexible beam using a shape memory alloy, and Keane (1995) optimized the geometry of structures for vibration control. Dimeo and Lee (1995) controlled a boiler and turbine using the genetic algorithm. Sharatchandra et al. (1998) used the GA for shape optimization of a micropump. Kaboudan (1999) used genetic algorithms for time-series prediction. Luk et al. (1999) developed a GA-based fuzzy logic control of a solar power plant using distributed collector fields. Additional applications of GAs combined with other SC techniques have been used for optimization of the control process [Matsuura et al., 1995; Trebi-Ollennu and White, 1997; Rahmoun and Benmohamed, 1998; Ranganath et al., 1999; Lin and Lee, 1999].

16.3.4 Final Remarks

There are two main advantages when using a genetic or evolutionary approach to optimization. One is that the methods seek the global optimum. The other advantage is that they can be used in discrete systems, in which derivatives do not exist or are meaningless. Examples of this are piping networks and positioning of electronic components. As with all tools, the reader must evaluate the advantages and disadvantages in terms of specific applications.

16.4 Fuzzy Logic and Fuzzy Control

16.4.1 Introduction

Fuzzy sets and fuzzy logic date back to Lotfi Zadeh's [Zadeh, 1965, 1968a, 1968b, 1971] work concerning complex systems. Fuzzy sets and fuzzy logic have been present in controls applications since the late 1970s [Mamdani, 1974; Mamdani and Assilian, 1975; Mamdani and Baaklini, 1975]. Fuzzy logic and its application to feedback control is comprised of two components. First, fuzzy logic is not model based so it can be applied to systems for which developing analytical models, either from first principles or from some identification techniques, is impractical or expensive. Second, it provides a convenient mechanism for application to feedback control of human (or expert) intuition regarding how a system should be controlled. This section outlines basic fuzzy set definitions, fuzzy logic concepts, and their primary application to control systems. First, an illustrative controls application of fuzzy logic is presented in complete detail. The example is followed by a more complete exposition of the mathematics of fuzzy logic intended to provide the reader with a complete set of tools with which to approach a fuzzy control problem.

16.4.2 Example Implementation of Fuzzy Control

This section first introduces a typical structure of fuzzy controllers by presenting an example of a common fuzzy control application — namely, to stabilize the inverted pendulum system illustrated in Figure 16.16 where the control input is a force of magnitude *u*. In this problem, only the pendulum angle is stabilized. This is accomplished via linguistic variables and fuzzy if–then rules such as:

1. If the pendulum angle is zero and the angular velocity is zero, then the control force should be zero.
2. If the pendulum angle is positive and small and the angular velocity is zero, then the control force should be positive and small.
3. If the pendulum angle is positive and large and the angular velocity is zero, then the control force should be positive and large.

TABLE 16.3 Fuzzy Logic Rules to Determine Control Force

	Angular Velocity				
Error	Negative Large (1)	Negative Small (2)	Zero (3)	Positive Small (4)	Positive Large (5)
(1) Negative large	Negative large	Negative large	Negative large	Negative small	Zero
(2) Negative small	Negative large	Negative large	Negative small	Zero	Positive small
(3) Zero	Negative large	Negative small	Zero	Positive small	Positive large
(4) Positive small	Negative small	Zero	Positive small	Positive large	Positive large
(5) Positive large	Zero	Positive small	Positive large	Positive large	Positive large

4. If the pendulum angle is positive and small and the pendulum angular velocity is negative and small, then the control force should be zero.

The linguistic variables are the angle error and the angular velocity. These rules are better expressed in tabular form in Table 16.3. The first enumerated rule is expressed in the third column and third row of the table. The second rule is in the third column and fourth row. The third rule is in the third column and fifth row. The fourth rule is in the second column and fourth row. These rules were determined by intuition. For example, whether the second column and second row should be "negative small" or "negative large" is determined by experience, guesswork, or tuning.

The next basic element of the fuzzy controller is the fuzzy set, which basically encapsulates the notion of to what degree the angle is "zero," "negative small," etc. Figure 16.17 illustrates the fuzzy sets that define the fuzzy state of the angle of the pendulum system. In the figure, if the pendulum angle is −7.5°, then the degree of membership in the "negative small" fuzzy set is 0.5, and the degree of membership in the "zero" fuzzy set is also 0.5. The degree of membership in the other fuzzy sets is 0. Figures 16.18 and 16.19 illustrate similar fuzzy sets that are defined for the angular velocity and the control force, respectively.

Figure 16.20 illustrates the overall control structure. First, a sensor measures the state (θ, $\dot{\theta}$). Second, the state is "fuzzified" by computing the degree of membership of the state in each of the fuzzy sets, A_i, used in the if–then rules. Third, the if–then rules in the rule base are evaluated in parallel, and the output of each rule is the fuzzy set (control force), which has the shape of the fuzzy set associated with the output of the if–then rule but is "capped" or "cut off" at the degree of membership of the state in the associated

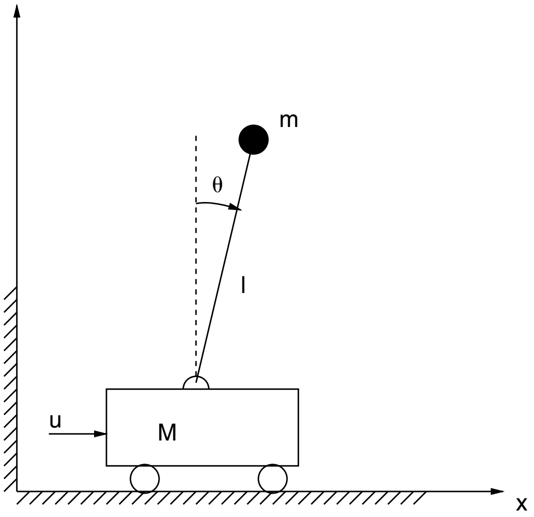


FIGURE 16.16 Pendulum system.

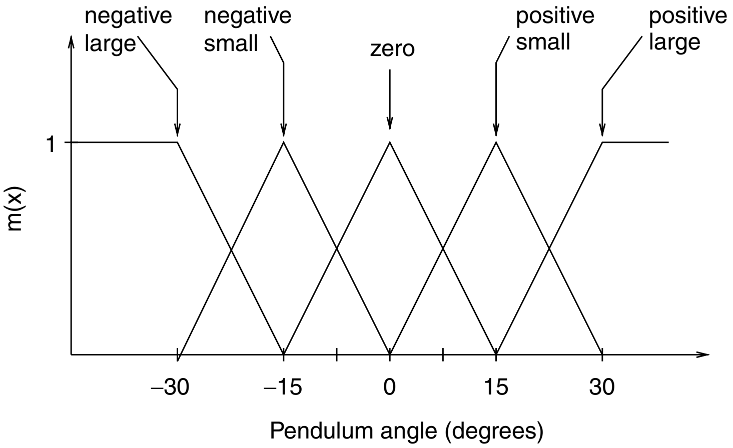


FIGURE 16.17 Pendulum angle fuzzy set.

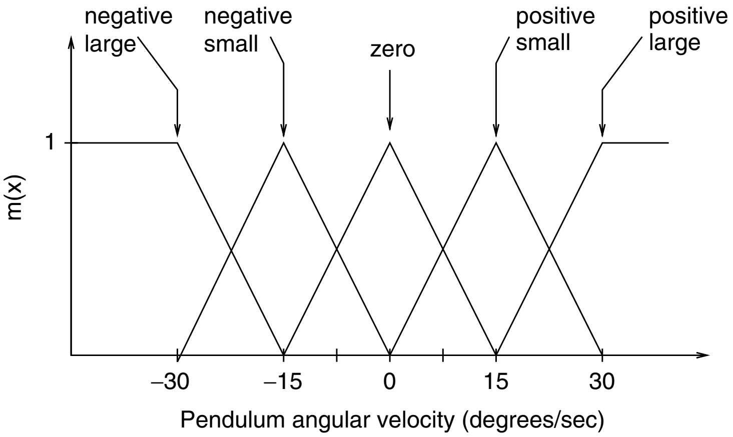


FIGURE 16.18 Pendulum velocity fuzzy set.

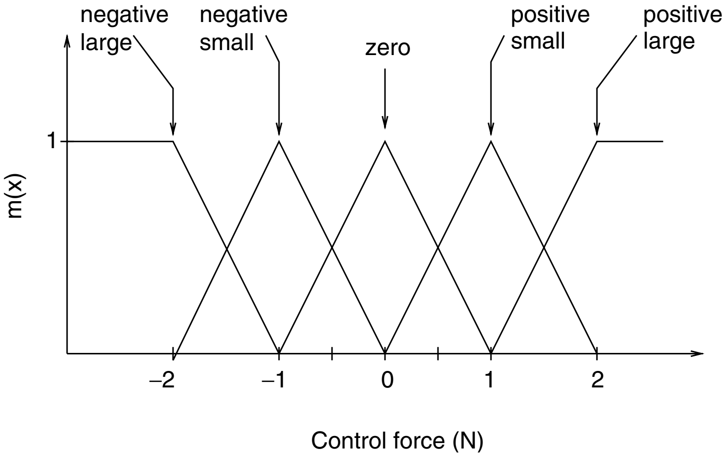


FIGURE 16.19 Pendulum force fuzzy set.

fuzzy set. If there is a logical operation, such as "and" in the antecedent (the "if" part) of the rule, then the minimum of the degree of membership in each of the fuzzy sets is used.

As a concrete example of this "fuzzy inference," consider the case where the pendulum angle is −20° and the angular velocity is +22.5°/s. The fuzzy state of the angle of the system is determined according to

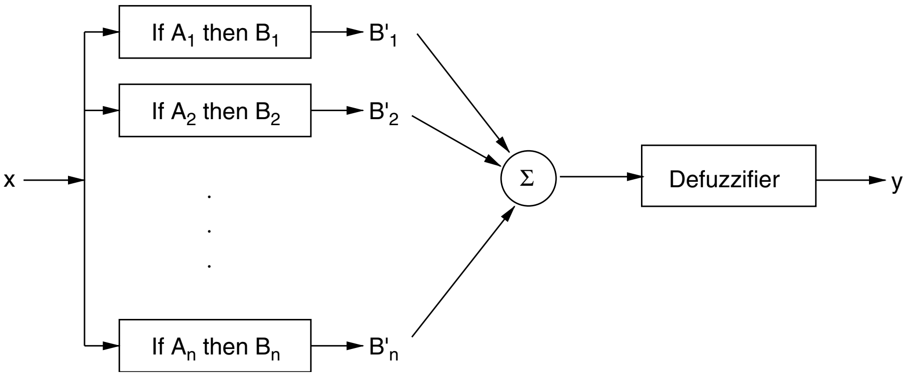


FIGURE 16.20 Fuzzy control structure.

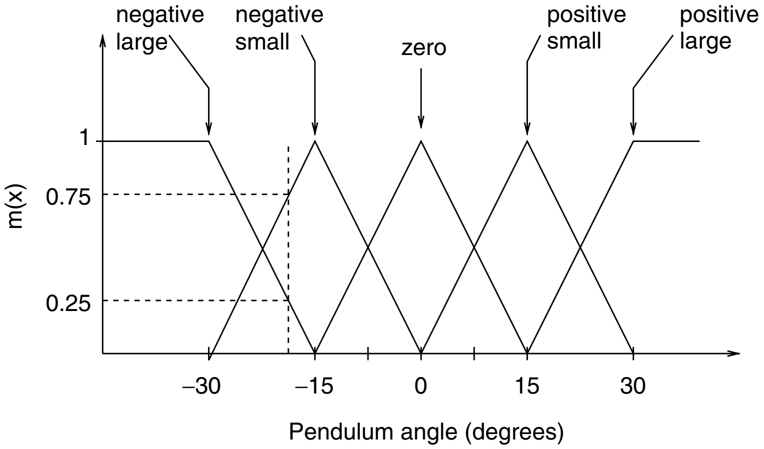


FIGURE 16.21 Fuzzification of pendulum angle.

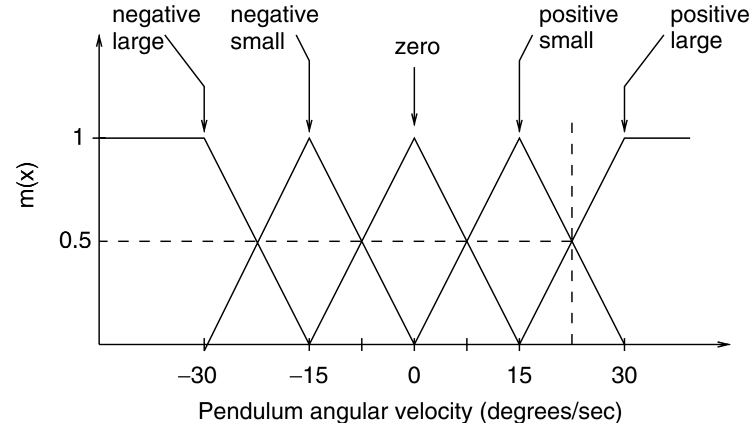


FIGURE 16.22 Fuzzification of pendulum angular velocity.

Figure 16.21, where the state of the system is represented by a 0.25 degree of membership in the "negative large" fuzzy set, and a 0.75 degree of membership in the "negative small" fuzzy set. Figure 16.22 shows the velocity is characterized by a 0.5 degree of membership in both the "positive large" fuzzy set and the "positive small" fuzzy set.

Now, the output of each rule will be the corresponding force fuzzy set, but modified so that its maximum value is capped to be the minimum degree of membership of the two elements of the antecedent

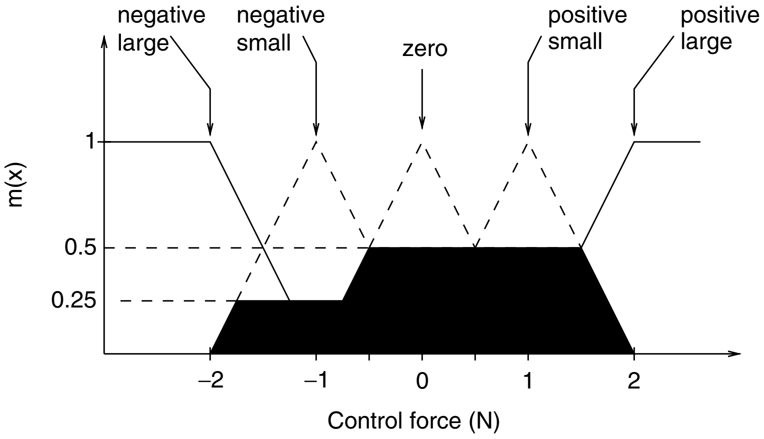


FIGURE 16.23 Aggregation of fuzzy output sets.

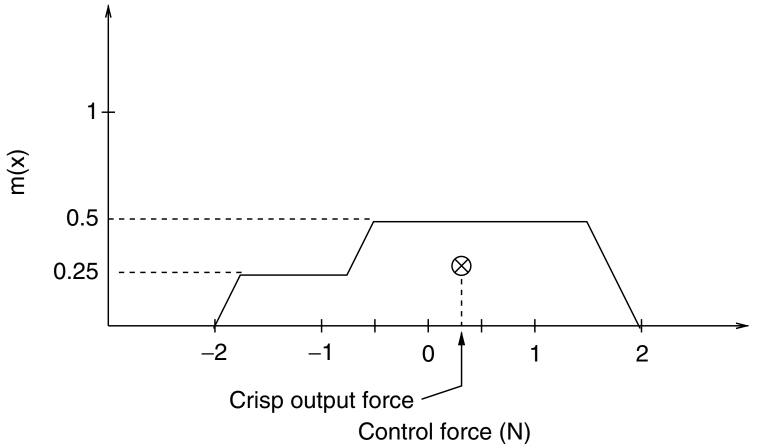


FIGURE 16.24 Defuzzification of output by computing centroid.

part of each rule. In particular, only four of the rules listed in the table will evaluate to nonzero values — namely, the top two rows in the last two columns of Table 16.3. Considering the "negative large" position and "positive small" velocity first, the "negative small" force output will be capped at 0.25, which is the degree of membership in the "negative large" position fuzzy set which is less than the 0.5 membership of the angular velocity in the "positive small" fuzzy set. In the "negative large" position and "positive large" velocity, the output will again be capped at 0.25, as similarly, it is less than the 0.5 membership of the angular velocity in the "positive large" fuzzy set. In the cases of "negative small" position and "positive small" velocity, as well as "negative small" position and "positive large" velocity, the output of the "zero" and "positive small" output force fuzzy sets will both be capped at 0.5. Once the outputs from each if–then rule are computed, they are aggregated into one large fuzzy set. In this aggregation, if two of the fuzzy outputs overlap, then (opposite to the "and" combination for the fuzzy rules) the maximum of the two sets is taken. Returning to the example, Figure 16.23 illustrates the aggregation of the four rules for the angle of $-20°$ and angular velocity of $+22.5°/s$. "Defuzzification" is necessary to have a crisp output force, and Figure 16.24 demonstrates a common technique to compute the value of the crisp output as the centroid of the aggregated fuzzy output set.

Simulating such a system is straightforward using Matlab. If the pendulum mass is 0.1 kg, the cart mass 2.0 kg, the length of the pendulum 0.5 m, and the values of the membership functions are as illustrated in Figure 16.25, the response of the cart and pendulum system is illustrated in Figures 16.26 and 16.27. Figure 16.26 illustrates the response of the pendulum angle, and Figure 16.27 illustrates the velocity of the pendulum. Figure 16.28 illustrates the control effort. Because the cart position was not controlled, its steady-state response is actually a constant, nonzero velocity. Figure 16.29 illustrates the "response surface"

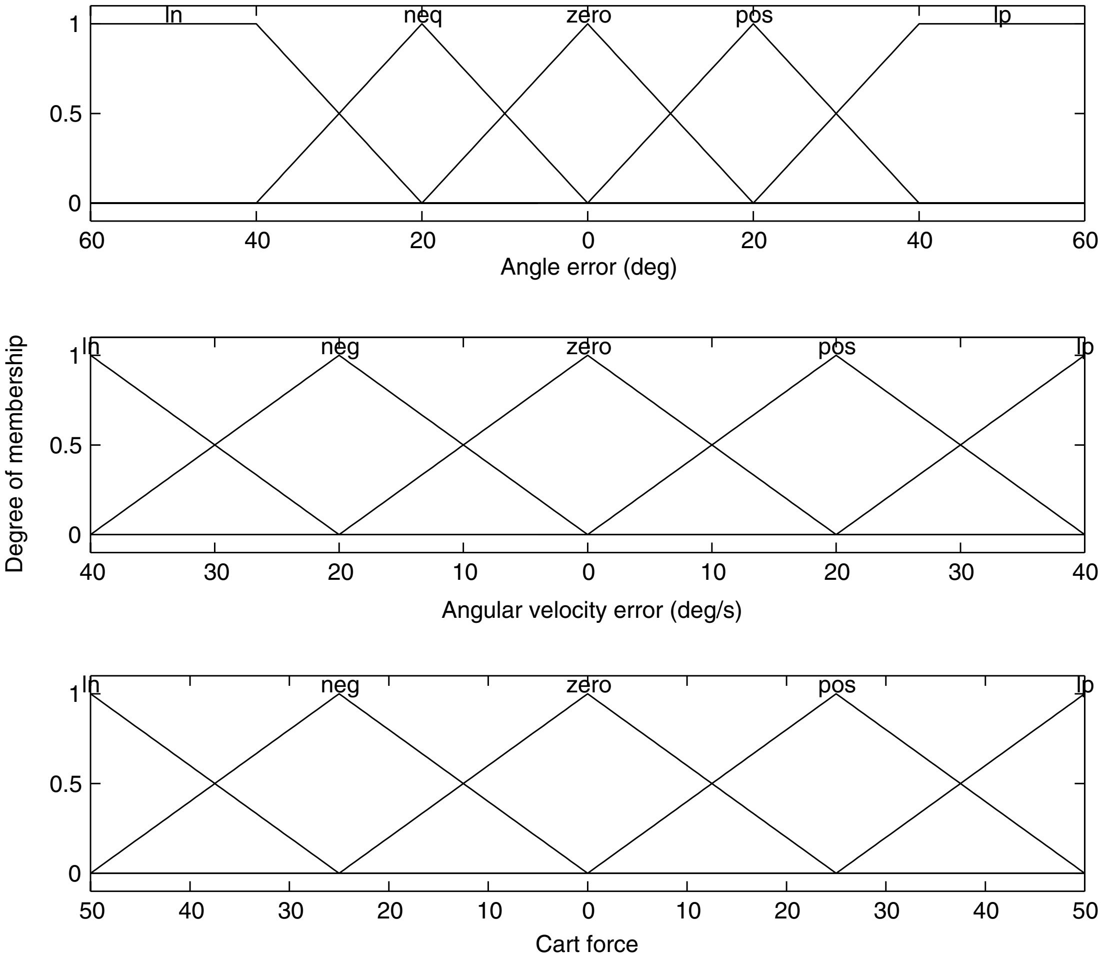


FIGURE 16.25 Membership functions for cart and pendulum simulation.

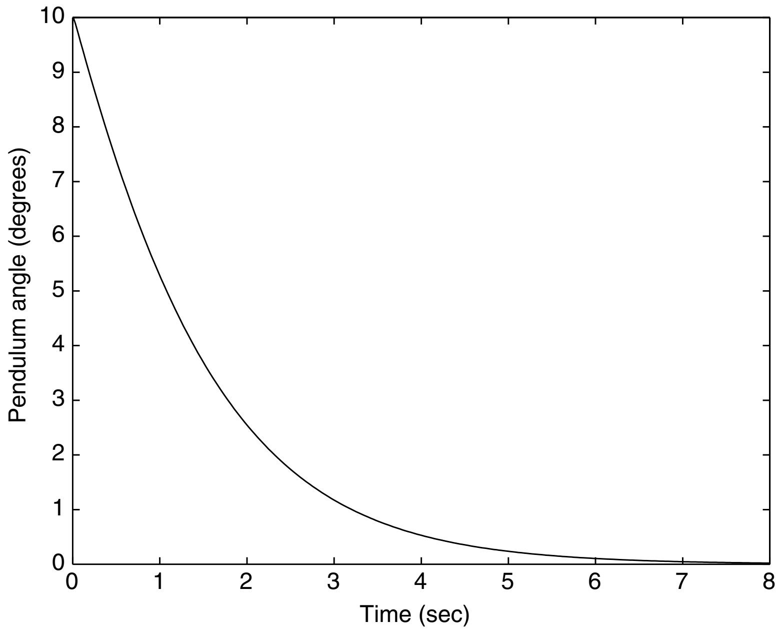


FIGURE 16.26 Pendulum position.

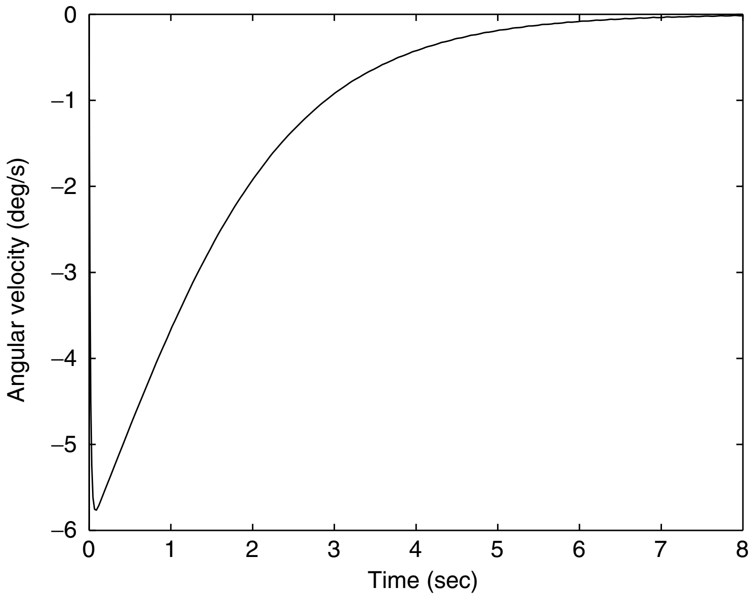


FIGURE 16.27 Pendulum velocity.

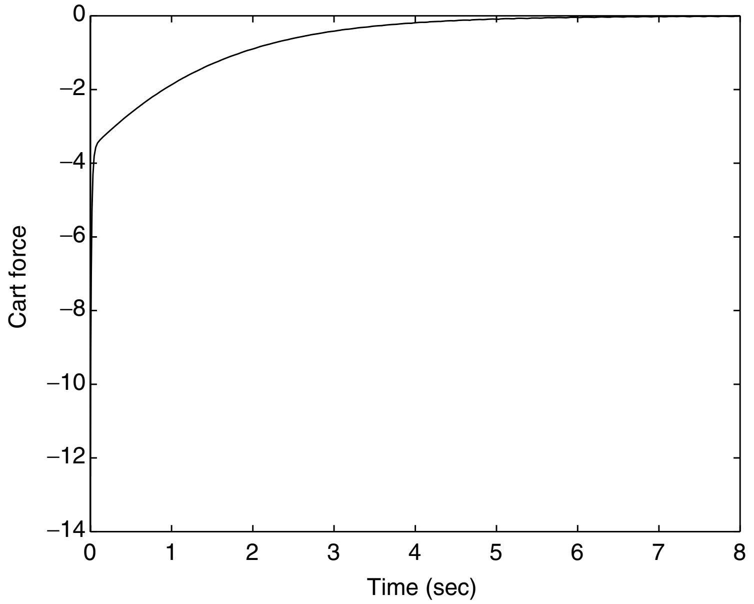


FIGURE 16.28 Control effort required to stabilize inverted pendulum.

(i.e., the plot of the function defining the control force computed by the fuzzy controller as a function of the two input variables).

The remainder of this section outlines the mathematical foundations of fuzzy logic which allow the reader to adapt this example for a particular application. Note that in the pendulum example, the "and" conjunction, the aggregation of the outputs, and the means to defuzzify the output were all implemented in certain, specific ways. These are not necessarily the only or best implementations. The mathematical outline will consider in more general terms fuzzy statements such as, "If A and B, then C" or "If A or B, then C," which will lead to a list of possible alternative implementations of such a fuzzy inference system. Which type of implementation is best may be application dependent, although the previous procedure is the predominant approach to fuzzy control.

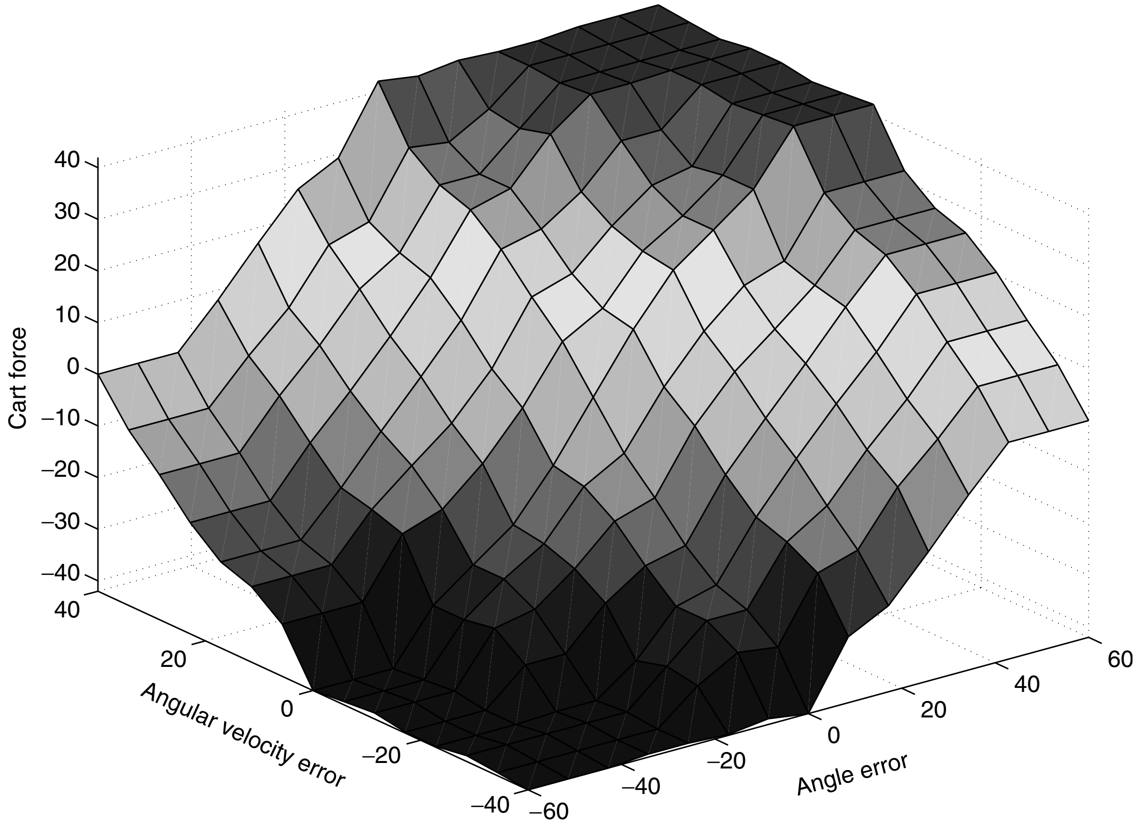


FIGURE 16.29 Response surface for pendulum fuzzy controller.

16.4.3 Fuzzy Sets and Fuzzy Logic

16.4.3.1 Introduction

This section introduces fuzzy sets, fuzzy logic, and their mathematical foundations. First, this section considers the concept of a membership function, and more specifically, whether an element belongs to a set or whether membership in a set is a matter of degree. Instead of either belonging or not belonging to a crisp set, an element can partially belong to a "fuzzy" set. Several examples of fuzzy sets are provided, and the properties of traditional crisp sets are compared with the analogous properties of fuzzy sets. There is a "crisp" aspect to the normal definition of fuzzy sets because the membership function returns a crisp value. Fuzzy sets can be generalized to have fuzzy-valued membership functions. After defining fuzzy sets and outlining their properties, operations on fuzzy sets such as the complement, intersection, etc. are defined and contrasted with the analogous operations on crisp sets. Finally, fuzzy arithmetic and fuzzy logic are introduced as well as the notion of an additive fuzzy system, which is the basic framework used in most fuzzy controls (in fact, the pendulum example above used this type of inference system).

16.4.3.2 Fuzzy vs. Crisp Sets

The traditional notion of a set is called a crisp set. Examples of crisp sets include:

1. The set of integers $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$
2. The set of all people taller than $5'8''$
3. Closed or open intervals of real numbers between a and b: $[a, b]$, (a, b), respectively
4. A set defined by explicitly listing its elements, such as the set containing the letters a, b, and c: $\{a, b, c\}$.

Unless otherwise indicated, crisp sets are not considered ordered. Crisp sets can be distinguished from fuzzy sets because in crisp sets an element either is a member of the set or is not a member of the set.

Mathematically, one can define a membership function m which maps from a universal set U which is the set of all possible elements, to the set $\{0, 1\}$, where for set A and element $x \in U$:

$$m : U \rightarrow \{0, 1\} \tag{16.9}$$

That is, the membership function returns a 1 if x is a member of A, and returns 0 if x is not a member of A.

Crisp sets have a list of standard properties related to concepts in classical logic. In particular, if the following operations are defined:

1. Complement: $\bar{A} = U - A = \{x \in U | x \notin A\}$
2. Union: $A \cup B = \{x \in U | x \in A \text{ or } x \in B\}$
3. Intersection: $A \cap B = \{x \in U | x \in A \text{ and } x \in B\}$

then verifying the following partial list of fundamental properties of crisp sets is straightforward:

1. Involution: $\bar{\bar{A}} = A$
2. Contradiction: $A \cap \bar{A} = \phi$
3. Excluded middle: $A \cup \bar{A} = U$

Having defined the membership function as a mapping from the universal set to the set containing zero and one, it is natural to consider a generalization of the mapping. Instead of considering the membership function as a binary mapping, the membership function for a fuzzy set is a mapping to the interval [0, 1]:

$$m : U \rightarrow [0, 1] \tag{16.10}$$

Now the mapping returns a value anywhere in the range between and including zero and one which encapsulates the notion that membership can be a matter of degree. This notion of degree enables fuzzy sets to express transitions between membership in sets where the transition is gradual (as opposed to crisp).

A prototypical example is temperature and whether the temperature on any given day is hot or cold. There is the set of hot days and the set of cold days. If these sets were crisp, they would require sharp boundaries. For example, if the temperature is above 80°F, it is hot; otherwise, it is not hot. Similarly, if the temperature is below 45°F, it is cold; otherwise, it is not cold. Such a rigid mathematical treatment of the notions of hot and cold is not appealing because humans are inclined to treat the transition to and from the set of hot and cold temperatures as gradual. A more appealing notion is that a given temperature may have a degree of membership in the set of hot days having a value of zero, one, or some value between zero and one. These values in between zero and one represent the transition from a day being not hot to the day being hot.

Membership functions have been described only as a mapping from the universal set to the interval from zero to one. Figure 16.30 illustrates several examples of typical membership functions. The membership function illustrated in the upper left figure is an example of a membership function that may model cold where the variable x represents temperature. For low temperatures, the value of the membership function is one, illustrating that the temperature is cold. High temperatures do not belong to the set of cold days, hence the value of the membership function is zero. Between the two extremes is a transition period where the temperature only partially belongs to the set of cold days. The figure in the upper right-hand corner is the analogous membership function for the set of hot days. Other fuzzy sets may require that only values within a certain range have a significant degree of membership in the fuzzy set. Possible examples of such membership functions are illustrated in the bottom two figures, which could represent warm days.

An interesting feature of all the examples of fuzzy sets presented above is that the membership functions are crisp values; that is, $m(x)$ is a crisp number. Depending on the application, requiring m to return a crisp value may be overly precise. Fuzzy sets can be generalized by defining membership functions to return a range of values instead of a crisp value. In particular,

$$m : U \rightarrow I([0, 1]) \tag{16.11}$$

where I represents the family of all closed intervals of real numbers in [0, 1] that the shaded portion in Figure 16.31 illustrates. Note that further generalization is possible because interval valued membership functions

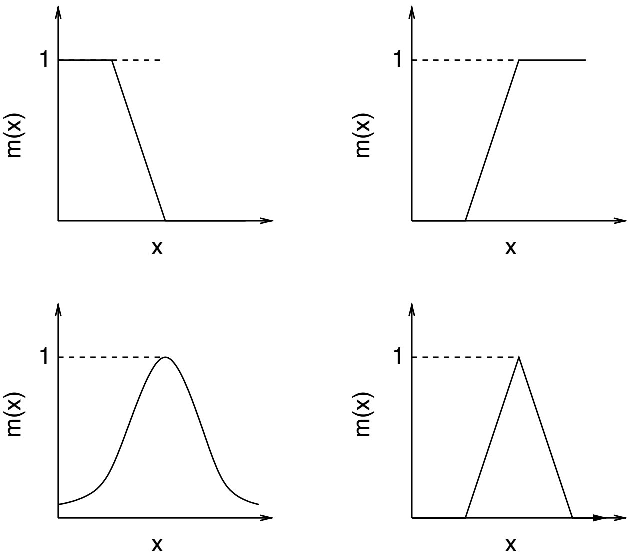


FIGURE 16.30 Examples of membership functions. (Adapted with permission from Klir, G.J., and Yuan, B., 1995.)

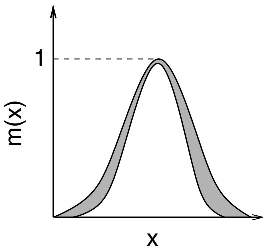


FIGURE 16.31 Fuzzy set defined by a fuzzy membership function.

can be generalized to have their intervals be fuzzy. Further generalizations are subsequently possible in a recursive fashion. Refer to Klir and Yuan (1995) for complete details.

16.4.3.3 Operations on Fuzzy Sets

Analogous to operations on crisp sets, a variety of operations can be defined on fuzzy sets. Adopting the standard notational shortcut where:

$$A(x) = m(x) \tag{16.12}$$

where $m(x)$ is the membership function that defines the fuzzy set A. We define the "standard" fuzzy complement, intersection, and union as follows:

1. Complement: $\bar{A}(x) = 1 - A(x)$
2. Intersection: $(A \cap B)(x) = \min[A(x), B(x)]$
3. Union: $(A \cup B)(x) = \max[A(x), B(x)]$
4. Subsethood: $A \subseteq B \Leftrightarrow A(x) \leqslant B(x)$

where each operation holds for all x. It is important to note that these are not the only ways to define these operations, although they are the typical ways. The intersection can also be defined in other common ways:

$$\begin{aligned}
(A \cap B)(x) &= A(x) \cdot B(x), \\
(A \cap B)(x) &= \max[0, A(x) + B(x) - 1] \\
(A \cap B)(x) &= \begin{cases} a & \text{if } b = 1 \\ b & \text{if } a = 1 \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{16.13}$$

The union also can be defined by:

$$(A \cup B)(x) = A(x) + B(x) - A(x) \cdot B(x),$$
$$(A \cup B)(x) = \min[1, A(x) + B(x)],$$
$$(A \cup B)(x) = \begin{cases} a & \text{if } b = 0 \\ b & \text{if } a = 0 \\ 0 & \text{otherwise} \end{cases} \tag{16.14}$$

For a more complete, axiomatic development, and a list of further possible definitions of intersections and unions of fuzzy sets, see Klir and Yuan (1995). In the more mathematical literature, intersections may be called t-norms, and unions may be called t-conorms. Most properties associated with crisp sets still hold for fuzzy sets, except for the properties of contradiction and excluded middle. The equality conditions of contradiction and excluded middle for crisp sets are replaced by subset conditions for fuzzy sets:

1. Contradiction: $A \cap \bar{A} \supset \phi$
2. Excluded Middle: $A \cup \bar{A} \subset U$

16.4.4 Fuzzy Logic

Fuzzy sets and their operations and properties provide the mathematical foundation for fuzzy logic, which is the basis for fuzzy control and other applications of fuzzy logic. Because feedback control is based upon measuring state variables, an important type of fuzzy set for fuzzy control is defined by a membership function whose domain is the set of real numbers:

$$m : \Re \rightarrow [0, 1] \tag{16.15}$$

which provides the degree to which a given variable is "close" to a specified value. Arithmetic operations on fuzzy numbers can then be defined as follows:

1. Addition: $(A + B)(z) = \sup_z \min[A(x), B(y)],$
   $z = x + y$
2. Subtraction: $(A + B)(z) = \sup_z \min[A(x), B(y)],$
   $z = x - y$
3. Multiplication: $(A + B)(z) = \sup_z \min[A(x), B(y)],$
   $z = x \cdot y$
4. Division: $(A + B)(z) = \sup_z \min[A(x), B(y)],$
   $z = x/y$

This arithmetic basis provides the foundation for the application of linguistic variables in fuzzy control algorithms. A linguistic variable is a fuzzy number that represents some sort of linguistic concept such as "very cold," "cold," "chilly," "comfortable," "warm," "hot," or "very hot." An example of a linguistic variable was previously illustrated in the pendulum example where the elements of the state of the pendulum $(\theta, \dot{\theta})$ were described in linguistic terms such as "negative large," "positive small," etc. Linguistic variables, or fuzzy numbers, allow linguistic terms to represent the approximate condition of the state of the system. As illustrated in the pendulum example, linguistic variables are an effective means to "translate" human expertise germane to a controls application into appropriate fuzzy rules used in a fuzzy controller.

Developing the standard additive model [Kosko, 1997] using the Mamdani inference system illustrates best the inference system typically used in fuzzy controllers. This model is the framework underlying most fuzzy controllers and is the framework of the previous pendulum controller example. Figure 16.20 illustrates the standard additive model [Kosko, 1997].

A set of if–then rules, which require some basic fuzzy logic and inference, are central to this system. Considering the linguistic variables that correspond to the fuzzy numbers representing the state of the pendulum, there are basic (or primary) terms, "negative," "zero," and "positive," and two hedges, "small" and "large." For other applications, different primary terms can be used, as well as different hedges, such as "very," "more," "less," "extremely," etc.

Several operators on fuzzy numbers are useful for implementing a fuzzy inference system. In particular, a fuzzy number can be concentrated or dilated according to:

$$A^k(x) = (A(x))^k \tag{16.16}$$

where A is the concentration operator if $k > 1$ or the dilation operator if $k < 1$ that can be used to represent the linguistic hedges "very" and "more or less," respectively. The operator "not" and the relations "and" and "or" are related to the definitions of complement, intersection, and union as follows:

1. *Not A* $\neg A(x) = \bar{A}(x) = 1 - A(x)$
2. *A and B* $(A\ \textbf{and}\ B)(x) = (A \cap B)(x)$
3. *A or B* $(A\ \textbf{or}\ B)(x) = (A \cup B)(x)$

Note that the definitions of "and" and "or" are not unique, as the definitions of the complement, intersection, and union are not unique. Thus, any of the possible definitions of intersection and union can be used to implement the logical "and" or logical "or."

An example of one way to evaluate the multiconditional approximate reasoning inference system in the standard additive model typical for fuzzy controllers is as follows: given a measured state variable, x, it may be "fuzzified" to account for measurement uncertainty. (Such a fuzzification was not considered in the pendulum example — in that case, the degree of membership of the crisp state value was used). As Figure 16.32 illustrates, if a measurement from a sensor is x, then the fuzzified set $X(x)$ may be defined to account for sensor uncertainty, where the shape of the membership function defining the fuzzy set $X(x)$ depends upon the type of uncertainty expected from the sensor. The degree of consistency between the fuzzified state measurement and a fuzzy set A_i is computed as the height of the intersection between $X(x)$ and $A_i(x)$. This is essentially determining the degree to which "if X is A_i" is satisfied. Because there are various means to compute the intersection of two fuzzy sets, the value of this degree of consistency will depend upon the definition of intersection used. In particular, if the standard intersection is used, then the degree of consistency is given by:

$$r_i(X) = \sup_x \min[X(x), A_i(x)] \tag{16.17}$$

where the "min" function computes the standard intersection, and the "sup" function determines its maximum value, as Figure 16.33 illustrates for two arbitrary fuzzy sets. Note that this is a generalization of using the degree of membership of a crisp value. The degree of membership is the supremum of the intersection of the line representing the crisp value of the variable and the fuzzy set, as Figure 16.21 illustrates.

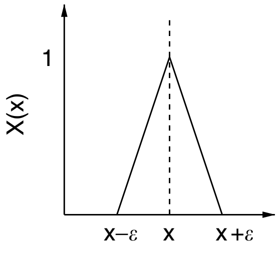


FIGURE 16.32 Fuzzifying a crisp variable.

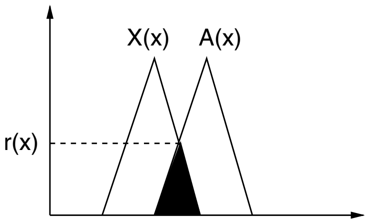


FIGURE 16.33 Degree of consistency between fuzzy sets $X(x)$ and $A(x)$.

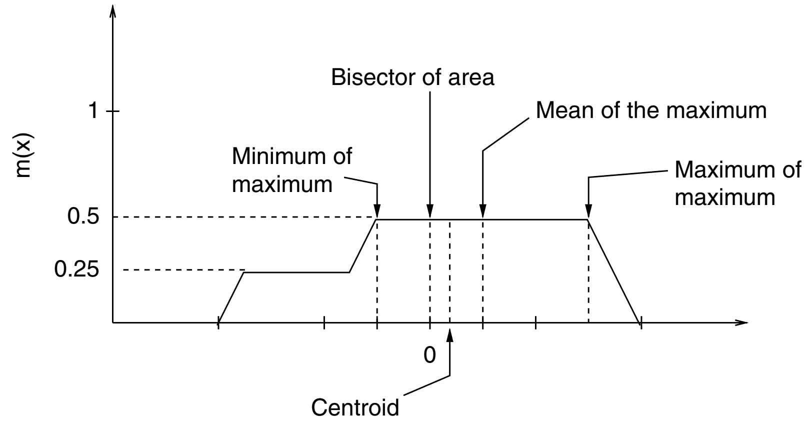


FIGURE 16.34 Defuzzification methods.

Having determined the degree to which "if X is A_i" is satisfied, the result of "then Y is B_i" must be determined. The most common (and most effective) technique was illustrated in the pendulum example. This technique lets the resulting fuzzy set, B', be determined according to $B' = \min[r_i, B]$ which is simply the "clipping" approach illustrated in the pendulum example.

The formulation to do so is as follows: given an if–then rule, if "X is A, then Y is B," where X and Y are fuzzy sets representing the state of linguistic variables, the task is to determine the application of this rule to a fuzzy set A' which is not necessarily identical to A to determine the appropriate conclusion, B', as illustrated in the following list:

Rule:	If X is A, then Y is B.
Fact:	X is A'.
Conclusion:	Y is B'.

The "min" operator used to determine the degree of consistency neither satisfies the rules of classical (Boolean) logic when reduced to the crisp case [Terano, 1992], nor does it satisfy all the axioms that may be generated as reasonable extensions of the classical case [Klir and Yuan, 1995]. Possibilities other than the "min" operator as fuzzy implications include $\max[1 - A(x), \min[A(x), B(y)]]$ (due to Zadeh), or $\min[1, 1 - A(x) + B(x)]$ (the Lukasiewicz implication). A list of such fuzzy implications, as well as a full exposition regarding their properties, can be found in Klir and Yuan (1995) or Jang et al. (1997). A more basic presentation is in Terano (1992) or Jang et al. (1997). From a controls perspective, note that "very good results are obtained" from the more general implications, but that Mamdani (1974), attempting to actually control a steam engine, "obtained excellent results from the max–min compositions" illustrated. A complete and rigorous exposition of fuzzy logic is based upon considerations of fuzzy relations and fuzzy implications, which are beyond the scope of this section.

The final step is defuzzification, where there are various alternative approaches to the centroid method presented in the pendulum example. In addition to the centroid, the following are possible methods for defuzzification:

1. Bisector of area
2. Mean of the maximum
3. Smallest of maximum
4. Largest of maximum

Figure 16.34 illustrates these concepts.

16.4.5 Alternative Inference Systems

The Mamdani inference system considered so far in this presentation is not the only inference system used in fuzzy control applications. In particular, the so-called TSK fuzzy model (named for Takagi, Sugeno and

Kang [Jang et al., 1997]) is an alternative model which has an advantage because it does not require defuzzification of the output, which can be computationally costly.

In particular, in the TSK model, fuzzy rules are of the form "if X is A and Y is B, then $z = f(x, y)$." In contrast to the Mamdani model, the output of the rules is a function, as opposed to a fuzzy set. For the pendulum example, possible TSK rules may include:

1. If the pendulum angle is zero and the angular velocity is zero, then $u = 0$.
2. If the pendulum angle is positive and small and the angular velocity is zero, then $u = 0.5\theta$.
3. If the pendulum angle is positive and large and the angular velocity is zero, then $u = 0.7\theta$.
4. If the pendulum angle is positive and small and the pendulum angular velocity is negative and small, then $u = 0.4\theta + 0.6\dot{\theta}$.

Defuzzification of the outputs is not required, but the outputs from each of the rules still need to be combined. Two possible alternatives are often employed: weighted average and weighted sum.

For the weighted average, if z_1 and z_2 are the output functions for two rules, and r_1 and r_2 are the degrees of consistency between the input data and antecedent fuzzy sets, A_1 and A_2, then the output is computed as:

$$u = \frac{r_1 z_1 + r_2 z_2}{r_1 + r_2} \tag{16.18}$$

If the weighted sum is used, then simply:

$$u = r_1 z_1 + r_2 z_2 \tag{16.19}$$

A final control paradigm briefly summarized here is model-based fuzzy control, which considers the design of fuzzy rules given the (nonlinear) model of the system to be controlled, which is in contrast with the heuristic approach of the traditional fuzzy logic control paradigm outlined above. The advantage of this approach is that it makes use of analytical model information that may be available but is completely ignored in the standard fuzzy control paradigm.

At least two different forms of model-based fuzzy control paradigms exist: the so-called Takagi–Sugeno fuzzy logic controllers (TSFLCs) and sliding-mode fuzzy logic controllers (SMFLCs). For TSFLCs, rules are determined by considering the dynamics of the system in various "fuzzy regimes" of the state space and then determining appropriate (linear) control laws at the center of each of these fuzzy regimes. SMFLC rules are determined by considering the distance between the state vector and a desired "sliding surface." For further details, refer to Palm et al., 1997.

16.4.6 Other Applications

Although feedback control is the primary application of fuzzy logic, it certainly is not the exclusive application. Other applications include identification and classification techniques such as handwriting recognition, robotics, intelligent agents, and database information retrieval [Yen and Langari, 1999]. Additional identification and classification techniques include nonlinear system identification and adaptive noise cancellation [Jang et al., 1997], modeling [Babuska, 1998], PID controller tuning [Yen and Langari, 1999], process control and analysis [Ruan, 1997], and traffic control [Dubois, 1980].

16.5 Conclusions

We reviewed some of the major soft computing (SC) techniques used for complex systems. Due to limitations of space, SC is described only in outline. The purpose is to show the way the methods work, the possible range of applications, and to introduce these new technologies. SC techniques are not model based so they are most suitable for applications in which first-principles-based approaches either are not possible or are too slow. There are many such instances in the control area for which soft computing is especially appropriate. As MEMS devices are in the frontiers of hardware, many of the issues are still not completely clear, and the model equations cannot always be computed quickly enough for real-time control purposes. It is possible, that SC techniques could lend a hand to the use of these devices in real applications.

Acknowledgments

M.S. wishes to thank his colleagues K.T. Yang and R.L. McClain and his students X. Zhao, G. Díaz, A. Pacheco-Vega, and W. Franco for collaboration on artificial neural networks and genetic algorithms. He also thanks D.K. Dorini of BRDG-TNDR for sponsoring the research and CONACyT (Mexico) and the Organization of American States for support of the students. B.G. would like to thank Neil Petroff for his many thoughtful comments and suggestions concerning the section on fuzzy logic and fuzzy control.

References

Aminzadeh, F., and Jamshidi, M. (1994) *Soft Computing: Fuzzy Logic, Neural Networks, and Distributed Artificial Intelligence*, Prentice-Hall, Englewood Cliffs, NJ.

Angeline, P.J., Saunder, G.M., and Pollack, J.B. (1994) "Complete Induction of Recurrent Neural Networks," in *Proc. of the Third Annual Conf. on Evolutionary Programming*, A.V. Sebald and L.J. Fogel, eds., World Scientific, Singapore, pp. 1–8.

Babuska, R. (1998) *Fuzzy Modeling for Control*, Kluwer Academic, Boston.

Berlin, A.A., Chase, J.G., and Jacobsen, S.C. (1998) "MEMS-Based Control of Structural Dynamic Instability," *J. Intel. Mater. Syst. Struc.* **9**(7), pp. 574–86.

Bouchon-Meunier, B., Yager, R.R., and Zadeh, L.A., eds. (1995) *Fuzzy Logic and Soft Computing*, World Scientific, River Edge, NJ.

Buckley, J.J., and Feuring, T. (1999) *Fuzzy and Neural: Interactions and Applications*, Physica-Verlag, New York.

Chan, H.L., and Rad, A.B. (2000) "Real-Time Flow Control Using Neural Networks," *ISA Trans.* **39**(1), pp. 93–101.

Díaz, G. (2000) Simulation and Control of Heat Exchangers Using Artificial Neural Networks, Ph.D. dissertation, Department of Aerospace and Mechanical Engineering, University of Notre Dame.

Díaz, G., Sen, M., Yang, K.T., and McClain, R.L. (1999) "Simulation of Heat Exchanger Performance by Artificial Neural Networks," *Int. J. HVAC&R Res.* **5**(3), pp. 195–208.

Díaz, G., Sen, M., Yang, K.T., and McClain, R.L. (2001a) "Dynamic Prediction and Control of Heat Exchangers Using Artificial Neural Networks," *Int. J. Heat Mass Transf.* **44**(9), pp. 1671–9.

Díaz, G., Sen, M., Yang, K.T., and McClain, R.L. (2001b) "Stabilization of Thermal Neurocontrollers," submitted for review.

Díaz, G., Sen, M., Yang, K.T., and McClain, R.L. (2001c) "Adaptive Neurocontrol of Heat Exchangers," *ASME J. Heat Transf.* **123**(3), pp. 556–62.

Dimeo, R., and Lee, K.Y. (1995) "Boiler-Turbine Control System Design Using a Genetic Algorithm," *IEEE Trans. Energy Convers.* **10**(4), p. 752.

Drago, G.P., and Ridella, S. (1992) "Statistically Controlled Activation Weight Initialization," *IEEE Trans. Neural Networks* **3**(4), pp. 627–31.

Dubois, D. (1980) *Fuzzy Sets and Systems — Theory and Applications*, Academic Press, New York.

Eeckman, F.H., ed. (1992) *Analysis and Modeling of Neural Systems*, Kluwer Academic, Boston, MA.

Flood, I., and Kartam, N. (1994) "Neural Networks in Civil Engineering. I. Principles and Understanding," *ASCE J. Comp. Civil Eng.* **8**(2), pp. 131–48.

Fogel, L.J. (1999) *Intelligence Through Simulated Evolution*, John Wiley & Sons, New York.

Gad-el-Hak, M. (1994) "Interactive Control of Turbulent Boundary Layers: A Futuristic Overview," *AIAA J.* **32**, pp. 1753–65.

Gad-el-Hak, M. (1999) "The Fluid Mechanics of Microdevices — The Freeman Scholar Lecture," *J. Fluid Eng. Trans. ASME* **121**(1), pp. 5–33.

Gagarin, N., Flood, I., and Albrecht, P. (1994) "Computing Truck Attributes with Artificial Neural Networks," *ASCE J. Comput. Civil Eng.* **8**(2), pp. 179–200.

Gaudenzi, P., Fantini, E., Koumousis, V.K., and Gantes, C.J. (1998) “Genetic Algorithm Optimization for the Active Control of a Beam by Means of PZT Actuators,” *J. Intel. Mater. Syst. Struc.*, **9**(4), pp. 291–300.

Goldberg, D.E. (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, Reading, MA.

Haykin, S. (1999) *Neural Networks, A Comprehensive Foundation*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ.

Ho, C.-M., and Tai, Y.-C. (1996) “Review: MEMS and Its Applications for Flow Control,” *J. Fluid Eng. Trans. ASME* **118**(3), pp. 437–47.

Ho, C.-M., and Tai, Y.-C. (1998) “Micro-Electro-Mechanical-Systems (MEMS) and Fluid Flows,” *Annu. Rev. Fluid Mech.* **30**, pp. 579–612.

Jacobson, S.A., and Reynolds, W.C. (1993) “Active Control of Boundary Layer Wall Shear Stress Using Self-Learning Neural Networks,” AIAA Paper No. 93-3272, American Institute of Aeronautics and Astronautics, Washington, D.C.

Jain, L.C., and Fukuda, T., eds. (1998) *Soft Computing for Intelligent Robotic Systems*, Physica-Verlag, Heidelberg, Germany.

Jang, J.S.R., Sun, C.T., and Mizutani, E. (1997) *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*, Prentice-Hall, Englewood Cliffs, NJ.

Kaboudan, M.A. (1999) “A Measure of Time Series’ Predictability Using Genetic Programming Applied to Stock Returns,” *J. Forecasting* **18**(5/6), p. 345.

Kamarthi, S., Sanvido, V., and Kumara, R. (1992) “Neuroform — Neural Network System for Vertical Formwork Selection,” *ASCE J. Comp. Civ. Eng.* **6**(2), pp. 178–99.

Kao J.J. (1999) “Optimal Location of Control Valves in Pipe Networks by Genetic Algorithm — Closure,” *J. Water Res. Planning Manage. Div. ASCE* **125**(1), pp. 68–9.

Karmin, E.D. (1990) “Simple Procedure for Pruning Back Propagation Trained Neural Networks,” *IEEE Trans. Neural Networks* **1**(2), pp. 239–42.

Katisikas, S.K., Tsahalis, D., and Xanthakis, S.A. (1995) “A Genetic Algorithm for Active Noise Control Actuator Positioning,” *Mech. Syst. Signal Proc.* **9**(6), p. 697.

Keane, A.J. (1995) “Passive Vibration Control via Unusual Geometries the Application of Genetic Algorithm Optimization to Structural Design,” *J. Sound Vib.* **185**(3), p. 441.

Klir, G.J., and Yuan, B. (1995) *Fuzzy Sets and Fuzzy Logic: Theory and Applications*, Prentice-Hall, Englewood Cliffs, NJ.

Kosko, B. (1997) *Fuzzy Engineering*, Prentice-Hall, Englewood Cliffs, NJ.

Lee, C., Kim, J., Babcock, D., and Goodman, R. (1997) “Application of Neural Networks to Turbulence Control for Drag Reduction,” *Phys. Fluids* **9**(6), p. 1740.

Lehtokangas, M., Saarinen, J., and Kaski, K. (1995) “Initializing Weights of a Multilayer Perceptron Network by Using the Orthogonal Least Squares Algorithm,” *Neural Comput.* **7**, pp. 982–99.

Lin, L.C., and Lee, G.Y. (1999) “Hierarchical Fuzzy Control for C-axis of CNC Turning Centers Using Genetic Algorithms,” *J. Intel. Robotic Syst.* **25**(3), pp. 255–75.

Löfdahl, L., and Gad-el-Hak, M. (1999) “MEMS Applications in Turbulence and Flow Control,” *Prog. Aerosp. Sci.* **35**(2), pp. 101–203.

Luk, P.C.K., Low, K.C., and Sayiah, A. (1999) “GA-Based Fuzzy Logic Control of a Solar Power Plant Using Distributed Collector Fields,” *Renew. Energy* **16**(1–4), pp. 765–68.

Mamdani, E.H. (1974) “Application of Fuzzy Algorithms for Control of Simple Dynamic Plant,” *IEEE Proc.* **121**(12), pp. 1585–8.

Mamdani, E.H., and Assilian, S. (1975) “An Experiment in Linguistic Synthesis with a Fuzzy Logic Controller,” *Int. J. Machine Stud.* **7**(1), pp. 1–13.

Mamdani, E.H., and Baaklini, N. (1975) “Perspective Method for Deriving Control Policy in a Fuzzy-Logic Controller,” *Electron. Lett.* **11**, pp. 625–6.

Man, K.F., Tang, K.S., and Kwong, S. (1999) *Genetic Algorithms*, Springer-Verlag, Berlin.

Matsuura, K., Shiba, H., Hirotsune, M., and Hamachi, M. (1995) “Optimizing Control of Sensory Evaluation in the Sake Mashing Process by Decentralized Learning of Fuzzy Inference Using a Genetic Algorithm,” *J. Ferment. Bioeng.* **80**(3), pp. 251–258.

Michalewicz, Z. (1992) *Genetic Algorithm + Data Structure = Evolution Programs*, Springer-Verlag, Berlin.

Michalewicz, Z., Janikow, C.Z., and Krawczyk, J.B. (1992) "A Modified Genetic Algorithm for Optimal Control Problems," *Comp. Math. Appl.* **23**(12), pp. 83–94.

Mitchell, M. (1997) *An Introduction to Genetic Algorithms*, MIT Press, Cambridge, MA.

Mordeson, J.N., and Nair, P.S. (1988) *Fuzzy Mathematics: An Introduction for Engineers and Scientists*, Physica-Verlag, New York.

Nagaoka, Y., Alexander, H.G., Liu, W., and Ho, C.M. (1997) "Shear Stress Measurements on an Airfoil Surface Using Micro-Machined Sensors," *JSME Int. J. Series B — Fluids Thermal Eng.* **40**(2), pp. 265–72.

Nagaya, K., and Ryu, H. (1996) "Deflection Shape Control of a Flexible Beam by Using Shape Memory Alloy Wires Under the Genetic Algorithm Control," *J. Intel. Mater. Syst. Struc.* **7**(3), p. 336.

Nakashima, M., Maruyama, Y., and Hasegawa, T. (1998) "Basic Experiments on Robot-Based Vibration Control of the Hot-Line Work Robot System Using Genetic Algorithm," *J. Electr. Eng. Jpn.* **123**(2), p. 40.

Nelson, B.J., Zhou, Y., and Vikramaditya, B. (1998) "Sensor-Based Microassembly of Hybrid MEMS Devices," *IEEE Contr. Syst. Mag.* **18**(6), p. 35.

Nordin, P., Banzhaf, W., and Brameier, M. (1998) "Evolution of a World Model for a Miniature Robot Using Genetic Programming," *Robot Autonomous Syst.* **25**(1–2), pp. 105–16.

Pacheco-Vega, A., Sen, M., Yang, K.T., and McClain, R.L. (1998) "Genetic-Algorithm-Based Prediction of a Fin-Tube Heat Exchanger Performance," *Proc. 11th Int. Heat Trans. Conf.* **6**, pp. 137–42.

Pacheco-Vega, A., Diaz, G., Sen, M., Yang, K.T., and McClain, R.L. (2001) "Heat Rate Predictions in Humid Air-Water Heat Exchangers Using Correlations and Neural Networks," *ASME J. Heat Transf.* **123**(2), pp. 348–54.

Pal, S.K., and Mitra, S. (1999) *Neuro-Fuzzy Pattern Recognition*, John Wiley & Sons, New York.

Palm, R., Driankov, D., and Hellendoorn, H. (1997) *Model Based Fuzzy Control*, Springer-Verlag, Berlin.

Perhinschi, M.G. (1998) "Optimal Control System Design Using a Genetic Algorithm," *ZAMM* **78** (suppl. 3), p. S1035.

Rahmoun, A., and Benmohamed, M. (1998) "Genetic Algorithm Based Methodology to Generate Automatically Optimal Fuzzy Systems," *IEE Proceedings — Control Theory and Applications*, **145**(6), pp. 583–6, 1988.

Ranganath, M., Renganathan, S., and Rao, C.S. (1999) "Genetic Algorithm Based Fuzzy Logic Control of a Fed-Batch Fermenter," *Bioprocess and Biosystems Engineering* **21**(3), pp. 215–8.

Reis, L.F.R., Porto, R.M., and Chaudhry, F.H. (1997) "Optimal Location of Control Valves in Pipe Networks by Genetic Algorithm," *J. Water Res. Planning Manage.* **123**(6), p. 317.

Ruan, D., ed. (1997) *Intelligent Hybrid Systems*, Kluwer, Norwell, MA.

Rumelhart, D.E., Hinton, D.E., and Williams, R.J. (1986) "Learning Internal Representations by Error Propagation," in *Parallel Distributed Processing: Exploration in the Microstructure of Cognition*, Vol. 1, D.E. Rumelhart and J.L. McClelland, eds., MIT Press, Cambridge, MA.

Rzempoluck, E.J. (1998) *Neural Network Data and Analysis Using Simulnet*, Springer, New York.

Schalkoff, R.J. (1997) *Artificial Neural Networks*, McGraw-Hill, New York.

Sharatchandra, M.C., Sen, M., and Gad-el-Hak, M. (1998) "New Approach to Constrained Shape Optimization Using Genetic Algorithms," *AIAA J.* **38**(1), pp. 51–61.

Sen, M., and Yang, K.T. (2000) "Applications of Artificial Neural Networks and Genetic Algorithms in Thermal Engineering," in *The CRC Handbook of Thermal Engineering*, F. Kreith, ed., CRC Press, Boca Raton, FL, pp. 620–61.

Seywald, H., Kumar, R.R., and Deshpande, S.M. (1995) "Genetic Algorithm Approach for Optimal Control Problems with Linearly Appearing Controls," *Journal of Guidance, Control and Dynamics* **18**(1), pp. 177–182, 1995.

Subramanian, H., Varadan, V.K., Varadan, V.V., and Vellekoop, M.J. (1997) "Design and Fabrication of Wireless Remotely Readable MEMS Based Microaccelerometers," *Smart Mater. Struct.* **6**(6), pp. 730–8.

Suzuki, Y., and Kasagi, N. (1997) "Active Flow Control with Neural Network and Its Application to Vortex Shedding," in *Proc. 11th Symp. on Turbulent Shear Flows*, pp. 9.18–9.23, Grenoble, France.

Tang, K.S., Man, K.F., and Chu, C.Y. (1996a) "Application of the Genetic Algorithm to Real-Time Active Noise Control," *Real-Time Syst.* **11**(3), p. 289.

Tang, K.S., Man, K.F., and Gu, D.W. (1996b) "Structured Genetic Algorithm for Robust H Control Systems Design," *IEEE Trans. Ind. Electron.* **43**(5), p. 575.

Terano, T. (1992) *Fuzzy Systems Theory and Its Applications*, Academic Press, San Diego, CA.

Thibault, J., and Grandjean, B.P.A. (1991) "Neural Network Methodology for Heat Transfer Data Analysis," *Int. J. Heat Mass Transf.* **34**(8), pp. 2063–70.

Trebi-Ollennu, A., and White, B.A. (1997) "Multiobjective Fuzzy Genetic Algorithm Optimisation Approach to Nonlinear Control System Design," *IEE Proceedings — Control Theory and Applications*, **144**(2), pp. 137–42, 1997.

Vandelli, N., Wroblewski, D., Velonis, M., and Bifano, T. (1998) "Development of a MEMS Microvalve Array for Fluid Flow Control," *J. Microelectromech. Syst.* **7**(4), pp. 395–403.

Varadan, V.K., Varadan, V.V., and Bao, X.Q. (1995) "Comparison of MEMS and PZT Sensor Performance in Active Vibration and Noise Control of Thin Plates," *J. Wave-Mater. Interact.* **10**(4), p. 51.

Warwick, K., Irwin, G.W., and Hunt, K.J. (1992) *Neural Networks for Control and Systems*, Short Run Press, Ltd., Exeter.

Wessels, L., and Barnard, E. (1992) "Avoiding Fake Local Minima by Proper Initialization of Connections," *IEEE Transactions on Neural Networks* **3**(6), pp. 899–905.

Yager, R.R., and Zadeh, L.A. (1994) *Fuzzy Sets, Neural Networks, and Soft Computing*, Van Nostrand-Reinhold, New York.

Yen, J., and Langari, R. (1999) *Fuzzy Logic*, Prentice-Hall, Englewood Cliffs, NJ.

Zadeh, L.A. (1965) "Fuzzy Sets," *Inf. Control* **8**, pp. 338–53.

Zadeh, L.A. (1968a) "Probability Measures and Fuzzy Systems," *J. Math. Anal. Appl.* **23**(2), pp. 421–27.

Zadeh, L.A. (1968b) "Fuzzy Algorithm," *Inf. Control* **12**, pp. 94–102.

Zadeh, L.A. (1971) "Toward a Theory of Fuzzy Systems," in *Aspects of Network and System Theory*, R.E. Kalman and N. Dellaris, eds., Holt, Rinehart and Winston, New York.

Zeng, P. (1998) "Neural Computing in Mechanics," *AMR* **51**(2), pp. 173–97.

Zhao, X. (1995) Performance of a Single-Row Heat Exchanger at Low In-Tube Flow Rates, M.S. thesis, Department of Aerospace and Mechanical Engineering, University of Notre Dame.