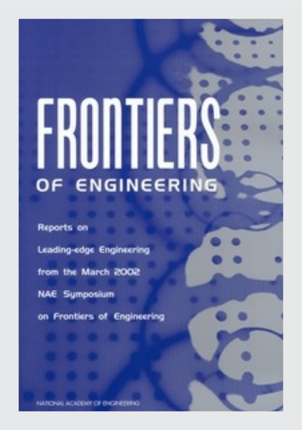


Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2001 NAE Symposium on Frontiers of Engineering

DETAILS

138 pages | 6 x 9 | PAPERBACK
ISBN 978-0-309-08498-7 | DOI 10.17226/10494

AUTHORS

National Academy of Engineering

SEVENTH ANNUAL SYMPOSIUM ON FRONTIERS OF ENGINEERING

NATIONAL ACADEMY OF ENGINEERING

NATIONAL ACADEMY PRESS
Washington, D.C.

NATIONAL ACADEMY PRESS - 2101 Constitution Ave., N.W. - Washington, D.C. 20418

NOTICE: This publication has been reviewed according to procedures approved by a National Academy of Engineering report review process. Publication of signed work signifies that it is judged a competent and useful contribution worthy of public consideration, but it does not imply endorsement of conclusions or recommendations by the NAE. The interpretations and conclusions in such publications are those of the authors and do not purport to represent the views of the council, officers, or staff of the National Academy of Engineering.

Funding for the activity that led to this publication was provided by Cummins, Inc., Defense Advanced Research Projects Agency, Department of Defense (DDR&E-Research), Microsoft Corporation, National Aeronautics and Space Administration, and United Technologies Corporation.

International Standard Book Number 0-309-08498-9

Additional copies of this report are available from National Academy Press, 2101 Constitution Avenue, N.W., Lockbox 285, Washington, D.C. 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu.

Printed in the United States of America

THE NATIONAL ACADEMIES

National Academy of Sciences
National Academy of Engineering
Institute of Medicine
National Research Council

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Bruce M. Alberts is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Wm. A. Wulf is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Bruce M. Alberts and Dr. Wm. A. Wulf are chairman and vice chairman, respectively, of the National Research Council.

ORGANIZING COMMITTEE

MICHAEL L. CORRADINI (Chair), Chair, Engineering Physics Department; Professor of Nuclear Engineering and Engineering Physics, University of Wisconsin-Madison

DAVID J. BEEBE, Associate Professor, Department of Biomedical Engineering, University of Wisconsin-Madison

SUE McNEIL, Director, Urban Transportation Center, University of Illinois at Chicago

PRISCILLA P. NELSON, Director, Division of Civil and Mechanical Systems, Directorate of Engineering, National Science Foundation

DONALD R. NILSON, Director, Engineering Technology & Strategic Planning, Lockheed Martin Aeronautics Company

JOHN D. NORTON, Staff Scientist, Medtronic, Inc.

SHARON L. NUNES, Director, Life Sciences Solution Development, Corporate Technology Group, IBM Corporation

ALBERT P. PISANO, FANUC Chair of Mechanical Systems, Electronics Research Lab, University of California, Berkeley

VENUGOPAL V. VEERAVALLI, Associate Professor, Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign

MINERVA M. YEUNG, Principal Engineer and Manager of Media Technology Research, Microprocessor Research Labs, Intel Corporation

Staff

JANET R. HUNZIKER, Program Officer

MARY W. L. KUTRUFF, Administrative Assistant

LANCE R. TELANDER, Senior Project Assistant

Preface

This is the seventh volume highlighting the presentations at the National Academy of Engineering (NAE) annual Frontiers of Engineering Symposium program, which brings together 100 outstanding young leaders in engineering to share their cutting-edge research and technical work. The 2001 symposium was originally scheduled for September 13–15, but was cancelled and rescheduled for March 1–3, 2002, because of the events of September 11. Approximately 85 participants attended the rescheduled symposium at the Beckman Center in Irvine, California. The papers included in this volume are extended summaries of the presentations prepared by the speakers. The intent of this volume, and of the preceding volumes in the series, is to describe the philosophy behind this unique meeting and to highlight some of the exciting developments in engineering today.

GOALS OF FRONTIERS OF ENGINEERING

The practice of engineering is changing. Engineers must be able to adapt and thrive in an environment of rapid technological change and globalization. In addition, engineering is becoming more interdisciplinary. The frontiers of engineering often occur at the intersections between engineering disciplines or at the intersection between what has traditionally been labeled "science" and "engineering." Thus, both researchers and practitioners must be aware of developments and challenges in areas other than their own.

At the three-day Frontiers of Engineering Symposium, 100 of this country's best and brightest engineers, ages 30 to 45, can learn from their peers about developments at the leading edge of engineering. This broad overview of current developments in many fields of engineering often leads to insights of cross-disciplinary applications. In addition, the symposium brings together engineers in academia, industry, and government, enabling them to establish contacts with and learn from people they would probably not meet in the usual round of professional meetings. We hope this networking will lead to collaborative work that facilitates the transfer of new techniques and approaches from one field of engineering to another.

The number of participants at each meeting is kept at 100 to maximize the opportunities for interactions and exchanges among the attendees, who are invited to attend after a competitive nomination and selection process. The topics and speakers for each meeting are chosen by an organizing committee of engineers in the same age group as the participants. Different topics are covered each year, and, with a few exceptions, different individuals are invited to participate.

Each speaker at the Frontiers of Engineering Symposium faces a unique challenge—to convey the excitement of his or her field to a technically sophisticated but nonspecialist audience. To meet this challenge, speakers are asked to provide brief overviews of their fields that include a definition of the frontiers of the field; a brief description of current experiments, prototypes, and design studies; a description of new tools and methodologies; identification of limitations on advances and controversies; a brief description of the most exciting results and most difficult challenges of the past couple of years; and a summary statement of the theoretical, commercial, societal, and long-term significance of the work.

CONTENT OF THE MARCH 2002 SYMPOSIUM

The presentations of the symposium covered four broad areas: leading edge aerodynamics technologies, civil systems, wireless communications, and technology and the human body. In the session called "Flight at the Leading Edge," speakers discussed next-generation jet propulsion, miniature unmanned air vehicles (mini-UAVs), and artificial flying insects. Developments ranged from enabling technologies for familiar aircraft to leading-edge concepts involving autonomous aircraft and micromechanical flyers that are changing the way we travel, provide national defense, and collect and share electronic information. The second session, "Civil Systems," focused on the systems that provide communications and information, the delivery of water and power, the removal of waste products, and travel from home to work and other activities. A common theme in these talks was the inherently interdisciplinary process required to operate, maintain, and replace civil systems. The speakers addressed the role of engineers in the development and deployment of decision support systems, computer-based supervisory controls, robotics, and new sensors and materials to relieve pressure on existing civil systems and meet the demand for new ones. In the third session, "Wireless Communications," speakers described the critical issues and state-of-the-art technologies for cost-effective, ubiquitous, and integrated personal communications networks that provide multimedia services; another topic was the interaction of wireless communications with the physical world. Each of the speakers in this session covered one aspect of the subject: a fundamental theoretical framework for wireless communications to guide the development of new technologies; a robust telecommunications infrastructure that will make ubiquitous mobile wireless access possible; service architectures that will enable emerging wireless applications; and wireless sensor systems that will link the physical world to communication networks. The concluding session, "Technology and the Human Body," addressed the importance of science and technology to our health and well-being. This session began with an overview of modeling and simulation of the human body and how they can be used to improve our understanding of chronic diseases. The second talk presented research on electronic devices that enable mobility in patients who are paralyzed. The final talk was on recent advances in tissue engineering, including microfabrication techniques and cellular interactions. (See Appendixes for complete program.)

It is traditional to invite a distinguished engineer to address the participants at dinner on the first evening of the symposium. This year, Nicholas Donofrio, senior vice president and group executive for technology and manufacturing, IBM Corporation, spoke about technology innovation. He described the challenges posed by information overload and limits to Moore's Law, addressed privacy concerns, and pointed out the necessity for computer modeling and simulation to become standard tools for all engineers. Mr. Donofrio emphasized the importance of mentoring and maintaining diversity in a strong engineering workforce. The full text of Mr. Donofrio's remarks are included in this volume.

NAE is deeply grateful to the following organizations for their support of the Seventh Annual Symposium on Frontiers of Engineering: Defense Advanced Research Projects Agency; Department of Defense–DDR&E-Research; National Aeronautics and Space Administration; Microsoft Corporation; United Technologies Corporation; and Cummins, Inc. NAE would also like to thank the members of the Symposium Organizing Committee (see p. *iv*), chaired by Michael Corradini, for planning and organizing the event.

Contents

FLIGHT AT THE LEADING EDGE: EXTREME AERODYNAMICS FROM THE MEGA TO THE MICRO

CIVIL SYSTEMS

WIRELESS COMMUNICATIONS

TECHNOLOGY AND THE HUMAN BODY

DINNER SPEECH

APPENDIXES

Flight at the Leading Edge: Extreme Aerodynamics from the Mega to the Micro

Active Flow Control: Enabling Next-Generation Jet Propulsion Aerodynamics

JEFFREY W. HAMSTRA AND DANIEL N. MILLER
Lockheed Martin Aeronautics Company
Fort Worth, Texas

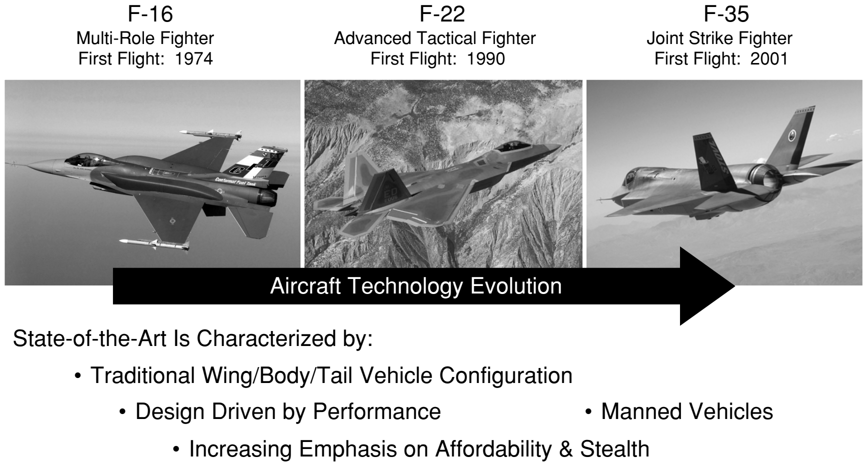
INTRODUCTION

Jet engine inlet and exhaust systems will play a major role in determining the configuration and capability of tomorrow's military air vehicles. To support advances in vehicle design, these systems must deliver higher aerodynamic performance, as well as enhanced functionality (such as thrust vectoring), and at the same time be lighter in weight, less expensive, and smaller than current state-of-the-art systems. Traditionally, the physical laws governing high-speed viscous flow have limited the implementation of the exotic inlet and exhaust flowpaths that will be required for the future. The emerging technology of active flow control (AFC) could provide a breakthrough in aeronautical science that would enable the engineering design of next-generation inlet and exhaust systems.

BACKGROUND

Combat aircraft have continued to evolve since the introduction of the jet engine during World War II. The F-16, first flown in 1974, the F-22 (1990), and the F-35 (2001), exemplify the last 25 years of this evolution (Figure 1). These aircraft are characterized by a traditional wing/body/tail arrangement, are all commanded by an on-board pilot, and are all driven by the requirement for superb aerodynamic performance. Throughout the evolutionary process, two additional design characteristics, affordability and stealth, have become increasingly important.

The propulsion system for a combat air vehicle is critically important in terms of cost, weight, volume, stealth, performance, and overall configuration integration. Major propulsion system components, such as the engine inlet



F-16
Multi-Role Fighter
First Flight: 1974

F-22
Advanced Tactical Fighter
First Flight: 1990

F-35
Joint Strike Fighter
First Flight: 2001

Aircraft Technology Evolution

State-of-the-Art Is Characterized by:

- Traditional Wing/Body/Tail Vehicle Configuration
- Design Driven by Performance
- Manned Vehicles
- Increasing Emphasis on Affordability & Stealth

FIGURE 1 Today's combat aircraft are driven by requirements for superb aerodynamic performance. Source: 2002 by Lockheed Martin. Published with permission.

system and the engine exhaust system, are also critical. The inlet system captures outside air and delivers it to the engine. Inlet systems are typically 10 to 20 feet long and weigh on the order of 500 to 1200+ pounds. Major performance figures-of-merit for the inlet system are *pressure recovery* (a measure of the overall efficiency of the system) and *distortion* (a measure of the pressure nonuniformity at the inlet/engine interface). Severe distortion can cause stalling or even flameout of the jet engine. The purpose of the exhaust system is to convert the engine's high-temperature thermodynamic energy into net propulsive force. Performance figures-of-merit for the exhaust system include *gross thrust coefficient* (a measure of the efficiency of the system) and *thrust vector angle* (a measure of the system's ability to divert or steer the exhaust thrust in a nonaxial direction). Thrust vectoring is used in conjunction with the wing and tail flaps to control the air vehicle.

As combat aircraft continue to evolve, they must retain current levels of aerodynamic performance and, at the same time, become more affordable and more stealthy. These improvements will require changes in the design of inlets and exhausts; both systems will have to be shorter and more compact, simpler and lighter in terms of mechanical complexity and moving parts, and shaped to conform to the advanced, all-wing, tailless vehicle configurations of future aircraft (Figure 2). Internal studies at Lockheed Martin (LM) have shown that many of the necessary design characteristics of future inlet and exhaust systems are not achievable with current aerodynamic technology. Researchers in the

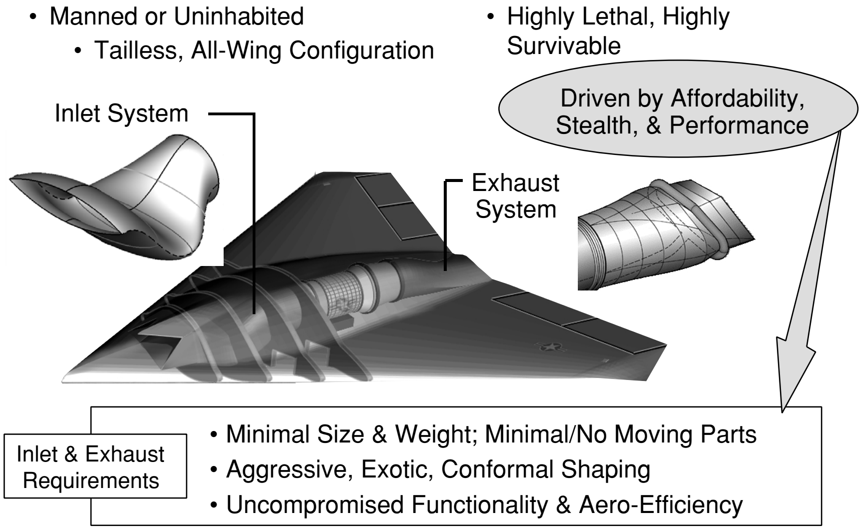

• Manned or Uninhabited
• Tailless, All-Wing Configuration
• Highly Lethal, Highly Survivable
Driven by Affordability, Stealth, & Performance
Inlet System
Exhaust System
Inlet & Exhaust Requirements
• Minimal Size & Weight; Minimal/No Moving Parts
• Aggressive, Exotic, Conformal Shaping
• Uncompromised Functionality & Aero-Efficiency

FIGURE 2 Advanced combat aircraft must retain high aeroperformance while emphasizing affordability and stealth. Source: 2002 by Lockheed Martin. Published with permission.

government, at LM, at other companies, and at universities are all investigating an emerging technology, AFC, to address these problems. AFC is defined as the ability to control large-scale aerodynamic flow phenomena with very small-scale (or microscale) perturbations to the flow field near the wall.

FLOW CONTROL APPLIED TO THE ENGINE INLET SYSTEM

One application for AFC is in an advanced inlet system that features conformal shaping and extreme serpentine wall curvature (Figure 2) (Anderson et al., 1999; Bender et al., 1999; Hamstra et al., 2000). With current technology, high-speed flow entering the inlet duct is unable to negotiate the extreme internal wall curvature and thus detaches or "separates" from the wall. This behavior, which has its genesis in the very thin "boundary layer" of air next to the wall, results in massive pressure loss and flow distortion at the inlet/engine interface. These flow field characteristics greatly reduce net thrust and, sometimes, even result in engine stall. AFC can be used to energize/restructure the boundary layer in a way that would prevent flow separation and greatly diminish distortion. Key design considerations include using the proper type of actuation/energization device and identifying the proper "receptive zones" for device placement because even very small perturbations near the wall may cause a global change in the entire flow field.

Numerous design variables must be properly chosen in designing an AFC actuation system. These variables include physical size, orientation, location, and the number of actuators near the inlet wall. Because the size of the AFC design space is so large, researchers at LM have helped pioneer the use of a coupled design of experiments (DOE)/computational fluid dynamics (CFD) process for optimization. This process uses CFD to evaluate each element in the design matrix and repetitive applications of DOE methods to search through the design space. The process is occasionally checked and validated through testing of large-scale (~30 percent) models.

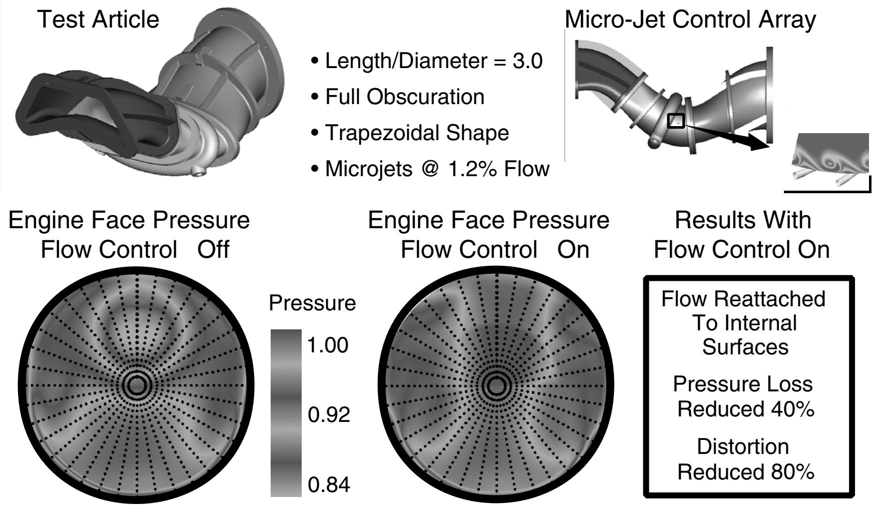
A joint LM/NASA Glenn Research Center (GRC) team has conducted several validation tests at the GRC W1B test facility. The tests featured models fabricated from resin using a "rapid prototyping" laser stereo-lithography process. The results of one test (Figure 3) show that, without flow control, the inlet produces zones of massive pressure loss and resultant high distortion, yielding a pressure contour pattern unacceptable for turbofan engine operation. With flow control, pressure losses were decreased by 40 percent, distortion was decreased by 80 percent, and an acceptable pattern was produced. These tests demonstrate the viability of flow control under realistic conditions on a relevant, large-scale inlet-system configuration.

Continuing research is focused on making the inlet flow-control suite robust across the range of maneuver, speed, and airflow settings envisioned for future inlet systems. Control schemes that incorporate distributed feedback sensors and reactive, closed-loop control logic are also under study. The goal of this research is to produce a flow-control system that allows unprecedented freedom in inlet design while simultaneously optimizing engine inflow conditions across the entire operating range of the aircraft.

FLOW CONTROL APPLIED TO THE ENGINE EXHAUST SYSTEM

A second application for AFC is in the engine exhaust system. Modern exhaust systems incorporate significant complexity to provide thrust vectoring and jet area control (Bender et al., 2000; Miller and Catt, 1995; Miller et al., 1997, 1999, 2001; Vakili et al., 1999; Yagle et al., 2001). These mechanical subsystems incur weight and cost impacts to the vehicle and limit the exhaust system to shapes that can be easily and efficiently mechanized, namely, simple round designs or simple rectangular designs. With AFC, researchers hope to design systems that achieve the same functionality without mechanical flowpath variations, thus reducing cost and weight while enabling more exotic cross-sectional shapes that can conform to the body of an advanced aircraft.

The fundamental approach to flow control in the exhaust system is shown in Figure 4. High-pressure air is injected into the nozzle's divergent section with



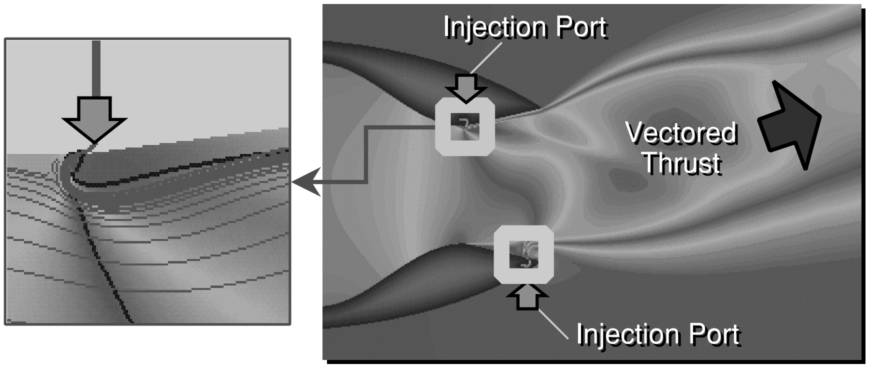
- Length/Diameter = 3.0
- Full Obscuration
- Trapezoidal Shape
- Microjets @ 1.2% Flow

Flow Reattached To Internal Surfaces

Pressure Loss Reduced 40%

Distortion Reduced 80%

FIGURE 3 Inlet flow control has been demonstrated through large-scale aerodynamic testing. Source: 2002 by Lockheed Martin. Published with permission.



- Symmetric Injection To Control Jet Flow Area
- Asymmetric Injection To Control Thrust Vector Angle

FIGURE 4 Flow control creates "virtual surfaces" to provide classic nozzle functionality without mechanical variation. Source: 2002 by Lockheed Martin. Published with permission.

the goal of creating "virtual aerodynamic surfaces" that provide the same functionality as current variable-geometry mechanical flaps. By injecting symmetrically about a given axial location, the minimum flow area of the jet plume can be changed, thereby allowing for changes in the engine power setting. By injecting asymmetrically, the direction of the jet plume can be changed, thereby allowing for thrust vectoring.

The process used to develop the exhaust system flow-control suite is identical to the one used for the inlet system. DOE methods are used to search through the design space, and CFD is used to evaluate the thrust coefficient and vectoring capability of each design element. Large-scale tests are occasionally conducted to validate the design process.

OTHER APPLICATIONS

AFC technology is also under development for a number of other air-vehicle system applications. These applications include the control and stabilization of aircraft wakes in the vicinity of directed-energy weapons; augmentation or replacement of conventional aircraft-control effectors, such as trailing edge flaps; and suppression of acoustic loads for internal weapons bays.

SUMMARY

AFC is an emerging technology that could enable a breakthrough in traditional aerodynamic design limitations for a wide range of advanced combat aircraft systems. For the engine inlet and exhaust systems, flow-control technology has the potential to enable unprecedented freedom to incorporate exotic flowpath designs, enable optimization of engine inflow conditions regardless of aircraft condition, and provide superior exhaust system functionality with reduced weight and cost.

REFERENCES

Anderson, B., D. Miller, P. Yagle, and P. Truax. 1999. A Study on MEMS Flow Control for the Management of Engine Face Distortion in Compact Inlet Systems. ASME Paper No. FEDSM99-6920.

Bender, E., B. Anderson, and P. Yagle. 1999. Vortex Generator Modeling for Navier-Stokes Codes. ASME Paper No. FEDSM99-69219.

Bender, E., D. Miller, B. Smith, P. Yagle, P. Vermeulen, and S. Walker. 2000. Simulation of Pulsed Injection in a Crossflow Using 3-D Unsteady CFD. AIAA Paper No. 2000-2318.

Hamstra, J., D. Miller, P. Truax, B. Anderson, and B. Wendt. 2000. Active inlet flow control technology demonstration. Aeronautical Journal 104(1040):473–480.

Miller, D., and J. Catt. 1995. Conceptual Development of Fixed-Geometry Nozzles Using Fluidic Throat-Area Control. AIAA Paper No. 95-2603.

Miller, D., J. Catt, and S. Walker. 1997. Extending Flow Control of Fixed Nozzles Through Systematic Design: Introducing Assisted Reinjection. ASME Paper No. FEDSM97-3680.

Miller, D., P. Yagle, and J. Hamstra. 1999. Fluidic Throat Skewing for Thrust Vectoring in Fixed-Geometry Nozzles. AIAA Paper No. 99-0365.

Miller, D., P. Yagle, E. Bender, and P. Vermeulen. 2001. A Computational Investigation of Pulsed Injection into a Confined Expanding Cross Flow. AIAA Paper No. 2001-3026.

Vakili, A., S. Sauerwein, and D. Miller. 1999. Pulsed Injection Applied to Nozzle Internal Flow Control. AIAA Paper No. 99-1002.

Yagle, P.J., D.N. Miller, K.B. Ginn, and J.W. Hamstra. 2001. Demonstration of fluid throat skewing for thrust vectoring in structurally fixed nozzles. Journal of Engineering for Gas Turbines and Power 123(3):502–507.

Miniature Spy Planes: The Next Generation of Flying Robots

STEPHEN J. MORRIS
MLB Company
Palo Alto, California

INTRODUCTION

Affordable airborne access to information has been on the wish list of the military, businesses, and individuals for a long time, but nearly 100 years after the Wright brothers' flight at Kitty Hawk this dream has not yet come true. Anyone who wants to check out events on the ground at a particular moment still needs a full-size airplane or helicopter, a pilot's license, and a nearby airfield. The barriers to obtaining information from airborne platforms are the technological difficulties of controlling and navigating the aircraft and the remote sensing of information. In the last three decades, advancements in digital computers, satellite position sensing, solid-state inertial sensing, and video imaging have made possible the first generation of small, affordable, robotic aircraft that require only moderate operator skills and can provide useful image data. These vehicles typically have a wingspan of less than 6 feet and weigh less than 10 pounds, so they pose minimal safety hazards to the public. Because they fly at low altitudes (less than 500 feet), they operate below the airspace where full-size aircraft operate and are, therefore, allowed to fly freely within the framework of local city ordinances.

A typical miniature unmanned air vehicle (mini-UAV) has a quiet engine and is difficult to spot in the air. It can cruise the skies day or night in a wide range of weather conditions gathering information for law enforcement, traffic monitoring, air-pollution control, farming, fire spotting, power line inspection, search and rescue, and weather monitoring. In the near future, individuals will be able to log on to the Web site of a mini-UAV service and request image data for a specific region. This request will prompt teams of mini-UAVs to fly to the appropriate location, gather the requested data, and send the data to the customer over the Internet for a reasonable price. The mini-UAVs doing the work will be all but invisible.

In this paper, I will describe some successful mini-UAV designs, the technical challenges of miniaturizing UAVs, and future applications.

BACKGROUND

The military has used UAVs to gather intelligence since World War II. In the Vietnam War, drones built by Teledyne Ryan flew regular reconnaissance missions gathering valuable photographic data under autonomous flight control. The Israeli airforce pioneered the use of smaller, low-cost drones for reconnaissance and decoy missions. In all of these examples, UAVs were flown in the same airspace as man-carrying aircraft, which was necessary for military reasons.

The Federal Aviation Administration (FAA) has not yet decided how UAVs will be incorporated into civilian airspace, which has kept their commercial use in this country to a minimum. Mini-UAVs offer a solution to this problem because they can operate safely at altitudes below the operating altitude of full-size air traffic and, because of their size, speed, and weight, they pose little threat to the public.

Until the introduction of the Global Positioning System (GPS) in the 1980s, UAVs required expensive inertial guidance systems that were often adapted from man-carrying aircraft or long-range missiles. Thus, UAVs were large, costly aircraft only the military could operate. Most of the current military UAVs are still large and costly, primarily because they are designed for long-range, high-altitude missions and because they carry heavy mil-spec sensor payloads.

MINI-UAVS

In the last few years, mini-UAVs have become effective surveillance systems. A few examples of notable mini-UAVs are described below.

Pointer

In the late 1980s, AeroVironment, Inc., of Simi Valley, California, developed a small, low-cost, remotely piloted drone for the Marines that could be carried in two backpacks and could be flown with moderate pilot skills. With a 9-foot wingspan, Pointer was the first true mini-UAV system to be commercially produced (Figure 1). Pointer carries a forward-looking color camera and is powered by an electric motor that provides up to 90 minutes of operation with present lithium battery technology.

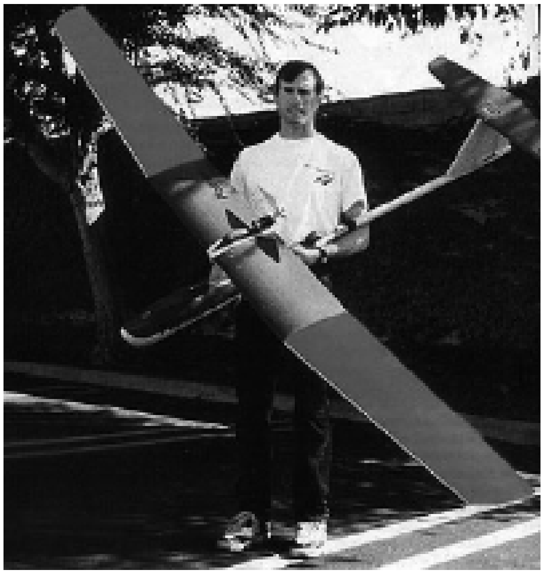


FIGURE 1 Pointer UAV by AeroVironment. Source: AeroVironment, Inc., Simi Valley, California.

Microair Vehicles

In 1997 the Defense Advanced Research Projects Agency funded the development of reconnaissance microaircraft (i.e., the largest dimension of 6 inches). Lockheed-Sanders and AeroVironment have studied these microair vehicles (or MAVs), and each has developed a successful flying example. The typical MAV mission requires a modest duration of 30 minutes, a range of 1 mile, and operation in winds up to 20 mph. AeroVironment's Black Widow (Figure 2) has a 6-inch wingspan, uses an electric motor for propulsion, and weighs 80 grams (Grasmeyer and Keennon, 2001). The Lockheed-Sanders MicroStar MAV has a 12-inch wingspan, uses electric propulsion, and is capable of autonomous flight with GPS navigation. Both MAVs carry color cameras and use lithium batteries for propulsion. Black Widow and MicroStar have demonstrated the possibility of significant miniaturization in reconnaissance aircraft, as long as mission duration and range remain short.

Aerosonde

In 1998, a 30-pound aircraft with a 9-foot wingspan flew 2,500 miles in 27 hours across the Atlantic Ocean under autonomous control, mimicking the flight

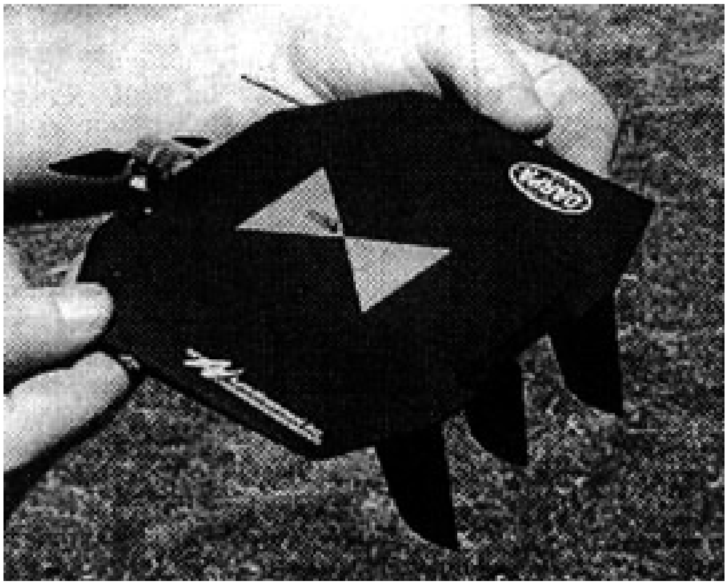


FIGURE 2 Black Widow by AeroVironment. Source: AeroVironment, Inc., Simi Valley, California.

made by Charles Lindbergh 51 years earlier. Aerosonde, by Aerosonde Ltd., demonstrated that the flight performance of smaller size UAVs need not be restricted (Figure 3). Using a modified model-aircraft engine, composite model-airplane wings, and GPS navigation, Aerosonde's transatlantic flight required only 1.5 gallons of gasoline.

Bat

The low-cost Bat by MLB (Figures 4 and 5a,b) is designed to operate in populated areas, can be easily transported by a single person, and delivers high-quality image data. Bat has a 5-foot wingspan, weighs 9 pounds, and folds up for transport. The entire aircraft and ground station fits easily into the trunk of a car and can be readied for flight in 10 minutes.

Using a miniature flight-control system (Figure 6), Bat can fly autonomously between specified points and can take off and land unaided. The aircraft can operate for one hour and can transmit video and flight data over a 2-mile radius. The sensor package is a 3-axis gimbal mount with two video cameras, each with a different field of view lens that can be switched in flight for closer views of desired areas. The gimballed camera is inertially stabilized using the flight-

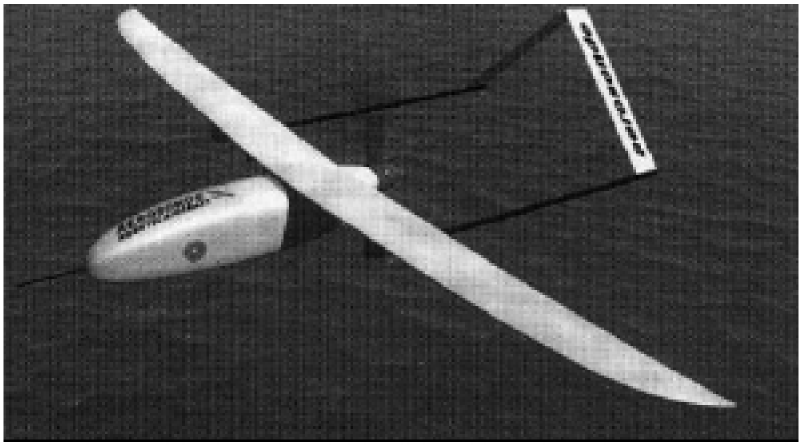


FIGURE 3 Aerosonde by Aerosonde Ltd. Source: Shephard's Unmanned Vehicle Systems Handbook 2001 (Shephard Press, 2000).

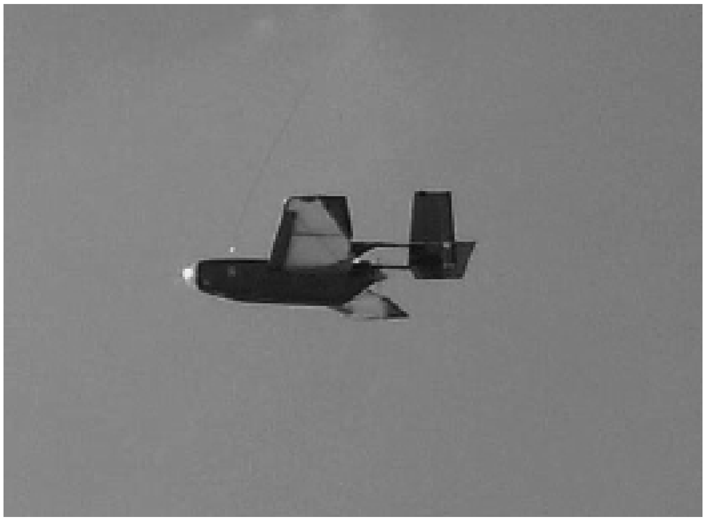


FIGURE 4 Bat by MLB. Source: Stephen Morris, MLB Company, Palo Alto, California.

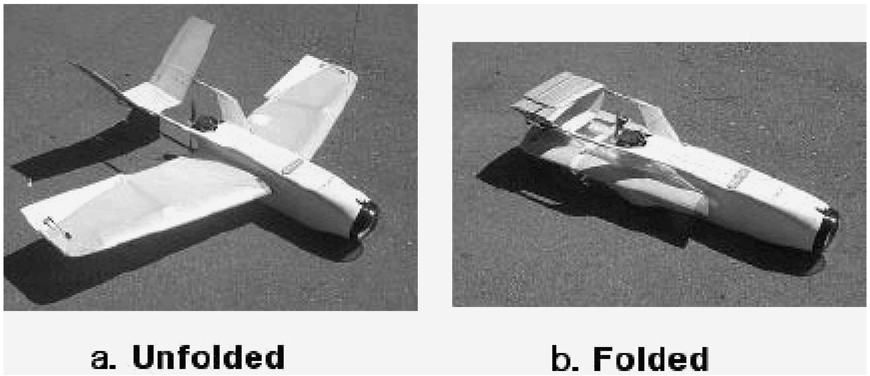


FIGURE 5 Bat by MLB shown (a) with wings open and (b) with wings folded to 4-foot size. Source: Stephen Morris, MLB Company, Palo Alto, California.

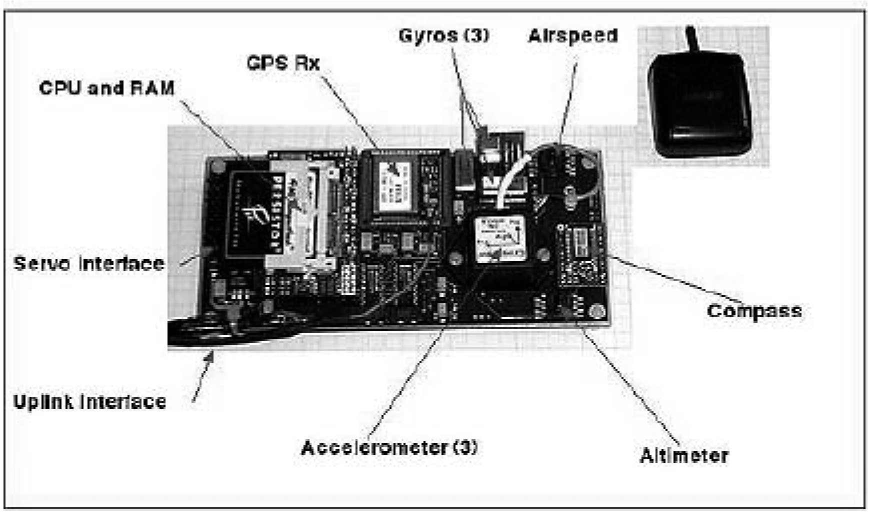


FIGURE 6 Miniature flight-control hardware. Source: Stephen Morris, MLB Company, Palo Alto, California.

control computer, and the UAV operator can aim the camera remotely. In the event of a structural failure or a failure of the flight-control system, a separate system automatically deploys a parachute to slow the aircraft to a safe descent speed.

A ground station based on a PC laptop makes the Bat mini-UAV easy to operate. Figure 7 shows a screen snapshot of the display, which includes Bat's track on a moving map, altitude, speed, system status, and other flight data parameters. The geographic location of the image being viewed by the camera is also shown and recorded so that all images are georeferenced. The operator specifies flight plans by clicking on the map and inserting new way-points. Once the aircraft is launched, it follows the course defined by the way-points at the speed and altitude specified for each leg of the course.

Bat has been used in research projects on wildlife habitat to provide image mosaics of wetland areas and to deploy a miniature sensor network that transported data back to a remote ground station via Bat. Figure 8 shows Bat dropping sensors along a roadway during a military demonstration at 29 Palms Marine Base in California. The sensors, developed by Professor Kris Pister at the University of California at Berkeley (Kahn et al., 1999), are designed to form data networks automatically, detect moving vehicles, and relay the data back to the UAV.

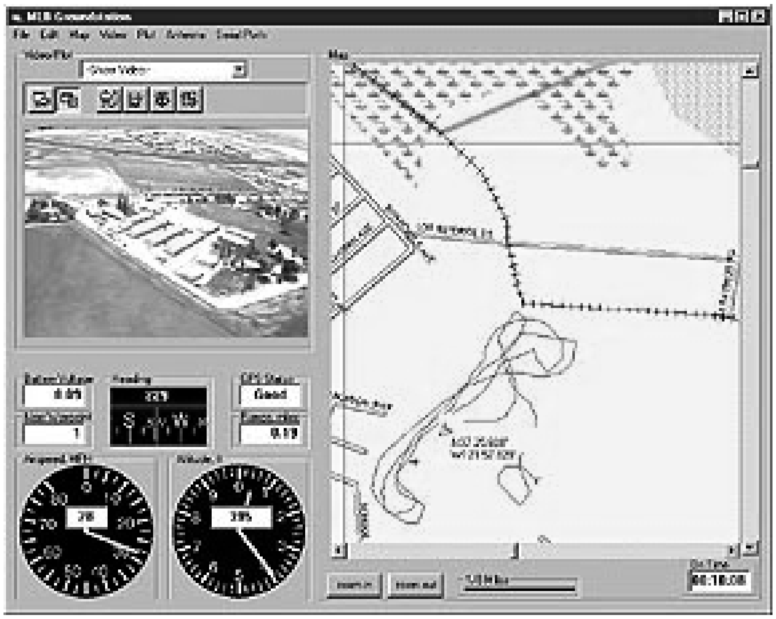


FIGURE 7 Ground station screen image. Source: Stephen Morris, MLB Company, Palo Alto, California.

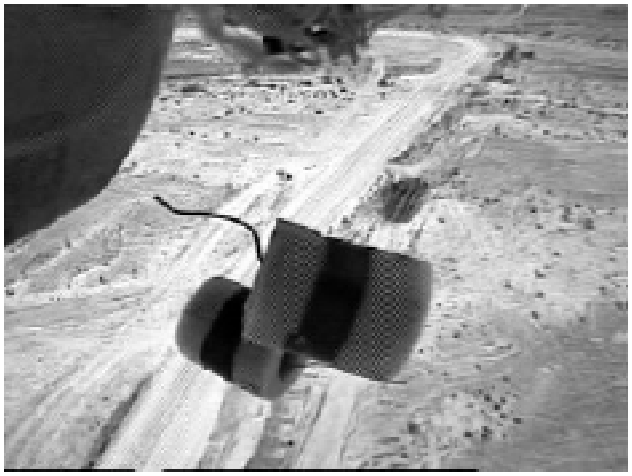


FIGURE 8 Bat deploying microsensors along a roadway. Source: Stephen Morris, MLB Company, Palo Alto, California.

SCALING ISSUES AND TECHNICAL CHALLENGES

How small can a UAV be? Nature provides some readily observable answers to this question. Small birds are capable of sensing images and can migrate thousands of miles. Even insects can operate in a radius of several miles. Logically, surveillance aircraft could be just as small if engineers could design systems as adaptable and capable as the systems found in nature. The major technical challenges facing engineers in scaling UAVs to smaller sizes are discussed below.

Aerodynamics

As wings become smaller, the nature of the airflow over them changes, because the interaction between the wings and air particles scales with size, speed, and air viscosity. The Reynolds number, a dimensionless parameter that characterizes the flow regime over bodies, involves the ratio of inertial to viscous forces in a fluid flow. As the Reynolds number decreases, the viscous effect becomes dominant, and it becomes difficult to generate lift while maintaining low drag. Many birds, insects, and model aircraft operate at a low Reynolds number (approximately 200,000 to 20,000) and still perform adequately for their purposes. At a very low Reynolds number, the best strategy for generating lift with low-energy input shifts from smooth, shaped airfoils to rougher wing surfaces and unsteady movement through the fluid. The wing motion of small insects changes rapidly to generate unsteady vortices in the viscous flow, producing lift with minimal energy.

Initially, researchers believed that Reynolds number effects would produce the greatest challenge to miniaturizing UAVs, but aircraft like Black Widow and Aerosonde have demonstrated that a high level of performance can be obtained with proper aerodynamic design. The impact of increased drag, indicated by a low Reynolds number, on MAV performance was quantified by Morris (1997), who showed that it was much less important than propulsion efficiency and lift-generating capability. For mini-UAVs, a lift to drag ratio of 20:1 could be achieved, if necessary, for long-range, long-duration flights.

Propulsion

As the weight of UAVs decreases to less than 20 pounds, the options for efficient propulsion systems are also reduced dramatically. Modified model-aircraft engines that use methanol or gasoline as fuel are popular on mini-UAVs, but they are often inefficient and unreliable. Nevertheless, because fossil fuels have a high-energy density, these engines are still useful for most mini-UAV missions. However, efficient internal combustion engines of less than one horsepower have not been fully researched, and much progress could be made in this area. Electric power has been used successfully for short-range, short-duration aircraft, such as MAVs, but has been limited by the low-energy density of present battery technologies. Batteries with the highest energy densities (e.g., lithium) often have limited power density, which has further complicated the design of electric-powered UAVs.

Flight Control and Sensing

Birds and insects have evolved complex brains and musculature and can process a great deal of sensory input. Living creatures that fly have distributed actuator, sensor, and computation capabilities, which make them very agile, efficient, and adaptable. The agility of mini-UAVs is often limited because extremely small, lightweight, high-quality sensors and actuators have not yet been developed. Scaling to smaller sizes compounds the problem of flight control because the dynamics of the aircraft increase in frequency as size decreases, in the same way a small pendulum has a higher natural frequency than a large one. Therefore, mini-UAVs require higher bandwidth actuators than their larger counterparts. Recent advances in microelectromechanical systems (MEMS) technology have produced microscopic sensors (e.g., gyroscopes, accelerometers, pressure transducers, etc.) that are highly suitable for mini-UAVs. If advances in MEMS continue, an entire flight avionics system will soon be available on a single chip. This would greatly reduce the weight and volume of flight avionics, which tend to be a higher percentage of the total as UAV size shrinks. MEMS devices could also provide the bandwidth and accuracy necessary for flight-control sensing and computing, which would make mini-UAV flight much more agile.

Telemetry

The power required to transmit data varies with the square of the distance between the transmitter and the receiver, independent of the size of the aircraft sending the data. Therefore, one of the greatest scaling challenges for mini-UAVs is sending data over great distances without requiring excessive power (and weight). Possible solutions to this problem include high-gain antennas, reduced data rates, and burst-transmission communication schemes. In general, telemetry range is reduced as UAV size decreases unless the vehicle is linked to a communication network. In a recent demonstration, Bat was used to “truck” data from a deployed sensor network back to a remote location. This eliminated the need for either system to have a long-range telemetry link and mimicked the solution, often seen in nature, of storing the data and retransmitting at close range.

Data Quality

Most mini-UAVs carry a fixed camera as the primary image sensor. Flight tests have shown that image quality is degraded by aircraft motion caused by maneuvering or wind gusts. Imaging could be improved with an inertially stabilized servocontrolled gimbal camera mount that adjusts for aircraft motion and allows the operator to position the field of view. Small gimbal camera systems are being developed for the latest mini-UAVs.

Data quality is closely linked to the postprocessing of the image data collected by UAVs. Raw video images must be adjusted for camera alignment, combined into larger image maps (mosaics), and possibly have features identified and extracted. The data may then be combined with other databases (e.g., satellite data, digital maps, etc.) to maximize their value. Currently, little commercial software has been developed to process the vast amount of video data that would be generated by fleets of mini-UAVs. The full potential of UAV fleets will only be realized when the data-fusion bottleneck is eliminated.

Complete UAV System

UAV size is further influenced by the ground-based systems needed to operate them. As an example, consider a UAV so small that it cannot transmit good-quality data over great distances. For acceptable telemetry range, a large directional antenna would be required. Thus, the complete system might be bigger than for a larger UAV carrying a more powerful data transmitter. When transportation and cost are factored in, system size can be more important than UAV size. Realizing the smallest system size will require that the UAV and the ground-support equipment be considered simultaneously in the design process.

FUTURE APPLICATIONS

An example of a near-term commercial application for mini-UAVs is as an aid to precision agriculture. The amount of water, insecticide, and fertilizer used in agriculture is highly regulated and must be dispensed efficiently. High-value crops, such as strawberries and vineyards, require almost daily monitoring at specific points in the growing season to ensure a high quality and high yield per acre. Mini-UAVs could deliver the higher quality data more efficiently and at lower cost than satellites and light aircraft. Image data taken in the near-infrared and color spectra could be processed to make vegetation growth-index maps that show where fertilizer, water, or insecticide is needed. In the future, automated farming operations could use the data gathered by mini-UAVs to direct unmanned robotic tractors and harvesters for optimal crop management.

Current satellite imagery is limited to 10-meter pixel resolution, and images are affected by cloud cover and satellite trajectories. Because mini-UAVs fly at low altitudes, they could supplement satellite image databases with high-resolution imagery collected during periods of cloud cover or whenever desired.

Mini-UAVs will surely evolve into increasingly capable (and almost unnoticeable) surveillance platforms that will be economical to operate and safe to use over populated areas. Teams of mini-UAVs will be able to communicate with each other and organize themselves for optimal data gathering. In the not too distant future, mini-UAVs will become an essential part of information gathering systems that can supply near-real-time data to customers through the Internet. These transportable, inexpensive aircraft will also be used in remote locations that are currently too costly to monitor.

REFERENCES

Grasmeyer, J.M., and M.T. Keennon. 2001. Development of the Black Widow Micro Air Vehicle. AIAA Paper No. 2001-0127.

Kahn, J.M., R.H. Katz, and K.S.J. Pister. 1999. Next century challenges: Mobile networking for smart dust. Pp. 271-278 in Proceedings of the Fifth Annual ACM/IEEE International Conference on Mobile Computing and Networking. New York, N.Y.: ACM Press.

Morris, S. 1997. Design and flight test results for micro-sized fixed-wing and VTOL aircraft. Presented at the First International Conference for Emerging Technologies of Micro Air Vehicles, Atlanta, Ga., February 19–20, 1997. Available online at <http://www.spyplanes.com/Background/MAV_97/Mavpaper.htm>.

Shephard Press. 2000. Shephard's Unmanned Vehicle Systems Handbook 2001. Bucks, England: The Shephard Press.

Toward Micromechanical Flyers

RONALD S. FEARING
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley

In this paper I describe some of the microflyers being developed around the world and the engineering challenges they present. Advances in the late 1980s in microelectromechanical systems (MEMS), especially the surface micromachined electrostatic micromotor of Fan et al. (1988), inspired roboticists to think small. In 1987, Flynn proposed building inexpensive, disposable, autonomous milligram (mg)-mass microrobots that could be deployed in massive swarms. In my back-of-the-envelope calculations, the cost per unit would be $1.00 to $10.00, perhaps two to three orders of magnitude cheaper than conventional trash-can-size mobile robots. In this new paradigm, huge numbers of moderately intelligent and robust microrobots could perform tasks, such as searching an area, more effectively than a single expensive macrorobot.

Research on mobile microrobots is being done on legged devices (Eberfors et al., 1999; Yeh et al., 1996) and flying devices (Crary et al., 1992; Shimoyama et al., 1994). Legged devices can be made statically stable, and hence easy to control, but path planning can be difficult for an ant-sized device with six 5-millimeter (mm)-high legs. Flying devices have much more difficult power and control requirements, but avoiding obstacles while flying around a room seems easier than negotiating a shag carpet with six 5-mm-high legs.

For high speed, flying is the way to go. Small flying insects, such as the hoverfly, can reach peak speeds of 10 meters per second (about 1,000 body lengths per second). Rotary or flapping wings provide hovering capability but drastically increase power requirements. For example, the flyer reported in Flynn (1987) was a fixed-wing device with a 12-centimeter wingspan that weighed 80 mg. The device required only 5 watts per kilogram (W/kg) to fly using a propeller without flapping its wings. (It should be noted that this device was powered by a rubber band and had no control electronics.) By comparison, flapping flyers require 100 to 200 W/kg. The extra power of flapping flight greatly improves maneuverability.

Perhaps surprisingly, small, subcentimeter flyers may turn out to be easier to construct than larger flyers, not because the aerodynamics are easier, but because the actuator power density increases with higher operating frequencies. In addition, the surface area-to-volume ratio improves for small devices. Thus, small flyers might be driven by solar cells. With piezoelectric actuators running in resonance at the wing beat frequency, power density is proportional to frequency. In addition, at smaller scales, one can generally avoid conventional joints and bearings, which are heavy, and use flexural joints, which are lighter and scale down well in size.

Designing a micromechanical flyer, a device with maximum dimensions of 25 mm and a mass of 100 mg, for example, is a challenge on many fronts—aerodynamics, actuation, transmission, power supply, sensing, control algorithms, compact low-power electronics, and flight behavior. As far as I know, no flyer smaller than the Caltech microbat has flown freely. Only the 10-gram Caltech microbat ornithopter (Figure 1) has flown under its own power with passive stabilization.

ORNITHOPTER MICROAIR VEHICLES

Several groups, notably Caltech, SRI, and Vanderbilt, have developed components for small bird-size ornithopters. Interestingly, all of these designs use only 1 degree of freedom wings and rely on passive or coupling mechanisms to control wing rotation. The Caltech design (Figure 1) uses a standard direct

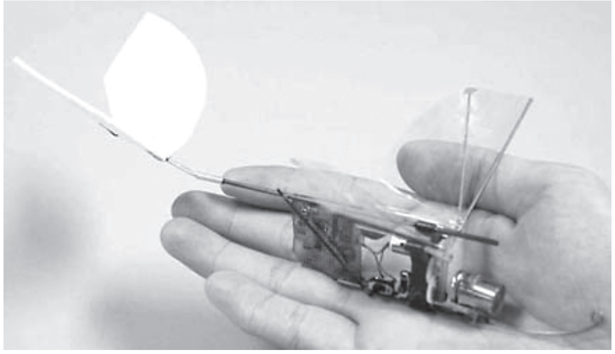


FIGURE 1 Caltech microbat. Source: Reprinted with permission from Elsevier Science (Pornsin-Sirirak et al., 2001).

current (DC) motor and gear box. The SRI device uses electrostrictive polymer actuators (Figure 2). The Vanderbilt device uses piezoelectric actuators (Figure 3).

ROTARY AND FLAPPING WING MICROFLYERS

Based on our macroscale experience, we assume that a rotary wing device with fixed angle of attack would be easier to fabricate than a beating wing device. We are right—to a point. The assumption doesn't hold up, however, when we reach the lower limit of magnetic motor size, which is currently 1.9-mm diameter, 5.5-mm length, and 91-mg mass (Faulhaber Group, 2001). Mock-ups of helicopters using these motors have been constructed by Institut für

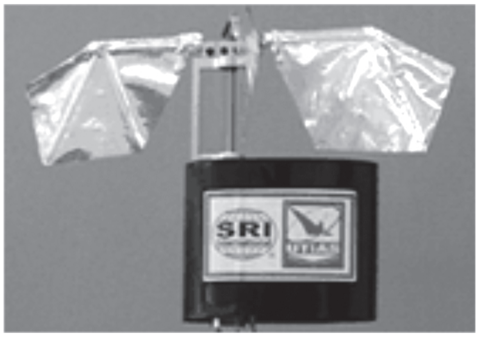


FIGURE 2 SRI flapper. Source: Reprinted with permission from SRI International (SRI, 2001).

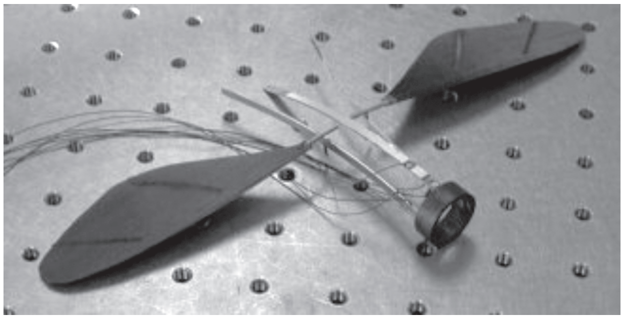


FIGURE 3 Vanderbilt University mesoscale flying robot. Source: Reprinted with permission from The American Society of Mechanical Engineers (Cox et al., 1999).

Mikroteknik Mainz in Germany (Figure 4a) and Stanford University (Figure 4b). Keeping the mass of the device under 100 mg will require more compact actuators. I expect that shrinking bearings further to create an efficient, high energy density, submillimeter magnetic motor will be difficult. Thus, in terms of miniaturization, one is encouraged to think of beating wings rather than rotary wings. Flapping wings can change the direction of applied torques in a wing beat, potentially improving maneuverability. A mock-up of the Berkeley micromechanical flying insect (MFI) is shown in Figure 5.

AERODYNAMICS

As a design target for the MFI, we chose the blowfly (*Calliphora*), which has a mass of 100 mg, a wing length of 11 mm, a wing beat frequency of 150 Hz, and actuator power of about 8 milliwatts (mW). At this size scale, our current

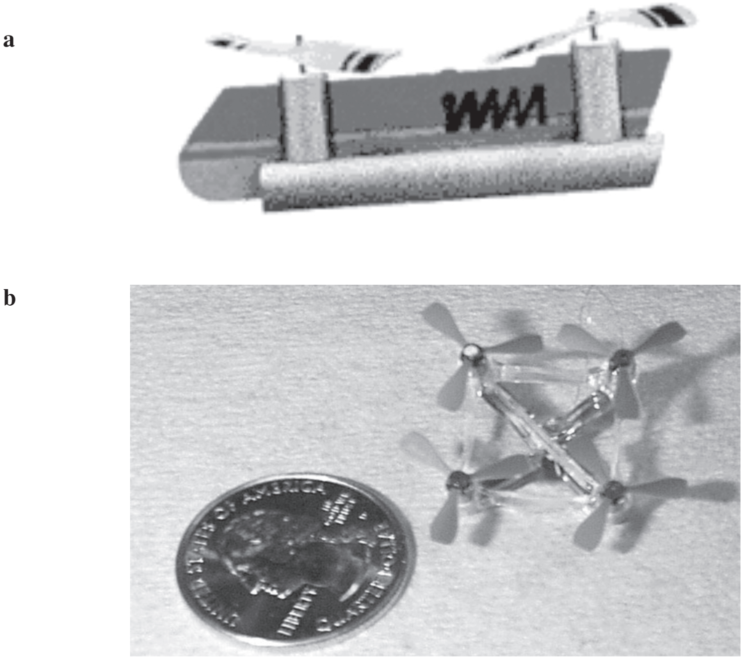


FIGURE 4(a) Institut für Mikroteknik Mainz's uncontrolled helicopter, length 24 mm, mass 0.4 grams. Source: Reprinted with permission from the Institut für Mikroteknik Mainz (IMM, 2001). **(b).** Mockup of the Stanford Mesicopter with four 1.5-cm rotors and a mass of 3 grams. Source: Reprinted with permission from Ilan Kroo (Department of Aeronautics and Astronautics, Stanford University, 2001).

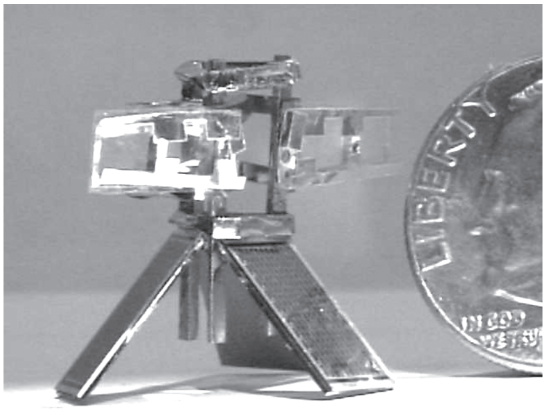


FIGURE 5 Mock-up of the Berkeley micromechanical flying insect (wingspan 25 mm and final mass target 100 mg).

understanding of nonsteady-state aerodynamics comes from experimental observations of real insects and kinematically similar mock-ups (Dickinson et al., 1999; Ellington et al., 1996).

The Robofly apparatus (Dickinson et al., 1999) consists of a two-winged system driven by three stepping motors, which can closely mimic the stroke kinematics of a fruit fly (*Drosophila*) or other arbitrary kinematics. Strain gauges are used to measure instantaneous wing forces, and the integral of forces around a closed wing beat cycle can be measured to determine net flight forces. Robofly running with a wing beat of 1/6 Hz in oil has the same Reynolds number as a *Drosophila* with a wing beat of 220 Hz in air. Flow was visualized using air bubbles in the oil tank and particle image velocimetry.

The Robofly apparatus has enabled Dickinson and colleagues to identify the three key aerodynamic mechanisms used by insects: delayed stall and wake capture (Figure 6) and rotational circulation (Figure 7). Dickinson et al. (1999) found wing trajectories that generate peak lift forces of four times the equivalent insect weight. Due to rotational lift, the timing of an equivalent of a backspin motion at the bottom of the wing stroke can change the net lift from positive to negative. The second key mechanism is the significant force generated by wake capture at the top and bottom of the wing stroke. Applying these results to the wing kinematics of the MFI, we realized that a rapid wing rotation of 90 degrees is necessary before the end of the downstroke to create adequate lift.

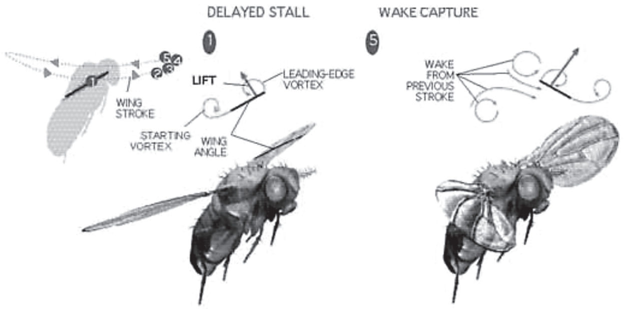


FIGURE 6 Delayed stall and wake capture effects. Source: Reprinted with permission from Bryan Christie (Dickinson, 2001).

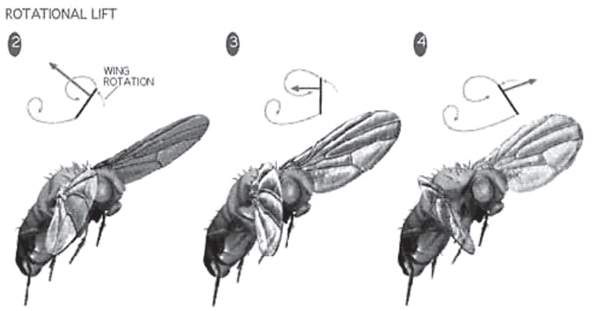


FIGURE 7 Lift due to wing rotation and ends of stroke. Source: Reprinted with permission from Bryan Christie (Dickinson, 2001).

THORAX DESIGN

We know that insect flight at the centimeter scale requires both large stroke amplitude and large wing rotation (Dickinson et al., 1999). *Drosophila* has a wing stroke of 160 degrees combined with wing rotation of more than 90 degrees (*Calliphora* has similar kinematics). Wing rotation is the challenging part of the design. The insect thorax has a complicated arrangement of linkages and cams, which is not yet fully understood and is likely to be too difficult to replicate (Nachtigall et al., 1998). The electromechanical design of the thorax poses some interesting challenges. The actuators combined with the wing transmission should weigh less than 50 mg and provide 10 mW of mechanical power to a wing being driven in flapping and rotation at 150 Hz with large amplitudes.

For actuators, we use piezoelectric unimorphs, which have a DC displacement of about 1 degree. Two stages of mechanical amplification using planar fourbars bring the output motion at DC to 50 degrees. Separate actuators drive the leading and trailing edges of a differential assembly, as shown in Figure 8a,b. The thorax is designed to be run in resonance with a quality factor (Q) of 2.5. A higher Q would reduce the response speed of the wing to controlling rotation; a lower Q would require an even higher transmission ratio. It turns out that even a very low wing inertia is not sufficient for proper operation; wing inertia ratios must be chosen for dynamic decoupling of the differential. The overall thorax

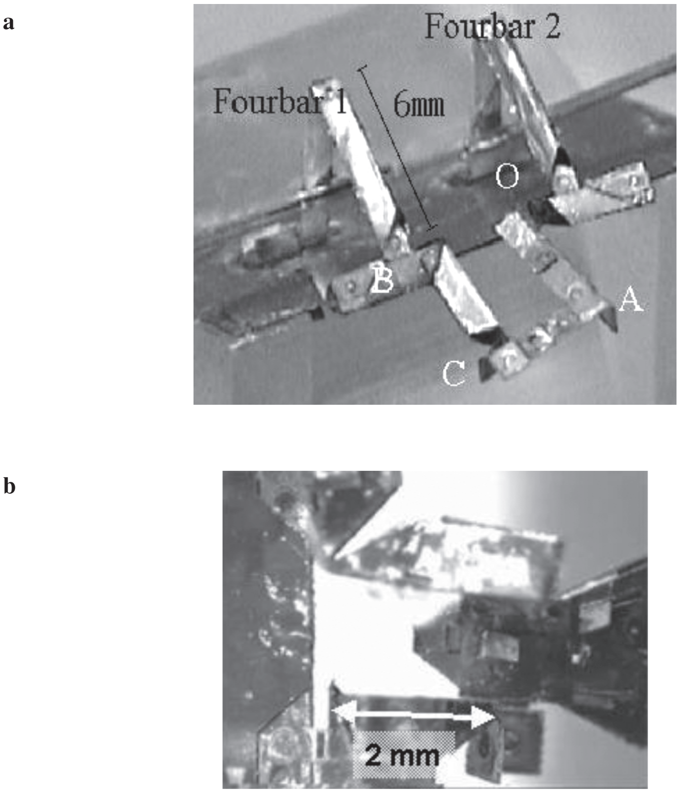


FIGURE 8(a) MFI thorax for driving one wing consisting of a pair of fourbars driving a wing differential mechanism, OACB. **(b)** Detail of compact, low-inertia, wing differential. Source: Reprinted with permission from Yan et al., 2000. Copyright 2001 by IEEE.

kinematics is a closed chain manipulator with 17 joints and 2 degrees of freedom. For light weight and strength, the structure is assembled from 12.7 micron stainless steel sheet folded into hollow beams with polyester flexure joints between links.

SUMMARY

Research on micromechanical flyers has led to an understanding of how to generate high-frequency, high-amplitude wing motions in a low-mass, compact device. The challenges ahead will be to integrate sensing and control devices into this inherently very unstable, but potentially high-performance, flying device. Taking our inspiration from real insects and our tools from MEMS, we will work on integrating optical flow and gyroscopic sensing on the MFI to control attitude and bring the device closer to its first free flight.

ACKNOWLEDGMENTS

The Berkeley MFI research summarized here is supported by the Office of Naval Research, DARPA, and the National Science Foundation. The contributions of the interdisciplinary MFI research team are gratefully acknowledged.

REFERENCES

Cox, A., D.J. Monopoli, and M. Goldfarb. 1999. Development of piezoelectrically actuated elastodynamic flapping micro-aerial vehicles. Pp. 257–262 in ASME International Mechanical Engineering Congress and Exhibition. AD-Vol. 59/MD-Vol. 87 Adaptive Structures and Materials Systems. New York: ASME.

Crary, S.B., G.K. Ananthasuresh, and S. Kota. 1992. Prospects for microflight using micromechanisms. Pp. 273–276 in Proceedings of the International Symposium on Theory of Machines and Mechanisms, International Federation of Theory of Machines and Mechanisms-Japan Council, Nagoya, Japan, September 24–26, 1992.

Dickinson, M. 2001. Solving the mysteries of insect flight. Scientific American 284(6):49–57.

Dickinson, M.H., F.-O. Lehmann, and S.P. Sane. 1999. Wing rotation and the aerodynamic basis of insect flight. Science 284(5422):1954–1960.

Eberfors, T., J. Mattson, E. Kalvesten, and G. Stemme. 1999. A walking silicon micro-robot. Pp. 1202–1205 in 10th International Conference on Solid-State Sensors and Actuators (TRANSDUCERS '99). New York: Elsevier.

Ellington, C.P., C. van den Berg, A.P. Willmot, and A.L.R. Thomas. 1996. Leading edge vortices in insect flight. Nature 384:626–630.

Fan, L.-S., Y.-C. Tai, and R.S. Muller. 1988. IC-processed electrostatic micro-motors. Pp. 666–669 in Proceedings of the International Electron Devices Meeting. New York: IEEE.

Faulhaber Group. 2001. Mensch and Technik. Available online at <*http://www.faulhaber.de*>.

Flynn, A.M. 1987. Gnat robots (and how they will change robotics). Pp. 221–225 in Proceedings of the IEEE MicroRobots and Teleoperators Workshop. New York: IEEE.

IMM (Institut für Mikroteknik Mainz). 2001. Available online at <*http://www.imm-mainz.de/images/crhubi.gif*>.

Nachtigall, W., A. Wisser, and D. Eisinger. 1998. Flight of the honey bee. VIII. Functional elements and mechanics of the 'flight motor' and the wing joint—one of the most complicated gear-mechanisms in the animal kingdom. Journal of Comparative Physiology B 168:323–344.

Pornsin-Sirirak, T.N., Y.-C. Tai, H. Nassef, and C.M. Ho. 2001. Titanium-alloy MEMS wing technology for a micro aerial vehicle application. Sensors and Actuators, A: Physical 89(1–2):95–103.

Shimoyama, I., Y. Kubo, T. Kaneda, and H. Miura. 1994. Simple microflight mechanism on silicon wafer. Pp. 148–152 in Proceedings of the IEEE Micro Electro Mechanical Systems. New York: IEEE.

SRI (SRI International). 2001. Artificial Muscle Transducers. Available online at <*http://www.erg.sri.com/automation/actuators.html*>.

Stanford University. 2001. Mesicopter Image Gallery. Available online at <*http://adg.stanford.edu/mesicopter/imageArchive/*>.

Yan, J., R.J. Wood, S. Avadhanula, D. Campolo, M. Sitti, and R.S. Fearing. 2001. Towards flapping wing control for a micromechanical flying insect. Pp. 3901–3908 in Proceedings of the IEEE International Conference on Robotics and Automation. Piscataway, N.J.: IEEE.

Yeh, R., E.J.J. Kruglick, and K.S.J. Pister. 1996. Surface-micromachined components for articulated microrobots. Journal of Microelectromechanical Systems 5(1):10–17.

SUGGESTED ADDITIONAL READINGS

Miki, N., and I. Shimoyama. 1998. Study on micro-flying robots. Advanced Robotics 13(3):245–246.

Ramamurti, R., and W.C. Sandberg. 2001. Computational study of 3D flapping foil flows. Presented at the 39th AIAA Aerospace Sciences Meeting, AIAA-2001-0605, Reno, Nevada, January 8–11, 2001.

Wang, Z.J. 2000. Two dimensional mechanisms for insect hovering. Physical Review Letters 85(10):2216–2219.

Civil Systems

Dynamic Planning and Control of Civil Infrastructure Systems

Feniosky Peña-Mora
Civil and Environmental Engineering Department
Massachusetts Institute of Technology
Cambridge, Massachusetts

Construction processes for civil infrastructure involve inherently complex interactions among variables, including but not limited to physical attributes, logistics, resource availability, budget restrictions, and management techniques. Poorly planned interactions among these variables lead to inefficiencies and uncertainties in project execution, a deterioration of planned construction sequences, schedule delays, and increased costs. The impact of such unbalanced interactions is greater in concurrent construction, a technique widely used in modern construction projects and a critical capability for returning infrastructure to service after a major disaster. In this paper, I will present a strategy for simulation-based reliability buffering that can contribute to more robust construction plans, reduce uncertainties, and mitigate the impact of changes.

RELIABILITY BUFFERING

The purpose of a reliability buffer is to avoid disruptions in a project schedule by failures in individual activities. Reliability buffering involves the systematic pooling, relocating, resizing, and recharacterizing of contingency buffers (Figure 1). Reliability buffering begins by eliminating contingency buffers that are explicitly or implicitly based entirely on an individual activity. In this way, these activities are subject to appropriate schedule pressure and a rubber band effect is avoided. In establishing precedences, the reliability buffer, which is inserted before the downstream activity, can be characterized as a time to identify problems or finish upstream activities and ramp up resources for completing downstream activities. Because the reliability buffer is put at the beginning of an activity instead of at the end, it addresses the issue of poorly defined tasks that

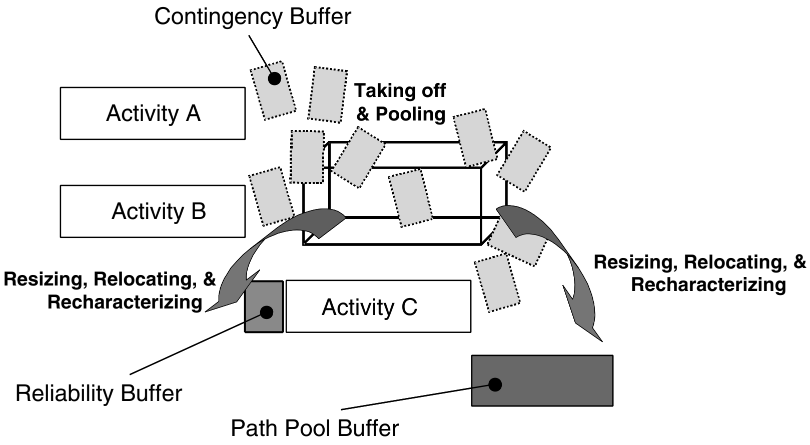


FIGURE 1 Steps in reliability buffering.

require time for better definition. Reliability buffering makes it possible to focus on activities with problems before they activate a domino effect, as might happen with traditional buffers.

Because reliability buffering is based on a simulation approach, it provides a systematic way of sizing a buffer. The buffer must be long enough to ensure that downstream activities will be performed according to plan (i.e., "reliable"). If a buffer is too long, it can create unproductive, idle time. Therefore, a buffer should be sized in a systematic way (based on simulation and analysis) rather than in an ad hoc way based only on individual experience. Moreover, with a dynamic buffering process, the initial size and location of a reliability buffer can be changed at any time during the construction, which might also change the initial precedence relationships.

FOCUSING ON FEEDBACK

Successful reliability buffering requires paying attention to feedback processes, which contribute to indirect and/or unanticipated events during a project and make the construction process dynamic and unstable. These inherent instabilities cannot be captured with traditional planning tools. Sometimes steps taken to reduce variations from the planned performance can fix problems and improve performance but, at the same time, worsen performance in another area as a result of unanticipated side effects. For example, when a construction project is behind schedule, one possible way to meet the original schedule is to replace current equipment with high-performance equipment. At first glance, this change should facilitate the construction process. However, it may take time for workers to learn how to operate the new equipment, or it may be difficult to coordinate use of the new equipment with subsequent processes. If the "fix" reduces productivity and increases coordination problems, changing equipment can actually increase the delay in the construction schedule. These feedbacks must be identified and analyzed before a change is made. Once major feedbacks have been identified, construction processes can be simulated more realistically before actual resources are committed.

CAPTURING CONSTRUCTION DYNAMICS

Dynamics should be captured in the construction schedule. Factors that trigger feedbacks in construction are changes, dependencies among activities, construction characteristics, and human responses to the work environment and policies. Normally, changes refer to work state, processes, or methods that deviate from the original plan or specification. Changes are major contributors to dynamics and instabilities; changes also create non-value-adding iterations.

As Figure 2 shows, changes are usually made to improve the quality of work or working conditions or to accommodate changes in scope. In addition, changes that have already been made can lead to other necessary changes in concurrent, succeeding, or preceding tasks. For example, changes in design that have been made by mistake can cause subsequent changes in construction. In this case,

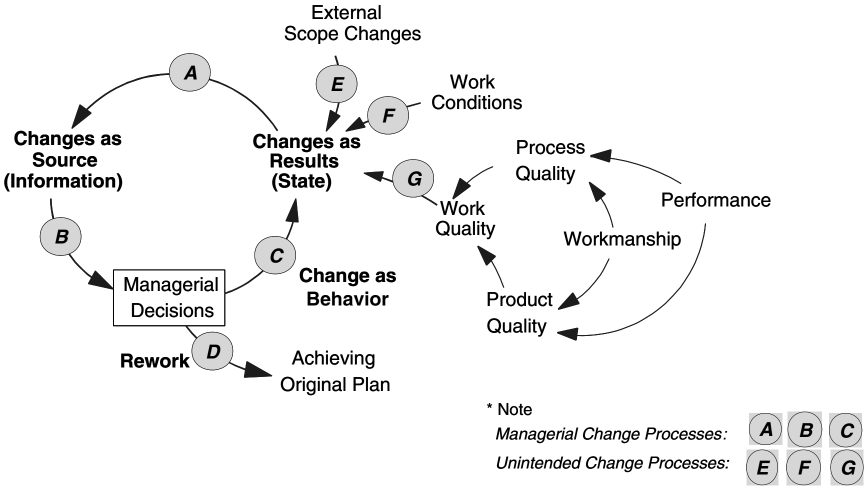


FIGURE 2 Changes as triggers of iteration.

even though the designers are responsible for the design changes, those changes can require changes by the construction crew. Changes can be categorized as unintended changes (such as the changes just described) and managerial changes, which are intentional decisions made during quality management or project monitoring and control. Both kinds of changes can lead to subsequent changes or require rework.

Sometimes, by adopting managerial changes, rework on problematic tasks that would require more resources can be avoided. However, even these changes can lead to subsequent changes that might have more of a negative impact on construction performance than the rework option. For example, if some piles have not been correctly positioned, it may be possible to proceed with the superstructure without correcting the position of the piles by changing the position of columns. However, this change option may require unplanned cantilever construction to preserve the original floor layout. The impact of this option will have to be compared to the impact of redriving the piles. The decision must be based on a good understanding of how changes evolve to non-value-adding iterations, which can have unanticipated, indirect side effects. This is particularly important for concurrent construction. In the case of reconstruction after a disaster, for example, a rushed schedule with limited resources can lead to managerial changes that may actually delay putting infrastructure back in service. In short, because construction changes combined with other factors have different impacts on the construction system, it is important to understand how they can affect the planning and control of construction projects.

CHANGE IMPACT VS. RELIABILITY BUFFERING

The impact of changes on construction performance can vary depending on whether changes are managerial (intentional) or unintentional. As Figure 3 shows, managerial changes can create subsequent non-value-adding iterations both at the point of change (C_{up}) and in downstream activities (C_{dn}). Thus, managerial changes might have more of an impact on construction performance than rework, depending on the sensitivity of associated tasks to the change and how much work has been done by the time the change is introduced. The impact of a managerial change on the upstream activity (C_{up}) is in proportion to the sensitivity of the upstream work to internal changes and the progress of the upstream work. The impact of the change on the downstream activity (C_{dn}) can be measured as a function of sensitivities to changes in the upstream work and progress in the downstream work at the time the change is made.

Unintended changes have more complex impact patterns. Normally, the impact of a change increases with the length of time before discovery and the distance of the discovery from the location of the original change. Unintended, undiscovered changes can create a ripple effect that affects all subsequent work. The impact of an unintended change can vary depending on whether the change

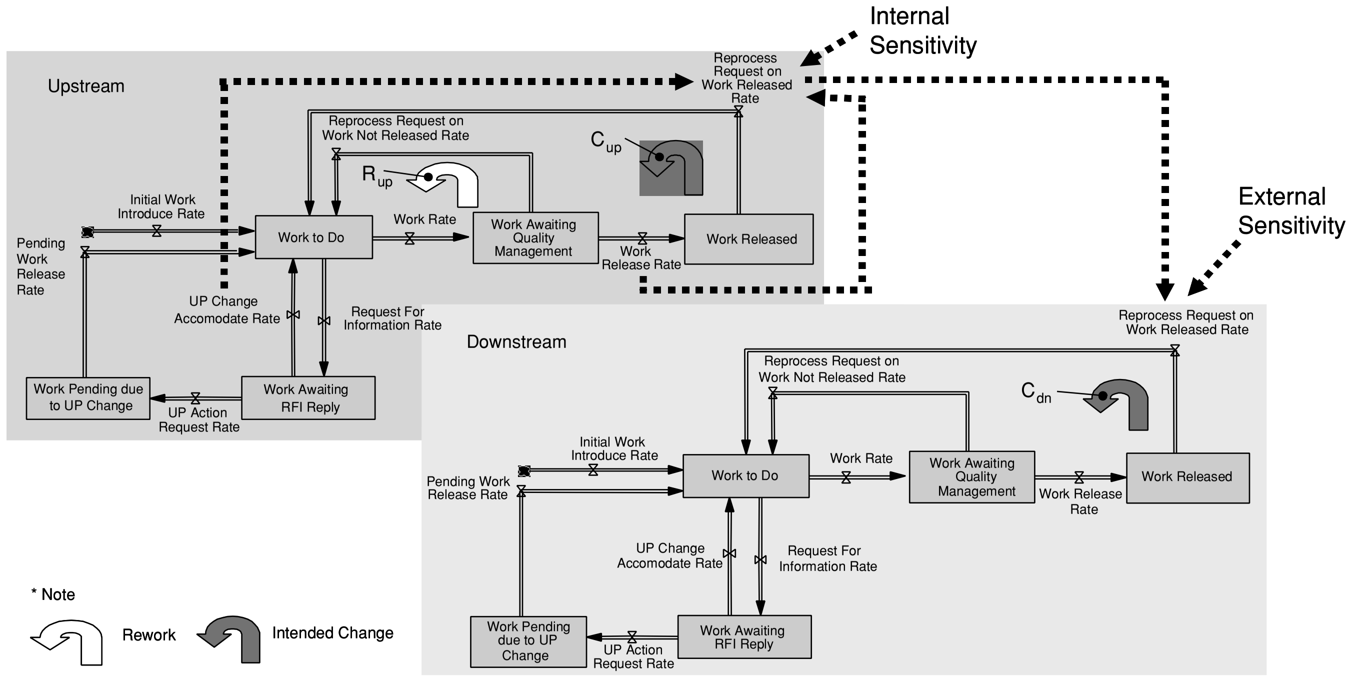

Internal Sensitivity
External Sensitivity
Upstream
Reprocess Request on Work Released Rate
Reprocess Request on Work Not Released Rate
C_up
R_up
Initial Work Introduce Rate
Pending Work Release Rate
Work to Do
Work Rate
Work Awaiting Quality Management
Work Release Rate
Work Released
UP Change Accomodate Rate
Request For Information Rate
Work Pending due to UP Change
UP Action Request Rate
Work Awaiting RFI Reply
Downstream
Reprocess Request on Work Released Rate
Reprocess Request on Work Not Released Rate
C_dn
Initial Work Introduce Rate
Pending Work Release Rate
Work to Do
Work Rate
Work Awaiting Quality Management
Work Release Rate
Work Released
UP Change Accomodate Rate
Request For Information Rate
Work Pending due to UP Change
UP Action Request Rate
Work Awaiting RFI Reply
* Note
Rework
Intended Change

FIGURE 3 Impact of managerial changes.

was the result of poor workmanship or the result of undiscovered changes in upstream work. The work quality of an activity is in proportion to its reliability, while the effect of an upstream change on work quality is a nonlinear function. By incorporating all of these determinants into the plan, the change impact on downstream activities can be categorized in terms of type, path, and magnitude.

If reliability buffering is used in the cases described above, the impact of subsequent changes in downstream activity can be absorbed by the systematical assigning of time buffers. Systematically and strategically located and sized time buffers can help reduce the domino effect of changes on downstream work by effectively controlling the start time and progress of the downstream work.

CONCLUSIONS

Reliability buffering is an effective technique for more robust construction planning and for addressing inherent uncertainties and the impact of intended and unintended changes. Research results thus far have shown that appropriately pooled, located, sized, and characterized reliability buffers could reduce the impacts of change on construction processes. In addition, case studies on bridge construction projects have demonstrated the applicability of reliability buffering for critical infrastructure construction. More research will have to be done to analyze fully the dynamics of construction projects in chaotic environments, such as disaster recoveries, before reliability buffering will be widely accepted by the construction industry.

ACKNOWLEDGMENTS

I would like to thank my graduate students for making this research possible: Moonseo Park, Margaret Fulenwider, and Sanghyun Lee. I would also like to acknowledge the contribution to this research by David N. Ford, assistant professor at Texas A&M University; and Joseph Peck, corporate planning and scheduling manager, William Lemoine, vice president, and John Foster, senior project manager, at Modern Continental Co.; and Philip Helmes, vice president at InteCap, Inc. Finally, I would like to acknowledge the support for this research received from InteCap, Inc., the National Science Foundation CAREER Award, and the White House PECASE Award CMS-9875557.

FURTHER READING

Fazio, P., O. Moselhi, P. Theberge, and S. Revay. 1988. Design impact of construction fast-track. Construction Management and Economics 6(2):195–208.

Goldratt, E.M. 1997. Critical Chain. Great Barrington, Mass.: North River Press.

Huovila, P., L. Koskela, and M. Lautanala. 1994. Fast or concurrent: the art of getting construction improved. Pp. 143–158 in Proceedings of the 2nd Workshop on Lean Construction, Santiago, Chile, L. Alarcón, ed. Rotterdam, Netherlands: A. A. Balkema.

Park, M. 2001. Dynamic Planning and Control Methodology for Large-Scale Concurrent Construction Projects. Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge.

Peña-Mora, F., and M. Park. 2001. Dynamic planning for fast-tracking building construction projects. Journal of Construction Engineering and Management 127(6):445–456.

Russell, A., and M. Ranasinghe. 1991. Decision framework for fast-track construction: a deterministic analysis. Construction Management and Economics 9(5):467–479.

Tighe, J. 1991. Benefits of fast tracking are a myth. International Journal of Project Management 9(1):49–51.

Williams, G. 1995. Fast-track pros and cons: considerations for industrial projects. Journal of Management in Engineering 11(5):24–32.

Improbable Is Not Impossible: Decision Making Under Uncertainty

Linda K. Nozick
School of Civil and Environmental Engineering
Cornell University
Ithaca, New York

INTRODUCTION

Risk is the result of chance and negative outcome. An earthquake strikes. A bomb planted by a terrorist explodes on a crowded street. A recession hits. Risk is inherent in every aspect of our lives. Personal decisions, such as which job to take or which investments to make, involve risk, as do decisions by corporations and governments. Consider, for example, residential building codes, which are developed and enforced by governments to promote health and safety. In general, code changes that improve safety add to the cost of construction. What level of safety should be mandated? If building codes are strengthened significantly, the costs of construction could make new homes less affordable. More expensive homes draw money that otherwise might be spent on other goods and services. High costs could make it more difficult for people to purchase new homes thereby causing them to remain in older homes that may have been built to lower standards (Hammitt et al., 1999).

Risk analysis and decision making under uncertainty are inherently cross disciplinary. Risk is primarily concerned with measuring the probability and severity of potentially negative outcomes. Risk is a scientific concept that can be qualitative or quantitative or both. By contrast, decision making, which is about the acceptability of risk, has political, ethical, and personal dimensions (Lowrance, 1976). Some will argue that risk is primarily subjective, that the notion of risk is a response to the human need to cope with uncertainty and that we develop models whose structures and uses are riddled with subjective judgments (Slovic, 1999). We must acknowledge that all models have an element of subjectivity; but the key difference between the "scientific" and "subjective" views of risk is the extent of this subjectivity. For the purposes of this paper, I will focus on risk analysis as a scientific enterprise and simply acknowledge that the alternate view exists and has some merits worth discussing.

The focus of this paper is on making decisions when there are multiple objectives and significant uncertainties associated with the possible outcomes. I will illustrate these ideas with a case study of hazardous materials transportation.

MEASURING RISK AND COMPARING ALTERNATIVES

The most common measure for estimating risk is expected value. Expected value is defined as the sum of the multiplication of each potential consequence by its probability of occurrence. This measure has advantages and disadvantages. The advantages include: (1) it is easy to understand, and (2) the risk of an action or alternative can be summarized in a single value. The disadvantage is that it gives equal weight to high-probability, mild-consequence outcomes and low-probability, high-consequence outcomes. For example, in an expected value calculation, an earthquake that destroys a collection of homes, which has a low probability of happening but catastrophic consequences, is treated the same as damage to the exterior of homes that has a high probability of occurring but minor consequences. Clearly, this is a distortion of the true impacts. To overcome this limitation, descriptions of the probability distribution of consequences are often used. These descriptions can be conditional, if the consequences are in a given portion of the distribution, such as the upper quartile; or unconditional, if they are applied to the entire distribution.

Choosing between alternatives, each of which can have a number of different consequences, is often a difficult task. The task is considerably simpler if the cumulative distribution function (CDF) of consequences for one alternative, $F(x)$, is equal to that of another alternative, $G(x)$, for some consequences and is larger for others. That is, if $F(x) \geq G(x)\ \forall x$, F stochastically dominates G. Notice that in Figure 1a the probability that any outcome is x or less is greater for $F(x)$ than for $G(x)$. Unfortunately, this is seldom the case. Usually the CDFs cross. If they cross at a level of consequence that is sufficiently rare, we might still be willing to say that F stochastically dominates G or, equivalently, that F partially stochastically dominates G. Consider the example in Figure 1b. Only very rare negative events are slightly more probable under alternative F.

MULTIPLE OBJECTIVES

Making trade-offs between conflicting, noncommensurate objectives is at the core of decision making. Uncertainty simply adds to the difficulty. For example, consider making decisions about aircraft safety. There is an inherent trade-off between the thoroughness of inspections of people and baggage and delays. The more carefully people and baggage are examined, the less likely a

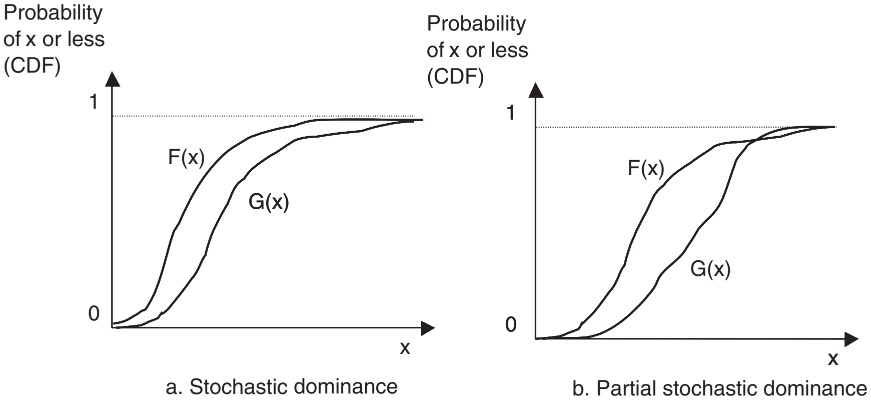


FIGURE 1 Stochastic dominance and partial stochastic dominance.

potentially dangerous item is to be smuggled aboard a plane. However, the more complete the inspections, the greater the delay.

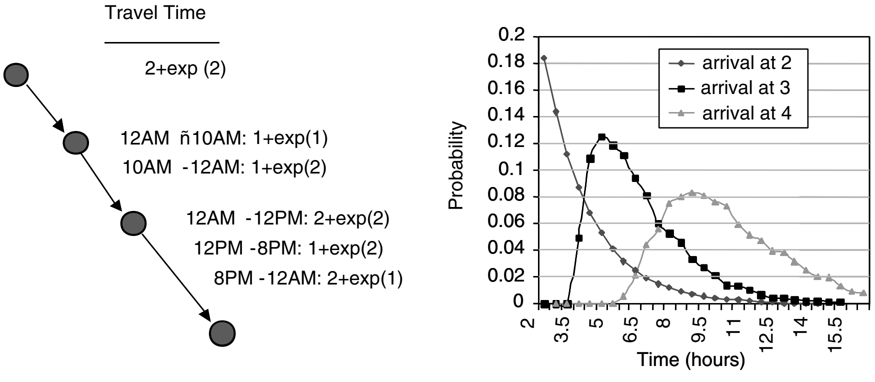
The concept of an optimal choice when there are multiple objectives is significantly different than when there is a single objective. With a single objective, the performance of one alternative or one set of alternatives is optimal in terms of the objective. With multiple objectives, there can be many solutions, commonly referred to as a nondominated set, each of which represents different trade-offs. All of the solutions in this set are equally bad with respect to all objectives. The core question is finding an acceptable trade-off between achieving one objective and failing to achieve another, or partly achieving both to an acceptable degree. Typically, there are no “right” or “wrong” answers, and different individuals may make different decisions.

CASE STUDY: ROUTING HAZARDOUS MATERIALS

More than 800,000 hazardous material (hazmat) shipments occur daily in the United States, and the total volume of hazmat shipped annually is about 3.9 billion tons. The U.S. Department of Commerce estimates that approximately one truck in five on U.S. highways is carrying hazmat. Even though the safety record of hazmat transportation is very good, significant public concerns have been raised about the risks to people and the environment from hazmat shipments.

This case study concerns the application of an algorithm for routing a hazmat shipment when there are multiple objectives and the performance of each facility with respect to each of the objectives is stochastic and varies over time. Consider the possible consequences of an accident and the necessity of minimizing travel time for a truck carrying a shipment of flammable liquid from Wilmington, Delaware, to Portland, Maine. The major hazard is assumed to be fire and/or vapor explosion in the event of an accident. Thus, the major at-risk population is other travelers on the highway near the shipment, rather than people residing some distance from the road. As a measure of consequence, we have used vehicle-minutes of exposure within *x* distance of the truck multiplied by the accident rate. That is, as the truck moves along the road, the vehicles potentially exposed to the risk of fire and/or explosion are (1) vehicles traveling in the same direction as the truck and less than *x* distance behind the truck and (2) vehicles traveling in the opposite direction at a distance of *x* or less ahead of the truck. In this case, *x* is assumed to be 0.5 miles. We then multiply 0.5 by the probability of an accident to calculate consequence.

An essential element of finding multiobjective shortest paths in stochastic dynamic networks is having a means of constructing the probability distribution of an attribute along a path. Because the probability distributions of the arc attributes are assumed to vary with time, a core step in the process is constructing a probability distribution of travel time from the point of origin to each node along a potential path. Because this is similar to adding together random variables, after a few links, the travel time distribution will be approximately normally distributed, so a procedure based on propagating means and variances can be expected to work well. Figure 2 illustrates a network with four links and estimated probability distributions for the time of arrival at each node. Notice that even with only four links, the arrival time at node four looks almost normally distributed. Therefore, all that is necessary is a procedure for propagating the mean and variance of arrival time at successive nodes.



a. Network with four links

b. Arrival time

FIGURE 2 Example of calculating travel time along a path.

Consider a path p from an origin node s to a destination node v in a stochastic dynamic network that includes a link connecting node i to node j. Let the random variable Y_i denote the arrival time at node i (which is also the departure time from node i, because of the assumption of no waiting); let the random variable $d_{ij}(Y_i)$ be the travel time on link (i, j). It is clear that $Y_j = Y_i + d_{ij}(Y_i)$; and the conditional mean and variance of link (i, j) travel time are denoted by $\mu(y_i)$ and $v(y_i)$. The mean and variance of Y_j are given by:

$$\mathrm{E}[Y_j] = \mathrm{E}[Y_i] + \mathrm{E}[\mu(Y_i)] \tag{1}$$

$$\mathrm{Var}[Y_j] = \mathrm{Var}[Y_i] + \mathrm{E}[v(Y_i)] + \mathrm{Var}[\mu(Y_i)] + 2\,\mathrm{E}[Y_i\,\mu(Y_i)] - 2\,\mathrm{E}[Y_i]\mathrm{E}[\mu(Y_i)] \tag{2}$$

Equations (1) and (2) allow propagation of the mean and variance of the arrival time at a node along the links of a path. Because traffic data are generally collected and summarized over discrete intervals (e.g., average speed over 1-hour intervals, or for a peak period of 3 to 4 hours), the data to support estimates of $\mu(Y_i)$ and $v(Y_i)$ as continuous functions of time are generally not available. Therefore, in the analysis here, the estimates of $\mathrm{E}[Y_i]$ and $\mathrm{Var}[Y_i]$ are propagated using discrete intervals. These calculations are embedded in a label-correcting algorithm along with the concept of partial stochastic dominance discussed above. This allows us to efficiently search for "good" paths. For more information on the algorithm, see Chang et al. (2001a, 2001b).

A key decision in this case is determining the best path through the New York City area. One option is to pass through the city. That is, to stay on I-95 from Wilmington, Delaware, cross the George Washington Bridge, and take the Cross Bronx to the New England Thruway. Another alternative is to go around the city by taking I-287 across the Tappan Zee Bridge to the Cross Westchester Expressway to the New England Thruway. Figure 3 illustrates these choices.

Figure 4 illustrates differences in the distributions for the travel time and consequences for these two choices. Path A, which goes through the city, is shorter on average but more variable. In addition, the consequences on Path A are more severe than on Path B.

CONCLUSIONS

This paper illustrates the critical role of decision making under uncertainty in all aspects of our lives. There are at least two major challenges in this field. The first is improving the scientific enterprise of decision making under uncertainty, by improving methods of estimating probability distributions and methods of optimizing actions considering these distributions. The second challenge is to resolve the philosophical debate raging in the risk-based decision-making community. In this paper, I have focused on risk as the result of chance and negative outcome based on the underlying assumption that outcomes are pro-

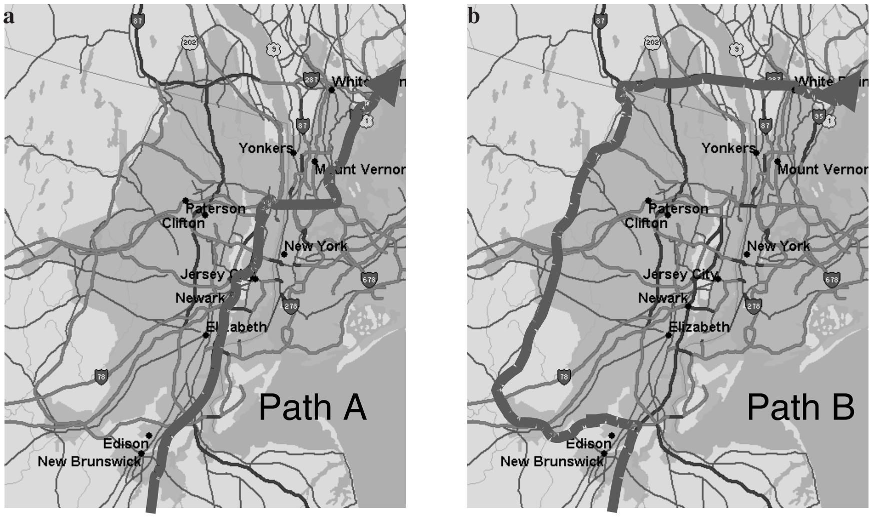


FIGURE 3 Two paths through the New York City area.

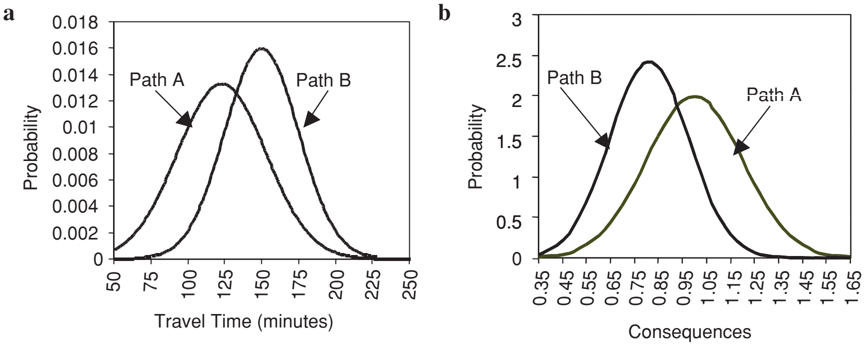


FIGURE 4 Travel times and consequences for Path A and Path B.

duced in ways that can be objectively analyzed. As evidence of subjectivity, those who argue that risk is inherently subjective often point out the great disparities in expenditures by different government agencies to save lives. For example, EPA spends about $7.6 million per year of life saved, whereas NHTSA spends about $78,000 per year of life saved (Tengs et al., 1995). Another example is the extremely difficult and complex decision required to address the storage of spent fuel or high-level radioactive waste, which has been pending for many years (North, 1999).

ACKNOWLEDGMENTS

This work has benefited from the contributions of Drs. Mark Turnquist and Tsung-sheng Chang. The support of the National Science Foundation is gratefully acknowledged.

REFERENCES

Chang, T., L. Nozick, and M. Turnquist. 2001a. Routing hazardous materials with stochastic dynamic facility attributes. Working paper, Cornell University.

Chang, T., L. Nozick, and M. Turnquist. 2001b. Routing hazardous materials with stochastic dynamic facility attributes: A case study. Working paper, Cornell University.

Hammitt, J., E. S. Belsky, J. I. Levy, and J. D. Graham. 1999. Residential building codes, affordability, and health protection: A risk-trade-off approach. Risk Analysis 9(6):1037–1059.

Lowrance, L. 1976. Of Acceptable Risk. Los Altos, Calif.: William Kaufman.

North, D. W. 1999. A perspective on nuclear waste. Risk Analysis 19(4):751–758.

Slovic, P. 1999. Trust, emotion, sex, politics, and science: Surveying the risk-assessment battlefield. Risk Analysis 19(4):689–701.

Tengs, T. O., M. E. Adams, J. S. Pliskin, D. G. Safran, J. E. Siegel, M. Weinstein, and J. D. Graham. 1995. Five hundred life-saving interventions and their cost-effectiveness. Risk Analysis 15(3):369–390.

Interdependencies in Civil Infrastructure Systems

MIRIAM HELLER
Infrastructure and Information Systems
National Science Foundation
Washington, D.C.

INTRODUCTION

Information systems hold the key to the efficient planning, design, construction, operation, maintenance, and retirement of our nation's very valuable civil infrastructure assets. Information systems are already being integrated into infrastructure operations to exploit new technologies, compensate for capacity limitations, address regulatory changes, increase efficiency, and protect against natural, accidental, and deliberate threats. Integrated information-infrastructure systems drive traffic signals and variable message signs on roadways and bridges, monitor potable water quality at treatment plants, pump water and wastewater, and activate switches in telecommunications systems that command transportation and water networks. All of these capabilities are enabled by energy and power infrastructures, which, in turn, depend on even more information infrastructure. In short, information systems can make or break civil infrastructure.

Integrated information systems have substantially improved unit-level and component-level operating efficiencies in transportation, water, telecommunications, and power infrastructures, just to name a few. The benefits include increased accuracy, expanded and improved services and products, reduced capacity needs, higher utilization, and lower costs. Theory suggests that further efficiencies are achievable by integrating information systems at increasingly higher levels: the subsystem level, the system level, and even across infrastructure systems. History suggests, however, that further efficiencies might be difficult to realize because of trade-offs with induced vulnerabilities.

As the automation of infrastructure systems increases, system behaviors are becoming complex beyond comprehension and more far-reaching than was ever anticipated. Interdependencies can reverberate perturbations globally. In 2001, for example, the real-life restructuring of California's electricity industry demonstrated linked and unexpected effects. Fuel production, refining, and distribution were disrupted, sometimes cutting off fuel supplies to the very plants that should have been generating their electricity. Interruptions in water distribution affected the state's agribusiness. Soaring wholesale power prices had rippling regional effects. In Washington state, salmon-protection and air-quality regulations had to be relaxed and aluminum mills shut down. Idaho farmers curtailed potato production to exploit Idaho Power Company's electricity buy-back program.

This paper focuses on the tension between the need to push our civil infrastructure systems to higher levels of efficiency and competitiveness and the need to ensure minimum levels of service, reliability, and security, even under critical conditions. To set the scene, some recent history is given, and infrastructure systems are described in terms of their performance, interdependencies, and vulnerabilities. This is followed by a description of some emerging frameworks that promise to capture these "systems of systems" and their interdependencies. A case study is presented highlighting the benefits of exploiting interdependencies, and research challenges are identified.

INFRASTRUCTURE INTERDEPENDENCIES

Infrastructure interdependencies appeared on the radar screens with Presidential Decision Directive 63 (PDD-63) on Critical Infrastructure Protection. Prompted by the Oklahoma City bombing in 1995 and the 1996 Defense Science Board Task Force on Information Warfare, PDD-63 was the culmination of a 15-month study by the President's Commission on Critical Infrastructure Protection, which revealed the rapidly growing capability of exploiting energy, banking and finance, transportation, vital human services (water, wastewater, and health services), and telecommunications infrastructures, especially through digital infrastructures (PDD-63, 1998). The directive acknowledged that our national and economic security depend on the critical infrastructures and information systems that support them. To ensure their reliability and protection, committees were established for each infrastructure sector and paired with their agency counterparts to study sector-specific problems. These initiatives have focused on protecting information systems against malicious intrusions (cyber attacks) that could cause the banking, finance, power systems, and other critical infrastructures to fail.

Infrastructure systems are also vulnerable to myriad stresses and failures as a result of everyday interdependencies, insufficiencies, and inefficiencies. Cascading power blackouts in the United States in July and August 1996 cost an estimated $1.5 billion, including related infrastructure and environmental impacts (Amin, 2000). On a grander scale, recent estimates of the annual cost to the U.S. economy from non-cyber power disturbances exceed $119 billion, most of which is related to disruptions to discrete manufacturing and electricity-dependent utilities

(Lineweber and McNulty, 2001). Traffic congestion costs the nation an additional $78 billion annually in 4.5 billion hours of extra travel time and 6.8 billion gallons of fuel idled away in traffic jams (TTI, 2001).

The longer we neglect these problems, the more they will create new and exacerbate old infrastructure vulnerabilities. The estimated cost of maintaining the *status quo* of existing infrastructure systems is $1.3 trillion over the next five years (ASCE, 2001). Although this figure seems high at first glance, it seems reasonable considering that the total U.S. investment in infrastructure is more than $7 trillion (CERF, 1997).

Natural interventions test the robustness and reliability of infrastructure design. The cost of earthquakes averages $4.4 billion per year (FEMA, 1999). Another intervention, space weather, was the culprit in 1998. When the *Galaxy 4* satellite's attitude control system failed, radio, television, pager, bank machine, and other satellite-linked services across North America were disrupted. As an example of the cost, two pager companies that did not have backup systems in place lost $5.8 million. Indirect and intangible costs included lost credit card sales, missed market trades, inability to contact doctors and emergency medical services, and many others.

The tragedies of September 11, 2001, have given us new data on the costs of physical infrastructure catastrophe, their interdependencies, and their resiliency. In the first weeks after the attacks, losses to the air transportation industry were estimated at $320 million per day. The direct and indirect costs of the closure of Reagan National Airport, the drastic decrease in tourism, lower consumer spending, and bankruptcies will probably never be tallied. The disaster relief package of $40 billion from the federal government provides, at best, a lower bound. The structural changes to the U.S. economy and the American life style have yet to be fully realized, much less assessed.

As a result of these events, questions about how to manage the life cycle (i.e., the design, construction, operation, maintenance, and retirement) of civil infrastructure systems and their digital infrastructure adjuncts have become urgent.

- What methods and tools can capture, clarify, and predict the complex behaviors and interdependencies of infrastructure systems?
- How can maximal efficiency during normal operations be balanced with resiliency, sustainability, and minimal vulnerability to common and catastrophic failures?
- Which measures of performance adequately capture system(s) complexity?
- Who are the decision makers and stakeholders, and what are their goals and objectives?
- How can risk and uncertainty be incorporated into the design and management of infrastructure systems?

Even before September 11, these questions had taken on greater urgency as infrastructure systems were being pressed to meet or surpass the levels of efficiency of the systems that create the demand for their services (e.g., just-in-time manufacturing, e-commerce sales and procurement, and overnight delivery). The vulnerabilities intrinsic to interdependent, slack-free, deteriorating, or externally threatened systems must be understood, predicted, sensed, and engineered to meet multiple performance measures. Optimizing these systems for normal conditions without considering the costs, risks, and uncertainties of "abnormal" conditions would be shortsighted and even dangerous. But even the horrors of September 11 must not blind us to the ongoing need for investing in the design and operation of infrastructures that can, and do, cost billions of unnecessary dollars and lead to many deaths every year.

EMERGING FRAMEWORKS

Developing a model of a single infrastructure system, with its own patterns of use, its interactions with associated natural and economic systems, and its reactions to technological and natural interventions, poses serious challenges. Many infrastructure systems (e.g., power, transportation, and telecommunications) are complex adaptive systems (CASs), that is, their collective, systemic behavior is emergent (i.e., it follows patterns that result, yet are not analytically predictable from, dynamic, nonlinear, spatiotemporal interactions among a large number of components or subsystems [Coveney and Highfield, 1995]). Because a CAS is greater than the sum of its parts, the system can only be described at levels higher than the components. The size and frequency of electricity disturbances, for instance, obey the power law, a characteristic of complex systems at the critical edge between order and chaos (Amin, 2001). CASs are adaptive in that the capabilities of components and decision rules change over time in response to interactions with other components and external interventions (Gell-Mann, 1994).

Despite the challenges, modeling systems of infrastructure systems, whether CAS or not, is necessary for optimal life-cycle management of civil infrastructure systems. Much remains to be done to develop models and merge individual models of coupled systems, including formalizing theories and conceptual frameworks for meta-infrastructure systems to support them. Although new methods and tools for individual infrastructure system models have been evolving, fewer attempts have been made, and even fewer successes attained, at modeling meta-infrastructure systems.

Rinaldi et al. (2001) have proposed a general framework for characterizing infrastructure interdependencies. The framework identifies infrastructure systems as CASs and provides details for developing agent-based simulations (ABSs) of complex systems. The authors identify six dimensions of infrastructure interdependencies: infrastructure environment, coupling, response behavior, failure types, infrastructure characteristics, and state of operation. Analyzing infrastructure in these terms yields new insights into infrastructure interdependencies. They also identify four *types* of interdependencies: physical, cyber, logical, and geographical. In a physical interdependency, the states of two infrastructures (e.g., a coal-transporting rail network and a coal-fired electrical plant that supplies the power to that rail network) depend on the material output of both. Other interesting issues are also raised, including requirements for an information architecture; data capture, storage, and privacy; and model metrics.

Haimes and Jiang (2001) extend Leontief's economic input-output models to evaluate the risk of inoperability in interconnected infrastructures as a result of one or more failures subject to risk management resource constraints. Interdependence is captured in Leontief's production coefficients, which here represent the probability of an interconnected infrastructure component propagating inoperability to another component. Infrastructure components are also subject to independent risks of failure. Finally, each component has an associated coefficient reflecting the amount of some resource (e.g., funds or personnel) required to manage the risk of inoperability. Thus, infrastructures are interdependent through failure propagation, specified in geographical, functional, temporal, and political dimensions, and through the allocation of limited resources for risk management. The authors also propose a hierarchical adaptation of this model to avoid over-aggregation and reductionism, reduce the dimensions of problems, provide more realistic systems models (both static and dynamic), and enable multi-objective analyses.

Another approach based on economics by Friesz et al. (2001) defines a spatial computable general equilibrium (SCGE) model of an economy comprised of spatially separated markets interconnected by a generalized transportation network. Each infrastructure model is conceptually extended to capture interdependencies using a multilayer network of SCGE models with interlayer coupling constraints. The authors first identify five sources of interdependency with the aim of devising a mechanism to express them mathematically. Interdependencies can be physical, budgetary, market-based or spatio-economically competitive, information-based, or environmental. The static model yields equilibrium values for the supply price of the commodity (e.g., good, passenger, message, data, water, or energy), flow quantity and path, levels and locations of commodity production, and transport costs. Methods of modeling the system dynamically, including ABS, are suggested for evaluating and enhancing infrastructure systems design and capital budget allocations for operations, maintenance, and replacement.

ABSs are emerging as the most promising modeling techniques for predicting, controlling, and optimizing infrastructure systems. Like CASs, agents in ABSs execute relatively simple decision rules within the structural definition and constraints of the infrastructure system(s). Agents' decisions are responses to the information they have about the system, some of which may be sensed.

ABS has two advantages. First, ABS can represent CAS without resorting to inappropriate analytical models; at the same time, it can enable predictions of the desirability of different policy options. North (2000) developed a series of ABSs to explore pricing and various levels of competition with deregulated electric utilities. The simulations addressed the effects of price swings for natural gas, such as those that would follow a pipeline interruption; the number of companies needed for truly competitive markets; and the identification of companies colluding to drive up electricity prices.

Second, ABS may offer an improved control paradigm that can be implemented at the hardware level. With centralized control, infrastructure systems are vulnerable to the weakest link; distributed control can limit, localize, and allocate risk. Some models have been proposed whereby infrastructure system agents could automatically reconfigure a system to "heal" failures (Amin, 2000). Distributed control also enables distributed power generation, as well as the control of multiple infrastructure systems.

The meta-infrastructure system approaches described above are reasonably representative of the current state of the art. It is interesting to note that none of these frameworks deals explicitly with interdependencies induced by sharing input resources. Physical interdependencies in Rinaldi come the closest; Friesz et al. and Haimes and Jiang both use implicit notions of activity levels.

Interdependencies from resource sharing arise when improved efficiency is achieved by reducing redundancy across systems. When systems use resources completely independently of one another to provide their respective services, the systems are independent with respect to that resource, assuming perfect market competition. If the resources could have been shared but were not, the resources were redundant. Every reduction in redundancy in these systems through resource sharing creates a certain class of system interdependency.

Reduced redundancy, the elimination of a redundant power generator and high utilization of remaining generators, for example, can render a system more vulnerable. A beneficial example of resource sharing would be a hydropower facility and drinking water plant that use and reuse the same river flow to generate their respective services. In fact, the chief of the Bureau of Reclamation recently stated that to use water stored by 457 dams in the western United States as efficiently as possible water should be passed through the dams multiple times for recreation, power generation, and irrigation (WaterTech Online, 2001). Finally, tracking resource quantities explicitly would make possible more accurate assessments of the external costs (e.g., environmental impact) of using those resources.

CASE STUDY

Colorado Springs Utilities, an innovative western water utility that has been researching multiple uses of water resources, estimates the benefits would be worth more than $500,000 per year, not including windfalls from high electricity prices (Jentgen, 2001). Their energy and water quality management system (EWQMS) is conceptually an extension of electric utilities' energy management systems (EMSs), which include power generation control and real-time power systems analysis. Some aspects of EWQMS can be substantially more complicated than EMS. For example, in EWQMS where hydropower is an option, decisions about pumped storage are coupled with the selection of electricity sources to exploit time-of-day electricity pricing. Alternatively, if spot market prices are exorbitant, hydropower might best be used to generate electricity for sale. Whereas EMS' power generation control has a short-term load-forecasting component, EWQMS has two sets of demands to predict and satisfy: one for electricity and one for water. In addition, scheduling decisions must also consider (1) what quantity of raw water from which source is subject to water rights and quantity and quality constraints, given variable pumping costs; (2) what quantity of water to treat at which plant, given variable treatment costs; and (3) what pumps to use for distribution, collection, and wastewater treatment and which ones to take off line for maintenance.

This case study shows how shared resources can simultaneously improve efficiency and reduce vulnerability through resource reuse. Heller et al. (1999) discuss the concept of shared resources as a means of achieving regional eco-efficiency. In this context, information system boundaries are extended to coordinate the shared production, consumption, treatment, or reuse of electricity, water, and wastewater resources among regional utilities and manufacturing facilities.

FUTURE RESEARCH

Interdependencies in civil infrastructure systems require much more attention and study. As long as we treat infrastructure systems in isolation, we will perpetuate suboptimal systems operations, inefficient resource use, and vulnerability to the risks and uncertainties of failure.

We need new frameworks for understanding systems of infrastructure systems as a basis for modeling the complex behaviors of individual infrastructure systems as well as coupled systems. Specifically, research should be focused on meta-infrastructure systems models in the context of multiple large-scale complex adaptive systems. We also need methodologies for designing and operating these systems of systems in a way that provides the best trade-offs in terms of efficiency, vulnerability, resiliency, and other competing objectives, under normal and disrupted conditions. Another area for research is the development of multiple performance measures and economic models that accommodate them to capture multiple stakeholders' interests and decision makers' missions, constituencies, resources, and schedules. Design and operations must be performance-based. Metrics and economic models must address organizational and human errors and threats, as well as the risks and uncertainties of extreme events. New paradigms for distributed control should be investigated and compared with centralized control options. To provide more and better information, research could focus on the design and development of infrastructure-level sensor systems and data management systems. Finally, efforts must be directed toward educating and training a workforce for research in infrastructure interdependencies.

Infrastructure systems, which were engineered to facilitate the competitive flow of people, goods, energy, and information, have expanded far beyond their original design specifications. To meet the exigencies of our greatly changed world, we must rethink and reengineer infrastructure systems life cycles to serve their original purposes under new conditions, such as globalization, deregulation, telecommunications intensity, and increased customer requirements. We must make sure information system interdependencies contribute to solutions and do not exacerbate, or even become, the problem.

ACKNOWLEDGMENTS

Many thanks to P. Nelson, T. Kelly, M. Amin, K. Sullivan, E. Casman, G. Marino, and M.E. Porter for contributing to this paper and to the National Science Foundation for the compensated time.

NOTE

The opinions expressed in this paper are those of the author only and do not necessarily represent those of the National Science Foundation or any other entity with which the author has been or is now affiliated.

REFERENCES

Amin, M. 2000. Toward self-healing infrastructure systems. IEEE Computer 33(8): 44–53.

Amin, M. 2001. EPRI/DoD Complex Interactive Networks/Systems Initiative. Workshop on Mitigating the Vulnerability of Critical Infrastructures to Catastrophic Failures, Alexandria Research Institute, September 10–11, 2001. Available online at <http://www.ari.vt.edu/workshop>.

ASCE (American Society of Civil Engineers). 2001. The 2001 Report Card for America's Infrastructure. Available online at <http://www.asce.org/reportcard/>.

CERF (Civil Engineering Research Foundation). 1997. Partnership for the Advancement of Infrastructure and Its Renewal through Innovative Technologies (PAIR) White Paper. Available online at <http://www.cerf.org/research/pair/issues.htm>.

Coveney, P., and R. Highfield. 1995. Frontiers of Complexity: The Search for Order in a Chaotic World. New York: Fawcett.

FEMA (Federal Emergency Management Agency). 1999. HAZUS®: Estimated Annualized Earthquake Losses for the United States. Washington, D.C.: Federal Emergency Management Agency.

Friesz, T., S. Peeta, and D. Bernstein. 2001. Multi-layer Infrastructure Networks and Capital Budgeting. Working Paper TF0801A. Department of Systems Engineering and Operations Research, George Mason University, Fairfax, Virginia.

Gell-Mann, M. 1994. Complex adaptive systems. Pp. 17–28 in Complexity: Metaphors, Models and Reality, G.A. Cowan, D. Pines and D. Meltzer, eds. Reading, Mass.: Addison-Wesley.

Haimes, Y.Y, and P. Jiang. 2001. Leontief-based model of risk in complex interconnected infrastructures. ASCE Journal of Infrastructure Systems 7(1):1–12.

Heller, M., E.W. von Sacken, and R.L. Gerstberger. 1999. Water utilities as integrated businesses. Journal of the American Waterworks Association 91(11):72–83.

Jentgen, L. 2001. Implementation Prototype Energy and Water Quality Management System. Presentation at the American Waterworks Association International Conference, Washington, D.C., June 22, 2001.

Lineweber, D., and S. McNulty. 2001. The Cost of Power Disturbances to Industrial and Digital Economy Companies. Available online at *<http://ceids.epri.com/ceids/Docs/outage_study.pdf>*.

North, M.J. 2000. An Agent-Based Tool for Infrastructure Interdependency Policy Analysis. Presentation at the Rand Workshop on Complex Systems and Policy Analysis: New Tools for a New Millennium, Washington, D.C., September 28, 2000. Available online at *<http://www.rand.org/scitech/stpi/Complexity/>*.

PDD63 (Presidential Decision Directive 63). 1998. The Clinton Administration’s Policy on Critical Infrastructure Protection: Presidential Decision Directive 63. Available online at *<http://www.terrorism.com/homeland/pdd63.htm>*.

Rinaldi, S.M., J.P. Peerenboom, and T. Kelly. 2001. Identifying, understanding, and analyzing critical infrastructure interdependencies. IEEE Control Systems Magazine 21(6):11–25.

TTI (Texas Transportation Institute). 2001. The 2001 Urban Mobility Study. College Station, Texas: Texas Transportation Institute.

WaterTech OnLine. 2001. US Official Calls for More Water Recycling. Available online at *<http://waternet.com/News.asp?mode=4&N_ID=24622>*.

Wireless Communications

Design Challenges for Future Wireless Systems

ANDREA GOLDSMITH
Department of Electrical Engineering
Stanford University
Stanford, California

A wireless communications system that provides a high-speed, high-quality exchange of information between portable devices located anywhere in the world is the communications frontier of the next century. The potential uses of wireless technology include multimedia, Internet-enabled cell phones; home entertainment networks; smart homes and appliances; remote learning and telemedicine; autonomous sensor and robot networks; and automated highways. The exponential growth of wireless communications worldwide, coupled with the potential impact of future wireless systems on people everywhere, has captured the attention of the media and the imagination of the public. However, many technical challenges will have to be overcome for the promise of wireless communications to become a reality. In this paper I will outline a vision for future wireless systems and then describe some of the inherent technical challenges to realizing this vision.

WIRELESS VISION

Wireless networks will enable people on the move to communicate with anyone, anywhere, at any time, via a range of multimedia services. Wireless networks will form virtual offices anywhere in the world from small hand-held devices that provide seamless telephone, modem, fax, and computer communications. Wireless local area networks (LANs) will connect palmtop, laptop, and desktop computers anywhere in an office building or on a campus, as well as in the corner cafe. In the home, LANs will enable a new class of intelligent home electronics that can interact with each other and with the Internet. Video teleconferencing will be possible between buildings that are blocks or continents apart; teleconferences can include travelers as well, from salespeople who miss their plane connections to CEOs sailing in the Caribbean. Wireless video will create remote classrooms, remote training facilities, and remote hospitals anywhere in the world. Wireless technology will also enable developing countries to build up telephone and computer networks quickly and cheaply, effectively bypassing the high cost and delays associated with building wired infrastructures. Sensors with wireless transceivers will have the ability to self-configure into networks, make localized decisions, and relay information to centralized decision centers. Potential applications for sensor networks include environmental monitoring, energy-efficient temperature control, security systems, and automated highways.

THE WIRELESS CHANNEL

The wireless communications channel poses the biggest obstacle to building high-performance wireless networks. Because the fundamental characteristics and capacity limits of the wireless channel affect all aspects of the wireless network design, designers cannot just borrow from the networking technologies evolving to support the explosive demand for wired networks, which typically operate over fiber-optic channels with very high data rates and very low probabilities of signal corruption. Neither of these characteristics is enjoyed by wireless channels.

One key aspect of radio-wave propagation that affects system design is path loss, which dictates that the power in a radio wave diminishes with the distance it travels. Thus, the coverage area of a wireless system is dictated by how far wireless devices can be separated and still maintain reasonable levels of received signal power. In addition, signals propagating out from a transmitter may be reflected off surrounding buildings or objects and arrive at the receiver after some delay and with phase shift relative to the original transmission. This causes self-interference between the signal transmission and its reflection, which can give rise to large amplitude fluctuations (fading) in the received signal and interference between subsequently transmitted bits. Self-interference can severely degrade the quality of the received signal. A signal propagating through a wireless environment may also be blocked by buildings or other objects, resulting in a very weak signal at the receiver. In addition, because radio is a broadcast medium, all transmissions that share the same signal bandwidth interfere with each other, and this interference must be managed through the system design. Moreover, as devices move around, their signals and interference propagation characteristics are subject to random changes. In short, the wireless radio channel as a medium for reliable high-speed communication poses severe challenges to designs of wireless systems. Not only is it susceptible to noise, interference, signal blockage, and reflections, but the impediments also change over time in unpredictable ways.

CAPACITY LIMITS OF WIRELESS CHANNELS

The nature of wireless signal propagation, coupled with limited available bandwidth, severely limits the rate at which data can be sent over wireless channels. Although the fiber-optic cable that supports most wired network traffic can send data at a rate measured in gigabits (10^9 bits) per second, indoor wireless systems can only send data at the rate of megabits (10^6 bits) per second, and outdoor wireless systems at the rate of kilobits (10^3 bits) per second. Thus, the data rates for wireless systems today are from three to six orders of magnitude lower than the data rates of their wired counterparts, and the gap is growing. Given that end-users will expect the same performance for both types of networks and that applications for wireless systems will be mostly designed relative to wired network constraints (because wired networks are more prevalent), overcoming and adapting to the data rate limitations of wireless channels will be essential for the success of high-performance wireless systems.

Channel capacity is a fundamental characteristic of a wireless channel that defines the channel's maximum possible data rate. One way to increase the capacity of a wireless channel is to use more bandwidth—a larger portion of the radio spectrum—because channel capacity is typically proportional to the bandwidth of the channel. Unfortunately, bandwidth is a very expensive and tightly regulated commodity, so increasing capacity by increasing channel bandwidth typically comes at a very high price. A less costly approach is to develop techniques that maximize channel capacity in a given bandwidth. However, the random propagation and interference characteristics of the wireless channel conspire to reduce this capacity. In fact, these characteristics make it difficult to derive the fundamental capacity limits in the first place. Thus, determining the capacity limits of wireless channels and developing strategies to achieve transmission data rates close to capacity remains an area for active research.

WIRELESS SYSTEM DESIGN CHALLENGES

Challenges to making the wireless vision a reality arise at all levels of system design, including the hardware, link, network, and application levels. Future wireless networks will require tightly integrated, adaptive designs that transcend the boundaries of these design layers to deliver the best possible performance given the random variations in the constraints of the overall system.

The challenge for designers of wireless hardware is to enable terminals with multiple modes of operation to support different applications. Desktop computers currently have the capability of processing voice, image, text, and video data for small, lightweight, hand-held devices; however, breakthroughs in circuit design will be necessary before multimode operation can be implemented. Because most people will not carry around a 20-pound battery, the signal processing and communications hardware of portable terminals must consume very little power. Many of today's signal processing techniques that increase channel capacity and mitigate channel impairments require a lot of processing power. Thus, breakthroughs in efficient algorithms, as well as in battery design, will be necessary to overcome the limitations of hardware power.

Another major design challenge will be overcoming the capacity limits, interference levels, and random variations of the wireless channel. Significant breakthroughs have been made in this arena over the last decade, driven mainly by commercial cellular technology. These breakthroughs include: multiple antennas at the transmitter and receiver to increase channel capacity; sophisticated coding strategies to correct channel-induced bit errors; multiuser detection techniques to reduce interference; equalization, spread-spectrum, and multicarrier modulation to reduce self-interference from signal reflections; adaptive modulation to optimize performance over time-varying channels; and dynamic resource allocation for sharing power and bandwidth among multiple users in the system as channel conditions and requirements change. Although these breakthroughs represent significant accomplishments, much more work will be necessary to improve wireless channels.

Wireless networking also presents a significant challenge. The network must be able to locate a given user among millions of mobile terminals and route a call to that user, which could be moving at speeds of up to 100 mph. The finite resources of the network must be allocated fairly and efficiently to meet changing user demands and locations. Today, a tremendous infrastructure has been developed for wired networks: the telephone system, the Internet, and fiber-optic cable, which should also be used to connect wireless systems into a global network. However, because wireless systems with mobile users are not likely to be competitive with wired systems in terms of data rate and reliability, the design of protocols to provide interfaces between wireless and wired networks with vastly different performance capabilities remains a challenge.

Wireless systems must support wireless applications, which may have very different requirements (e.g., voice mail and email). It is impossible to design a "one-size-fits-all" wireless network that can support all of the applications that exist today, let alone the applications that will evolve in the future. Moreover, it is impossible to guarantee fixed performance metrics (e.g., data rate or a hard-delay constraint) for a wireless network because of the underlying random channel and network dynamics. Thus, wireless applications will have to be adapted to these dynamics to deliver the best end-to-end performance. For example, a wireless video application might require a data rate of 10 megabits per second for very high picture and sound fidelity. However, if the underlying network cannot support this rate, the resolution could be scaled back to a rate commensurate with system capabilities. Under extremely poor network conditions, the video might revert to a sound-only mode. Although ideally applications could always deliver high performance, the ability to scale back quality on the fly in response to degraded network conditions will be an essential characteristic of robust wireless system design.

Perhaps the most significant technical challenge to wireless network design is an overhaul of the design process itself. Most wired network designs are based on a layered approach; that is, the system hardware design, link design, network design, and application design are all created independently of one another with baseline mechanisms that interface between the design layers. Although this methodology leads to some inefficiencies and performance loss because the global design is not optimized, it also greatly simplifies the overall system design. For wired networks, because they have large capacity and good reliability, and because the performance loss resulting from layering is fairly low, the layered design works well. The situation is very different for wireless networks. Not only can wireless links exhibit very poor performance, but this performance, along with user connectivity and network topology, changes over time. In fact, the very notion of a wireless link is somewhat fuzzy because of the nature of radio propagation, and because of the dynamic nature and poor performance of the underlying wireless communication channel. High-performance wireless systems will have to be optimized for this channel and must adapt to its variations as well as to user mobility. Thus, wireless systems will require a tightly integrated and adaptive design that transcends hardware, link, network, and application layers. Given the underlying constraints and dynamics of the channel and network, as well as the application requirements, each layer of the system design, as well as across layers, will have to adapt for the system to deliver the best end-to-end performance.

SUMMARY

Wireless communications could have an enormous impact on our daily lives, not only in the way we communicate, but also as a technology enabler for other large-scale systems. However, many technical challenges will have to be overcome for us to realize this potential. This will require an interdisciplinary approach to wireless system design and creative thinking about the large-scale systems that wireless technology could enable.

FURTHER READING

Fasbender, A., R. Reichert, E. Geulen, J. Hjelm, and T. Wierlemann. 1999. Any network, any terminal, anywhere. IEEE Personal Communications 6(2):22–30.

IEEE Communications Magazine. 2001. QoS and resource allocation in the 3rd generation. IEEE Communications Magazine (Special Issue), February 2001.

Katz, R.H. 1994. Adaptation and mobility in wireless information systems. IEEE Personal Communications 1(1):6–17.

Negus, K., A. Stephens, and J. Lansford. 2000. HomeRF: wireless networking for the connected home. IEEE Personal Communications 7(1):20–27.

Noble, B. 2000. System support for mobile, adaptive applications. IEEE Personal Communications 7(1):44–49.

Next-Generation Mobile Wireless Internet Technology

Rajiv Laroia
Flarion Technologies
Bedminster, New Jersey

Next-generation wireless data systems are expected to offer broadband Internet connectivity. The North American and European standards for these are based on code division multiple access (CDMA) technology, which was originally developed for circuit-switched voice systems and is not as suitable for packet-data transmission. In this paper, I describe flash-OFDM, a new air-interface technology designed specifically for packet-data transmission. Flash-OFDM is a mobile, wide-area-network cellular technology that works well with all existing Internet protocols and delivers a seamless extension to the Internet over the air. Flash-OFDM base stations are access routers that plug directly into the edge routers of an all-IP infrastructure. Mobility is managed using the mobile IP protocol.

Unlike third-generation and other conventional wireless technologies that are first designed primarily at the physical layer, flash-OFDM is a data-oriented technology designed and optimized across all layers of the protocol stack, including the networking layers. As its name suggests, the underlying multiple-access technology for flash-OFDM is orthogonal-frequency division multiple access (OFDMA or OFDM). However, flash-OFDM is much more than OFDM; it is an innovative, system-level technology that exploits the unique properties of OFDM to enable efficient packet-data transmission in a cellular network.

INTRODUCTION

The potential for a mobile wireless data technology that offers cost effective, seamless connectivity to the Internet is virtually unlimited. One of the big drawbacks of conventional cellular wireless data technologies, such as third-generation (3G) technologies, is that they evolved from circuit-switched voice technology, which is inherently inefficient and in many ways unsuitable for handling data traffic. The problem is compounded because the Internet evolved in the wired world, which offers much more reliable (error-free) connections than the inherently unreliable and error-prone wireless channel. As a consequence, the protocols of the Internet misinterpret channel errors for network congestion resulting in very poor link efficiency.

Flash-OFDM, an air-interface technology designed specifically for packet-data, addresses most of the problems of conventional wireless data systems. Flash-OFDM offers reliable packet-data connectivity with very small delays, thereby supporting interactive data applications, such as voice, video, games, and more. Flash-OFDM also enables complete control over quality of service (QOS) on both uplink and downlink. Thus, with flash-OFDM, operators can offer simplified pricing plans, such as flat-rate, “all-you-can-eat” billing, and can maximize revenue by offering tiered services at tiered prices. Enterprise users, for example, could pay more than consumers for higher quality service.

Because flash-OFDM is data-oriented technology, base stations are access routers designed to connect directly into the edge routers of a managed, all-IP network with all standard components and no special wireless-specific boxes. In contrast, 3G data systems require many expensive, wireless-specific, non-IP boxes in a circuit-oriented access network to manage mobility.

PACKET-SWITCHED AIR INTERFACE

The telephone network, which is designed basically for voice transmission, is an example of a circuit-switched system. Circuit-switched systems exist only at the physical layer and use the channel resource to create a bit pipe, a dedicated resource that requires no control once it is created (some control may be required in setting up or bringing down the pipe). Circuit-switched systems, however, are very inefficient for transmitting burst-data traffic.

Packet-switched systems are very efficient for transmitting data traffic but require control layers in addition to the physical layer that creates the bit pipe. A media-access control (MAC) layer is required for data users to share the bit pipe. A link layer is also necessary to create a reliable link from the error-prone pipe to the network layers so overflows of packet data can be transmitted. The Internet is a good example of a packet-switched network.

Because all conventional cellular wireless systems, including 3G systems, were fundamentally designed as circuit-switched, voice-transmission systems, they were designed and optimized primarily at the physical layer. The choice of code-division multiple access (CDMA)[1] as multiple-access technology at the physical layer was also dictated by voice-transmission requirements. Flash-OFDM,

[1] 3G systems in Europe (WCDMA) and the United States (CDMA 2000) are based on code division multiple access or CDMA technology.

however, a packet-switch system designed for data transmission, is optimized across the physical, MAC, link, and network layers. The choice of OFDM as the multiple-access technology is based not just on physical layer requirements but also on MAC, link, and network layer requirements.

ADVANTAGES OF FLASH-OFDM

Physical Layer

OFDM is a robust, multiple-access technology that can address problems with the wireless channel, such as multipath fading, delay spread, and Doppler shifts. OFDM, also called multitone modulation, divides the available spectrum into a number of equally spaced tones. For each OFDM symbol duration, information-carrying symbols (e.g., QPSK or QAM)[2] are loaded on each tone. Flash-OFDM is a spread-spectrum technology that fast-hops across all tones in a pseudorandom, predetermined pattern. With fast-hopping, a user assigned one tone does not transmit every symbol on the same tone but uses a hopping pattern to jump to different tones for every symbol duration. Different base stations use different hopping patterns and can use the entire available spectrum (frequency reuse of 1). Thus, in a cellular deployment, flash-OFDM has all the advantages of CDMA systems, including frequency diversity[3] and out-of-cell (intercell) interference averaging—a spectral efficiency benefit that narrow-band systems like conventional TDMA (time division multiple access) do not have.

In addition, because the various tones in OFDM are orthogonal, different users in the same cell use different resources (tones) and hence do not interfere with each other. This is similar to TDMA, where different users in a cell transmit at different time slots and do not interfere with one another. In contrast, CDMA users in a cell do interfere with each other, thereby increasing the total interference in the system. At the physical layer, therefore, flash-OFDM has the advantages of both CDMA and TDMA and is at least three times as efficient as CDMA. In other words, at the physical layer, flash-OFDM creates the fattest pipe of all cellular technologies.

MAC AND LINK LAYERS

In addition to the huge advantage of flash-OFDM at the physical layer, the most significant advantages of flash-OFDM for data transmission are at the MAC and link layers. Flash-OFDM exploits the granular nature of resources in OFDM

[2] Quadrature phase shift keying, quadrature amplitude modulation.

[3] Frequency diversity counteracts fading when a user's signal spans a wide spectrum, and, thus, usually does not all fade at the same time.

to create extremely efficient control layers. With appropriately designed OFDM, it is possible to send a very short amount (as little as one bit) of information from the transmitter to the receiver with virtually no overhead. In other words, a transmitter that has not been transmitting can begin transmitting, transmit as little as one bit of information, and stop without causing any resource overhead. With CDMA or TDMA, the granularity is much coarser and initiating a transmission wastes a significant resource. TDMA, for example, has a frame structure, and whenever a transmission is initiated a minimum of one frame (a few hundred bits) of information is transmitted. The frame structure does not cause any significant inefficiency in user data transmission because data traffic typically consists of a large number of bits. However, for transmitting control-layer information, the frame structure is extremely inefficient because control information typically consists of only one or two bits but requires a whole frame. A nongranular technology can, therefore, be a disadvantage at the MAC and link layers.

Flash-OFDM takes advantage of the granularity of OFDM in its control-layer design enabling the MAC layer to perform efficient packet switching over the air and, at the same time, providing all of the hooks to handle QOS. Flash-OFDM also supports a link layer that uses local (as opposed to end-to-end) feedback to create a very reliable link from an unreliable wireless channel with very low delays. Therefore, there are few delays at the network layers and no significant delay jitter. Hence, the system can support interactive applications, such as (packet) voice transmissions. Moreover, Internet protocols like TCP/IP (transport control protocol) run smoothly and efficiently over a flash-OFDM air link. By contrast, TCP/IP performance on 3G networks is very inefficient because the link layer introduces significant delay jitter that TCP interprets as network congestion and, therefore, responds by backing off to the lowest rate.

Packet switching leads to efficient statistical multiplexing of data users and enables wireless operators to support many more users for a given user experience. Combined with QOS support and a pipe three times as fat as with CDMA or TDMA, flash-OFDM enables operators to maximize revenue per Hertz of spectrum.

SUMMARY

Flash-OFDM is an air-interface technology designed to meet the needs of packet-data transmission. It provides seamless mobile connectivity to the Internet and can handle all current applications, including interactive data applications and peer-to-peer applications. Flash-OFDM would also impose the fewest constraints on projected future applications.

Service Architectures for Emerging Wireless Networks

S. Muthukrishnan
AT&T Laboratories, Florham Park, New Jersey
Rutgers University, New Brunswick, New Jersey

As the next generation of wireless telecommunications systems (e.g., 3G) emerges, vendors and service providers are conducting technical and market trials. The challenge now is to design and deploy an infrastructure for services on an (inter)national scale that combines voice and data transmission, integrates multiple devices and networks, and provides a wide range of services for wireless users—enhanced messaging, location-aware services, tracking and notification services, etc. We address the design issues and challenges involved to deploy this communication and service infrastructure, unique in scope and scale, which would be a great engineering achievement.

CONTEXT

Wireless communication has been around for a long time (e.g., commercial radio, satellite broadcasting, two-way radio systems, etc.). The first commercial wireless telecommunications networks were analog cellular systems; in the past decade, most of these systems have become digital (so called 2G) systems, and cellular traffic has increased dramatically. We now have four major cellular companies (cellcos) (Verizon Wireless with 27.9 million customers, Cingular with 20 million, AT&T Wireless with 18.8 million, and Sprint PCS with 11.2 million) and a dozen other large companies. There are an estimated 100 million wireless subscribers in this country, which is still fewer than the number of wire-line customers. Cellcos offer telephone service, voice mail, one- or two-way short messaging service (SMS), enhanced directory services (e.g., concierge service by AT&T Wireless), and E911 (rough locating of emergency callers). In addition, most cellcos offer basic data service (e.g., PocketNet by AT&T Wireless and Wireless Web by Sprint PCS), which give cell-phone users access to a limited amount of information about movies, stock quotes, news, and sports. There is a data network—Cellular Digital Packet Data (CDPD)—that powers wireless modems for PC, laptop, and personal digital assistant (PDA), such as Palm or PocketPC, users and supplies data at roughly 10kbps (9.6, 13.3, etc.). There is also a fledgling convergence of PDA and phone systems—Kyocera produces PDAs that are also cell phones, and Handspring is a PDA with a module extension that makes it a cell phone. Overall, estimates of data users do not exceed 5 million nationwide. Internationally, technology and data rates are much the same as in the United States, but usage capacity and usage patterns differ. Wireless customers outnumber wire-line customers in a few European countries, and SMS is widely popular. Developing countries are adopting cell phones aggressively. China has more than 80 million subscribers; almost 40 percent of people in the Czech Republic subscribe to cell phones; and DoCoMo in Japan has 25 million subscribers to its data/information service called i-mode, which was launched in February 1999 (DoCoMo, 2001).

Most people believe that the cell phone market has just begun to take hold and that it will grow significantly; the enormously successful Internet is expected to be a new market for wireless technology. The world of the future is envisioned as a world in which users will be able to communicate anywhere and everywhere by voice (e.g., telephone calls) and data (e.g., text, audio, images, etc). Governments are providing incentives to encourage the growth of wireless communications to realize this vision, which calls for increased high-speed capacity. The governments of various countries in Asia, Europe, and North America have provided frequency spectrum for commercial development and auctioned off frequency licenses. Major cellcos are in various stages of building next-generation wireless networks (3Gs, with a possible evolution path via 2.5G) using this spectrum. Some debate has arisen about the advantages of different physical-layer technologies (e.g., TDMA, GSM, GPRS/EDGE, variants of CDMA, etc.). Vendors are in the process of producing the network elements (i.e., base stations, switches, etc.) to support the technologies service providers will need, and many technology trials have been conducted. Japan's DoCoMo is doing a market trial of 3G systems, and AT&T Wireless is doing a market trial of 2.5G systems. Many cellcos in Asia (e.g., Japan, Korea, and Taiwan), Europe, and North America have announced plans for market trials and the deployment of 3G (2.5G+) systems over the next five years. These systems promise capacities of a few tens of kbps (50+ for GPRS) to a few hundred kbps (350+ for 3G) per cell, which will be significantly larger than the capacity of existing 2G systems. In short, we are on the threshold of building a new telecommunications infrastructure for large-scale use.

Major communication and services infrastructures (e.g., roads, bridges, railroads, telephone networks, radio broadcasting networks, cable networks, the Internet, etc.) are built only once every few decades. It is helpful to keep this perspective in mind when thinking about the next-generation wireless network.

Cost is an important consideration. Cellcos worldwide have spent hundreds of billions of dollars *up front* to acquire licenses to use spectrum, a surprisingly expensive "access cost" for developing a service in the historical context. Hence, market share will be a crucial guide to the nature and development of services. Technologically, wireless access (transport as well as backbone networks) will handle both real-time voice (telephone calls) as well as data (real-time streaming or nonreal-time) traffic. This will require more sophisticated networking than the existing Internet and, consequently, will pose new challenges. Emerging wireless networks will have to handle a large number of devices, and, because customers are likely to own more than one wireless device (e.g., cell phones, PDAs, laptops, and other appliances with wireless modems, etc.), 3G networks will have to handle many more customers than existing networks. Thus, scalability will be a central concern. One result will be that next-generation wireless networks may need larger address space than the space currently available in the Internet and, hence, may deploy IPv6[1] ; this will present another set of challenges because IPv6 is an emerging technology. Furthermore, end devices are likely to be diverse. Phones, computing devices, media, and gaming devices networked by wireless means will have different form factors, displays, and computing power and will understand varied protocols and data formats. Providing suitable service across heterogeneous wireless devices presents another challenge. In addition, the wireless channel is unpredictable, with variations that depend on the mobility of the user and the terrain, which affects perceived bandwidth and may lead to intermittent connectivity. Special care will be necessary in handling the application as well as the user to ensure that service is adequate with any application. Finally, security and privacy concerns will be critical, perhaps even more urgent than in existing Internet and telecommunications infrastructure because the air interface is more accessible to snoopers than other access methods, and users will have a large "footprint" on the network because of their mobility.

To summarize, emerging wireless networks will involve the deployment of cutting-(bleeding?)-edge technologies that will be converged and complex, a marvelous engineering achievement. In general, the deployment of new networks will provide opportunities for initiating new value-added services. In the current context, massive investments that have been (and will be) made will provide additional incentives for companies to develop viable revenue sources, and hence, to focus on services and applications.

APPLICATIONS AND SERVICES

I will approach the subject of applications and services for wireless users in emerging networks primarily from the point of view of large, cellular

[1] See <*http://playground.sun.com/pub/ipng/html/ipng-main.html*> for more details on IPv6, which, among many other things, provides 128 bits of address space as opposed to the 32-bit address space on the current Internet.

telecommunications companies (cellcos), which will play a major role in the development of the next-generation wireless infrastructure and have already made a sizeable investment. Realistically, however, no one cellco will become the one-stop portal for all services of interest to wireless users. It is reasonable to expect that third-party companies will provide niche services. In the future, the industry will have to resolve the tension between services managed by cellcos (within the so called "walled garden") and services managed by third-party companies and find revenue-sharing mechanisms for third-party services. The following description of services and applications in emerging wireless networks will address the technological, as well as the historical context. I begin with detailed descriptions of two services—enhanced messaging and location-aware services—as a way of identifying some of the issues a services architecture will have to address. Then, I will provide an overview of other considerations for cellcos designing a services architecture.

ENHANCED MESSAGING

Current wireless systems provide simple, nonreal-time messaging services, such as voice mail and one- or two-way text messaging using SMS and paging; instant messaging is a popular real-time application. Emerging wireless telecommunication systems envisage messaging applications with much richer media, involving audio, video, Web, and text messages. Here is a sample scenario for multimedia messaging.

Alice gets off a plane and turns on her phone (the use of 3G systems will continue to be prohibited on airplanes!). She has an email from Bob containing the system drawing in image format and a note that all changes have been made as previously agreed upon. She trusts Bob, so without checking the image, she forwards it to the product group with a voice memo. As she waits for her baggage to arrive (3G networks will not reduce baggage delivery time!), she checks her personal email and sees a message from her supermodel sister, Carole, containing a cute animation video clip of "Dilbert." She watches the clip and forwards it to her colleagues and friends. She also responds to her sister with an archived audio clip of the "Top Ten Reasons Why Models are Clueless" from David Letterman. Meanwhile, she receives an SMS message notifying her of a birthday e-card from Erin that was sent to her home PC earlier; she connects to the associated Web site, listens to part of the audio, which happens to be her favorite song, and skips to the end. The e-card image cannot be rendered on her phone legibly, so she forwards it to her home PC. As she walks to the exit, she receives an instant message from the limo driver waiting for her. She responds and is directed to the limo. Finally, she sits back massaging her aching eyes, wrists, and shoulders—airplane rides and 3G handsets are still hard on your joints!

This simple, fictional scenario captures some of the expectations of future wireless systems. First, to engage users, applications will have to be highly relevant, simple to use, and fun. Current mailers would be challenged to meet these requirements. The first problem is to avoid redundant data transfers. For example, Alice forwards Bob's image attachment without viewing it. Therefore, there is no reason for the image attachment to be downloaded from the mail server to the handset and resent from the handset to the mail server for forwarding. This would waste uplink and downlink bandwidth. The second problem is data transcoding. Wireless handsets are likely to remain heterogeneous, with different abilities to decode data in different formats or to render them meaningfully at different granularity. Current mailer systems do not provide a way for clients to negotiate from one level of detail to another or to request conversion from one format to another. This is likely to be a major problem. Either fairly uniform handsets will have to be introduced, or the content will have to be customizable for any of the available devices, neither of which is a clean solution to the problem. Negotiating the desired data format is a network-level challenge; transcoding from one format to another is a data-management challenge.

Third, the problem at the back end is duplicate data, or the "curse of forwarding"; data is duplicated many times over because personal users tend to forward their favorite attachments to friends, family, and colleagues. In commercial settings, a manager may send a notice to several employees, or a portal may send an advertisement clip to several customers. In current mailer systems, the forwarded documents are predominantly text documents. If we extrapolate the phenomenon to bulky multimedia content (e.g., audio and video clips) over the wireless networks, where multimedia content is expected to be extensive, and mail management is performed by cellcos with tens of millions of users, the problem takes on a different dimension. Finally, there is the problem of "atomic content." Current mailer systems do not allow the mobile user to see a "synopsis" of the multimedia attachment (e.g., a summary of the content, rather than a display of content type) or to perform light editing (e.g., searching to an interesting portion of the text and splicing in a figure or a Web link) without downloading the entire attachment. Thus, the multimedia content is essentially "atomic," that is, it allows for no flexible, client-server manipulation. Unless this changes, this will result in a very heavy handset (e.g., all software and resources for multimedia manipulation must be on the handset) and could possibly lead to a poor overall user experience.

The problems I have described above are mainly data management problems, and they are not difficult to solve, in principle, but they will require clean engineering solutions. For details on how to store duplicate, multimedia dataspace efficiently, how to perform on-demand content-type negotiation for transcoding, how to enable message sending by proxy to avoid redundant data transfers, and streaming (versus downloading) of multimedia data in the messaging context, see Nelson and Muthukrishnan (2001).

Now let us focus on a specific problem inherent in the scenario described above, namely, maintaining the presence of a user on the network. Current cell phone systems use home location registers (HLRs) and visiting location registers

(VLRs) to maintain device status information.[2] For every cell phone, the HLR stores account status, presence information, and the mobile switching center (MSC) currently serving the device. A VLR, a component of an MSC, stores status information about devices in the MSC's service area. For example, a VLR often stores an estimate of the device's location to reduce the paging required to complete an incoming phone call. Instant messaging uses subscriber presence information to update buddy lists and determine when instant messages can be delivered. For conventional Internet use, determining presence includes determining whether the user is logged in and whether the user is actively using the computer (e.g., see <*http://www.jabber.com/pdf/Jabber_ Server_White_Paper.pdf*>). A weakness of current technology for maintaining and serving users' status information is the focus on device status rather than on customer status. Many customers own more than one mobile wireless device (e.g., a cell phone, an advanced pager [such as a RIM Blackberry], and a PDA with wireless connectivity [such as a Palm VII]). Some customers own more than one cell phone (e.g., a work cell phone and a personal cell phone). Hence, wireless application infrastructure must store and serve information about the user, rather than the device.

Another specific problem, associated with the multiple presence of a user, is maintaining user preferences for notifications. Customers may want to use their devices in conjunction with other communications mechanisms. For example, users might desire SMS notifications when email is delivered. If the user is near one of his or her wire-line phones, he or she might prefer to receive voice messages on the wireline device. Similarly, if the user is active on an Internet connected computer, he or she might prefer to receive messages on the computer. Also, a user named Joe on hotmail who receives an instant message request might be called John on a 3G phone and would prefer to receive an SMS message as an alert. To deliver these services, the wireless applications infrastructure will have to provide a mechanism for users to set their preferences for handling messaging events as logical rules, which will have to be maintained and applied by the service provider.

LOCATION-AWARE SERVICES

Mobility is a unique feature of wireless networks. Therefore, a user's location at any given time is a unique resource. Offering user services based on current, posited, or desired location is a tantalizing idea. I will address three issues: location-aware services; determining a user's location; and an architecture for providing location-aware services.

The simplest location-aware service would be providing geopersonal information, that is, providing information about a user's interests based on location.

[2] For CDPD systems, the network elements are different, and device ID is a fixed IP address.

The information may be surreal (e.g., “Where am I?”) or more practical (e.g., “Where is my child, spouse, or FedEx delivery agent now?”). Other typical information would involve locating closest businesses (e.g., hardware stores), entertainment (e.g., movie theaters or restaurants), traffic (e.g., road conditions), and so on.

The second issue is location determination. One approach is to coopt the user to enter a certain location, which would be inconvenient; therefore, I will focus on automatic localization without user’s aid. The Global Positioning System (GPS) provides a satellite-based solution that is accurate but does not work indoors or in dense urban areas. Furthermore, GPS must be incorporated into the handset, an expense cellcos may prefer to avoid. Some progress has been made toward creating low-cost chip sets and assisted or differential GPS solutions to overcome current coverage problems. There are also network-centric location determination methods by which a user can query network elements and use radio-layer parameters. These methods include: basic cell-ID that identifies the cell region of the user; and refined methods, including AOA (angle of arrival), E-OTD (enhanced observed time difference), signal attenuation, and TDOA (time difference of arrival). These methods all require that network elements be accessible, be “queryable,” or be able to dump traffic logs frequently; they also require vendor support and would tax already overburdened network elements. In general, these methods are preferable only when handsets are “dumb,” that is, not equipped with a powerful processor and a sophisticated operating system. For smart handsets, a scalable solution is to rely on the back channel cell-ID notification and translate cell-ID to physical location with the help of the cellco. This works very well and suffices for most geopersonal information services (Muthukrishnan et al., 2001). In a related note, E911 is a government-enforced program that localizes emergency callers; whatever infrastructure cellcos put in place must meet this demand, but only for in-progress cell phone calls and only for a small number of calls (statistics put the number of 911 calls per year nationwide at 50 million, while the number of cell phone calls per day is likely to be in the hundreds of millions [CTIA, 2001]). Hence, the infrastructure need not provide a scalable solution for location-aware services for significantly more data users particularly for location tracking. See Hightower and Borriello (2001) for a recent comparison of location-determination technologies.

A third issue is the development of an architecture for providing location-aware services. Although I will focus on data users with smart handsets, the discussion also applies to voice users and data users with other handsets, with suitable restrictions. The first approach is a location-server architecture. In its simplest form, the user’s location information is stored in a database together with customer information. This approach is dynamically populated by the location-determination technologies. The mobile user has the ability to direct any information service provider (e.g., Yahoo, Mapquest, etc.) to access the location information from this server. A number of details will have to be resolved, including how requests can be authenticated at the location server, how adequate privacy and security can be provided, and appropriate interfaces for location information suited to different applications. The second approach is a proxy-based architecture. In its simplest form, this consists of a proxy with which mobile clients communicate. Mobile clients pass their "location" information to the proxy, which not only translates it into geographic location, but also acts as a conduit to third-party information service providers by converting user requests to geocoded requests (e.g., by rewriting the URLs). Again, a number of issues will have to be addressed: how security, such as privacy and authentication, can be provided; where the proxy should be located; and appropriate interfaces for location information suited to different applications and third-party information service providers. Other architectures are possible, including local proxies.

There are also more advanced location-aware services. Instant messaging has recently expanded into wireless communications. In this context, standards organizations have recognized that location may be used for other purposes, such as sending messages to someone only if they are less than a mile from the user, etc. (see <*http://www.instantmessaging.org/*>). This would require tracking mobile users. Some location-aware services could include changing the ringing tone/volume, based on where the user is or letting the user know when he/she is at an exit on the highway. Two enticing applications are geographic push, that is, notifying or messaging any user in a geographic area, and geographic bookmarking, wherein users "bookmark" locations, manage them, and navigate using them. Location tracking would tax the scalability of the infrastructure for location-aware services, but geographic push may be considered a basic component. Geographic bookmarks would entail careful data management systems in the back end. Finally, portals might be personalized to the user's location.

OVERALL ARCHITECTURE AND ENGINEERING ISSUES

Besides enhanced messaging and location-aware services, next-generation wireless systems would enable a suite of services: administrative services (e.g., profile, preferences, purchases, personalization, security), content services (e.g., providing a Web space for downloads, stored content, coupons), personal information management (e.g., schedules, contacts, synchronization with remote devices), and transaction services (e.g., prepayments, credit, cash withdrawals). Users' expectations and experiences with wireless applications will evolve over time (Henry and Ziang, 2000). In addition to the basic services described above, cellcos are likely to support "walled garden" services and applications and provide a branded portal with associated services, such as searching. Finally, cellcos are likely to provide mechanisms for corporate Intranets, including VPNs, and for external net and third-party services. Some of these issues are discussed in

3GPP (2001). The schematics of the services architecture and a description of the flow of controls and data for various services are provided below.

The entire services architecture will have to be deployed at the scale described above. Therefore, it will have to be distributed and duplicated with careful load balancing. Database management will be critical, and the entire system will have to be engineered and maintained: including provisioning, billing, hardware/software/configuration management, fault and performance management, and data centers and customer care. All of these functions are likely to be more challenging than they are for existing communication networks. The entire endeavor will have to be of "telco quality," that is, with tiny fault probability, which will necessitate failback options. Finally, the services architecture will have to be built for the short-term, as well as for long-term evolution. Roaming among carriers and (inter)national markets, as well as interoperability of systems, services, and interfaces, should be considered for the short term. Telematics (i.e., wireless access to car users) will be of short-term to midterm interest. In the long term, wide area telecommunication infrastructures will have to interface with local, home, and personal area networks to form unified networks. Finally, service providers are likely to develop novel applications, much as they have on the Web. Whatever services architecture is put into place, it will have to be capable of evolving, as necessary.

OTHER REMARKS

Much of the discussion thus far in the media and in industry has been about physical-layer access technologies, spectrum allotments, and "killer apps" in next-generation wireless systems. I have presented some issues related to the design and deployment of a (conservative) services infrastructure. The challenges are likely to be of the "computer science and engineering" variety, namely, database management, scheduling and load balancing in distributed systems, user interface design, security, privacy, and authentication, on a grand scale.

If we look beyond engineering, a new communication and services infrastructure is being put into place, which means we have an opportunity to specify public needs that must be addressed. For example, public safety, crisis and emergency management, services for the physically challenged, and so on can be championed and implemented. From a scientific point of view, a large number of mobile users with a variety of handsets equipped with high-quality wireless connectivity represent a unique capability of gathering vast quantities of data in a distributed, real-time manner that can be integrated with other data sources.

The best way to take advantage of this previously unimaginable opportunity has yet to be determined. Market and industry may have to evolve to isolate certain "free content" to provide third-party mobile application developers an opportunity to add services atop the basic infrastructure. For example, local governments could deploy traffic and weather monitoring nodes and network them by wired and/or wireless means, and then make the data feed available to all service providers. The new infrastructure will make it easier for small third parties to provide geography-specific applications.

New communication infrastructure is also a vehicle for innovative social use. For example, artists might plant "graffiti" in the wireless world that can only be seen by users on their devices based on geographic proximity, or a large-scale symphony might be orchestrated by appropriately dialing cell phones in the audience and using their ring tones. The possibilities are limited only by the imagination.

Finally, the question arises whether devices and users can be addressed based solely on their geographic attributes, such as location and direction of movement. Answering this question will require a higher level of abstraction akin to content-addressable networking on the Web. It remains to be seen what kind of applications this will lead to.

ACKNOWLEDGMENTS

Thanks to Rittwik Jana, Ted Johnson, and Andrea Vitaletti, my colleagues working on location-aware services; Darin Nelson and Daryl Bunce at AT&T Wireless for multimedia messaging system design; Nathan Hamilton for sharing his expertise; and Robin Chen, Paul Henry, and Rob Calderbank for their discussions and perspectives.

BIBLIOGRAPHY

CTIA (Cellular Telecommunications and Internet Association). 2001. Available online at <*http://www.wow-com.com/news/publications*>.

DoCoMo. 2001. Subscribers for i-mode Service of NTT DoCoMo. 2001. Available *online at* <*http://www.nttdocomo.com/i/i_m_scr.html*>.

Henry, P., and Z. Ziang. 2000. A Subjective Survey of User Experience for Data Applications in Future Cellular Wireless Networks. AT&T Technical Memorandum.

Hightower, J., and G. Borriello. 2001. Location systems for ubiquitous computing. IEEE Computer (Special Issue), August 2001:57–66.

Muthukrishnan, S., R. Jana, T. Johnson, and A. Vitaletti. 2001. Location based services in a wireless WAN using cellular digital packet data (CDPD). Pp. 74–83 in Proceedings of the ACM International Workshop on Data Engineering for Wireless and Mobile Access (Mobide). New York: Association for Computing Machinery.

Nelson D., and S. Muthukrishnan. 2001. Design issues in multimedia messaging for next generation wireless systems. Pp. 98–103 in Proceedings of the ACM International Workshop on Data Engineering for Wireless and Mobile Access (Mobide). New York: Association for Computing Machinery.

3GPP (Third Generation Partnership Project). 2001. Available online at <*http://www.3gpp.org*>.

Wireless Integrated Network Sensors (WINS): The Web Gets Physical

Gregory J. Pottie
Electrical Engineering Department
University of California, Los Angeles

INTRODUCTION

Wireless integrated network sensors (WINS) provide distributed network and Internet access to sensors, controls, and processors embedded in equipment, facilities, and the environment. WINS combine sensor technology, signal processing, computation, and wireless networking capability in integrated systems. With advances in integrated circuit technology, sensors, radios, and processors can now be constructed at low cost and with low power consumption, enabling mass production of sophisticated compact systems that can link the physical world to networks (Asada et al., 1998; Bult et al., 1996; Dong et al., 1997; Lin et al., 1998). These systems can be local or global and will have many applications, including medicine, security, factory automation, environmental monitoring, and condition-based maintenance. Because of their compactness and low cost, WINS can be embedded and distributed at a small fraction of the cost of conventional wire-line sensor and actuator systems. Designers of systems with hundreds, or even thousands, of sensors will face many challenges.

Centralized methods of sensor networking make impractical demands on cable installations and network bandwidth. The burden on communication system components, networks, and human resources can be drastically reduced if raw data are processed at the source and the decisions conveyed. The same holds true for systems with relatively thin communications pipes between a source and the end network or systems with large numbers of devices. The physical world generates an unlimited quantity of data that can be observed, monitored, and controlled, but wireless telecommunications infrastructure are finite. Thus, even as mobile broadband services become available, processing of raw data at the source and careful control of communications access will be necessary.

In this paper, I present two scenarios illustrating different aspects of the design trade-offs. The first example is an autonomous network of sensors used to monitor events in the physical world for the benefit of a remote user connected via the Web. The second scenario explores how sensor information from an automobile could be used. A general architecture for both is shown in Figure 1. The figures does not show in detail how services can actually be supported by Internet-connected devices, but two clusters of nodes, connected through separate gateways to the Internet, can supply some services. The nodes are assumed to be addressable either through an Internet protocol address or an attribute (e.g., location, type, etc.). Unlike pure networking elements, the nodes contain a combination of sensors and/or actuators. In other words, they interact with the physical world. The gateway may be a sensor node similar to other nodes in the cluster, or it may be entirely different, performing, for example, extra signal processing and communications tasks and having no sensors. In the cluster in the top left portion of Figure 1, nodes are connected by a multihop network, with redundant pathways to the gateway. In the bottom cluster, nodes may be connected to the gateway through multihop wireless networks or through other means, such as a wired local area network (LAN). The nodes in different clusters may be all one type or they may vary within or among clusters. In a remote monitoring situation, part of the target region may have no infrastructure; thus, the multihop network must be capable of self-organization. Other parts of the region may already have assets in place that are accessible through a preexisting LAN. There is no requirement that these assets be either small or wired. The

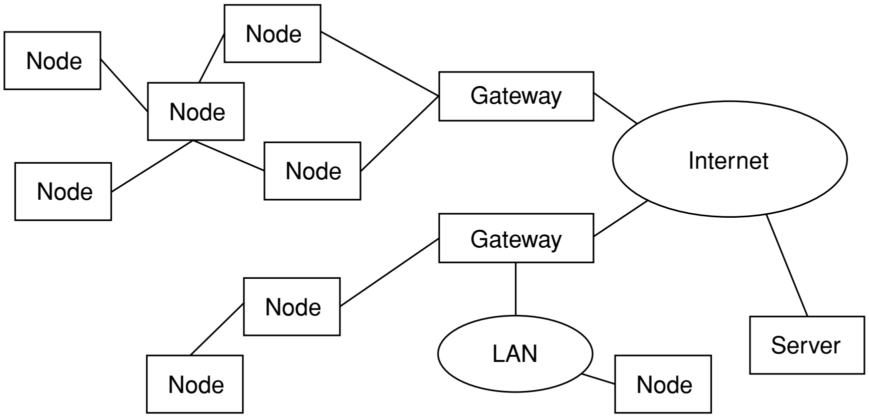


FIGURE 1 WINS network architecture.

point is to design a system that makes use of all available devices to provide the desired service.

In the next section, I briefly describe some design heuristics. This is followed by a discussion of current research on the deployment of large networks in areas without infrastructure support. The next section focuses on how sensor networks in vehicles can be linked with the Internet.

DESIGN HEURISTICS

Pottie and Kaiser (2000) described some of the fundamental physical constraints on the cost of sensing, detection, communication, and signal processing. They identified five basic design constraints:

1. For reliable detection in many situations, sensors must be in close proximity to a physical event (e.g., physical obstructions to cameras); thus large numbers of sensors may be needed. The type of information obtained with large numbers of sensors is qualitatively different from the information obtained with remote arrays.
2. The cost of sensors, radios, and signal processing will come down as the cost of integrated circuit technology comes down. The cost of batteries and other energy sources, however, will come down much more slowly.
3. The cost per bit for communications energy is often many orders of magnitude higher than for the energy required to make decisions at the source. Whereas processing cost is limited to first order only by current technology, the efficiency of communications has fundamental limits.
4. Networks must be self-organizing to be economical.
5. Scaling to larger numbers while maintaining physical responsiveness requires a hierarchy with distributed operation at lower levels and increasingly centralized control at higher levels.

Note that a hierarchy does not necessarily imply a need for heterogeneous devices. Consider, for example, a human organization. The processing abilities are roughly equal at all levels, but in progressing up the chain, different information is processed at different levels of abstraction and aggregation. Commands moving down the chain also differ in the level of abstraction, from policies to work directives, that require different levels of interpretation. This flexibility enables individuals at the lower levels to deal with local changes in the work situation much faster than if a central controller had to be consulted for each action; at the same time, global goals continue to be pursued. With machines of course, we can provide highly differentiated abilities to devices at different levels of the hierarchy. For example, a backbone long-range high-speed communications pipe can greatly reduce latency compared to multihop links. Thus, even though a logical rather than physical hierarchy is arguably much more important to scalability, the designer of large-scale systems must not be seduced by the siren song of homogeneity and should consider both. In any case, homogeneity is impractical in long-lived systems composed of integrated circuit components. For systems that use the Internet, the architecture must accommodate successive generations of more powerful components.

REMOTE MONITORING

I will now consider a more concrete example of a system that identifies particular classes of targets passing through a remote region. The targets could be military vehicles, species of animals, pollutants, seismic events on Mars, or, on a smaller scale, enzyme levels in the bloodstream. In any case, let's assume there is no local power grid or wired communications infrastructure, but that there are long-range communications for getting information to and from a remote user. In laying out a network like the one depicted in Figure 1, both energy and communications bandwidth can be critical constraints. If the network must scale in the number of elements, much of the signal processing will have to be performed locally. For example, in studying the behavior of animals in the wild, a dense network of acoustic sensors might be used. The nodes would contain templates for identifying the species emitting the call. Nodes that made a tentative identification could then alert their immediate neighbors so the location of the animal could be roughly determined by triangulation. Infrared and seismic sensors might also be used in the initial identification and location processes. Other nodes would then be activated to take a picture of the target location so a positive identification could be made. This hierarchy of signal processing and communications would be orders of magnitude more efficient in terms and energy and bandwidth than sending images of the entire region to the gateway. In addition, with the interaction of different types of nodes, most of the monitoring would be automated; humans would be brought into the loop only for the difficult final recognition of the visual pattern of preselected images. Upon positive identification, the audio and infrared files corresponding to the image would be added to a database, which could subsequently be mined to produce better identification templates. Note that with long-range communications links (via the gateway), the user could make the full use of web-accessible utilities. Thus the end user would not have to be present in the remote location, and databases, computing resources, and the like could all be brought to bear on interpreting the (processed) data.

Experimental apparatus for initial exploration of an application domain and the apparatus that will actually be needed for large-scale deployment may differ. Because networked sensors have hitherto been very expensive, relatively little array data are available for most identification purposes, and sensors have typically been placed much farther from potential targets than they will be with WINS. This means, paradoxically, that initially fairly powerful nodes will have to be constructed to conduct large-scale experiments to collect raw data and suitable identification algorithms developed from the resulting database. In experimenting with different networking algorithms, it is desirable, from the point of view of software development, to provide an initial platform with considerable flexibility. The DARPA SensIT Program has produced development platforms to support this kind of experimentation (Kumar, 2001). Other researchers have focused on specializing functions and miniaturizing components to demonstrate that large networks of small nodes can be produced. Sensoria Corporation's WINS NG 2.0, for example, nodes include ports for four sensors, a real-time digital signal processor, memory, a main processor running Linux, a battery and port for external power, the global positioning system (GPS), Ethernet, an RS-232 port, and two radios for convenient synthesis of multihop networks. Software interfaces have been created to enable programmers to control remotely a large number of physical attributes of nodes and to download new applications remotely. Thus, diverse users can produce algorithms for networking, target identification, and distributed database management. On another track, researchers at the University of California-Berkeley are engaged in producing very small nodes with limited sensing and communications abilities to demonstrate that sensing, signal processing, and communications can be combined in a miniature package.

AUTOMOTIVE APPLICATIONS

All automobiles produced recently include many processors and sensors, as well as a variety of networks for sensing, control, and entertainment systems. For example, hundreds of sensor parameters are accessible through the on-board diagnostic port. However, there are no connections between these networks and external communications systems, such as cellular phones. The Automotive Multimedia Interface Collaboration (AMI-C) has been working on ways to connect these networks and provide standardized buses in automobiles so a wide range of consumer electronics can be installed. This would create an automotive intranet that could then be conveniently accessed via the Internet (AMI-C, 2001). Ports on the bus could include any of a number of radios, so wireless devices in the vehicle could become part of the intranet, or short-range high-speed communications could be possible between a vehicle and a residence or service station.

A key component of the architecture envisioned in the AMI-C standard is a gateway that separates proprietary and safety-sensitive systems in the automobile from after-market consumer electronics. The gateway would have separate ports for interfacing with legacy networks and consumer buses. The gateway would also host software for managing the various services envisioned for internet-connected vehicles. For example, maintenance information would enable manufacturers to learn how their vehicles are actually used or enable consumers to evaluate the need for repairs and determine the effectiveness of repairs by comparing data before and after. Other potential uses could include uploading of entertainment information and locating nearby retail stores, restaurants, or service stations.

A vital function of the gateway in making such services economical is management of the communications links. Presently, cell phones have a much higher cost per bit delivered than other means of communication. However, if the automobile also has a short-range broadband link, such as IEEE 802.11b, then information might be processed and stored until it can be uploaded to a home computer when the car is parked near the residence. In a similar way, entertainment information or software upgrades could be downloaded overnight. Another approach would be to communicate over high-speed links at a gas station during refueling, for example, to receive updated information or complete a purchase of digital audio files. For very high-priority services, such as emergency assistance, the cell phone would be used, rather than waiting until a high-speed port comes into range. Based on the vehicle operator's preferences, the gateway could choose an appropriate mix of local processing, storage, and communications that would provide services at the desired costs.

The high-level requirements for the design of the gateway are surprisingly similar to the requirements for the development nodes described in the first scenario for conducting large-scale data collection experiments. Common requirements include real-time components, general purpose processors, wired and wireless network communication interfaces, application program interfaces that permit construction of software by third parties, and remote controllability via the Web. Although the devices are quite different, the same architecture applies. For a vehicle, the gateway may have some devices that respond to the physical world directly, or such devices may be accessible through local area networks in the vehicle. For the sake of economy, some of these devices and/or the gateway would perform local processing. Rather than sending a continuous record of engine temperature, for example, detailed reports might be stored only when temperatures cross a critical threshold or when the temperature is high and another sensor indicates possible problems. Further processing might even make a preliminary diagnosis, after which a query to an expert system located on the Web might be made. In this way, the vehicle would not have to host the complete diagnostics system.

Remote monitoring and control would also be attractive for other reasons. Vehicle owners will probably not want to program their preferences while operating the vehicle, and any sensible regulatory regime will surely discourage driver distractions. Scaling is also a concern. Providing services to millions of vehicles presents enormous challenges, both in terms of the huge volume of data that can be generated by vehicles and the quantity of entertainment information that may have to be transported to them. With a gateway and back-end web-server network that enables remote downloading of software, many different companies will be able to compete for providing information services to automobile owners.

CONCLUSION

Intertwined network processing is a central feature of systems that connect the physical and virtual worlds. Research is now proceeding on the design of small, specialized nodes that could potentially be deployed in very large numbers and on the creation of dense networks of larger nodes that can be used to learn more about the types of networking, sensing, and signal processing that will be needed in future systems. Because of constrained communications, design considerations for scalable networks will be similar even if data rates and processing capabilities vary greatly. Signal processing and communications must be considered together for a very broad range of systems that interface to the physical world.

REFERENCES

AMI-C (Automotive Media Interface Collaboration). 2001. Available online at *www.ami-c.org/home.htm.*

Asada, G., M. Dong, T.S. Lin, F. Newberg, G. Pottie, H.O. Marcy, and W.J. Kaiser. 1998. Wireless integrated network sensors: low power systems on a chip. Pp. 9–12 in Proceedings of the 24th IEEE European Solid-State Circuits Conference. Den Hague, The Netherlands: Elsevier.

Bult, K., A. Burstein, D. Chang, M. Dong, M. Fielding, E. Kruglick, J. Ho, F. Lin, T.-H. Lin, W.J. Kaiser, H. Marcy, R. Mukai, P. Nelson, F. Newberg, K.S.J. Pister, G. Pottie, H. Sanchez, O.M. Stafsudd, K.B. Tan, C.M. Ward, S. Xue, and J. Yao. 1996. Low power systems for wireless microsensors. Pp. 17–21 in Proceedings of International Symposium on Low Power Electronics and Design. Monterrey, Calif.: IEEE.

Dong, M.J., G. Yung, and W.J. Kaiser. 1997. Low power signal processing architectures for network microsensors. Pp. 173–177 in Proceedings of 1997 International Symposium on Low Power Electronics and Design. Monterrey, Calif.: IEEE.

Kumar, S. 2001. Sensor Information Technology. Available online at *www.darpa.mil/ito/research/sensit.*

Lin, T.-H., H. Sanchez, R. Rofougaran, and W.J. Kaiser. 1998. CMOS front end components for micropower RF wireless systems. Pp. 11–15 in Proceedings of the 1998 International Symposium on Low Power Electronics and Design. Monterey, Calif.: IEEE.

Pottie, G.J., and W.J. Kaiser. 2000. Wireless integrated network sensors. Communications of the ACM 43(5):51–58.

Technology and the Human Body

Applying Simulation Technology to the Life Sciences

THOMAS PATERSON
Entelos, Inc.
Menlo Park, California

Although simulation is considered a powerful, well accepted approach to understanding the behavior of complex engineered systems, it is not held in the same high esteem in the life sciences. The many differences between the fields of engineering and biology are the results of long-standing different traditions reinforced by undergraduate and graduate degree programs. Engineering embraces the principles of design, systems, and differential equations. By contrast, the life sciences embrace the principles of observation, reductionism, and classical statistics. Clearly, engineers and life scientists have very different ways of looking at the world that present both opportunities and challenges for cross-disciplinary groups engaged in the simulation of biological systems.

SIMULATING ENGINEERED SYSTEMS

For the most part, the components of engineered systems are well understood, their characteristics specified. The behavior of the integrated system, however, is not well understood. Simulation is used to move from component specification to system behavior. Based on iterative testing of changes in component specification, component designs can be refined to generate the desired behavior for the integrated system. This "bottom-up" process is primarily iterative optimization. The use of extensive simulation in Boeing's design of the 777 aircraft, for example, has been well documented in the press.

In some instances, this process is reversed to move from system behavior to component specification. Commonly referred to as "reverse engineering," this process has been used to determine the design of an industry or military competitor's system. Reverse engineering was used, for example, to clarify infraredseeker technology for antiaircraft missiles during the Cold War. The reverse engineering "top-down" approach is characterized by uncertainties about component specifications that could produce the observed system behaviors, especially when the system has only been partially observed, either with respect to internal states or temporally. One can think of possible component specifications as hypotheses to clear up these uncertainties. Simulation provides a means of testing these hypotheses.

Both bottom-up and top-down approaches are used in simulating biological systems. Each approach is motivated by different trends and is focused on different applications.

BOTTOM-UP SIMULATIONS OF BIOLOGICAL SYSTEMS

Many groups, particularly in academia, are pursuing the bottom-up approach to simulation. Some of this research is in areas where the "first principles" of the underlying processes have been well characterized, such as the physicochemical processes involved in fluid flow, tissue oxygenation, and basic metabolic pathways. Simulations in other areas, such as attempts to simulate the logic underlying the behaviors of bacteria, cells, and tissues, are based on less of a first-principles foundation. The goal of bottom-up simulations is often to elicit "emergent behaviors" that demonstrate how a relatively simple network of nonlinear components can give rise to very complex behaviors. Simulations of action potentials via fluctuations in ion channel function, which date back to the early 1950s, have yielded insights into cellular electrophysiology.

The advent of molecular biology in the late 1980s provided entirely new technologies for perturbing and observing intracellular processes. Today, academic groups around the globe are developing complex simulations of intracellular processes, but bottom-up simulation is coming up against significant limitations, particularly the problem of observability. Put simply, one can think of two levels of data related to the observability of a biological system: (1) structured-level data, and (2) dynamic-level data. Structural-level data provide information sufficient to affirm the existence of a pathway and describe how it fits into a larger network of pathways. Dynamic-level data provide information to describe quantitatively the nonlinear dynamics of that pathway in the multiplicity of boundary conditions specified by the surrounding network.

Molecular biology has provided a wealth of structural-level data for many systems but dynamic-level data for only a few cellular systems, including cardiac cells. The scarcity of dynamic-level data is likely to continue for some time. The rate at which structural-level data are being generated far surpasses, and is accelerating, relative to the rate at which dynamic-level data are being generated. In the absence of the latter, attempts are being made to use bottom-up simulations to support multiple hypotheses. Managing this gap is a key challenge for bottom-up simulation.

TOP-DOWN SIMULATIONS OF BIOLOGICAL SYSTEMS

The key advantage to top-down simulations of biological systems is that the process begins with the end, the full envelope of functional (phenotypic) behaviors of which an integrated biological system is capable, in mind. Top-down simulations typically begin with simple representations of the dynamics of the system phenotype, because these data are much more prevalent than data on internal states, and therefore, provide a firm foundation for guiding the development of the simulation.

These initial simulations, however, cannot fully reproduce the observed phenotypes. Therefore, more detail is added to more closely reproduce the phenotype envelope. The implementation of these details is constrained by lack of these data, which typically become like more structural-level data as the components of the simulation become more detailed. As the simulation goes deeper, it becomes increasingly necessary to explore multiple alternative hypotheses.

Top-down simulations follow the principle that simulations should be "as simple as possible, but no simpler." The addition of detail is motivated by testing alternative hypotheses to explain key phenotypic behaviors. By contrast, in bottom-up simulations the detail is motivated by the availability of data.

The several groups in academia that are pursuing the top-down approach at the cellular or tissue level typically have researchers who are both developing simulations and collecting data. This creates a tight iterative loop for the formulation and testing of hypotheses *in silico* and the confirmation of hypotheses *in vitro*. These top-down simulations help researchers design efficient, incisive experiments and then provide a framework for interpreting experimental results in the larger context of the integrated system. Entelos, Inc., uses this same top-down approach at the clinical (human) level. Collaborative relationships with our pharmaceutical partners provide the empirical capabilities that complement our simulation technology and the expertise in *in silico* research.

THE CENTRAL ROLE OF SIMULATION IN BIOLOGY

The hypothesis-data cycle just described is central to the advancement of science. The physical sciences have benefited from both theoretical and empirical research made possible by the common language of mathematics since their inception. By contrast, the classical statistics used in the life sciences help researchers describe *that* certain phenomena occur but not *how* they occur. Prior to the advent of molecular biology in the 1980s, research in quantitative physiology had made significant progress using engineering-like approaches to understanding biology. Since then, however, molecular biology, with its high-throughput data collection (rather than formal hypotheses of integrated biological systems), has taken center stage.

Until recently, empirical technologies have completely outpaced technologies for managing hypotheses because there is simply too much data for researchers to keep in their heads. This is why simulation, particularly top-down simulation, can play a central role in biology. With this kind of simulation, hypotheses of biological function (or dysfunction) can be made explicit and open to critique and review. Top-down simulations can also address fundamental issues of complex biological systems that cannot be managed in mental models, including redundancy, feedback, and multiple timescales.

Simulation also has risks, of course. Black-box simulations cannot serve as vehicles for dialogue and inquiry; for that they must be made into "clear cube" simulations. Simulations must be considered as a key technology in the discipline of *in silico* research and not accepted on blind trust. Simulations also have organizational risks. Top-down simulations have the potential to show that some research projects are likely to be ineffective in the clinic. Although this would be tremendously valuable to the organization, individual researchers will feel challenged.

In the past 10 years, Entelos has addressed these issues through integrated development of methodologies for *in silico* research and the development of proprietary software technology. All of our researchers, engineers and life scientists, scientific advisors, and software engineers have contributed to the advancement of our top-down simulation.

THE FUTURE OF SIMULATION TECHNOLOGY IN THE LIFE SCIENCES

Many projections have been made concerning the impact of research and health care in the postgenomic era, including predictions of innovative new therapies and personalized treatment regimens based on individual genetic factors. All of these projections depend on an understanding of the role of specific genes and proteins in integrated human physiology. For the reasons described above, top-down simulation will be pivotal to making these possibilities a reality. Research in the life sciences will change accordingly, with dual, yet integrated, communities for theoretical and empirical research much as we find in the physical sciences. With an impressive array of high-throughput empirical technologies, the paradigm will shift from a shotgun, "data for data's sake" paradigm to fill databases for later investigation to a focused, iterative, "data to confirm detailed top-down hypothesis" paradigm. All of these changes will be led by industry rather than academia because the pharmaceutical and health care industries are facing acute challenges that are motivating them to advance *in silico* technologies, methodologies, and applications.

Reengineering the Paralyzed Nervous System

P. Hunter Peckham
Department of Biomedical Engineering
Case Western Reserve University
Cleveland, Ohio

INTRODUCTION

Damage to the central nervous system is the major cause of disability in the United States. In some cases, such as in spinal cord injuries or strokes, connectivity has been lost because the pathway has been severed. In other cases, such as in Parkinson's disease, the neural circuits behave in a disordered fashion. Whether the origin of the damage is congenital, traumatic, or age-related, improving neural connectivity and restoring function has a major impact on the lives of people with these injuries. Many approaches to restoring the connectivity of neural elements are being explored (e.g., gene therapies, stem cell transplants, tissue engineering). One of the most promising is engineering, which can provide an interface with the nervous system to restore functions.

Through the delivery of low levels of electrical current in precise ways, control of the nervous system can be regained and function restored. Understanding how such an interface works requires a fundamental appreciation of the structure of nerves and how they work. First, consider a single nerve fiber. From the cell body, or soma, at one end, hundreds of dendrites emerge, through which input is provided to the cell. Only one axon leaves the cell. The axon delivers information to another structure, such as another nerve cell or a muscle cell. Electrical stimulation is usually delivered to the axon somewhere along its length. The electrical current causes the permeability of the membrane to change causing an efflux/influx of sodium, potassium, calcium, and other ions. When the difference across the membrane reaches a sufficient level, an action potential is generated that propagates along the axon in both directions from its point of origin. This fundamental principle, called "gating" the membrane potential, is the basis for restoring function to the nervous system by electrical activation.

The action potential generated by an electrical current causes events analogous to the events that occur in the normal generation of nerve impulses.

Using electrical current to restore neural function has many advantages. First, most events involving the nervous system are communicated naturally by electrical means. Second, electrical stimulation has the capacity (1) to *activate* a single nerve fiber or multiple nerve fibers to generate movement and sensation, (2) to *inhibit* the firing of nerve fibers to reduce spasticity and pain, and (3) to *activate or inhibit complex neural circuits*, called neuromodulation, to change the firing of entire circuits of cells so it could be used to restore a wide range of different functions. Third, the effect of electrical stimulation can be *localized*, and turning off the current can eliminate the effect. Currents could also be delivered in such a way as to prolong the effect by taking advantage of the inherent plasticity of the nervous system. Fourth, electrical stimulation is incredibly *efficient*. A very small amount of current can generate enough muscle activation to lift the body. Electrical stimulation also acts very rapidly; the effect can be observed in seconds. Finally, electrical stimulation *can be applied safely*. Methods of delivering electrical current to biological tissue have already been developed through careful research and testing. Safe, stimulating waveforms that use bidirectional pulses with charge densities below established limits are well tolerated by biological tissues. Thus, electrical stimulation is an extraordinarily versatile, effective, and safe tool for manipulating the activity of the nervous system.

Electrical activation of the nervous system is applicable to virtually every disorder involving the central nervous system (i.e., the brain and spinal cord). Some devices have already been granted regulatory approval and are commercially available in the United States. These include devices for restoring hand function, controlling bladder and bowel function, controlling respiration in spinal cord injuries, suppressing seizures in epilepsy, suppressing tremors in Parkinson's disease, and restoring audition for people with hearing loss. Clinical research is being done on human subjects to enable patients to stand and walk, swallow, control the anal sphincter, and see. Basic research is also continuing on all of these applications to improve function and extend their applicability. For example, electrical stimulation has had limited success in restoring function in individuals with stroke, brain injuries, multiple sclerosis, and cerebral palsy, although theoretically their neurological disabilities can be overcome. For patients with spinal cord injuries, for example, the technique must be operable for extended periods of time, perhaps for 50 years or more. In addition, these injuries affect more than one organ system, the limbs and bladder, for instance. Ideally, therefore, the technology will be applicable to multiple systems.

IMPLEMENTATION OF NEUROPROSTHESES

Several factors must be considered in the clinical implementation of neuroprostheses. The use of a neuroprosthesis always involves trade-offs between physiological, technological, and clinical factors.

Physiological Considerations

Physiological factors are associated with the creation of a safe, effective interface between the prosthesis and the nervous system. First and foremost, the delivery of the electrical stimulus must be safe. A sufficient charge must be directed across the nerve membrane to depolarize it and generate action potentials, without generating toxic species in sufficient quantities to cause damage. Destruction (necrosis) or damage to the nerve tissue would exacerbate the problem. To understand the complexity of the problem, consider a device that could restore respiration. Biphasic (bidirectional or AC) current regulated *pulses* with charge reversal has been found to be effective. Eliciting an action potential in a compound nerve may require 10–20 mA at 30 V at a frequency of 20Hz 24 hours per day for up to 50 years.

Another physiological consideration is the control and coordination of activation of the muscle. The physiologic control of muscles is graded, and this must be duplicated in the reengineered system. There are only two fundamental mechanisms for controlling muscle force, (1) activating more muscle fibers (recruitment) or (2) activating muscle fibers faster. The latter leads to fatigue. Therefore, the preferable rate of stimulation is 20Hz or less. Controlling force by recruitment requires that the number of nerve fibers activated be increased as the controlling current is increased. The resulting activation is a nonlinear function, generally sigmoidally shaped. High-gain regions of the relationship may cause difficulties in control because small changes in current can cause large changes in the number of activated nerve fibers, as can small movements between the electrode and the nerve. In addition, a fundamental characteristic of muscle is that its force is dependent on its length; therefore, muscle length must also be considered in artificial control. Generally, an action is not caused by the “simple” generation of force from a single muscle but is the result of many muscles working together to produce the desired movement. Even for a simple movement, this means that one muscle (an agonist) increases in strength as a second muscle (an antagonist) works in opposition and decreases in strength. When one considers a complex action, such as walking or moving an arm, one can begin to appreciate the complexity of restoring movement through electrical activation.

The stability of the electrically activated response must also be considered. Muscles become fatigued with sustained contraction, whether naturally or electrically induced. With electrical stimulation, however, muscles become fatigued faster for two reasons. First, in an electrically stimulated contraction, there is less rotation of activated fibers than in a natural, voluntary contraction. Second, paralyzed muscles are generally less able to sustain force because their metabolic properties have been compromised since the injury. Electrical activation can effectively reverse this “disuse atrophy” to increase the fatigue resistance of paralyzed muscles.

Technological Considerations

The fundamental technology in systems for neuroprosthetic devices includes stimulators, electrodes, sensors, and the lead wires or communication channels that connect them. The form of the technology depends on the application. In the examples given above, which must be used for a substantial portion of a person’s life, the most effective devices would be implanted. The specificity and reliability afforded by implantation results in vastly improved function and convenience for the user. Therefore, the device must be thoroughly reliable, designed to accommodate enhancements, and be repairable without compromising the remaining components.

The requirements for an electronic device that can operate in the body for 50 years are stringent. For example, the current technology used to control the motor system consists of a multichannel, implantable stimulator with multiple leads that extend from the implanted electronics to the terminal electrodes placed adjacent to the nerve-muscle connection in the distal limb. The implantable stimulator contains hybrid microelectronics to provide the stimulation and control functions. The battery is not implanted because power consumption is too high for this to be practical. (To get an idea of power consumption, consider a device with eight channels of stimulation activated at 10–20 mA at 30 V at a frequency of 20 Hz 24 hours per day.) Currently, the electronics are powered and controlled by a radio-frequency signal transmitted through the skin with tuned coils (transmission frequency approximately 6.7MHz).

The implanted electronics are protected from moisture by a titanium package with glass-metal feedthroughs for the leads. The configuration of the package depends on the application; generally 8 to 16 feedthrough pins are used for the stimulation and control functions. The leads present a difficult mechanical challenge because they are subject to repeated cycles of both bending and stretching. In addition, each lead must have a midline connector so repairs can be made in the event of failure. Stress concentrations are created both at these connectors and at the junction where the leads exit the feedthroughs. In addition, the passage of current through the electrodes causes electrochemical reactions at the interface to the tissue, which can cause degradation of the electrode, as well as the tissue. The biological compatibility of the materials with the surrounding tissue is essential in all types of implanted devices because any weakness in the design will be exploited by the environment. The problem is even more difficult for neuroprosthetic applications in terms of protecting the implanted electronics and ensuring the long-term continuity of the lead electrode.

Clinical Considerations

In developing a neuroprosthetic device, it is particularly important to understand the function that is to be restored and how this aspect of the disability is treated medically. The technology must be not only functional, but must also be deployable by clinical practitioners (physicians, therapists, and nurses) whose appreciation of the complexity of the technology may be limited. The design must also meet the requirements of the user, such as an acceptable level of risk, time commitment, and the effort required for implementation and training. The neuroprosthesis must not only function acceptably, but it must also be easy and natural to use and easy to put on. Acceptable function may be less than full, normal function.

RESTORING UPPER LIMB FUNCTION

The focus of our work has been on a neuroprosthesis to restore hand and arm function (Figure 1) for people with cervical-level spinal cord injuries. These individuals have lost control of their hands and lower extremities but retain

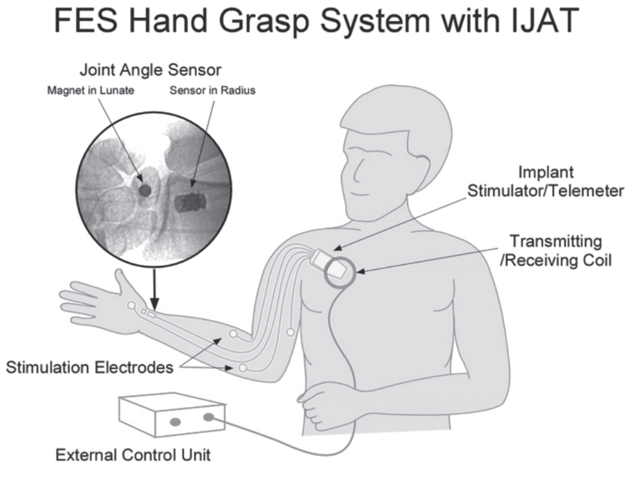


FIGURE 1 Sample implementation of neuroprosthesis for restoration of hand-arm control. SOURCE: Reprinted with permission from the Journal of Rehabilitation Research and Development (Bhadra et al., 2002).

control of their upper arms. The neuroprosthesis we have developed incorporates an implantable sensor that transduces joint angle (IJAT), a multichannel stimulator-telemeter, and an external control unit. Movements of the wrist are transduced by the IJAT and used to control the stimulation applied to the paralyzed finger and thumb muscles. Two grasp patterns are provided: (1) lateral pinch-release, in which the thumb contacts the side of the index finger; and (2) palmar prehension-release, in which the index and long fingers oppose the thumb. The former grip is typically used for picking up or holding small objects and the latter for grasping larger objects. Grasp is proportional; flexion of the wrist corresponds to full hand opening, and wrist extension corresponds to maximum grasping strength. Intermediate positions of the wrist correspond to intermediate grasp positions between these two extremes.

The system operates in the following manner. Contacting an external switch turns the system on, which transmits the radio frequency to the implant from the external controller, thus powering the transducer. This also establishes the "zero" command position of the wrist, corresponding to full hand extension, which is achieved by stimulating each of the target muscles at the appropriate level. For example, for hand extension, the finger and thumb muscles are maximally stimulated, and the finger and thumb flexors are inactive. These values are stored in a look-up table, in which any given wrist position corresponds to stimulus levels for each muscle. From the position of wrist extension, the user maneuvers the hand around the object and extends the wrist, causing the flexor muscles to be stimulated to a higher level and the extensor stimulation to decrease. Activating the external switch again sets a hold command, which maintains the stimulus level even if the wrist position changes. Other switch commands allow the user to regain control, reset zero, reset hold, or turn the system off. This system also enables users to regain control of elbow extension, which has been lost because of paralysis of the triceps. The switch enables the user to select alternative modes in which the triceps is either on or off.

This system is a second-generation neuroprosthesis, five of which have been implemented in human subjects. The first-generation neuroprosthesis, which has an external sensor on the opposite shoulder for control and eight channels of stimulation, has completed clinical trials (Peckham et al., 2001), has been approved by the Food and Drug Administration, and is commercially available (NeuroControl Corporation, Vallee View, Ohio). Approximately 200 first-generation devices have been implanted worldwide. Both systems enable people with spinal cord injuries to grasp and release common objects and thus perform many everyday activities, such as eating, writing, and grooming, These functions, which are essential for independence and self-sufficiency, often lead to dramatic changes in patients' lives.

Future Development

Many new tools, such as sensors, electrodes, stimulators, and detailed "instruction sets" of how to use them, are expected to become available in the future. By describing how these tools interact with the underlying neural tissue and modeling this performance, the instruction set allows us to predict how the tools will perform in various situations. Sensors that detect physical movement, pressure, or electrical activity may be used for control or feedback.

Advances in microsensors and bioMEMS are likely to yield great dividends. Current triaxial accelerometers and micropressure transducers are small enough and low-power enough to be implanted in the body. With advances in electrode technology, we will be able to stimulate selected fascicles of a whole nerve and create unidirectional impulses on the nerve. This will make complete and selective activation of nerves possible, as well as the inhibition of neural activity, such as the blocking of spastic activity or pain. These electrodes will also make it possible to record the natural activity of afferent nerve fibers for feedback and control. The development of a microelectrode will make possible the stimulation of spinal circuitry and cortical centers and selective recording from these regions. Complex high-density circuitry could be incorporated into the electrodes themselves, which could lead to direct access to the central nervous system and direct interfaces with the neural circuitry that controls complex coordinated functions at the spinal or cortical level. It could also enable us to extract control information from cortical neurons and, eventually, to translate the intention to move into signals that could be used to control movement. Finally, high-density stimulation and transmitting devices are under development that will enable the activation of more channels of stimulation in a smaller volume; this would greatly facilitate the development of complex visual prostheses.

New technology will provide tools for the development of more precise interfaces with the damaged nervous system leading to even more significant clinical results. We have already made progress in this direction by showing that afferent signals recorded from the nerves innervating the bladder during filling could be used to help control bladder activity. The neuroprosthesis for hand control described above, which uses both implantable sensors and stimulators, is undergoing clinical evaluation. This device could eliminate much of the external hardware and provide natural control of the hand that is easy for the user to learn. Systems that provide more than one function are not far away.

In the future, neuroprostheses may be used independently or in conjunction with other approaches, which may ultimately provide the best effect. For example, the plasticity of the nervous system is being revealed in clinical trials for body-weight supported walking and constraint-induced arm therapy. Function probably improves because residual spinal and cortical circuits have the capacity to alter their functions in an activity-dependent way. These adaptations are driven by the individual's remaining voluntary function but could also be triggered or reinforced by an electrical stimulus.

Using these tools effectively and developing new tools will require continued progress in our understanding of the pathophysiology of neural injury and how to interact with disordered control. As these technologies mature and become more available, advances can be expected to accelerate. New devices will almost certainly address a wider range of problems and benefit a growing number of individuals. Electrical stimulation is a powerful tool that will continue to be an essential aspect of new devices to mitigate the effects of disabling central nervous system conditions.

ACKNOWLEDGMENTS

The research described in this article was supported in large part by the National Institutes of Health, the National Institute of Neurological Disease and Stroke Neural Prosthesis Program, and the Department of Veterans Affairs Rehabilitation Research and Development Service. Support was also received from the Food and Drug Administration Orphan Products Division, the Paralyzed Veterans of America Spinal Cord Research Foundation, the Movement Disorder Foundation, and the Shapiro Foundation.

REFERENCES

Bhadra, N., P.H. Peckham, M.W. Keith, K.L. Kilgore, and F. Montague. 2002. Implementation of an implantable joint angle transducer. Journal of Rehabilitation Research and Development 39(3):411–422.

Brindley, G.S., and D.N. Rushton. 1995. The sacral anterior root stimulator as a means of managing the bladder in patients with spinal cord lesions. Bailliere's Clinical Neurology 4(1):1–13.

Chapin, J.K., and K.A. Moxin, eds. 2000. Neural Prostheses for Restoration of Sensory and Motor Function. Boca Raton, Fla.: CRC Press.

Creasey, G.H., and P.H. Peckham, eds. 1999. Functional Electrical Stimulation. Topics in Spinal Cord Injury Rehabilitation, Vol. 5, no. 1.

Peckham, P.H., M.W. Keith, K.L. Kilgore, J.H. Grill, K.S. Wuolle, G.B. Thorpe, P. Gorman, J. Hobby, M.J. Mulcahey, S. Carroll, V.R. Hentz, and A. Wiegner. 2001. Efficacy of an implanted neuroprosthesis for restoring hand grasp in tetraplegia: a multicenter study. Archives of Physical Medicine and Rehabilitation 82:1380–1388.

Triolo, R.J., ed. 2000. Electrical Stimulation. Assistive Technology (Special Issue), Vol. 12, no. 1.

Merging Living Cells and Microsystems Engineering

Mehmet Toner
Harvard Medical School
Massachusetts General Hospital
Boston, Massachusetts

Microfabrication techniques that have revolutionized the electronics industry are about to revolutionize the pharmaceutical, biotechnology, and biomedical device industries. Photolithography, etching techniques, and deposition methods can create large numbers of microscopic features on silicon or glass substrates. Among these features are biochemical reaction chambers, bioseparation channels, arrays of biological molecules, integrated microelectronics, micropumps and microvalves to control fluid movement, and many other components. These features can also be combined to create fully integrated devices that perform sample preparation, separation, detection and/or analysis, as well as drug delivery and in situ mechanical sensors. The two leading applications of microfabrication in biology are (1) "genes-on-a-chip" to monitor the expression level of potentially all genes in humans or various model systems and organisms simultaneously and (2) "lab-on-a-chip"-type devices to perform biochemistry in microchambers (Voldman et al., 1999).

Equally exciting is that the biomedical application of microfabricated devices is no longer limited to nonliving systems, such as genes-on-a-chip or lab-on-a-chip. Recent advances in understanding cellular behavior in microenvironments are paving the way toward the development or living microdevices. Cellular behavior is significantly influenced or dictated by the characteristics of the local environment: the presence and location of other cells; the chemistry, composition, and texture of the extracellular matrix; and the composition and temporal variation of soluble factors. As Figure 1 shows, the merger of living cells and microdevices involves exquisite control over these three fundamental determinants of cellular behavior.

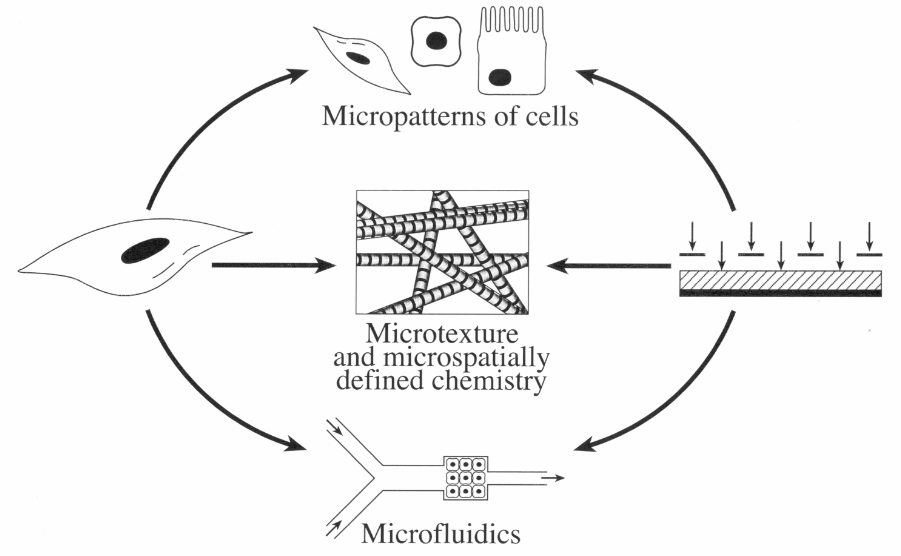


FIGURE 1 The merger of living cells and microsystems engineering involves control of three key ingredients of cellular environment, namely, cells, extracellular matrix, and soluble factors.

MICROSYSTEMS ENGINEERING IN BIOMEDICAL SCIENCES

This section is a brief overview of microsystem tools as applied in biology, especially in living systems. Emerging living cell-based devices are expected to become key technologies in twenty-first century medicine with a broad range of applications, from diagnostic, tissue engineered products to cell-based, high-throughput drug screening tools to basic cell biology tools. Achieving these challenging goals will depend heavily on processing techniques at the micron scale.

Microfabrication techniques are based on photolithography, which enables the creation of precise physical structures with micron scale dimensions (Voldman et al., 1999). Combined with the deposition of metal films, micromachining, etching, and bonding, microfabrication has been widely used for a large number of microdevices, some of them, such as semiconductors, very complex. In some of the early applications of microfabrication in biology, in fact, these processes were used to produce microtextured surfaces to determine the response of cells to well defined micron-scale disturbances in their environment, as well as to confine cells to micron to millimeter regions to probe cellular motility.

Despite the powerful advantages of traditional tools of microfabrication for the development of biological and medical microdevices, the usefulness of these techniques when merged with living cells is limited because of the nonbiological nature of many of the processing steps associated with photolithography. To address these limitations, over the last decade a number of critical advances have been made in the use of microfabrication in cellular systems. Collectively, these advances have dramatically redefined the landscape of microfabrication in biology and medicine and have resulted in the emergence of a new field called "bioMEMS" (bio-microelectromechanical systems).

Because living cells respond to their environment, the chemical nature of the surface is of the utmost importance in how cells behave in contact with artificial surfaces or scaffolds. Surface chemistry techniques have been developed to tailor the interaction of cells with surfaces (Kleinfeld et al., 1988). Self-assembled monolayers (SAMs) with various end-group functionalities to which cells attach provide an excellent tool for creating engineered surfaces for living cells. Various functional groups, such as extracellular matrix proteins (including specific peptides sequences), are used to control the preferential attachment of cells to surfaces. Conversely, fluoroalkylsilane SAMs provide surfaces that are nonadhesive to cells and proteins. These techniques have quickly become ubiquitous in bioMEMS and have greatly influenced our ability to create cellular micropatterns (Whitesides et al., 2001).

A major breakthrough in bioMEMS has been the development of so-called "soft lithography," which uses poly(dimethylsiloxane) (PDMS) to generate micropatterns (Whitesides et al., 2001). PDMS is essentially a transparent rubber that is ideally suited for biological applications because it is chemically unreactive. Soft lithography is based on a PDMS negative replica mold lifted-off from a microfabricated master wafer. The PDMS replica is then dipped in an ink solution containing SAMs and microstamped onto gold or silver-coated surfaces to transfer the SAMs to the regions where the PDMS stamp contacts the substrate. Because SAMs with many different functional end-groups can be generated easily, this approach enables engineering of surfaces at the micron scale with a large assortment of chemistries containing adhesive and nonadhesive microdomains. These engineered surfaces provide excellent platforms for cell attachment and culture.

In another application, the microstructured PDMS master can be brought together with a large variety of surfaces to seal points of contact hermetically; this is possible because of the highly conformal nature and hydrophobicity of PDMS. Using this technique, complex microfluidic systems can be designed. The microfluidic chambers can then be used either to perform surface chemistry on exposed areas or to pattern cells directly on surfaces (Folch and Toner, 2000). The generation of three-dimensional microfluidic structures provides flexibility and variety in the types of cellular systems that can be generated using these tools (Whitesides et al., 2001).

Another variation of PDMS technology involves curing PDMS in microfluidic networks to generate very thin self-sealing stencils that contain complex geometric arrangements of holes. When these stencils are contacted with a surface, the exposed regions of the surface can thereafter be modified, either through various surface chemistry techniques or by directly attaching living cells. The biologically compatible nature of PDMS and its ease of fabrication (e.g., much less stringent clean-room requirements than photolithography) have made soft lithography a very widely used and, in most instances, the preferred approach in the development of cellular microsystems.

SOME CURRENT AND FUTURE APPLICATIONS OF LIVING MICRODEVICES

The ability to integrate cells with microdevices is important for controlling cellular interactions on a subcellular level, for obtaining highly parallel, statistically meaningful readouts over large cell populations, and for miniaturizing instrumentation for minimally invasive, portable, fast, inexpensive devices. The benefits of integrated, miniaturized systems are high surface-area-to-volume ratio, high-throughput screening capabilities, smaller required volumes of reagents and samples, integration with electronics, and potential automation with a consequent increase in reliability and decrease in cost. The rich assortment of living microsystems that can be built using bioMEMS techniques have important applications in fundamental cell biological studies, tissue engineering, cell separation and culture devices, and drug discovery.

By precisely controlling the shape and type of the extracellular matrix of cells using micropatterned islands of cell adhesive and nonadhesive molecules, it is now possible to investigate how cells respond to their environments and determine what controls cellular phenotype. Micropatterning techniques can also be used to control the position of multiple cell types with respect to each other to investigate how homotypic and heterotypic interactions between cells regulate differentiation and the functioning of engineered tissues (Folch and Toner, 2000).

Microfabrication enables the creation of complex tissue-engineered products that can be used to replace or augment failing organs and tissues. The number and variety of microfabricated tissue-engineered products have increased rapidly in the last several years: microporous biocapsules to provide immunological protection to pancreatic islets; micropatterned hepatocytes and mesenchymal cells as key components of liver assist devices; microtextures basement-membrane-type structures to create skin with proper surface texture; and microspatial control of the distribution of vitronection or other extracellular matrix proteins to promote the formation of mineralized bone-like tissue in vivo (Bhatia and Chen, 1999).

With recent advances in stem cell biology, it has become clear that the cell source for many, if not all, tissue-engineered products will be obtained from stem cells. The primary challenge for stem cell biology is to determine biological and physical cues leading to the differentiation of these cells into functional tissue units. Microfabricated cell-culture devices will provide tools for engineering well controlled cellular environments, including microtexture, chemistry, soluble stimuli, and physical forces, to help in generating information about the conditions that lead to the differentiation of stem cells into functional tissue units.

Microfabricated structures are also used to measure individual cell mechanics in a population of cells as they pass through microbarriers. The same concepts can also be used to separate cells, based on size and shape or based on the expression of various surface receptors, by coating the microbarriers with different cell-attachment molecules. Another recent innovative application of microfluidic systems is the development of the so-called "reproductive chip" (Glasgow et al., 2001). As this technology evolves, it will become possible to move mammalian oocytes and embryos through various microchannels to perform a multitude of processing steps, such as microinjection of a sperm for fertilization of an oocyte or drilling of the zona pellucida to enhance embryo hatching, all without the need for cumbersome handling and manipulation procedures in a clinical setting.

Biosensors that incorporate living cells have the added advantage of rapidly monitoring the presence of toxic molecules or environmental pollutants, including biowarfare and chemical warfare agents (Pancrazio et al., 1999). Micropatterning provides the capability of establishing a network of cells, such as neurons, to investigate the effects of pollutants on the collective behavior of cells. Cell-based biosensors enable monitoring of the physiological behavior of an analyte of interest.

As we enter the postgenome era during which a massive amount of data concerning the intricate genetic machinery has been assembled, we are poised to decipher the complexity of hierarchical interactions that keep an organism at homeostasis. The best approach to accomplish this goal is a holistic, rather than reductionist, approach that monitors the behavior of cells and their genes in real time under multiple conditions, such as mimicking disease, trauma, development, and so forth. Cell chips are being developed to monitor simultaneously the behavior of living cells exposed to a multitude of environmental conditions using microfluidic and micromixing techniques. The same systems could be used to monitor the metabolic activity of cells. Thus, it is foreseeable that scientists will soon be able to observe simultaneously gene expression levels and the metabolism of massively parallel arrays of cells.

Massively parallel cell-based sensors will also be used to screen large quantities of drugs rapidly, which will dramatically expedite the drug discovery process, which will rely much less on animal models (Kapur et al., 1999). These systems will inevitably involve genetically engineered cells to monitor gene expression profiles in response to exposure to various drugs and toxins. They will also include multiple cell types to investigate complex interactions between cells from different tissues. The ability to use living cells will also dramatically elevate the biological content of the screening. A large library of molecular probes is already available for monitoring cell function, gene expression, protein levels, and subcellular structures. It is foreseeable that a good part of preclinical and clinical trials will soon be performed using so-called "mouse-on-a-chip" or "human-on-a-chip"-type devices. These sophisticated microdevices will contain microengineered tissue units coupled to each other by complex microfluid-handling networks. Microfluidic mixing systems will also precisely regulate the composition and concentration of drugs to be tested and reduce the cost of drug development in the pharmaceutical industry.

This concise overview describes some of the key advances and challenges related to the coming merger of microfabrication technology and living cells and a sampling of a broad range of exciting opportunities and promises in biology and medicine.

REFERENCES

Bhatia, S.N., and C.S. Chen. 1999. Tissue engineering at the micro-scale. Biomedical Microdevices 2:131–144.

Folch, A., and M. Toner. 2000. Microengineering of cellular interactions. Annual Review of Biomedical Engineering 2:227–256.

Glasgow, I.K., H.C. Zeringue, D.J. Beebe, S.-J. Choi, J. Lyman, N.G. Chan, and M.B. Wheeler. 2001. Handling individual embryos using microfluidics. IEEE Transactions on Biomedical Engineering 48:570–578.

Kapur R., K.A. Giuliano, M. Campana, T. Adams, K. Olson, D. Jung, M. Mrksich, C. Vasudevan, and D.L. Taylor. 1999. Streamlining the drug discovery process by integrating miniaturization, high-throughput screening, high-content screening, and automation on the CellChip™ system. Biomedical Microdevices 2:99–109.

Kleinfeld D., K.H. Kahler, and P.E. Hockberger. 1988. Controlled outgrowth of dissociated neurons on patterned substrates. Journal of Neuroscience 8:4098–4120.

Pancrazio, J.J., J.P. Whelan, D.A. Borkholder, W. Ma, and D.A. Stenger. 1999. Development and application of cell-based biosensors. Annals of Biomedical Engineering 27:697–711.

Voldman, J., M.L. Gray, and M.A. Schmidt. 1999. Microfabrication in biology and medicine. Annual Review of Biomedical Engineering 1:401–425.

Whitesides, G.M., E. Ostuni, S. Takayama, X. Jiang, and D.E. Ingber. 2001. Soft lithography in biology and biochemistry. Annual Review of Biomedical Engineering 3:335–373.

Dinner Speech

Technology Innovation in the New Era

NICHOLAS M. DONOFRIO
IBM Corporation
Armonk, New York

I'm delighted to join you, and I'm grateful to Bill Wulf for thinking that, after 35 years at IBM, I still have it in me to stand among such accomplished engineering leaders. Not only that, but for his confidence that I can communicate with you in a coherent way! Of course, that remains to be seen. I'm even more delighted that he's asked me to impart some words of wisdom that may actually be helpful to you. I must admit, though, that I'm a little nervous about dispensing advice. As you know, Socrates did a lot of that, and they poisoned him! So, it's a good thing that dinner came *before* the speech!

The aim of this symposium is something I'm passionate about—exposure to thinking and ideas from a range of technically driven disciplines. If I'm remembered for anything when my time at IBM is through, I hope it will be for my commitment to an *integrated* technical community that is multidisciplinary and multicultural. The days of operating in isolation are history, and the day of the interdisciplinary engineer is here. This will be a fundamental requirement from now on.

For those of you who don't know me, let me tell you a little bit about myself. I'm a technologist at heart, an electrical engineer educated at Rensselaer and Syracuse. I actually did honest work at IBM once upon a time, designing circuits and chips for almost half of my career. Today, I am very fortunate to be leading IBM's technical strategy, as well as a technical community of some 170,000 people worldwide.

I've seen some amazing things over the course of my career at IBM. I've seen an improvement of more than six orders of magnitude in fundamental information technologies—semiconductors, storage technologies, and communications technologies—in size, speed, density, capacity, and price/performance.

This industry continues to find a way to make things faster, better, less expensive, and more productive.

When I started at IBM in 1967, I was convinced that technology innovation had reached its pinnacle. What could be more state-of-the-art than the vacuum tube? How could you ever improve on the 80-column punch card as the user-friendly input device? What could be more advanced than storing five megabytes of data by stacking 50 24-inch discs on top of each other? And what could possibly replace the thousands of tiny ceramic ferrite cores that were strung together to form a computer's memory?

That was the way it was back then. That was the state of the art. Most people today—in fact, most engineers today—don't even know what I'm talking about when I reminisce about those things. But ever since its birth, the information technology industry has been all about change, neverending change. And the changes I've seen—even helped create in some cases—have been mind boggling. I submit that if there's any value to you in my being here tonight, that would be it. I've seen it all, I've worked through it all, I've managed through it all, I've led a number of teams who've done it all, and, if nothing else, I have a point of view!

The question I keep asking myself is this. Where have all of the incredible advancements and progress this industry has made gotten us? Are we on the right track? Especially as new business and computing models continue to evolve? In some ways, I think yes. In other ways, I'm convinced that those of us who create technologies for the IT industry need to rethink some assumptions.

Despite our advancements, the core of what we do is still called "computing," and the basic approach hasn't changed much in a hundred years. The processor takes input, performs mathematical functions, uses some memory and logic, and produces some form of output. Certainly, we've added storage to the mix, we can display output much more graphically, and we've sped up or increased the capacity of the fundamental components of the centralized computing machine. But the approach misses the boat in two big ways. First, it ignores the real needs of the users of IT who need IT for more than just "computing." And second, it doesn't take advantage of the amazing implications of today's enabling technologies.

Information technology is supposed to be the great enabler helping us tackle scientific problems, run businesses, automate processes and enable the collection, spread, and creation of knowledge and unprecedented collaboration and teamwork. In many ways, IT *is* the great enabler, and it will continue to be. At the same time, we are suffering from information glut, outages and reboots, system administration nightmares, security issues, privacy issues. So-called natural interfaces and natural-language queries are not natural at all. And, IT requires far too much human intervention, management, and maintenance. Our world needs exactly the opposite. The industry is working hard—and I know

IBM is committed—to developing an autonomic computing approach. We'll get there someday—but not without your help, regardless of your discipline.

What do I see as your crucial challenges and opportunities? And what wisdom can I impart to help you along the way? Here are a few suggestions.

First, information overload. The fact is, we're all drowning in data, and much of it will continue to be unstructured in the form of databases, web pages, images, audio files, and the like. Unstructured information is useless from a real knowledge perspective. To create knowledge, information has to be managed, sifted, mined, and interpreted. The evolution, and eventual marriage, of the Internet and IT demand that this potential source of knowledge be tapped. We need your thinking, your talent, your intelligence to help us find ingenious ways of mining and personalizing data so it comes to the requester in a usable form. As engineering leaders, you must be at the forefront of new innovations. Success will require collaboration and teamwork across the computer sciences.

Second, privacy. This is not a technology issue, but a policy issue, and it's a very thorny issue. The ability to reach "markets of one" is both efficient and positive, but there's a dark side to this kind of marketing that threatens personal privacy. Protecting people's privacy will require personal responsibility and a privacy policy. We can't let ourselves, or anyone else, go down that dark path. Every one of us must participate in finding a balance between the needs of the institution and the needs of the consumer. Security does not automatically create privacy. You can have the Fort Knox of data management, but if you don't have rock-solid policies in place, you're not going to meet the individual's expectations about privacy.

Third, the Moore's Law challenge. You will face many of these challenges before your careers are over! Are you thinking about what's next? You should. But I'm an optimist. I'm not worried about what lies beyond silicon because I'm convinced we've just begun a new 30-year period of rapid transformation. Engineers and researchers always have, and always will, come up with fascinating substitutions.

Fourth, computer modeling and simulation. Engineering today is a broad discipline with ill-defined boundaries. Traditional disciplines are being complemented by new ones, such as mathematical engineering, people skilled in mathematical techniques who use them the way engineers use a set of tools, not to prove a theory, but as techniques to achieve big results. We can now run a broad set of specifications and a broad range of scenarios to improve the design of just about anything. We can model things far more precisely than ever before, and thus meet the challenges of multiple-scale phenomena. Simulations of all kinds can now eliminate any assumptions and lead to incredible results. The tools available today will soon become standard tools for engineers. I urge you to get on board.

Finally, the next generation of engineers. You probably know this better than anyone else, but there won't be one without your help. The apathy, even antipathy, of young students toward science and engineering is alarming. Despite vast opportunities, our pipeline of engineers, scientists, and technically driven people is drying up. And the problem among young women and underrepresented minorities is acute.

How can we fill the pipeline? First, we must make a commitment to diversity, to tapping into underserved, underused, highly talented segments of our population. Diversity of thought is as important to the future of technical professions as diversity of composition.

Second, our nation needs role models, and our young people need mentors. No one is better qualified for that role than you. This issue *must* be taken personally. I urge you to inject yourselves into elementary and middle schools whenever you can to help young people become aware of the rewards of engineering and technology, to understand their options. More of us need to be involved in organizations like National Engineers Week and NACME (National Action Council for Minorities in Engineering). Like many of you, I firmly believe that engineers and scientists are the real wealth-generators of our world society. Everyone else lives off of what engineers produce!

Each of you is a leader, and the world looks to you for guidance and direction. As you assume and develop your mantle of leadership, I offer you some words of advice. At least I can tell you what has worked for me and the people I lead. First, expand your horizons. Being talented is not enough for true success or personal fulfillment. You must also be involved in outside endeavors and make contributions to enrich society. And be recognized for your contributions; communicate your innovations.

Next, take careful inventory of your skills. Build on your strengths, and work on your weaknesses. Make a difference! Third, embrace change. You must change as technology changes. During your careers, base technology will change by a factor of 10 million, and what you learn today will be irrelevant before you retire. A mentor of mine once told me that in the future people will be valued less for what they know than for their ability to deal with what they don't know. Be flexible. Learn to lead, and learn to follow. Lots of people will depend on your ability to rise to any challenge. Being a team player is essential. It is far better to be a member of a championship team than to stand in the winner's circle alone.

Finally, keep your senses. Keep your sense of *history*, who you are, where you came from, what you believe. Be proud of those things, and be proud of your achievements. Keep your sense of *balance* between work and family, school and life; these are tough juggling acts. And only you can strike the right balance. Keep your *sense of humor*. This is our best tiebreaker, the ultimate sanity check, the only thing that separates us from other mammals.

APPENDIXES

Contributors

NICHOLAS M. DONOFRIO leads the strategy for developing and commercializing advanced technology for IBM's global operations. His responsibilities include overseeing IBM research, the Global Integrated Supply Chain Team, the Integrated Product Development Team, and the Import Compliance Office. He also leads IBM's worldwide quality initiatives. He is chairman of IBM's Corporate Technology Council, chairman of the Board of Governors for the IBM Academy of Technology, a member of IBM's Corporate Development Committee, and a member of the IBM Chairman's Council. Mr. Donofrio spent the early part of his career in microprocessor development as a designer of logic and memory chips. Mr. Donofrio is a strong advocate of education, particularly in mathematics and science, the keys to economic competitiveness. His focus is on advancing education, employment, and career opportunities for underrepresented minorities and women. He is a member of the Board of Directors for the National Action Council for Minorities in Engineering, a fellow of the Institute of Electrical and Electronics Engineers (IEEE), and a member of the National Academy of Engineering. Mr. Donofrio earned a B.S. in electrical engineering from Rensselaer Polytechnic Institute and an M.S., also in electrical engineering, from Syracuse University. In 1999, he was awarded an honorary doctorate in engineering from Polytechnic University. In 2002, he was awarded the IEEE Mensforth International Gold Medal for his outstanding contribution to the advancement of manufacturing engineering. (*nmd@us.ibm.com*)

RONALD S. FEARING is a professor in the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley, which he joined in January 1988. His principle research interests are in microrobotics, tactile sensing, teletaction, and dextrous manipulation. He has a Ph.D. from Stanford in electrical engineering (1988), and an S.B. and S.M. in electrical engineering and computer science from the Massachusetts Institute of Technology (1983). He was the recipient of the Presidential Young Investigator Award in 1991. (*ronf@eecs.berkeley.edu*)

ANDREA GOLDSMITH is an associate professor of electrical engineering at Stanford University. She received her B.S. (1986), M.S. (1991), and Ph.D. (1994) in electrical engineering from the University of California, Berkeley. From 1986 to 1990 she was affiliated with Maxim Technologies, where she worked on packet radio and satellite communication systems; from 1991 to 1992 she was affiliated with AT&T Bell Laboratories, where she worked on microcell modeling and channel estimation. Before joining the faculty at Stanford, she was an assistant professor of electrical engineering at the California Institute of Technology. Dr. Goldsmith's research has been focused on capacity of wireless channels, wireless communication theory, adaptive modulation and coding, joint source and channel coding, cellular system design and resource allocation, and ad-hoc wireless networks. She is a Terman Faculty Fellow at Stanford and a recipient of the Alfred P. Sloan Fellowship, the National Science Foundation CAREER Development Award, the Office of Naval Research Young Investigator Award, a National Semiconductor Faculty Development Award, an Okawa Foundation Award, and the David Griep Memorial Prize from U.C. Berkeley. She is an editor for the *IEEE Transactions on Communications* and the *IEEE Wireless Communications Magazine*. (*andrea@ee.stanford.edu*)

JEFFREY W. HAMSTRA has 17 years of experience in propulsion/airframe integration and is currently senior manager of the Propulsion Systems Branch at Lockheed Martin Aeronautics Company in Fort Worth, Texas. Previous positions include group chief, Propulsion and Thermodynamics; lead engineer, Propulsion Technology; and lead engineer, Special Aircraft Programs. He has also acted as program manager for several propulsion integration R&D contracts. After graduating from the University of Michigan with an MSAE (1984) and BSAE (1983), Mr. Hamstra began his professional career as an inlet systems engineer on the F-22 program. His recent publications include, "Active Inlet Flow Control Technology Demonstration" (ICA 2000-6112) and "Fluidic Throat Skewing for Thrust Vectoring in Fixed Geometry Nozzles" (AIAA 99-0365). He holds three U.S. patents, is vice chair of the AIAA Air Breathing Propulsion Committee, and is a member of the ASME/IGTI Aircraft Engine Committee. (*jeffrey.w.hamstra@lmco.com*)

MIRIAM HELLER is currently the program director of the Infrastructure and Information Systems Program in the Civil and Mechanical Systems Program of the National Science Foundation. Prior to that she was an assistant professor in the Department of Industrial Engineering at the University of Houston. She has a Ph.D. in environmental systems analysis from the Johns Hopkins University. As a Fulbright Scholar during the 1999–2000 academic year, Dr. Heller conducted research in France in industrial ecology. Her interdisciplinary work in civil infrastructure and industrial ecology is based on her background in operations research, systems engineering, artificial intelligence, information technology, and environmental science. She has also held several positions in the private sector, including Citicorp Credit Services and Digital Equipment Corporation. Her research on building bridges between industry and academia through courses, projects, and professional service has been funded by both public and private sector sources. (*mheller@nsf.gov*)

RAJIV LAROIA is a founder of Flarion Technologies in Bedminster, New Jersey. As the company's chief technology officer, he is responsible for setting product direction and overseeing all R&D. Prior to launching Flarion Technologies, he was with Lucent Technologies Bell Laboratories. In 1997, he became head of Bell Labs' Digital Communications Research Department in the Wireless Research Center, where he and his team worked on the initial development of a flash-OFDM, technology-based, wireless data system. Dr. Laroia's years at Bell Labs have generated numerous publications and more than 15 patents (granted and pending) with a total patent licensing revenue of more than $25 million. From 1994 to 1997, he was the associate editor of *IEEE Transactions on Information Theory*. He received his Ph.D. (1992) and M.S. (1989) from the University of Maryland, College Park and a B.S. (1985) from the Indian Institute of Technology, Delhi—all in electrical engineering. His thesis contributed to V.34, the ITU voiceband modem international standard, and led to a patent that has generated more than $2 million for the University of Maryland, College Park. In 1992, he was the recipient of the Best Graduate Student of the Year Award at the University of Maryland. He is a fellow of the IEEE. (l*aroia@flarion.com*)

DANIEL N. MILLER has 15 years of experience with propulsion/airframe integration. For the last five years he has been the advanced-propulsion principle investigator for Lockheed Martin Advanced Development Programs (the "Skunk Works"). Before that he was assigned to several advanced aircraft projects, including the National Aerospace Plane Program. Mr. Miller has B.S.M.E. and M.S.M.E. degrees from the University of Wisconsin, holds four U.S. patents, is the author of numerous technical publications, and is a member of the AIAA Fluid Dynamics Technical Committee, the Flow Control Committee, and the Flow Control Working Group. (*daniel.n.miller@lmco.com*)

STEPHEN J. MORRIS, president of MLB Company in Palo Alto, California, has been designing, building and flight-testing small unmanned aircraft for more than 15 years. He conducted research on oblique-wing aircraft and led the design and construction of the oblique-all-wing demonstrator that was test flown in 1994. Dr. Morris was a member of the Stanford team that won the 1995 International Aerial Robotics Competition and winner of the 1997 and 1998 ISSMO Micro Air Vehicle Competition. He is also the chief of design for Bright Star Gliders, a manufacturer of hang gliders. Dr. Morris has a B.S. in mechanical engineering from Bucknell University and M.S. and Ph.D. degrees in aeronautical and astronautical engineering from Stanford University. (*mlbco@ mindspring. com*)

S. MUTHUKRISHNAN is a technical consultant at AT&T Labs and an associate professor at Rutgers University. He received his Ph.D. in computer science from the Courant Institute of Mathematical Sciences, New York University, in 1994 and was a postdoctoral visitor at DIMACS, a consortium of Princeton and Rutgers University funded by the National Science Foundation, working on algorithms for computational biology. He was on the faculty at the University of Warwick, U.K., and later joined Bell Laboratories, Lucent Technologies, where he worked on algorithmic issues in databases and wireless systems. His work at AT&T Research involved developing wireless applications and infrastructure solutions for the AT&T wireless network. His work on enabling location-based services has been tested in the AT&T network and was covered by CBS and MSNBC. Dr. Muthukrishnan has published more than 80 papers on a wide range of subjects, including networking, databases, computational biology, and theoretical computer science. He has been granted seven patents with several pending. He has served on the program committees of various Association for Computing Machinery (ACM) and Society for Industrial and Applied Mathematics (SIAM) conferences and is co-chair of a three-year project at DIMACS on next-generation networking and applications. (*muthu@research.att.com or muthu@cs.rutgers@edu*)

LINDA K. NOZICK, an associate professor in the School of Civil and Environmental Engineering at Cornell University, joined the faculty at Cornell after completing her doctorate at the University of Pennsylvania. She has also been a visiting associate professor in the Operations Research Department (1998–1999) at the Naval Postgraduate School in Monterey, California, and a visiting professor in the General Motors Global R&D Center in Warren, Michigan (1998). She is a member of INFORMS, the Production and Operations Management Society, and the Transportation Research Board of the National Research Council. Dr. Nozick has served on the Editorial Advisory Board of *Transportation Research (Part A)* and is currently an associate editor of *Naval Research Logistics*. She was recipient of a Presidential Early Career Award for Scientists and Engineers (PECASE) in 1997 and the James and Mary Tien Excellence in Teaching Award in 1996. Her primary research interest is the development of mathematical models for the management of complex systems. (*LKN3@cornell.edu*)

THOMAS PATERSON is chief scientific officer at Entelos, Inc., in Scotts Valley, California, where he leads the development of principles and methodologies for biologic systems modeling. He also pioneered the synthesis of system dynamics and decision analysis, which led to applications for pharmaceutical discovery and development in immune systems dysfunction, infectious diseases, and several other biologic areas. Before joining Entelos, Mr. Paterson held leadership positions at SDG, GTE Government Systems, and the Institute for Defense Analysis. He received an S.B. in aeronautics and astronautics from the Massachusetts Institute of Technology and an M.S. in decision analysis from Stanford University. (*paterson@entelos.com*)

P. HUNTER PECKHAM is professor of biomedical engineering and director of the Rehabilitation Engineering Center at Case Western Reserve University (CWRU). He received his B.S. in mechanical engineering from Clarkson College of Technology, Potsdam, New York, and his M.S. and Ph.D. in biomedical engineering from CWRU. He is also director of the Veterans Affairs Center of Excellence in Functional Electrical Stimulation (FES), a consortium involving the Cleveland VA Medical Center, CWRU, and MetroHealth Medical Center. The FES center focuses on the clinical development and implementation of systems to restore control of movement in paralysis. Dr. Peckham's research is focused on rehabilitation engineering and neuroprostheses, specifically functional restoration of the paralyzed upper extremity in individuals with spinal cord injuries. He and his collaborators have developed implantable neural prostheses that use electrical stimulation to control neuromuscular activation and have implemented procedures to provide control of grasp-release in individuals with tetraplegia; this function enables individuals with central nervous system disabilities to perform essential activities of daily living. He is currently working on the integration of technological rehabilitation and surgery to restore functional capabilities. (*pxp2@po.cwru.edu*)

FENIOSKY PEÑA-MORA, associate professor of information technology and project management in the Civil and Environmental Engineering Department's Intelligent Engineering Systems Group at the Massachusetts Institute of Technology (MIT), is the leader of the DaVinci Agent Society Initiative at MIT, and the author of several publications on computer-supported design, computer-supported engineering design and construction, project control, and management of large-scale engineering systems. He was the recipient of a 1999 National Science Foundation CAREER Award and a 1999 Presidential Early Career Award for Scientists and Engineers (PECASE). Dr. Peña-Mora has been a consultant for industry and government in Argentina, Colombia, Dominican Republic, Japan, Puerto Rico, and the United States and is the chief technology officer for Peña Alcántara Consultants, a firm that specializes in project management and information technology. He was the cofounder and chief technology officer for inMeeting, an Internet company that specializes in managing collaborative sessions on multiple devices. Dr. Peña-Mora was chief information technology consultant for the project director of the Boston Central Artery/Third Harbor Tunnel Project; his role was to provide information technology support for change management and process integration during the design and construction stages of this $13.6 billion, decade-long, regional engineering project. (*feniosky@mit.edu*)

GREGORY J. POTTIE received his B.Sc. in engineering physics from Queen's University, Kingston, Ontario, in 1984, and his M.Eng. and Ph.D. in electrical engineering from McMaster University, Hamilton, Ontario, in 1985 and 1988 respectively. From 1989 to 1991, he worked at Motorola/Codex on projects related to voice-band modems and digital subscriber lines. In 1991, he joined the faculty of the Electrical Engineering Department, University of California, Los Angeles (UCLA), where he is now professor and vice-chair for graduate programs. His research interests include channel-coding theory, wireless communication systems, and wireless sensor networks; current projects include the design of robust links in mobile networks and investigation of information theoretic issues in sensor networks. From 1997 to 1999, Dr. Pottie was secretary to the Board of Governors for the IEEE Information Theory Society. In 1998, he was named the Faculty Researcher of the Year for the UCLA School of Engineering and Applied Science. Dr. Pottie is cofounder of Sensoria Corporation. (*pottie@ee.ucla.edu*)

MEHMET TONER is associate professor of surgery and bioengineering at Harvard Medical School and Massachusetts General Hospital (MGH) and a member of the senior scientific staff at the Shriners Burns Hospital. At MGH, he is associate director of the Center for Engineering in Medicine and director of the Microsystems Bioengineering Core Facility. Dr. Toner received a B.S. from Istanbul Technical University and an M.S. from the Massachusetts Institute of Technology, both in mechanical engineering. He completed his Ph.D. in medical engineering at Harvard University-MIT Division of Health Sciences and Technology in 1989. Dr. Toner is associate editor of the *Journal of Biomechanical Engineering*, associate editor of the *Annual Reviews in Biomedical Engineering*, a member of the editorial boards of *Cryobiology* and *Cryo-Letters* and a member of the Scientific Advisory Board of Organogenesis, Inc., a tissue-engineering company. He is also a member of the National Program Committee of the Biomedical Engineering Society and president-elect of the Society for Cryobiology. His contributions to bioengineering have been recognized by the YC Fung Young Faculty Award in Bioengineering from the American Society of Mechanical Engineers in 1994, and the John F. and Virginia B. Taplin Faculty Fellow Award from Harvard and MIT. Dr. Toner recently became a fellow of the American Institute of Medical and Biological Engineering. (*mtoner@sbi.org*)

Program

NATIONAL ACADEMY OF ENGINEERING

Seventh Annual Symposium on
Frontiers of Engineering
March 1–3, 2002

FLIGHT AT THE LEADING EDGE: EXTREME AERODYNAMICS FROM THE MEGA TO THE MICRO

Organizers: Donald Nilson, Albert Pisano

Active Flow Control: Enabling Next-Generation Jet Propulsion Aerodynamics
Jeffrey W. Hamstra, Lockheed Martin Aeronautics Company

Miniature Spy Planes: The Next Generation of Flying Robots
Stephen J. Morris, MLB Company

Toward Micromechanical Flyers
Ronald S. Fearing, University of California, Berkeley

* * *

CIVIL SYSTEMS

Organizers: Sue McNeil, Priscilla Nelson

Dynamic Planning and Control of Civil Infrastructure Systems
Feniosky Peña-Mora, Massachusetts Institute of Technology

Improbable Is Not Impossible: Decision Making Under Uncertainty
Linda K. Nozick, Cornell University

Interdependencies in Civil Infrastructure Systems
Miriam Heller, National Science Foundation

* * *

WIRELESS COMMUNICATIONS

Organizers: Venugopal Veeravalli, Minerva Yeung

Design Challenges for Future Wireless Systems
Andrea Goldsmith, Stanford University

Next-Generation Mobile Wireless Internet Technology
Rajiv Laroia, Flarion Technologies

Service Architectures for Emerging Wireless Networks
S. Muthukrishnan, AT&T Laboratories and Rutgers University

Wireless Integrated Network Sensors (WINS): The Web Gets Physical
Gregory J. Pottie, University of California, Los Angeles

* * *

TECHNOLOGY AND THE HUMAN BODY

Organizers: David Beebe, John Norton, Sharon Nunes

Applying Simulation Technology to the Life Sciences
Thomas Paterson, Entelos, Inc.

Reengineering the Paralyzed Nervous System
P. Hunter Peckham, Case Western Reserve University

Merging Living Cells and Microsystems Engineering
Mehmet Toner, Harvard Medical School and Massachusetts General Hospital

Participants

NATIONAL ACADEMY OF ENGINEERING

Seventh Annual Symposium on
Frontiers of Engineering
March 1–3, 2002

David J. Beebe *(unable to attend)*
Associate Professor
Department of Biomedical Engineering
University of Wisconsin-Madison

Kaushik Bhattacharya
Professor
Applied Mechanics & Mechanical Engineering
California Institute of Technology

Ann M. Bisantz
Assistant Professor
Department of Industrial Engineering
State University of New York at Buffalo

Karl F. Böhringer
Assistant Professor
Department of Electrical Engineering
University of Washington

Brian S. Caruso
Senior Project Manager
Camp Dresser and McKee, Inc.

John M. Chapin
Chief Technical Officer
Vanu, Inc.

Christopher S. Chen
Assistant Professor
Biomedical Engineering and Oncology
Johns Hopkins University

Julie Chen
Associate Professor
Department of Mechanical Engineering
University of Massachusetts

Isaac Chuang
Associate Professor
Media Laboratory
Massachusetts Institute of Technology

Alice S. Chuck
Research Scientist 3
Amgen, Inc.

Michael L. Corradini
Chair, Engineering Physics Department
Professor of Nuclear Engineering and Engineering Physics
University of Wisconsin-Madison

Michael J. Daily
Senior Scientist, Department Manager
HRL Laboratories

Rachel A. Davidson
Assistant Professor
School of Civil and Environmental Engineering
Cornell University

Tejal A. Desai
Assistant Professor
Department of Bioengineering
University of Illinois at Chicago

Reginald DesRoches
Assistant Professor
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dennis E. Discher
Associate Professor
School of Engineering and Applied Sciences
University of Pennsylvania

Pedro Domingos
Assistant Professor
Department of Computer Science and Engineering
University of Washington

Christoph Erben
Bell Laboratories, Lucent Technologies

Ronald S. Fearing
Professor
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley

Steven E. Fredrickson
Project Manager
NASA Johnson Space Center

Andrea Goldsmith
Associate Professor
Department of Electrical Engineering
Stanford University

Susan C. Hagness
Assistant Professor
Department of Electrical and Computer Engineering
University of Wisconsin-Madison

Christopher D. Hall
Associate Professor
Department of Aerospace and Ocean Engineering
Virginia Polytechnic Institute and State University

Jeffrey W. Hamstra
Senior Manager
Propulsion Systems Branch
Lockheed Martin Aeronautics Company

Vassily Hatzimanikatis
Assistant Professor
Department of Chemical Engineering
Northwestern University

Miriam Heller
Program Director
Infrastructure and Information
Systems Program,
Division of Civil and Mechanical
Systems
National Science Foundation

Gavin Hendricks
Chief, Fluid Systems
Pratt & Whitney

Mark D. Hickman
Assistant Professor
Department of Civil Engineering and
Engineering Mechanics
University of Arizona

Jeffrey W. Holmes
Assistant Professor
Department of Biomedical
Engineering
Columbia University

Keng D. Hsueh
Manager, Material Planning and
Logistics
Global Bill Of Material (BOM)
Knowledge Management,
Process Leadership and Systems
Ford Motor Company

Luke Hwang
Mechanical Engineer
Procter & Gamble Co.

Victoria Interrante
McKnight Land-Grant Professor
Department of Computer Science and
Engineering
University of Minnesota

Abu Saeed Islam
Member, Research and Technical
Staff II
Xerox Corporation

Kenneth Kieffer
Engineering Manager
Eastman Kodak Company

Frederik C. M. Kjeldsen
Research Staff Member
T.J. Watson Research Center
IBM

Kara M. Kockelman
Clare Boothe Luce Assistant
Professor
Department of Civil Engineering
University of Texas, Austin

Willem Kools
Research Scientist
Millipore Corporation

Arun Krishnan
Project Manager
Siemens Corporate Research

Mukesh Kumar
Project Manager
Applied Technology/Biomaterials
Division
Biomet, Inc.

Sehoon Kwak
Technologist
DaimlerChrysler Corporation

Amit Lal
Assistant Professor
Department of Electrical and
Computer Engineering
University of Wisconsin-Madison

Rajiv Laroia
Founder and Chief Technology Officer
Flarion Technologies

Jacqueline Li
Assistant Professor
Department of Mechanical Engineering
Cooper Union

Peter Lorraine
Physicist
GE Corporate Research and Development

Garrick E. Louis
Assistant Professor
Department of Systems and Information Engineering
University of Virginia

Hideo Mabuchi
Associate Professor
Department of Physics
California Institute of Technology

Theresa A. Maldonado
Associate Professor
Department of Electrical Engineering
University of Texas, Arlington

Kathryn A. McCarthy
Department Manager
Idaho National Engineering and Environmental Laboratory

Sue McNeil
Director
Urban Transportation Center
University of Illinois at Chicago

Stephen Morris
President
MLB Company

Dennis A. Muilenburg
Vice President
Engineering/Air Traffic Management
Boeing Company

S. Muthukrishnan
Technology Consultant, AT&T Labs
Associate Professor, Rutgers University

Martin R. Myers
Technical Advisor
Metallurgical Engineering Material
Cummins, Inc.

Satish Narayanan
Research Engineer/Scientist
United Technologies Research Center

Priscilla P. Nelson
Director, Division of Civil and Mechanical Systems
Directorate of Engineering
National Science Foundation

Donald R. Nilson
Director
Engineering Technology & Strategic Planning
Lockheed Martin Aeronautics Company

John D. Norton
Staff Scientist
Medtronic, Inc.

Linda K. Nozick
Associate Professor
School of Civil and Environmental Engineering
Cornell University

Sharon L. Nunes
Director, Life Sciences Solution Development
Corporate Technology Group
IBM Corporation

Lynne E. Parker
Senior Research Staff
Intelligent and Emerging Computational Systems Section
Oak Ridge National Laboratory

Thomas Paterson
Chief Scientific Officer
Entelos, Inc.

P. Hunter Peckham
Professor and Director, Rehabilitation Engineering Center
Department of Biomedical Engineering
Case Western Reserve University

Feniosky Peña-Mora
Associate Professor
Department of Civil and Environmental Engineering
Massachusetts Institute of Technology

Per F. Peterson
Professor and Chair
Department of Nuclear Engineering
University of California, Berkeley

Patrick M. Pierz
Director
Thermal and Fluid Sciences Division
Cummins, Inc.

Albert P. Pisano *(unable to attend)*
FANUC Chair of Mechanical Systems
Electronics Research Lab
University of California, Berkeley

Gregory J. Pottie
Professor
Department of Electrical Engineering
University of California, Los Angeles

Sanjay Raman
Assistant Professor
Bradley Department of Electrical and Computer Engineering
Virginia Polytechnic Institute and State University

Brigette Rosendall
Engineering Specialist
Bechtel Corporation

Yong Rui
Researcher
Microsoft Research

Jennifer K. Ryan
Assistant Professor
School of Industrial Engineering
Purdue University

Dave W. Schroeder
Project Manager
Biomet, Inc.

Linda Slonksnes
Senior Systems Engineer
Science Applications International Corporation

Jeffry Sniegowski
Vice President, Production
MEMX, Inc.

Patrick T. Spicer
Technology Leader
Complex Liquids Research
Procter & Gamble Co.

Mark Stalzer
Director, Research Lab
HRL Laboratories

S. Kamakshi Sundaram
Senior Research Scientist II
Environmental Technology Division
Pacific Northwest National Laboratory

Mehmet Toner
Associate Professor of Surgery
Harvard Medical School
Massachusetts General Hospital

Bernhardt L. Trout
Assistant Professor
Department of Chemical Engineering
Massachusetts Institute of Technology

Loren Turner
Senior Transportation Engineer
Transportation Laboratory
Caltrans

Darrin Uecker
Vice President
Engineering Division
Computer Motion, Inc.

Venugopal V. Veeravalli
Associate Professor
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

Charles W. Whetsel
Chief Engineer, Mars Program
Jet Propulsion Laboratory

Joyce Y. Wong
Assistant Professor
Department of Biomedical Engineering
Boston University

Minerva M. Yeung
Principal Engineer and Manager of Media Technology Research
Microprocessor Research Labs
Intel Corporation

Zhuomin Zhang
Associate Professor
Department of Mechanical Engineering
University of Florida

David A. Zumbrunnen
Professor
Department of Mechanical Engineering
Clemson University

Guest Speaker

Nicholas M. Donofrio
Senior Vice President
Technology and Manufacturing
IBM Corporation

Guests

Malcolm J. Abzug
Retired President
ACA Systems, Inc.

Irving L. Ashkenas
Retired Vice Chairman of the Board
Systems Technology, Inc.

William F. Ballhaus, Sr.
President
International Numatics, Inc.

Yves Belanger
Vice President
LXLI International, Ltd.

Jean Kumagai
Senior Associate Editor
IEEE Spectrum Magazine

Claude Lajeunesse
President, Canadian Academy of Engineering
President, Ryerson University

Duane T. McRuer
Chairman
Systems Technology, Inc.

Laurel Sheppard
Contributing Editor
Society of Women Engineers Magazine

Robert H. Wertheim
Consultant
Science Applications International Corporation

National Academy of Engineering

Wm. A. Wulf
President

Lance A. Davis
Executive Officer

Janet Hunziker
Program Officer

Barbara Lee Neff
Executive Assistant to the President

Penny Gibbs
Program Associate

Lance R. Telander
Senior Project Assistant