FRONTIERS
OF ENGINEERING

Reports on Leading-Edge Engineering from the 2002 NAE Symposium on Frontiers of Engineering

NATIONAL ACADEMY OF ENGINEERING
OF THE NATIONAL ACADEMIES

EIGHTH ANNUAL SYMPOSIUM
ON
FRONTIERS OF ENGINEERING

NATIONAL ACADEMY OF ENGINEERING
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS - 500 Fifth Street, N.W. - Washington, D.C. 20001

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 the Air Force Office of Scientific Research, Defense Advanced Research Projects Agency, U.S. Department of Defense (DDR&E-Research), National Aeronautics and Space Administration, Microsoft Corporation, Ford Motor Company, IBM Corporation, and Cummins, Inc.

International Standard Book Number 0-309-08732-5

Additional copies of this report are available from The National Academies 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

Advisers to the Nation on Science, Engineering, and Medicine

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 chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

ORGANIZING COMMITTEE

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

ANN M. BISANTZ, Assistant Professor, Department of Industrial Engineering, State University of New York at Buffalo

ISAAC CHUANG, Associate Professor, Media Laboratory, Massachusetts Institute of Technology

PABLO G. DEBENEDETTI, Class of 1950 Professor and Chair, Department of Chemical Engineering, Princeton University

FREDERIK C. M. KJELDSEN, Research Staff Member, T.J. Watson Research Center, IBM

HIDEO MABUCHI, Associate Professor, Department of Physics, California Institute of Technology

KATHRYN A. McCARTHY, Department Manager, Idaho National Engineering and Environmental Laboratory

PER F. PETERSON, Professor and Chair, Department of Nuclear Engineering, University of California at Berkeley

BRIGETTE ROSENDALL, Engineering Specialist, Bechtel Corporation

Staff

JANET R. HUNZIKER, Program Officer

MARY W. L. KUTRUFF, Administrative Assistant

LANCE R. TELANDER, Senior Project Assistant

JENNIFER M. HARDESTY, Senior Project Assistant

Preface

This volume highlights the papers presented at the Eighth Annual National Academy of Engineering (NAE) Frontiers of Engineering Symposium. Every year the symposium brings together 100 outstanding young leaders in engineering to share their cutting-edge research and technical work. The 2002 symposium was held September 19–21 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 THE FRONTIERS OF ENGINEERING PROGRAM

The practice of engineering is changing. Engineers today must be able to adapt and thrive in an environment of rapid technological change and globalization. Engineering is becoming increasingly more interdisciplinary, and the frontiers often occur at intersections between engineering disciplines or at intersections between traditional "science" and engineering disciplines. 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, we invite 100 of this country's best and brightest engineers, ages 30 to 45, to join their peers to learn about cutting-edge developments in engineering. This broad overview of current developments in many fields of engineering often stimulates insights into cross-disciplinary applications. Because the engineers at the symposium work in academia, industry, and government, they can 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 to 100 to maximize opportunities for interactions and exchanges among the attendees, who have been chosen after a competitive nomination and selection process. The topics and speakers for each meeting are selected 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 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 (including 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 few years; and a summary statement of the theoretical, commercial, societal, and long-term significance of the work.

THE 2002 SYMPOSIUM

The presentations this year covered four broad areas: chemical and molecular engineering, human factors engineering, nuclear energy, and quantum information technology. Based on presentations given in the "Chemical and Molecular Engineering in the 21st Century" session, the field of chemical engineering is undergoing an exciting period of growth and transformation. Recent developments include: advances in computational power; the creation of powerful molecular simulation algorithms; advances in product engineering, as well as processes; and the emergence of new research at the interface between chemical engineering and biology. The speakers conveyed the scope of cutting-edge work in chemical and molecular engineering in talks on fuel cells, the computational design of materials, and state-of-the-art computational fluid dynamics and its applications in twenty-first century industry. The second session, "Technology for Human Beings," focused on the field of human factors (HR) and ergonomics, interactions between people and technology. The four talks in this session covered human factors interventions in complex, sociotechnical systems, such as nuclear power plants, the use of HR methods to reduce the number of crashes and driver errors in surface transportation, human-computer interactions in large panel displays, and brain-computer interactions that enable physically limited users to interact with computers. In the third session, "The Future of Nuclear Energy," speakers described the implications of trends in nuclear technologies for sustainability, safety, reliability, and economics of future nuclear energy systems. The speakers in this session covered different aspects of the subject: advanced nuclear reactor technologies; the licensing and building of new nuclear infrastructure in the United States; the potential of sustainable nuclear fission energy; and new applications of nuclear energy, specifically, space nuclear power and nuclear energy for microelectromechanical systems. The concluding session, "Engineering Challenges for Quantum Information Technology," described the great potential and enormous challenges in harnassing the quantum-mechanical behavior of nanoscale physical systems. In talks on quantum cryptography, ion-trap quantum computation, and scalable quantum computing using solid-state devices, the speakers identified the requirements for realizing large-scale quantum computers, when these systems will be available, and how they could be used (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, Andrew J. Viterbi, president of Viterbi Group and cofounder of Qualcomm, spoke on the development of digital communication and the wireless industry. The full text of Dr. Viterbi's remarks are included in this volume.

NAE is deeply grateful to the following organizations for their support of the Eighth Annual Symposium on Frontiers of Engineering: Air Force Office of Scientific Research, Defense Advanced Research Projects Agency, U.S. Department of Defense-DDR&E Research, National Aeronautics and Space Administration, Microsoft Corporation, Ford Motor Company, IBM 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

CHEMICAL AND MOLECULAR ENGINEERING IN THE 21ST CENTURY

TECHNOLOGY FOR HUMAN BEINGS

THE FUTURE OF NUCLEAR ENERGY

ENGINEERING CHALLENGES FOR QUANTUM INFORMATION TECHNOLOGY

DINNER SPEECH

APPENDIXES

Chemical and Molecular Engineering in the 21st Century

Fuel Cells That Run on Common Fuels

John M. Vohs
Department of Chemical and Biomolecular Engineering
University of Pennsylvania
Philadelphia, Pennsylvania

A fuel cell is a device that converts energy stored in chemical bonds in a fuel directly into electricity with high efficiency (Carette et al., 2000; Minh 1993). Unlike conventional methods of producing electricity, such as steam turbines, the efficiency of a fuel cell is not limited by the Carnot cycle; therefore, fuel cells can have energy-conversion efficiencies of 60 to 80 percent. This makes fuel cells environmentally friendly energy-conversion devices, and they have been proposed for use in applications ranging from large-scale power production to transportation to battery replacement.

Fuel cells were invented more than 150 years ago, but their commercialization has been very slow. To date, they have been used primarily in space vehicles, but these systems are quite costly and not suitable for commercial applications. In the last decade, however, there has been a dramatic increase in research, and it is now clear that fuel cells will enter the commercial marketplace in the not too distant future. Currently, most attention is focused on two types of fuel cells, polymer-electrolyte membrane (PEM) fuel cells and solid-oxide electrolyte fuel cells (SOFCs) (Carrette et al., 2000; Minh 1993). PEM systems use a proton-conducting polymer as the electrolyte and operate at low temperatures; SOFCs use an oxygen ion-conducting ceramic membrane as the electrolyte and operate at temperatures of 700 to 1,000°C.

In a PEM system, the charged species transported through the electrolyte are protons (H^+); thus, H_2 must be used as the fuel. This requirement presents many challenges for developers of PEM systems because hydrogen is difficult to store and costly to produce. The only viable source of hydrogen today is from the reforming of hydrocarbons (Ogden 2002). PEM systems also require high-puri-

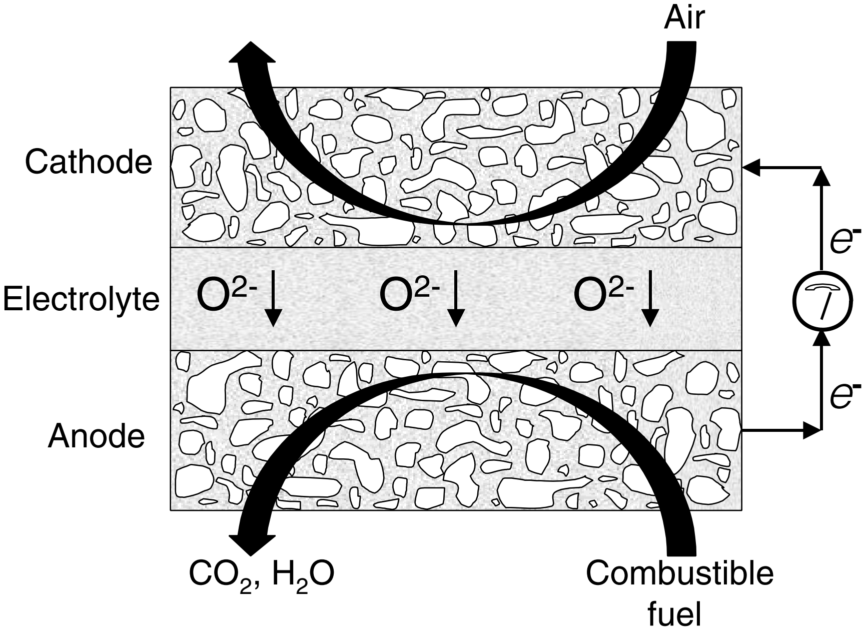

FIGURE 1 Schematic drawing of an SOFC.

ty H_2, because parts-per-million levels of carbon monoxide (CO) and sulfur poison the precious metal catalysts in the anode.

The species transported through the electrolyte in SOFCs are O^{2-} ions. This makes SOFCs more fuel flexible, and, in theory, any combustible gas could be used as the fuel. A schematic drawing of an SOFC is shown in Figure 1.

The cell is composed of a thin, dense layer of an oxygen ion-conducting electrolyte, typically yttria-stabilized zirconia (YSZ), and a porous anode and cathode. To obtain appreciable oxygen ion conductivity in the electrolyte, the system must operate at high temperatures. The cathode is composed of an electronically conducting oxide, such as strontium-doped $LaMnO_3$ (LSM). The cathode is exposed to air and reduces O_2 to O^{2-} ions using electrons supplied by the external circuit according to the following half-cell reaction:

$$1/2\ O_2 + 2\ e^- \rightarrow\ O^{2-} \quad (1)$$

The anode catalyzes the oxidation of the fuel using O^{2-} ions delivered through the electrolyte-producing electrons that flow through the external circuit to the cathode. If H_2 is used as the fuel, the anode half-cell reaction is as follows:

$$H_2 + O^{2-} \rightarrow H_2O + 2\ e^- \qquad (2)$$

In conventional SOFCs, the anode is almost always composed of a porous composite of nickel (Ni) and YSZ.

At equilibrium, the cell potential (i.e., the voltage difference between the anode and cathode) is given by the Nernst equation. For the two half-cell reactions shown above the Nernst equation in terms of the partial pressures of H_2 and H_2O at the anode, the partial pressure of O_2 at the cathode and Faraday's constant, F, is as follows:

$$E = E_o + \left(\frac{RT}{2F}\right) \cdot \ln\left(\frac{\left(P_{H_2,anode}\right) \cdot \left(P_{O_2,cathode}\right)^{1/2}}{\left(P_{H_2O,anode}\right)}\right) \qquad (3)$$

The open-circuit voltage for the cell, E_o, is determined by the standard Gibbs free-energy change for the overall reaction. Under operating conditions, when current flows from the anode to the cathode through an external load, the cell potential is less than the theoretical potential given by Eq. 3 because of various losses, such as the resistance to oxygen ion flow across the electrolyte. Because the work that can be performed by the electrons is directly proportional to the potential, the efficiency of the cell decreases as the amount of current used increases.

SOFCs have some fuel flexibility. The Ni current collector in the anode is a good hydrogen-oxidation catalyst and is not poisoned by CO. Indeed, CO can even be used as the fuel. Thus, the most common fuel for an SOFC is a hydrogen-rich synthesis gas (a mixture of CO and H_2), which can be produced by reacting methane with steam. Because Ni is an excellent catalyst for this so-called steam-reforming process, it is possible to perform this reaction within the anode by cofeeding steam with natural gas. It is much more difficult, however, to reform higher hydrocarbons, such as those found in common liquid fuels like gasoline or diesel. These fuels require a separate, partial oxidation (POX) reactor in which the fuel is partially oxidized using air to produce a CO/H_2 mixture. It is important to note that hydrocarbon reforming significantly decreases electrical efficiency. For example, the use of a POX reactor decreases the amount of electrical energy that can be extracted from the fuel by ~30 percent (this energy is converted to heat in the POX reactor).

Although SOFCs that rely on reforming of hydrocarbons to produce synthesis gas are a viable technology, it would be much simpler and more efficient if the reforming step could be avoided and the fuels used directly. Thus, one would prefer to have an anode half-cell reaction that involves the direct oxidation of a hydrocarbon fuel, such as the following:

$$C_nH_{2n+2} + (3n+1)\,O^{2-} \rightarrow n\,CO_2 + (n+1)\,H_2O + (6n+2)\,e^- \qquad (4)$$

The primary reason that this approach is not used in conventional SOFCs is that at high temperatures, decomposition of hydrocarbons to produce graphite or coke is thermodynamically favorable. Most metals catalyze this decomposition reaction. Ni is particularly active, and, when exposed to dry hydrocarbons at temperatures above 700°C, a fibrous layer of coke grows away from the surface of the metal. In a fuel cell, this produces rapid deactivation. This is demonstrated in Figure 2, which shows the current output of an SOFC with a Ni/YSZ cermet anode as a function of time. The cell was initially operated on H_2 and exhibited stable performance. After 60 minutes the fuel was switched to pure CH_4. Note that the cell rapidly deactivated, and after an additional 60 minutes, the cell was completely inactive. The deactivation was irreversible; the cell remained inactive when the fuel was switched back to H_2. Examination of the cell after this run revealed that the anode compartment was completely filled with coke.

Although the results presented in Figure 2 are discouraging, catalytic sci-

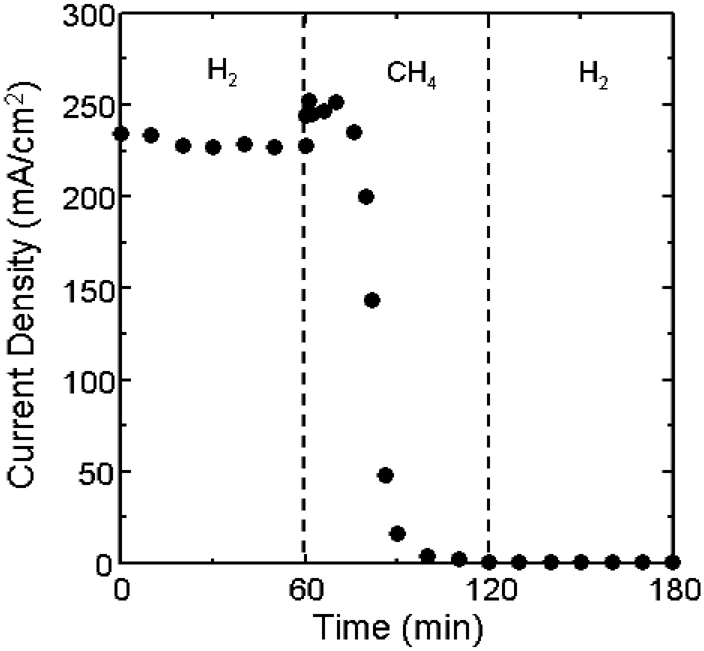

FIGURE 2 Performance of an SOFC with an Ni/YSZ anode as a function of time and fuel composition. Source: Park et al., 1999. Reprinted with permission.

ence provides clues to methods of overcoming the problem of anode fouling. Solving this problem has been the primary focus of our research for the past decade. Coke formation from hydrocarbons is thermodynamically favorable at high temperatures; however, this reaction proceeds very slowly in the absence of a catalyst. Thus, the key to producing an SOFC anode that is active for the direct oxidation of hydrocarbons is to use materials that do not catalyze carbon-deposition reactions. A catalyst for the oxidation of the hydrocarbons is still necessary, and the anode must be electronically conductive. Although most metals will catalyze carbon deposition from hydrocarbons, the noble metals, copper (Cu), silver (Ag), and gold (Au), are notable exceptions. We chose Cu to provide electrical conductivity in our anode design. Because Cu is inactive for hydrocarbon oxidation, a separate oxidation catalyst was also incorporated into the anode. Ceria (CeO_2) was chosen to provide this function. Ceria is an excellent hydrocarbon-oxidation catalyst, and unlike most metals, it does not catalyze coke formation.

Although the rationale for using a Cu/CeO_2/YSZ composite anode is relatively straightforward, synthesizing an anode with this composition presents several challenges. The most common method of fabricating Ni/YSZ anodes is to press a mixture of NiO and YSZ powders to form a wafer and then anneal the mixture at high temperatures in air to produce a dense NiO/YSZ composite. This is followed by annealing in H_2 to reduce the NiO to Ni. This step also makes the anode highly porous. Tape casting, tape calendaring, and similar ceramics-processing methods are often used in place of the pressing step for the NiO-YSZ mixture. The initial annealing step must be carried out at a temperature higher than 1,300ºC to sinter the YSZ component and achieve good ionic connectivity in the oxide phase. Unfortunately, because of the low melting temperatures of Cu_2O and CuO (1,235°C and 1,326°C, respectively), Cu/CeO_2/YSZ cermets cannot be produced in a similar way.

To avoid these problems, we have developed several novel fabrication methods in which the porous YSZ part of the anode cermet is prepared first, and the Cu and CeO_2 are added in separate steps that do not require high temperatures. The simplest method uses dual-tape casting, as shown in Figure 3 (Park et al., 2001). In this approach, an aqueous slurry containing YSZ powder and organic binders is cast into a film. Upon annealing in air at 1,500°C, the binders are burned out, and the oxide particles sinter to form a dense layer. Pore formers, such as graphite or polymer beads, can be added to the green tape to introduce porosity. The pore formers are oxidized during the annealing step and this gasification process produces pores in the sintered tape. With the appropriate choice of conditions, YSZ with porosity as high as 80 percent can be produced. In dual-tape casting, a second layer, which can have a different composition, is spread onto the first green layer. In Figure 3, the first layer becomes the dense electrolyte, while the second layer with pore formers becomes the anode. Cu and CeO_2 are added to the porous anode layer by impregnating it with aqueous

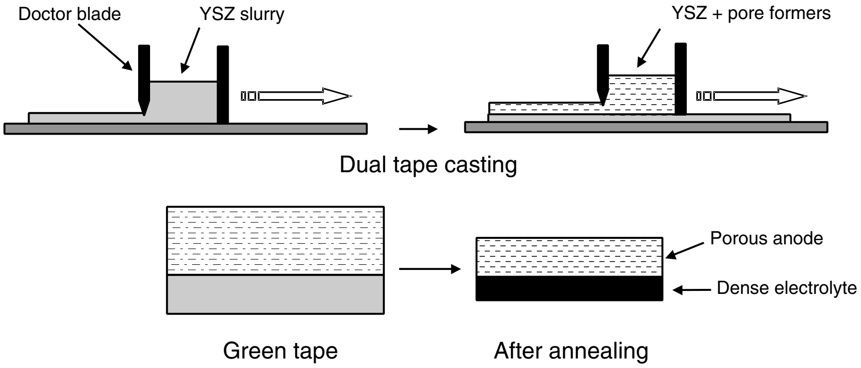

FIGURE 3 Schematic drawing of the tape-casting process used in the synthesis of an SOFC.

solutions of $Cu(NO_3)_2$ and $Ce(NO_3)_3$ followed by annealing at moderate temperatures (< 900°C) in H_2 and then air to produce metallic Cu and CeO_2. In a separate step, an LSM slurry is applied to the exposed surface of the electrolyte to form the cathode.

It has been shown that SOFCs with $Cu/CeO_2/YSZ$ anodes are highly resistant to fouling via carbon deposition and are remarkably fuel flexible (Kim et al., 2001; Park et al., 2000). This is apparent in Figure 4, which shows current density and voltage versus time for a fuel cell with a $Cu/CeO_2/YSZ$ anode operating on decane ($C_{10}H_{22}$), toluene (C_7H_8), and a synthetic diesel (Kim et al., 2001). Note that stable operation is achieved for all three fuels. Tests of up to three weeks in duration have been performed to demonstrate the long-term stability of these anodes.

The direct oxidation of hydrocarbon-based fuels in an SOCF has many significant advantages over H_2-powered fuel cell systems. One of the most important is increased efficiency. As noted above, reforming hydrocarbons to produce H_2 can decrease overall system efficiencies by as much as 30 percent. Direct oxidation of hydrocarbons also greatly simplifies fuel cell systems by eliminating the need to cofeed water when using methane, eliminating the need for external hydrocarbon reformers, and eliminating the need to store and transport H_2. Direct oxidation has the added advantage that the current infrastructure for the production and distribution of hydrocarbon-based fuels could be used. Although further development is necessary, fuels cells that rely on the direct oxidation of hydrocarbons may play an important role in the commercialization of fuel cell technologies.

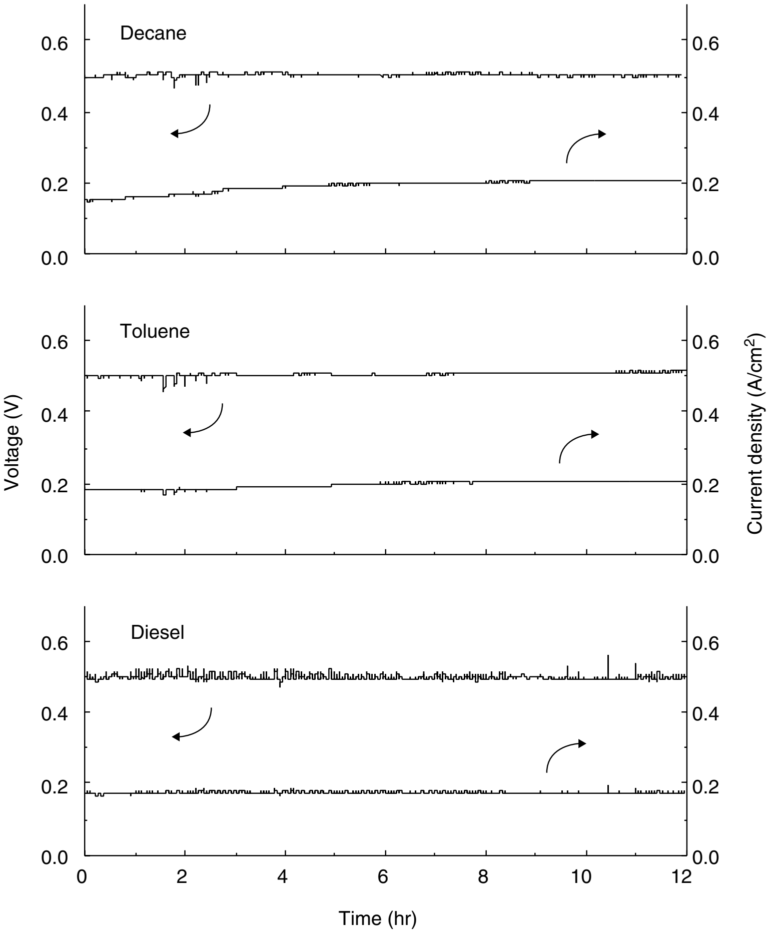

FIGURE 4 Voltage and current density vs. time for an SOFC with a Cu/CeO$_2$/YSZ anode operated at 700°C while using decane, toluene, and a synthetic diesel as the fuel. Source: Kim et al., 2001. Reprinted with permission.

ACKNOWLEDGMENTS

The research described in this paper was supported by the Defense Advanced Research Projects Agency and the Office of Naval Research. The author would also like to thank Prof. Raymond Gorte for his collaborations on the work described in this paper.

REFERENCES

Carrette, L., A. Friedrich, and U. Stimming 2000. Fuel cells: principles, types, fuels, and applications. ChemPhysChem 1(4): 162-193.

Kim, H., S. Park, J.M. Vohs, and R.J. Gorte. 2001. Direct oxidation of liquid fuels in a solid oxide fuel cell. Journal of the Electrochemical Society 148(7): A693-A695.

Minh, N.Q. 1993. Ceramic fuel-cells. Journal of the American Ceramic Society 76(3): 563-588.

Ogden, J.M. 2002. Hydrogen: the fuel of the future? Physics Today 55(4): 69-75.

Park, S., R. Craciun, J.M. Vohs, and R.J. Gorte. 1999. Direct oxidation of hydrocarbons in a solid oxide fuel cell: I. methane oxidation. Journal of the Electrochemical Society 146(10): 3603-3605.

Park, S., J.M. Vohs, and R.J. Gorte. 2000. Direct oxidation of hydrocarbons in a solid-oxide fuel cell. Nature 404: 265-267.

Park, S., R.J. Gorte, and J.M. Vohs. 2001. Tape cast solid oxide fuel cells for the direct oxidation of hydrocarbons. Journal of the Electrochemical Society 148(5): A443-A447.

Dimension-Dependent Properties of Macromolecules in Nanoscopic Structures

JUAN J. DE PABLO AND PAUL F. NEALEY
Department of Chemical Engineering
University of Wisconsin-Madison

Many materials exhibit size-dependent properties as system dimensions approach the atomic or molecular level. For example, metal and semiconducting nanoclusters with dimensions of a few nanometers exhibit remarkable optical, electrical, mechanical, catalytic, and magnetic properties (Murray et al., 2000). Nanocluster properties differ significantly from corresponding bulk properties; they depend on the quantum-level electronic structure of the ensemble and the ratio of surface to bulk atoms, and they provide the foundation for a wide range of innovative nanotechnologies. Unfortunately, much less is known about the properties of amorphous, polymeric materials in nanoscopic structures. Because the characteristic dimensions typically associated with polymeric molecules are on the order of 5- to 10-nm, it is natural to expect size-dependent properties in polymeric structures with dimensions from 10 to 100 nm.

Evidence for dimension-dependent properties of amorphous polymers has been observed in measurements of the glass transition temperature, T_g. Experiments by several research groups, including ours, report that the T_g of polymer films with nanoscopic dimensions can be significantly different from the corresponding bulk value (Forrest and Dalnoki-Veress, 2001). Based on seminal experiments by Forrest et al. (1996), it is now widely perceived that the T_g of freestanding, ultrathin films is substantially lower than that of the bulk material. Our recent simulations suggest that polymer chains or chain segments near the free surfaces of the films have greater mobility than the polymer in the interior of the film (Torres et al., 2000). We postulate that as the film thickness decreases, the fraction of the film with higher mobility increases, thereby resulting in the observed monotonic decrease with film thickness (Jain and de Pablo, 2000). In contrast, T_g for supported, ultrathin films has been observed to increase or decrease with respect to the bulk value, depending on the substrate. Recently, we have shown that the interfacial energy, γ_{sl}, between the substrate and the polymer is a significant parameter that governs the T_g of supported, ultrathin films (Fryer et al., 2001). The T_g of the films was characterized using local thermal analysis (a technique we developed at the University of Wisconsin [Fryer et al., 2000]), ellipsometry, and x-ray reflectivity. We have also demonstrated that the T_g of a 100-nm film of a model-resist resin, poly(4-hydroxystyrene) (PHS), is elevated above the bulk T_g by as much as 55°C by grafting the polymer to the substrate (Tate et al., 2001).

In the bulk, transport properties such as diffusion coefficients increase by several orders of magnitude in the narrow range of temperature over which the material undergoes a transition from a glass to a rubber (Nealey et al., 1993, 1994). Similarly, mechanical properties such as the Young's modulus decrease by 2 to 4 orders of magnitude over this same temperature range. Although the dimension dependence of T_g is now fairly well documented, less is known about mass transport and mechanical properties of polymers that are also likely to be dimension dependent. It is difficult to interpret consistently the current literature on diffusion in supported, thin films. Several studies report that chain diffusion slows in thin films (Frank et al., 1996) while dye diffusion has been reported to increase significantly (Tseng et al., 2000). We are not aware of any published reports on dimension-dependent mechanical properties of amorphous polymers.

In most commercial applications of films and coatings, the thickness of the polymer does not approach the sub-100 nm scale, and the dimension-dependent phenomena referred to above do not affect the properties, processing, or usefulness of the materials. However, in the microelectronics industry, the largest section of the U.S. economy, dimension-dependent properties of polymer nanostructures are anticipated to pose significant challenges, particularly with sub-100 nm patterning of photoresist materials (consisting of a polymer and photosensitive additives) by advanced lithography. To reach critical patterning dimensions of less than 100 nm, for example, the industry may be forced to use ultrathin films of polymer photoresist in conjunction with 157 nm and extreme ultraviolet (13.4 nm) lithography, due to the opacity of organic materials at these wavelengths (Brodsky et al., 2000; Stewart et al., 2000). In these systems, resist formulations and processing conditions may have to be optimized as a function of thickness for control over the transport of small molecules (e.g., photo-generated acids in films of chemically amplified photoresists), particularly during postexposure annealing (or "bake") (Fryer et al., 1999; Postnikov et al., 1999). The thermophysical and mass-transport properties of the films affect the sensitivity, resolution, contrast, and line-edge roughness of the photoresist. Optimization may be difficult because diffusion coefficients of probe molecules change by orders of magnitude as the temperature is varied within 15°C of T_g.

Dimension-dependent mechanical properties may pose the greatest challenges to the lithographic process and to nanofabrication techniques in general.

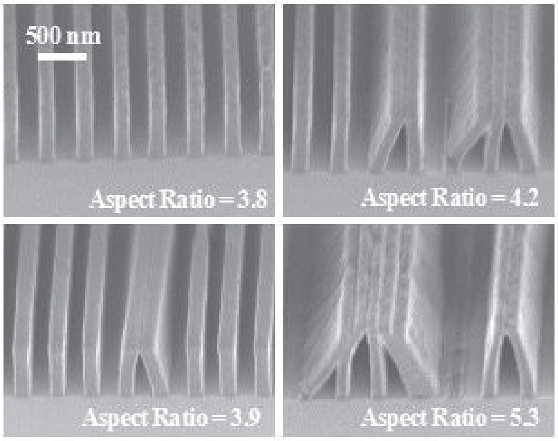

FIGURE 1 Collapse behavior of photoresist structures (Apex-E, pitch = 400 nm, linewidth = 200 nm).

In the bulk, the mechanical properties and the glass transition of a material are related; the shear modulus, for example, increases substantially as the temperature is decreased below T_g. It is therefore natural to anticipate similar relations in nanoscopic structures. When the distance between patterned resist structures (or MEMS [microelectromechanical systems] components) decreases, tremendous capillary forces are produced during drying of rinse liquids (e.g., water, after wet chemical processing) (Cao et al., 2000). These forces can cause the structures to collapse (see Figure 1) (Cao et al., 2000; Goldfarb et al., 2000; Namatsu et al., 1999). For arrays of dense (1:1) and semidense (1:2-1:5) lines with widths of 100 to 200 nm and aspect ratios of 3 to 4, we have shown that the susceptibility to collapse is strongly dependent on the aspect ratio and that the critical aspect ratio of collapse is inversely proportional to the distance between structures (Cao et al., 2000). These observations are consistent with models of capillary forces that predict that the force acting on the resist structures is inversely proportional to the distance between structures. They are also consistent with beam-bending models that predict that the maximum deformation of the resist structure in response to the imposed force is proportional to the aspect ratio cubed and is inversely proportional to the stiffness (Young's modulus).

The focus of our efforts over the last few years has been to fill a serious gap that hinders the development of nanotechnology and to develop a fundamental understanding of the properties of nanostructured polymeric materials. Through a combination of theoretical and experimental work, we have attempted to acquire a fundamental understanding of transport and mechanical properties of nanostructured polymers, because precisely these properties often determine their usefulness in coating, packaging, MEMS, microelectronic, and nanotechnology applications in general. We believe describing these systems with molecular-level models is essential to interpreting experimental results, guiding experimental design, and identifying new size-dependent phenomena. One broad objective of our research has been to determine the scale, if any, at which continuum-level representations of a material become inadequate.

MECHANICAL PROPERTIES OF NANOSTRUCTURED POLYMER SYSTEMS

One of the major aims of our research has been to quantify the mechanical properties of three-dimensional polymer nanostructures. Possible techniques to make these measurements include surface acoustic wave (SAW) analysis and scanning probe techniques. Unfortunately, the SAW techniques yield high-frequency mechanical properties that are difficult to extrapolate to the more relevant static or low-frequency regime. Scanning probe techniques, such as measuring the torsion of a cantilever required to push over a photoresist structure with dimensions of 100 to 200 nm are not convenient because of the gross mismatch in stiffness between standard cantilever beams and the polymer. To circumvent these problems, we have developed a formalism in which well defined forces are imposed on three-dimensional polymer nanostructures using simple and elegant principles from the classical thermodynamics of surface tension.

Test structures of poly-(methylmethacrylate) (PMMA) have been patterned using advanced lithographic techniques (electron-beam lithography). Images of sample structures and the principles of our experiments are depicted in Figure 2. The difference between the pressure in the liquid between structures and the pressure of the atmosphere is given by the Laplace equation; it is inversely proportional to the distance between the structures, assuming that gravitational effects are negligible at this scale and that the curvature of the meniscus can be modeled as a cylindrical surface (determined by the contact angle of the fluid in contact with the polymer). The net force on the structures is determined by the difference in pressure on the opposing walls of the structures times the surface area of the walls. Known forces can be applied by carefully choosing the dimensions and geometry of the system and slowly drying the structures initially immersed in liquid. The deformation behavior of the structures in response to the imposed force cannot be monitored in real time; the force increases as the distance between structures decreases (after deformation begins); and we don’t yet have analytical techniques to monitor the deformation of the nanostructures. Instead, we observe whether the structures are collapsed or not after complete drying and the severity of collapse as parameterized by the normalized collapse length (touching length of the beams/initial beam length) to calculate the mechanical properties of the structures with appropriate continuum-level and molecular-level models of the collapse process (see below). Our results confirm

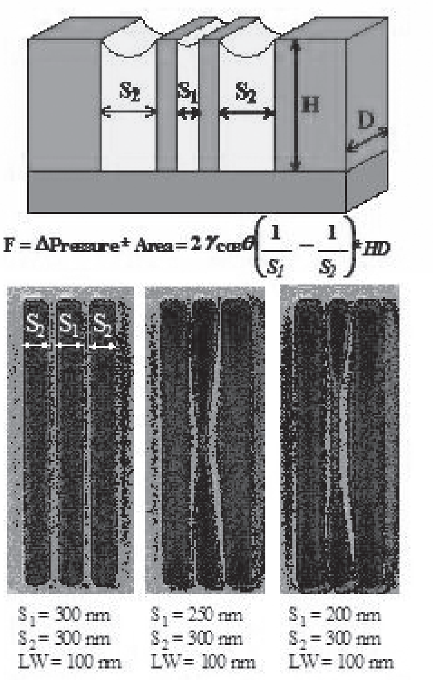

FIGURE 2 Schematic and scanning electron microscope (SEM) images of test structures in which well-defined capillary forces are applied to the walls of three-dimensional polymer nanostructures.

that the severity of collapse of 100-nm test structures scales with the initial capillary force for a number of different combinations of structure spacings.

Structures such as those depicted above have been patterned with linewidths from 300 nm to 90 nm, aspect ratios from 4 to 8, and distances or spacings between structures also in the 100- to 400-nm regime. A critical aspect ratio for collapse has been extracted from these experiments. Experimental critical aspect ratio data have been compared to those predicted using a simple continuum elastoplastic model of the polymeric material, with bulk values of the yield stress and the Young's modulus. Figure 3 shows some of our results for nanoscopic structures having widths as small as 100 nm; the agreement between continuum-level predictions using bulk material properties and experimental data is quantitative. Note, however, that below 100 nm, preliminary experimental results suggest that mechanical properties start to deviate from the bulk. Unfortunately experiments on smaller structures become increasingly difficult, and to anticipate the behavior of ultrasmall nanoscopic structures we must turn to molecular simulations.

Molecular simulations of virtual bending experiments of ultrasmall structures can be performed by applying a lateral force to a feature; the resulting deflection is measured to extract a modulus of elasticity. Simulations of a coarse-grained model of polystyrene have revealed that the Young's modulus of poly-

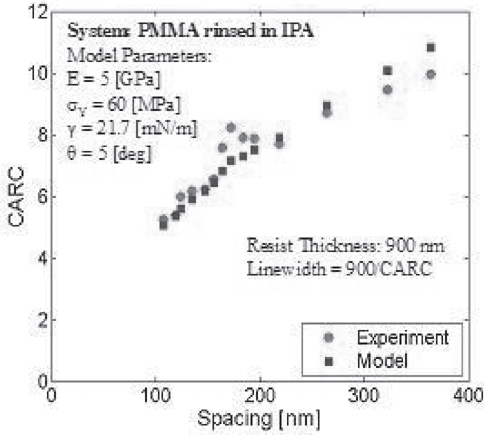

FIGURE 3 Critical aspect ratio for collapse (CARC) for PMMA rinsed in isopropyl alcohol. The material properties, surface tension, and contact angle were all determined from independent experiments on bulk samples.

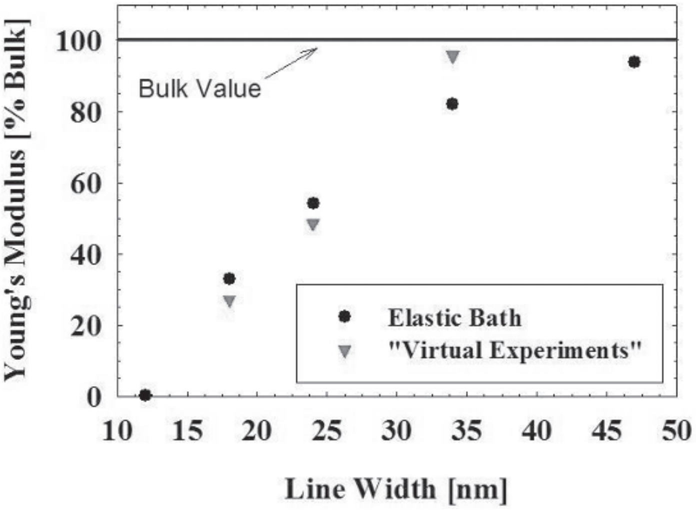

FIGURE 4 Young's modulus (normalized with respect to the bulk value) of nanoscopic structures (lines) as a function of width. The triangular symbols correspond to simulated bending and the circular symbols are from strain fluctuations in an elastic bath.

styrene exhibits a pronounced decrease below 50 nm (see Figures 4 and 5) (Boehm and de Pablo, 2002). The model used in our simulations consists of spherical Lennard-Jones interaction sites interconnected by harmonic bonds. This remarkable size-induced softening is reminiscent of a glass-to-rubber transition in the bulk, but it has not been considered before. A second intriguing result of our

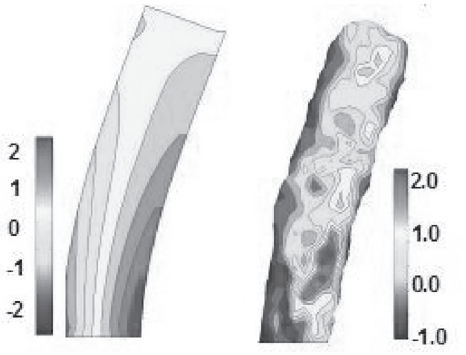

FIGURE 5 Cross-section view of the distribution of reduced normal stresses in a nanoscopic structure. The aspect ratio is 4, and the width is approximately 20 nm. The figure on the left corresponds to finite-element calculations, and that on the right corresponds to molecular simulations of a coarse-grained polymeric model.

simulations is that the spatial density and stress distributions in molecular nanostructures are markedly different from those predicted by continuum mechanics. Figure 5 shows the distribution of normal stresses in a polymeric line with a width of 20 nm and an aspect ratio of 4. On the left, we show numerical finite-element results from our elasticity-theory continuum analysis. On the right, we show the results of molecular simulations for a material with exactly the same bulk mechanical constants. Clearly, the stress distribution from continuum mechanics is unable to capture the phenomena at play in nanostructures; for the models considered in our work, the assumption of a mechanical continuum starts to deviate from the continuum observed in molecular models of nanoscopic polymeric structures between 50 and 70 nm.

The virtual bending experiments described above have the advantage that they mimic precisely a deformation experiment using an atomic force microscope. Unfortunately, these calculations are particularly demanding and are subject to large statistical uncertainty. To eliminate this problem, we have implemented novel simulation techniques in which nanoscopic lines are immersed in elastic media with well-defined mechanical properties (Van Workum, 2002). The elastic constants of the lines are inferred from simulations of the resulting composite material. These simulations have allowed us to determine all of the elastic constants of a nanoscopic structure from a single simulation and enabled the study of the development of anisotropy as we reach small dimensions. As illustrated in Figure 6, these calculations have shown that the elastic constants of nanoscopic structures are highly anisotropic.

Recently we have also developed a stress-fluctuation-based formalism to evaluate local elastic constants from molecular simulations. One of the assumptions implicit in continuum-level models of a pure material is that the moduli of elasticity are homogeneous throughout the system. This might not be the case, however, in glassy, amorphous polymeric systems. Figure 7 shows a cross section of a nanoscopic polymeric structure. The density distribution is shown on the left; density fluctuations are relatively small (a few percent) and are distribut-

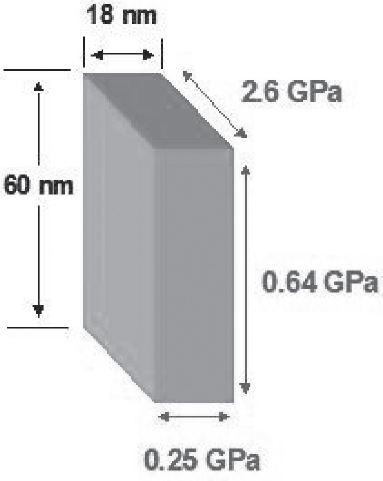

FIGURE 6 Elastic moduli of a nanoscopic polymeric structure determined from simulations of strain fluctuations. The bulk value is 3 Gpa.

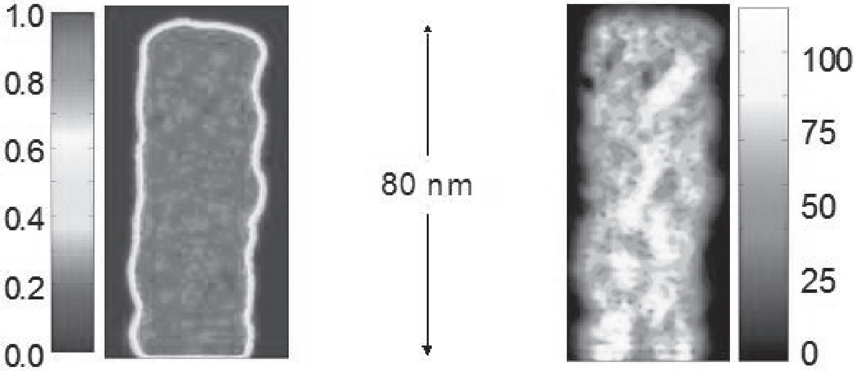

FIGURE 7 Cross-sectional view of the local density (left) and local elastic constant (right) distribution of a model polymeric nanostructure. The density is shown in reduced (Lennard-Jones) units, and the elastic constants are shown as a percentage of the bulk value.

ed uniformly throughout the structure. The local elastic modulus corresponding to the stress that arises upon vertical compression (in the direction normal to the substrate) of the structures is shown on the right. Near the air-polymer interface the material exhibits a gradual softening. Near the polymer-substrate interface the material undergoes a pronounced layering, which gives rise to large local elastic moduli. In the core of the structure, however, the local elastic constants of nanoscopic polymeric structures reveal distinct "soft" and "hard" domains. These domains have characteristic dimensions of approximately 2 to 4 nm and may be related to the so-called "dynamic heterogeneities" that have been postulated in the literature to explain experimental data for bulk, polymeric glasses (Ediger and Skinner, 2001).

CONCLUSIONS

Our experiments and simulations to date have revealed that, above approximately 100 nm, the mechanical behavior of nanoscopic polymeric structures can be described using continuum formalisms and bulk material properties. At smaller length scales, our results suggest that amorphous polymeric structures exhibit dimension-dependent, anisotropic, elastic constants. Furthermore, our calculations of local elastic constants indicate the existence of mechanical inhomogeneities on the scale of a few nanometers.

The calculations we have performed to date have been carried out on extremely simple polymer models, and our experiments have been limited to structures of approximately 100 nm. The extent to which specific interactions, temperature, molecular weight, and plasticizers influence the behavior described here remains unknown and is likely to become a fertile area of research in the years to come. Fundamental understanding of factors leading to size-dependent mechanical properties will result in strategies to improve the mechanical properties of nanoscopic polymer structures and eliminate collapse.

REFERENCES

Boehme, T., and J.J. de Pablo. 2002. Evidence for size-dependent mechanical properties from simulations of nanoscopic polymeric structures. Journal of Chemical Physics 116(9): 9939-9951.

Brodsky, C., J. Byers, W. Conley, R. Hung, S. Yamada, K. Patterson, M. Somervell, B. Trinque, H.-V. Tran, S. Cho, T. Chiba, S.-H. Lin, A. Jamieson, H. Johnson, T. Vander Heyden, and C.G. Wilson. 2000. 157 nm resist materials: progress report. Journal of Vacuum Science and Technology B 18(6): 3396-3401.

Cao, H.B., P.F. Nealey, and W.D. Domke. 2000. Comparison of resist collapse properties for deep ultraviolet and 193 nm resist platforms. Journal of Vacuum Science and Technology B 18(6): 3303-3307.

Ediger, M.D., and J.L. Skinner. 2001. Single molecules rock and roll near the glass transition. Science 292(5515): 233-234.

Forrest, J.A., K. Dalnoki-Veress, J.R. Stevens, and J.R. Dutcher. 1996. Effect of free surfaces on the glass transition temperature of thin polymer films. Physical Review Letters 77(10): 2002-2005.

Forrest, J.A., and K. Dalnoki-Veress. 2001. The glass transition in thin polymer films. Advances in Colloid and Interface Science 94(1-3): 167-196.

Frank, B., A.P. Gast, T.P. Russell, H.R. Brown, and C. Hawker. 1996. Polymer mobility in thin films. Macromolecules 29(20): 6531-6534.

Fryer, D.S., S. Bollepali, J.J. de Pablo, and P.F. Nealey. 1999. Study of acid diffusion in resist near the glass transition temperature. Journal of Vacuum Science and Technology B 17(6): 3351-3355.

Fryer, D.S., P.F. Nealey, and J.J. de Pablo. 2000. Thermal probe measurements of the glass transition for ultrathin polymer films as a function of thickness. Macromolecules 33(17): 6439-6447.

Fryer, D.S., R.D. Peters, E.J. Kim, J.E. Tomaszewski, J.J. de Pablo, P.F. Nealey, C.C. White, and W.-I. Wu. 2001. Dependence of the glass transition temperature of polymer films on interfacial energy and thickness. Macromolecules 34(16): 5627-5634.

Goldfarb, D.L., P.F. Nealey, J.P. Simons, W.M. Moreau, and M. Angelopoulos. 2000. Aqueous-based photoresist drying using supercritical carbon dioxide to prevent pattern collapse. Journal of Vacuum Science and Technology B 18(6): 3313-3317.

Jain, T.S., and J.J. de Pablo. 2002. Monte Carlo simulation of freestanding polymer films near the glass transition temperature. Macromolecules 35(6): 2167-2176.

Murray, C.B., C.R. Kagan, and M.G. Bawendi. 2000. Synthesis and characterization of monodisperse nanocrystals and close-packed nanocrystal assemblies. Annual Review of Materials Science 30: 545-610.

Namatsu, H., K. Yamazaki, and K. Kurihara. 1999. Supercritical drying for nanostructure fabrication without pattern collapse. Microelectronic Engineering 46(1-4): 129-132.

Nealey, P.F., R.E. Cohen, and A.S. Argon. 1993. Solubility and diffusion of polybutadiene in polystyrene at elevated temperatures. Macromolecules 26(6): 1287-1292.

Nealey, P.F., R.E. Cohen, and A.S. Argon. 1994. Effect of gas pressure on the solubility and diffusion of polybutadiene in polystyrene. Macromolecules 27(15): 4193-4197.

Postnikov, S.V., M.D. Stewart, H.V. Tran, M.A. Nierode, D.R. Medeiros, T. Cao, J. Byers, S.E. Webber, C.G. Wilson. 1999. Study of resolution limits due to intrinsic bias in chemically amplified photoresists. Journal of Vacuum Science and Technology B 17(6): 3335-3338.

Stewart, M.D., K. Patterson, M.H. Somervell, and C.G. Wilson. 2000. Organic imaging materials: a view of the future. Journal of Physical Organic Chemistry 13(12): 767-774.

Tate, R.S., D.S. Fryer, S. Pasqualini, M.F. Montague, J.J. de Pablo, and P.F. Nealey. 2001. Extraordinary elevation of the glass transition temperature of the polymer films grafted to silicon oxide substrates. Journal of Chemical Physics 115(21): 9982-9990.

Torres, J.A., P.F. Nealey, and J.J. de Pablo. 2000. Molecular simulation of ultrathin polymeric films near the glass transition. Physical Review Letters 85(15): 3221-3224.

Tseng, K.C., N.J. Turro, and C.J. Durning. 2000. Molecular mobility in polymer thin films. Physical Review E 61(2): 1800-1811.

Van Workum, K. 2002. Mechanical properties of polymeric nanostructures. PhD Thesis, University of Wisconsin.

The Role of Computational Fluid Dynamics in Process Industries

David Lee Davidson
Solutia, Inc.
Cantonment, Florida

Continuum mechanics, one of our most successful physical theories, is readily applicable to the process industries. In continuum mechanics, the existence of molecules is ignored, and matter is treated as a continuous medium. The continuum hypothesis is valid, provided the equations of continuum mechanics are applied at sufficiently large length scales and time scales that the properties of individual molecules are not noticed. The mapping of the laws of mass, momentum, and energy conservation to the continuum results in field equations that describe the dynamics of the continuum. These field equations, variously known as the equations of motion, the equations of change, or simply the conservation equations, are nonlinear, partial differential equations that can be solved, in principle, when combined with the appropriate constitutive information[1] and boundary conditions.

Continuum mechanics is the mechanical analog of classical electrodynamics, in which a set of field equations (Maxwell's equations) describe the dynamics of the relevant variables of the electrical and magnetic fields. Whereas Maxwell's equations are linear *unless* the constitutive behavior is nonlinear, the equations of continuum mechanics are nonlinear, *regardless* of the constitutive behavior of the materials of interest. The inherent nonlinearity of the conservation equations, which is due to convective transport of momentum, energy, and

[1] Examples of constitutive information are Newton's law of viscosity, which relates shear stress to shear rate, and Fourier's law of heat conduction, which relates heat flux to temperature gradient.

chemical species, is responsible for certain fluid mechanical phenomena, such as turbulence, that have no electrodynamic analog and that complicate solution of the conservation equations.

Analytical solutions (e.g., obtained by eigenfunction expansion, Fourier transform, similarity transform, perturbation methods, and the solution of ordinary differential equations for one-dimensional problems) to the conservation equations are of great interest, of course, but they can be obtained only under restricted conditions. When the equations can be rendered linear (e.g., when transport of the conserved quantities of interest is dominated by diffusion rather than convection) analytical solutions are often possible, provided the geometry of the domain and the boundary conditions are not too complicated. When the equations are nonlinear, analytical solutions are sometimes possible, again provided the boundary conditions and geometry are relatively simple. Even when the problem is dominated by diffusive transport and the geometry and boundary conditions are simple, nonlinear constitutive behavior can eliminate the possibility of analytical solution.

Consequently, numerical solution of the equations of change has been an important research topic for many decades, both in solid mechanics and in fluid mechanics. Solid mechanics is significantly simpler than fluid mechanics because of the absence of the nonlinear convection term, and the finite element method has become the standard method. In fluid mechanics, however, the finite element method is primarily used for laminar flows, and other methods, such as the finite difference and finite volume methods, are used for both laminar and turbulent flows. The recently developed lattice-Boltzmann method is also being used, primarily in academic circles. All of these methods involve the approximation of the field equations defined over a continuous domain by discrete equations associated with a finite set of discrete points within the domain and specified by the user, directly or through an automated algorithm. Regardless of the method, the numerical solution of the conservation equations for fluid flow is known as computational fluid dynamics (CFD).

CFD was initially done without automation because the need to solve these equations (e.g., in aircraft design) preceded the development of electronic computers by several decades. With the advent of electronic computers, more ambitious numerical calculations became possible. Initially, CFD codes were written for specific problems. It was natural to generalize these codes somewhat, and eventually, particularly as computational resources became more readily available, general-purpose CFD codes were developed. It was then recognized that a business could be built upon the development and licensing of these codes to industrial, academic, and government users. Today, many of the general-purpose commercial codes are quite sophisticated, cost a tiny fraction of their development cost, and are probably the mainstay of the industrial application of CFD.

Four steps are required to apply a general-purpose CFD code to an industrial problem. First, the domain must be defined. This amounts to constructing the geometry for the problem,[2] which is typically done using a computer-assisted design (CAD)-like preprocessor.[3] Within the preprocessor, relevant physics are defined, appropriate models are specified, boundary and initial conditions are applied, and solver parameters are specified. Because the conservation and constitutive equations must be discretized on the specified geometry, the domain discretization must be specified. This process, known as meshing or grid generation, is the second step in the application of a CFD code to an industrial problem. Meshing can be accomplished using two basic protocols: (1) structured meshing, which involves creating an assembly of regular, usually hexahedral (quadrilateral in two dimensions) elements or control volumes throughout the domain; and (2) unstructured meshing, which involves filling the geometry with control volumes, often tetrahedrons and prisms, in an irregular fashion. Unstructured mesh generators are usually simpler to use with complicated geometries and involve some degree of automation. For example, the user may specify one or more measures of surface grid density, and the mesh generator will fill the volume with elements according to some algorithm. In the third step, the equations are discretized over the specified grid, and the resulting nonlinear[4] algebraic equations are solved. The development of solvers is still an active area of research, the goal being to improve the likelihood and rate of convergence. The fourth step, after satisfactory convergence is obtained, is to interrogate the solution to obtain the desired information. That information may be a single number extracted from the solution data set, an animation illustrating the transient macroscopic behavior of the entire flow field, or anything in between. Because the data sets can be quite large,[5] robust tools for data set interrogation are often required. These are usually provided with the commercial CFD codes, but one leading commercial tool is a stand-alone CFD postprocessor (FIELDVIEW, 2002).

CURRENT INDUSTRIAL APPLICATIONS

CFD is routinely used today in a wide variety of disciplines and industries, including aerospace, automotive, power generation, chemical manufacturing, polymer processing, petroleum exploration, medical research, meteorology, and astrophysics. The use of CFD in the process industries has led to reductions in

[2] Specification of the time dependence (transient or steady state) and certain boundary conditions (periodic, symmetry) are also required to specify the domain completely.

[3] The development of graphical user interfaces for commercial CFD codes in the last 10 to 15 years has significantly increased their accessibility.

[4] The algebraic equations are linear if the associated partial differential equations are linear, for example, when the constitutive behavior is linear and diffusion dominates.

[5] Data sets of several hundred megabytes for industrial steady-state problems are not unusual, and many gigabytes are easily generated for typical transient problems.

the cost of product and process development and optimization activities (by reducing down time), reduced the need for physical experimentation, shortened time to market, improved design reliability, increased conversions and yields, and facilitated the resolution of environmental, health, and right-to-operate issues. It follows that the economic benefit of using CFD has been substantial, although detailed economic analyses are rarely reported. A case study of the economic benefit of the application of CFD in one chemical and engineered-material company over a six-year period conservatively estimated that the application of CFD generated approximately a six-fold return on the total investment in CFD (Davidson, 2001a).

CFD has an enormous potential impact on industry because the solution of the equations of motion provides everything that is meaningful to know about the domain. For example, chemical engineers commonly make assumptions about the fluid mechanics in process units and piping that lead to great simplifications in the equations of motion. An agitated chemical reactor may be designed on the assumption that the material in the vessel is completely mixed, when, in reality, it is probably not completely mixed. Consequently, the fluid mechanics may limit the reaction rather than the reaction kinetics, and the design may be inadequate. CFD allows one to simulate the reactor without making any assumptions about the macroscopic flow pattern and thus to design the vessel properly the first time. Similarly, the geometrically complicated parts required for melt spinning can be designed with CFD rather than rules-of-thumb or experiments, resulting in "right the first time" designs (Davidson, 2001b). Commercial publications (e.g., *CFX Update*, *Fluent News*, and *Applications from the Chemical Process Industry*) are filled with case studies illustrating how CFD was applied to the design of a particular unit, the optimization of a particular process, or the analysis of a particular phenomenon with good results.

AREAS OF RESEARCH

There are, of course, limitations to the application of CFD, and active research is being done to overcome them. The primary limitation is in the area of turbulent flow. Turbulent flows are solutions to the equations of motion and can be computed directly, at least in principle. This approach, known as direct numerical simulation, requires a spatial grid fine enough to capture the smallest length scale of the turbulent fluid motion (the Kolmogorov scale) throughout the domain of interest and a correspondingly small time step. In typical problems of industrial interest, the ratio of the length scale of the domain to the Kolmogorov length scale is so large that the required grid is prohibitively large. Available computational resources are usually inadequate for this task except for relatively simple problems.

Consequently, industrial practitioners of CFD use turbulence models, usually by solving the Reynolds-averaged equations, that is, equations generated by averaging the equations of motion over a time scale that is much larger than the time scale of the turbulent fluctuations but much smaller than the smallest time scale of interest in the application. This procedure results in a set of equations that have the same form as the original equations of motion, but with time-averaged quantities in place of instantaneous quantities, plus one additional term that arises from the nonlinear convective terms in the original equations of motion. In the Reynolds-averaged momentum-conservation equation, for example, this additional term has the form of an additional stress, known as the Reynolds stress. This term is modeled based on the time-averaged quantities of the flow field. A variety of turbulence models are available (Wilcox, 1998), but the workhorse model of industrial CFD is the so-called k-epsilon model, which was introduced several decades ago (Casey and Wintergerste, 2000; Launder and Spalding, 1974). These turbulence models can lead to significant inaccuracies, and CFD practitioners must use them carefully.

Large eddy simulation (LES) is an alternative approach to turbulence modeling. Turbulent flows are characterized by an eddy cascade, in which large eddies transfer their kinetic energy to smaller eddies, which in turn transfer kinetic energy to even smaller eddies, and so on until, at the Kolmogorov scale, the kinetic energy is transformed into heat. LES attempts to solve for the larger eddies directly while modeling the smaller eddies. Although LES is more computationally intensive than other kinds of turbulence modeling, it has been applied to industrial-scale problems (Derksen, 2001).

The second great limitation of CFD is dispersed, multiphase flows. Multiphase flows are common in industry, and consequently their simulation is of great interest. Like turbulent flows, multiphase flows (which may also be turbulent in one or more phases) are solutions to the equations of motion, and direct numerical simulation has been applied to them (Miller and Bellan, 2000). However, practical multiphase flow problems require a modeling approach. The models, however, tend to ignore or at best simplify many of the important details of the flow, such as droplet or particle shape and their impact on interphase mass, energy, and momentum transport, the impact of deformation rate on droplet breakup and coalescence, and the formation of macroscopic structures within the dispersed phase (Sundaresan et al., 1998).

ENTERPRISE-WIDE ACCESS

Although the commercial CFD industry has greatly simplified the use of CFD codes by providing CAD-like preprocessors, automatic mesh generation, graphical user interfaces for all aspects of model definition, and on-line documentation, the industrial practice of CFD is still primarily in the hands of specialists. Regular use of a general-purpose code requires significant expertise in transport phenomena, an understanding of the capabilities and limitations of the modeling approaches used to handle turbulence and dispersed multiphase flows, an understanding of the relationship between mesh quality, convergence, and solution accuracy, and proficiency with the various means of interacting with the CFD code, including the graphical user interface, advanced command languages (when available), and user-accessible FORTRAN subroutines.[6] For these reasons, attempts to train large numbers of engineers in the use of CFD have not been very successful (Davidson, 2001a). Nevertheless, the potential benefit of a much broader CFD user base is very great. In our opinion, CFD should be accessible to every person in the enterprise who makes decisions the outcomes of which are governed by the laws of physics, from the CEO who makes strategic business decisions based on business goals to the operator who adjusts valve positions to meet process goals.

We believe that this can be achieved through the development of so-called "digital experts," stand-alone CFD (and other) applications that would be integrated into commercial CFD codes (as appropriate) and wrapped in interfaces that speak the language of the industrial application, not the language of CFD. Digital experts would automate geometry construction, mesh generation, solver selection, and other processes behind the scenes. In addition, they would contain all of the algorithms necessary to nurse the CFD codes to solution automatically, without having to ask the user to define satisfactory convergence, for example. Finally, they would extract the essential ingredients from the complete CFD solution and present them to the user in a convenient and familiar format, so the user would not have to be concerned with interrogation of the flow field by computation or visualization. A discussion of one digital expert that has been developed for melt-fiber spinning has been published (Davidson, 2001b).

The decision to develop an industrial digital expert is based on the relationship between development cost and benefit. Recently, a commercial product has become available with the potential to change that relationship significantly. EASA™ (Enterprise Accessible Software Applications from AEA Technology), which was designed to help industrial practitioners develop digital experts, solves a number of problems for industrial developers, including construction of the graphical user interface, accessibility of the final product (the digital expert or *EASAp*) over the enterprise intranet, and the orchestration of computations on a heterogeneous computer network (Dewhurst, 2001). In essence, EASA allows an industrial CFD specialist to put bullet-proof digital experts in the hands of his or her coworkers, with a consistent interface tailored to the user.

[6] Although the graphical user interface (GUI) has greatly increased accessibility of the commercial codes, more experienced users sometimes avoid the use of the GUI altogether and rely on command languages, user FORTRAN, and ASCII files to integrate various tools and automate CFD tasks.

SUMMARY

CFD is a powerful tool for solving a wide variety of industrial problems. Commercial general-purpose codes have the potential to solve a very broad spectrum of flow problems. Current research is concentrated on overcoming the principle weaknesses of CFD, namely how it deals with turbulence and dispersed multiphase flows. Development work on solver algorithms, meshing, and user interface generation are ongoing, with the objectives of improving accuracy, reducing solution time, and increasing accessibility. In spite of the limitations of CFD, the economic value of industrial applications has been demonstrated in a variety of industries, and its value as a research tool has been accepted in many areas, such as meteorology, medicine, and astrophysics. In industry, CFD is presently primarily in the hands of specialists, but the development of digital experts and tools to facilitate the development of digital experts may revolutionize the way industry uses CFD by providing ready access throughout the enterprise. This would result in significant gains in productivity and profitability.

REFERENCES

Casey, M., and T. Wintergerste, eds. 2000. Best Practice Guidelines. Brussels: European Research Community on Flow, Turbulence and Combustion.

Davidson, D.L. 2001a. The Enterprise-Wide Application of Computational Fluid Dynamics in the Chemicals Industry. Proceedings of the 6th World Congress of Chemical Engineering. Available on Conference Media CD, Melbourne, Australia.

Davidson, D.L. 2001b. Spin*Expert*: The Digital Expert for the Design and Analysis of Fiber Spinning Operations. Pp. 219-226 in Proceedings of the 3rd International ASME Symposium on Computational Technology (CFD) for Fluid/Thermal/Chemical/Stress Systems and Industrial Applications. Atlanta, Ga.: American Society of Mechanical Engineers.

Derksen, J.J. 2001. Applications of Lattice-Boltzmann-Based Large-Eddy Simulations. Pp. 1-11 in Proceedings of the 3rd International ASME Symposium on Computational Technology (CFD) for Fluid/Thermal/Chemical/Stress Systems and Industrial Applications. Atlanta, Ga.: American Society of Mechanical Engineers.

Dewhurst, S. 2001. An exciting new way to deploy your CFD. CFX Update 21: 6.

FIELDVIEW. 2002. Intelligent Light Website. Available online at: <*http://www.ilight.com*>.

Launder, B.E., and D.B. Spalding. 1974. The numerical computation of turbulent flow. Computational Methods in Applied Mechanics and Engineering 3: 269–289.

Miller, R.S., and J. Bellan. 2000. Direct numerical simulation and subgrid analysis of a transitional droplet laden mixing layer. Physics of Fluids 12(3): 650–671.

Sundaresan, S., B.J. Glasser, and I.G. Kevrekidis. 1998. From bubbles to clusters in fluidized beds. Physical Review Letters 81(9): 1849–1852.

Wilcox, D.C. 1998. Turbulence Modeling for CFD. La Canada, Calif.: DCW Industries Inc.

SUGGESTED READINGS

Anderson, D.A., J.C. Tannehill, and R.H. Pletcher. 1984. Computational Fluid Mechanics and Heat Transfer. New York: Hemisphere Publishing Corp.

Batchelor, G.K. 1977. An Introduction to Fluid Dynamics. London: Cambridge University Press.

Berg, P.W., and J.L. McGregor. 1966. Elementary Partial Differential Equations. Oakland, Calif.: Holden-Day.

Bird, R.B., W.E. Stewart, and E.N. Lightfoot. 1960. Transport Phenomena. New York: John Wiley & Sons.

Carslaw, H.S., and J.C. Jaeger. 1959. Conduction of Heat in Solids. Oxford, U.K.: Clarendon Press.

Chandrasekhar, S. 1981. Hydrodynamic and Hydromagnetic Stability. New York: Dover Press.

Crochet, M.J., A.R. Davies, and K. Walters. 1984. Numerical Simulation of Non-Newtonian Flow. Amsterdam: Elsevier Science Publishers.

Jackson, J.D. 1975. Classical Electrodynamics. New York: John Wiley & Sons.

Schowalter, W.R. 1978. Mechanics of Non-Newtonian Fluids. Oxford, U.K.: Pergamon Press.

Van Dyke, M. 1975. Perturbation Methods in Fluid Mechanics. Stanford, Calif.: Parabolic Press.

Whitaker, S. 1984. Introduction to Fluid Mechanics. Malabar, Fla.: Krieger Publishing Co.

Technology for Human Beings

The Human Factor

Kim J. Vicente
Department of Mechanical & Industrial Engineering
University of Toronto
Toronto, Ontario

Many people find technology frustrating and difficult to use in everyday life. In the vast majority of cases, the problem is not that they are technological "dummies" but that the designers of technological systems did not pay sufficient attention to human needs and capabilities. The new BMW 7 series automobile, for example, has an electronic dashboard system, referred to as iDrive, that has between 700 and 800 features (Hopkins, 2001). An article in *Car and Driver* described it this way: "[it] may…go down as a lunatic attempt to replace intuitive controls with overwrought silicon, an electronic paper clip on a lease plan. One of our senior editors needed 10 minutes just to figure out how to start it" (Robinson, 2002). An editor at *Road & Track* agreed: "It reminds me of software designers who become so familiar with the workings of their products that they forget actual customers at some point will have to learn how to use them. Bottom line, this system forces the user to think way too much. A good system should do just the opposite" (Bornhop, 2002). As technologies become more complex and the pace of change increases, the situation is likely to get worse.

In everyday situations, overlooking human factors leads to errors, frustration, alienation from technology, and, eventually, a failure to exploit the potential of people and technology. In safety-critical systems, however, such as nuclear power plants, hospitals, and aviation, the consequences can threaten the quality of life of virtually everyone on the planet. In the United States, for example, preventable medical errors are the eighth leading cause of death; in hospitals alone, errors cause 44,000 to 98,000 deaths annually, and patient injuries cost between $17 billion and $29 billion annually (IOM, 1999).

DIAGNOSIS

The root cause of the problem is the separation of the technical sciences from the human sciences. Engineers who have traditionally been trained to focus on technology often have neither the expertise nor the inclination to pay a great deal of attention to human capabilities and limitations. This one-sided view leads to a paradoxical situation. When engineers ignore what is known about the physical world and design a technology that fails, we blame them for professional negligence. When they ignore what is known about human nature and design a technology that fails, we typically blame users for being technologically incompetent. The remedy would be for engineers to begin with a human or social need (rather than a technological possibility) and to focus on the interactions between people and technology (rather than on the technology alone). Technological systems can be designed to match human nature at all scales—physical, psychological, team, organizational, and political (Vicente, in press).

COMPUTER DISPLAYS FOR NUCLEAR POWER PLANTS

People are very good at recognizing graphical patterns. Based on this knowledge, Beltracchi (1987) developed an innovative computer display for monitoring the safety of water-based nuclear power plants. To maintain a safety margin, operators must ensure that the water in the reactor core is in a liquid state. If the water begins to boil, as it did during the Three Mile Island accident, then the fuel can eventually melt, threatening public health and the environment. In traditional control rooms, such as the one shown in Figure 1, operators have to go through a tedious procedure involving steam tables and individual meter readings to monitor the thermodynamic status of the plant. This error-prone procedure requires that operators memorize or record numerical values, perform mental calculations, and execute several steps.

Beltracchi's display (Figure 2) is based on the temperature-entropy diagram found in thermodynamic textbooks. The saturation properties of water are shown in graphical form as a bell curve rather than in alphanumeric form as in a steam table. Furthermore, the thermodynamic state of the plant can be described as a Rankine cycle, which has a particular graphical form when plotted in temperature-entropy coordinates. By measuring the temperature and pressure at key locations in the plant, it is possible to obtain real-time sensor values that can be plotted in this graphical diagram. The saturation properties of water are presented in a visual form that matches the intrinsic human capability of recognizing graphical patterns easily and effectively. An experimental evaluation of professional nuclear power plant operators showed that this new way of presenting information leads to better interactions between people and technology than the traditional way (Vicente et al., 1996).

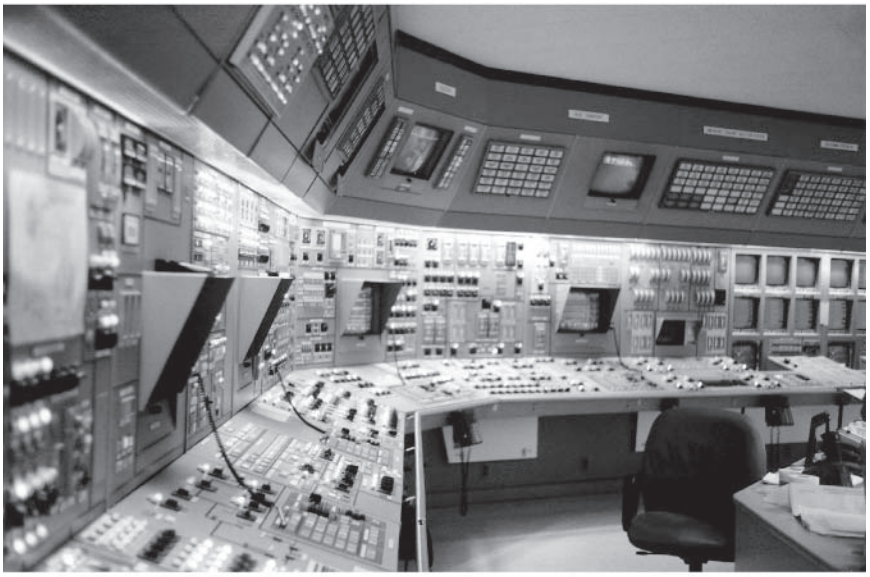

FIGURE 1 A typical control room for a nuclear power plant. Source: Photo courtesy of C.M. Burns.

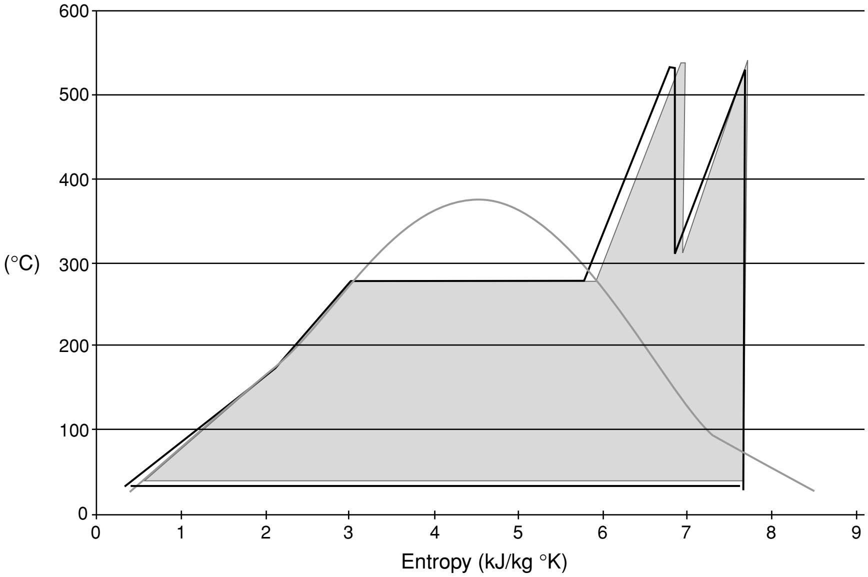

FIGURE 2 Beltracchi display. Source: Burns, 2000. Reprinted with permission.

FRAMEWORK FOR RISK MANAGEMENT

Public policy decisions are necessarily made in a dynamic, even turbulent, social landscape that is continually changing. In the face of these perturbations, complex sociotechnical systems must be robust. Rasmussen (1997) developed an innovative framework for risk management to achieve this goal (Figure 3).

The first element of the framework is a structural hierarchy describing the individuals and organizations in the sociotechnical system. The number of levels and their labels can vary from industry to industry. Take, for example, a structural hierarchy for a nuclear power plant. The lowest level usually describes the behavior associated with the particular (potentially hazardous) process being controlled (e.g., the nuclear power plant). The next level describes the activities of the individual staff members who interact directly with the process being controlled (e.g., control room operators). The third level from the bottom describes the activities of management that supervises the staff. The next level up describes the activities of the company as a whole. The fifth level describes the activities of the regulators or associations responsible for setting limits to the activities of companies in that sector. The top level describes the

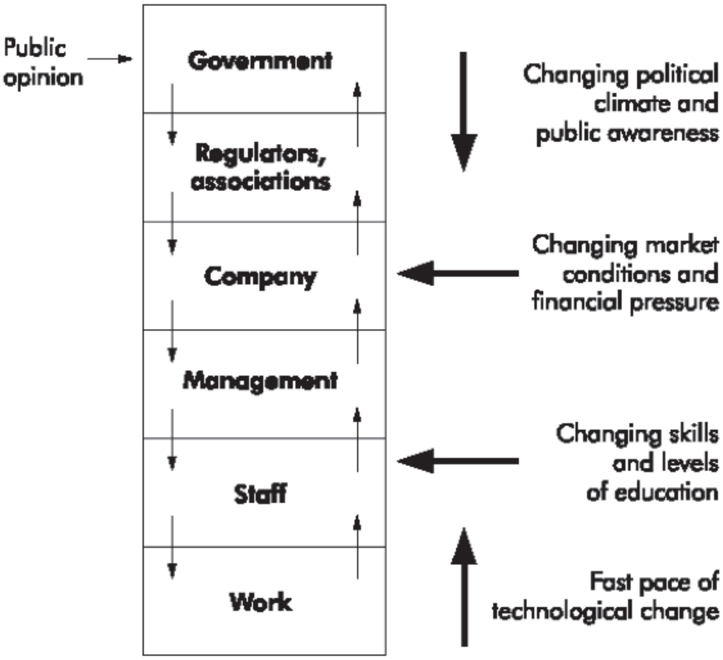
Public opinion
Government
Regulators, associations
Company
Management
Staff
Work
Changing political climate and public awareness
Changing market conditions and financial pressure
Changing skills and levels of education
Fast pace of technological change

FIGURE 3 Levels of a complex sociotechnical system involved in risk management. Source: Adapted from Rasmussen, 1997.

activities of government (civil servants and elected officials) responsible for setting public policy.

Decisions at higher levels propagate down the hierarchy, and information about the current state of affairs propagates up the hierarchy. The interdependencies among the levels of the hierarchy are critical to the successful functioning of the system as a whole. If instructions from above are not formulated or not carried out, or if information from below is not collected or not conveyed, then the system may become unstable and start to lose control of the hazardous process it is intended to safeguard.

In this framework, safety is an emergent property of a complex sociotechnical system. Safety is affected by the decisions of all of the actors—politicians, CEOs, managers, safety officers, and work planners—not just front-line workers. Threats to safety or accidents usually result from a loss of control caused by a lack of vertical integration (i.e., mismatches) among the levels of the entire system, rather than from deficiencies at any one level.

Inadequate vertical integration is frequently caused, at least partly, by a lack of feedback from one level to the next. Because actors at each level cannot see how their decisions interact with decisions made by actors at other levels, threats to safety may not be obvious before an accident occurs. Nobody has a global view of the entire system.

The layers of a complex sociotechnical system are increasingly subjected to external forces that stress the system. Examples of perturbations include: the changing political climate and public awareness; changing market conditions and financial pressures; changing competencies and levels of education; and changes in technological complexity. The more dynamic the society, the stronger these external forces are and the more frequently they change.

The second component of the framework deals with dynamic forces that can cause a complex sociotechnical system to modify its structure over time. On the one hand, financial pressures can create a cost gradient that pushes actors in the system to be more fiscally responsible. On the other hand, psychological pressures can create a gradient that pushes actors in the system to work more efficiently, mentally or physically.

Pressure from these two gradients subject work practices to a kind of "Brownian motion," an exploratory but systematic migration over time. Just as the force of gravity causes a stream of water to flow down crevices in a mountainside, financial and psychological forces inevitably cause people to find the most economical ways of performing their jobs. Moreover, in a complex sociotechnical system, changes in work practices can migrate from one level to another. Over time, this migration will cause people responding to requests or demands to deviate from accepted procedures and cut corners to be more cost-effective. As a result, over time the system's defenses are degraded and eroded.

A degradation in safety may not raise an immediate warning flag for two reasons. First, given the stresses on the system, the migration in work practices may be necessary to get the job done. That is why so-called "work-to-rule" campaigns requiring that people do their jobs strictly by the book usually cause complex sociotechnical systems to come to a grinding halt. Second, the migration in work practices usually does not have immediate visible negative impacts. The safety threat is not obvious because violations of procedures do not lead immediately to catastrophe. At each level in the hierarchy, people may be working hard and striving to respond to cost-effectiveness measures; but they may not realize how their decisions interact with decisions made by actors at other levels of the system. Nevertheless, the sum total of these uncoordinated attempts at adapting to environmental stressors can slowly but surely "prepare the stage for an accident" (Rasmussen, 1997).

Migrations from official work practices can persist and evolve for years without any apparent breaches of safety until the safety threshold is reached and an accident happens. Afterward, workers are likely to wonder what happened because they had not done anything differently than they had in the recent past.

Rasmussen's framework makes it possible to manage risk by vertically integrating political, corporate, managerial, worker, and technical considerations into a single integrated system that can adapt to novelty and change.

Case Study

A fatal outbreak of *E. coli* in the public drinking water system in Walkerton, Ontario, during May 2000 illustrates the structural mechanisms at work in Rasmussen's framework (O'Connor, 2002). In a town of 4,800 residents, 7 people died and an estimated 2,300 became sick. Some people, especially children, are expected to have lasting health effects. The total cost of the tragedy was estimated to be more than $64.5 million (Canadian).

In the aftermath of the outbreak, people were terrified of using tap water to satisfy their basic needs. People who were infected or who had lost loved ones suffered tremendous psychological trauma; their neighbors, friends, and families were terrorized by anxiety; and people throughout the province were worried about how the fatal event could have happened and whether it could happen again in their towns or cities. Attention-grabbing headlines continued unabated for months in newspapers, on radio, and on television. Eventually, the provincial government appointed an independent commission to conduct a public inquiry into the causes of the disaster and to make recommendations for change. Over the course of nine months, the commission held televised hearings, culminating in the politically devastating interrogation of the premier of Ontario. On January 14, 2002, the Walkerton Inquiry Commission delivered Part I of its report to the attorney general of the province of Ontario (O'Connor, 2002).

The sequence of events revealed a complex interaction among the various levels of a complex sociotechnical system, including strictly physical factors, unsafe practices of individual workers, inadequate oversight and enforcement by local government and a provincial regulatory agency, and budget reductions imposed by the provincial government. In addition, the dynamic forces that led to the accident had been in place for some time—some going back 20 years—but feedback that might have revealed the safety implications of these forces was largely unavailable to the various actors in the system. These findings are consistent with Rasmussen's predictions and highlight the importance of vertical integration in a complex sociotechnical system.

CONCLUSIONS

We must begin to change our engineering curricula so that graduates understand the importance of designing technologies that work for people performing the full range of human activities, from physical to political activities and everything in between. Corporate design practices must also be modified to focus on producing technological systems that fulfill human needs as opposed to creating overly complex, technically sophisticated systems that are difficult for the average person to use. Finally, public policy decisions must be based on a firm understanding of the relationship between people and technology. One-sided approaches that focus on technology alone often exacerbate rather than solve pressing social problems.

Because the National Academy of Engineering has unparalleled prestige and expertise, it could play a unique role in encouraging educational, corporate, and governmental changes that could lead to the design of technological systems that put human factors where they belong—front and center.

ACKNOWLEDGMENTS

This paper was sponsored in part by the Jerome Clarke Hunsaker Distinguished Visiting Professorship at MIT and by a research grant from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Beltracchi, L. 1987. A direct manipulation interface for water-based Rankine cycle heat engines. IEEE Transactions on Systems, Man, and Cybernetics SMC-17: 478-487.

Bornhop, A. 2002. BMW 745I: iDrive? No, you drive, while I fiddle with the controller. Road & Track 53(10): 74-79.

Burns, C.M. 2000. Putting it all together: improving display integration in ecological displays. Human Factors 42: 226-241.

Hopkins, J. 2001. When the devil is in the design. USA Today, December 31, 2001. Available online at: <*www.usatoday.com/money/retail/2001-12-31-design.htm*>.

IOM (Institute of Medicine). 1999. To Err Is Human: Building a Safer Health System, edited by L.T. Kohn, J.M. Corrigan, and M.S. Donaldson. Washington, D.C.: National Academy Press.

O'Connor, D.R. 2002. Report of the Walkerton Inquiry: The Events of May 2000 and Related Issues. Part One. Toronto: Ontario Ministry of the Attorney General. Available online at: <*www.walkertoninquiry.com*>.

Rasmussen, J. 1997. Risk management in a dynamic society: a modelling problem. Safety Science 27(2/3): 183-213.

Robinson, A. 2002. BMW 745I: the ultimate interfacing machine. Car and Driver 47(12): 71-75.

Vicente, K.J. In Press. The Human Factor: Revolutionizing the Way We Live with Technology. Toronto: Knopf Canada.

Vicente, K.J., N. Moray, J.D. Lee, J. Rasmussen, B.G. Jones, R. Brock, and T. Djemil. 1996. Evaluation of a Rankine cycle display for nuclear power plant monitoring and diagnosis. Human Factors 38: 506-521.

Human Factors Applications in Surface Transportation

THOMAS A. DINGUS
Virginia Tech Transportation Institute
Virginia Polytechnic Institute and State University
Blacksburg, Virginia

Our multifaceted surface transportation system consists of infrastructure, vehicles, drivers, and pedestrians. The hardware and software engineering subsystems, however, are generally traditional and not overly complex. In fact, the technology in the overall system is, by design, substantially removed from the "leading edge" to ensure that it meets safety, reliability, and longevity requirements. The complexities of the transportation system are primarily attributable to the human factors of driving, including psychomotor skills, attention, judgment, decision making, and even human behavior in a social context.

Many transportation researchers have come to the realization that human factors are the critical elements in solving the most intransigent problems in surface transportation (e.g., safety and improved mobility). That is, the primary issues and problems to be solved are no longer about asphalt and concrete, or even electronics; instead they are focused on complex issues of driver performance and behavior (ITSA and DOT, 2002).

A substantial effort is under way worldwide to reduce the number of vehicular crashes. Although the crash rate in the United States is substantially lower than it was, crashes continue to kill more than 40,000 Americans annually and injure more than 3,000,000 (NHTSA, 2001). Because "driver error" is a contributing factor in more than 90 percent of these crashes, it is clear that solutions to transportation problems, perhaps more than in any other discipline, must be based on human factors.

One of the hurdles to assessing the human factors issues associated with driving safety is the continuing lack of data that provides a detailed and valid representation of the complex factors that occur in the real-world driving environment. In this paper, I describe some new techniques for filling the data void and present examples of how these techniques are being used to approach important safety problems.

DATA COLLECTION AND ANALYSIS

There are two traditional approaches to collecting and analyzing human factors data related to driving. The first approach is to use data gathered through epidemiological studies (often collected on a national level). These databases, however, lack sufficient detail to be helpful for many applications, such as the development of countermeasure systems or the assessment of interactions between causal and contributing factors that lead to crashes.

The second approach, empirical methods, including newer, high-fidelity driving simulators and test tracks, are necessarily contrived and do not always capture the complexities of the driving environment or of natural behavior. For example, test subjects are often more alert and more careful in a simulation environment or when an experimenter is present in a research vehicle than when they are driving alone in their own cars. Thus, although empirical methods are very useful in other contexts, they provide a limited picture of the likelihood of a crash in a given situation or the potential reduction of that likelihood by a given countermeasure. State-of-the-art empirical approaches can only assess the relative safety of various countermeasures or scenarios. They cannot be used to predict the impact of a safety device or policy change on the crash rate.

Advances in sensor, data storage, and communications technology have led to the development of a hybrid approach to data collection and analysis that uses very highly capable vehicle-based data collection systems. This method of data collection has been used by some auto manufacturers since the introduction of electronic data recorders (EDRs) several years ago. EDRs collect a variety of vehicular dynamic and state data that can be very useful in analyzing a crash. However, they currently lack sufficient measurement capability to assess many human factors-related issues.

Recently, a handful of efforts has been started to collect empirical data on a very large scale in a "pseudonaturalistic" environment—subjects use the instrumented vehicles for an extended period of time (e.g., up to a year) for their normal driving, with no in-vehicle experimenter or obtrusive equipment. Unlike EDRs, these systems also use unobtrusive video and electronic sensors. The goal is to create a data collection environment that is valid, provides enough detail, and is on a large enough scale to reveal the relationship between human factors (e.g., fatigue, distraction, driver error, etc.) and other factors that contribute to crashes. Although these studies are not naturalistic in the strict sense, evidence collected thus far indicates that they can provide an accurate picture of the myriad factors involved in driving safety.

Three examples of pseudonaturalistic approaches are described below.

The Naturalist "100 Car" Study

The primary objectives of this study are to develop instrumentation, methods, and analysis techniques to conduct large-scale, pseudonaturalistic investigations. This pilot study will collect continuous, real-time data over a period of one year for 100 high-mileage drivers. The data set will include five channels of video data and electronic data from many sensors. Data-triggering techniques will be used to identify crashes, near crashes, and other critical incidents for further analysis. Thus, the data will provide video and quantitative information associated with any "event" of interest. These could include the use of cellular telephones (triggered via radio-frequency sensors), unplanned lane deviations (detected by lane-position sensors), short time-to-collision situations (detected by radar sensors), or actual crashes (detected by accelerometers), just to name a few. The data can then be analyzed to determine the exact circumstances for each event. Once the database is complete, the information can be used to develop and evaluate concepts for countermeasures of all types, everything from engineering solutions to enforcement practices.

Researchers anticipate that the database created as part of the 100 Car Study will provide a wealth of information, similar to the information provided by a crash database, but with a great deal more detail. Data analysis will be conducted for vehicle-following and reaction time, the effects of distractions (e.g., electronic devices), driver behavior in proximity to heavy trucks, and the quantitative relationship between the frequency of crashes and other critical incidents. However, a primary purpose of the 100 Car Study is to develop instrumentation and data collection and analysis techniques for much larger studies (e.g., a study of 10,000 cars).

Field Operational Tests of Human Factors and Crash Avoidance

Another use of large-scale data on naturalistic driving data is to evaluate the benefits of engineering-based safety measures. Pseudonaturalistic studies using actual vehicles are currently under way to assess collision-avoidance systems (which use forward-facing radar to warn drivers of an impending crash) as well as lane-position monitoring and driver-alertness monitoring systems for heavy trucks.

Unlike simulator or test-track studies, these studies can test devices in situ and monitor the interactions of all of the environmental factors described above. They can also assess how drivers adapt over time. This is an extremely important aspect of a safety evaluation because, if drivers rely too much on a safety device, the crash rate may actually *increase*.

Driver Fitness-For-Duty Studies

Fitness-for-duty studies are geared toward helping policy makers assess the effects of fatigue, alcohol, prescription drugs, and other factors that can affect a driver's behavior. For example, a recently completed pseudonaturalistic study assessed the quality of sleep that truck drivers obtain in a "sleeper berth" truck (Dingus et al., 2002). After epidemiological studies identified fatigue among truck drivers as a significant problem, it became important to determine the causes of the problem and possible ways to address it. Researchers hypothesized that a likely cause was the generally poor quality of sleep on the road. An additional hypothesis was that team drivers who attempted to sleep while the truck was moving would have the poorest quality sleep and, therefore, would be the highest risk drivers. A large-scale instrumented-vehicle study (56 drivers and 250,000 miles of driving data) was undertaken to assess sleep quality, driver alertness, and driver performance on normal revenue-producing trips averaging up to eight days in length. Using this methodology, it was determined that, although team drivers obtained a poorer quality of sleep than single drivers, the poor quality was offset by the efficient use of relief drivers. The results showed that single drivers suffered the worst bouts of fatigue and had the most severe critical incidents (by about 4 to 1). This (and other) important findings could only have been obtained from pseudonaturalistic studies.

SUMMARY

Significant progress in driving safety will require additional large-scale data from pseudonaturalistic studies that can complement existing data-gathering methods. Eventually, these data may improve our understanding of the causal and contributing factors of crashes so that effective countermeasures can be evaluated and deployed efficiently and quickly.

REFERENCES

Dingus, T., V. Neale, S. Garness, R. Hanowski, A. Keisler, S. Lee, M. Perez, G. Robinson, S. Belz, J. Casali, E. Pace-Schott, R. Stickgold, and J.A. Hobson. 2002. Impact of Sleeper Berth Usage on Driver Fatigue. Washington, D.C.: Federal Motor Carrier Safety Administration, U.S. Department of Transportation.

ITSA (Intelligent Transportation Society of America) and DOT (U.S. Department of Transportation). 2002. National ITS Program Plan: A Ten-Year Vision. Washington, D.C.: Intelligent Transportation Society of America and U.S. Department of Transportation.

NHTSA (National Highway Traffic Safety Administration). 2001. Traffic Safety Facts 2000: A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System. DOT HS 809 337. Washington, D.C.: National Highway Traffic Safety Administration.

Implications of Human Factors Engineering for Novel Software User-Interface Design

MARY CZERWINSKI
Microsoft Research
Redmond, Washington

Human factors engineering (HFE) can help software designers determine if a product is useful, usable, and/or fun. The discipline incorporates a wide variety of methods from psychology, anthropology, marketing, and design research into the development of user-centric approaches to software design. The ultimate goal is to ensure that software products are discoverable, learnable, memorable, and satisfying to use.

HFE is a discipline partly of science and partly of design and technological advancement. The scientific aspects of the discipline primarily bring together principles from psychology and anthropology. HFE professionals require a background in research on human cognitive abilities (e.g., attention, visual perception, memory, learning, time perception, categorization) and on human-computer interaction (HCI) (e.g., task-oriented methods, heuristic evaluations, input, visualization, menu design, and speech research). Although HFE is based on solid principles from basic research, it is constantly evolving in response to improvements in technology and changes in practices and values.

For software products to be successful in terms of ease of use, HFE practices must be incorporated early in the product development cycle. In areas of fierce competition or when innovation is necessary for product success, user-centered design practices have been shown time and again to make or break a product. There are too many principles of HCI and software design to include in this short paper, but many excellent books are available on the subject (e.g., Newman and Lamming, 1995; Preece, 1994; Shneiderman, 1998; Wickens, 1984). However, we can easily summarize the rule for incorporating HFE into the software design process with one golden principle—know thy user(s). HFE professionals must research end users' tasks, time pressures, working styles, familiarity with user interface (UI) concepts, and so forth. Many tools are available for determining these characteristics of the user base, including field work, laboratory studies, focus groups, e-mail surveys, remote laboratory testing, persona development, paper prototyping, and card sorts. Most, if not all, of these are used at various points in the product life cycle.

RATIONALE FOR USING HUMAN FACTORS ENGINEERING

The most obvious reason for including user-centered design practices during software product development is to ensure that the product will be useful (i.e., solves a real problem experienced by the target market) and usable (i.e., easy to learn and remember and satisfying to use). In addition, cost savings have been well documented (e.g., Bias and Mayhew, 1994; Nielsen, 1993). For example, a human factors engineer spent just one hour redesigning a graphical element for rotary-dial phones for a certain company and ended up saving the company about $1 million by reducing demand on central switches (Nielsen, 1993). A study of website design by Lohse and Spiller (1998) showed that the factor most closely correlated with actual purchases was the ease with which customers could navigate the website. In other words, ease of use is closely correlated with increased sales, especially on the web. Interestingly, the benefits can be measured not just in reduced costs or streamlined system design or sales figures. User-centered design also shortens the time it takes to design a product right the first time, because feedback is collected throughout the process instead of at the end of the process or after products have been shipped when reengineering is very costly. Finally, HFE saves money on the product-support side of the equation, because there are fewer calls for help from customers and fewer products returned. In the end, if users are satisfied with a company's products, they are likely to buy more products from that company in the future.

User expectations of ease of use have steadily increased, especially in the web domain, where it only takes one click for users to abandon one site for another. In this way, users are spreading the word virtually about their demands for web design. The same is true for packaged software. In addition, several countries (most notably Germany) have adopted standards for ease of use that software companies must demonstrate before they can make international sales. To meet these standards, many companies require the help of HFE professionals.

Last but not least, user-centered design is one of the few ways to ensure that innovative software solves real human problems. The following case study illustrates how one company uses HCI research in the software development process.

CASE STUDY: LARGE DISPLAYS

Microsoft Research has provided HCI researchers with access to very large displays, often 42 inches wide, running aspect ratios of 3072 × 768, using triple projections and Microsoft Windows XP support for multiple monitors. When we performed user studies on people using Windows OS software on 120-degree-field-of-view displays, it became immediately apparent that the UI had to be redesigned for these large display surfaces. For instance, the only task bar, which was displayed on the "primary" monitor of the three projections, did not stretch across all three displays. Similarly, there was only one Start menu, and one system tray (where notifications, icons, and instant messages are delivered) located in a far corner of only one of the displays. Even with these design limitations, our studies showed significant improvements in productivity when users performed everyday knowledge worker tasks using the larger displays. Our challenge was to create novel UIs that maintained or furthered the increased productivity but made using software programs easier. In addition, software innovations could not get in the way of what users already understood about working with Windows software, which already had a very large user base.

Following the principle of Know Thy Users, we hired an external research firm to do an ethnographical study of how 16 high-end knowledge workers multitasked and which software elements did or did not support this kind of work. In addition, we performed our own longitudinal diary and field studies in situ with users of large and multiple displays. This intensive research, which took more than a month, exposed a variety of problems and work style/task scenarios that were crucial during the design phase. For instance, we were surprised to see variations in the way multiple monitors were used by workers in different domains of knowledge. For example, designers and CAD/CAM programmers used one monitor for tools and menus and another for their "palettes"; other workers merely wanted one large display surface. We were also struck by the amount of task switching, and we realized that Windows and Office could provide better support for these users. After these studies had been analyzed and recorded, we had accomplished the following goals:

- We had developed a market profile of end users for the next two to five years and studied their working style, tasks, and preferences.
- We had developed personas (descriptions of key profiles in our user base) based on our research findings and developed design scenarios based on the fieldwork and the observed, real world tasks.
- We had developed prototypes of software UI design that incorporated principles of perception, attention, memory, and motor behavior and solved the real world problems identified in the research.
- We had tested ideas using tasks from real end-user scenarios and benchmarked them against existing software tools, iterated our software product designs, and retested them.
- We had learned a great deal about perception, memory, task switching,

and the ability to reinstate context and navigate to high-priority information and content.

The design solution we invented was a success with our target end users. After packaging the software technology for transfer to our colleagues, we moved on to research on creative, next-generation visualizations based on our studies. We will continue to do user studies on our new, more innovative designs to ensure that we remain focused on providing better solutions to real problems for large displays than are provided by UIs with standard software.

CONCLUSION

I have argued that innovation and technology transfer for successful software products must be guided by sound HFE, because customers today expect it and clearly favor usable systems. In addition, sound HFE saves time and money during software development because there are fewer calls to the help desk, fewer product returns, more satisfied and loyal customers, more innovative product solutions, and faster development life cycles.

REFERENCES

Bias, R.G., and D.J. Mayhew. 1994. Cost Justifying Usability. Boston: Academic Press.

Lohse, G., and P. Spiller. 1998. Quantifying the Effect of User Interface Design Features on Cyberstore Traffic and Sales. Pp. 211-218 in Proceedings of CHI 98: Human Factors in Computing Systems. New York: ACM Press/Addison-Wesley Publishing Co.

Nielsen, J. 1993. Usability Engineering. Boston: Academic Press.

Newman, W.M., and M.G. Lamming. 1995. Interactive System Design. Reading, U.K.: Addison-Wesley.

Preece, J. 1994. Human-Computer Interaction. Menlo Park, N.J.: Addison-Wesley.

Shneiderman, B. 1998. Designing the User Interface, 3rd ed. Menlo Park, N.J.: Addison-Wesley.

Wickens, C.D. 1984. Engineering Psychology and Human Performance. Glenview, Ill.: Scott, Foresman & Company.

Frontiers of Human-Computer Interaction: Direct-Brain Interfaces

MELODY M. MOORE
Computer Information Systems Department
Georgia State University
Atlanta, Georgia

A direct-brain interface (DBI), also known as a brain-computer interface (BCI), is a system that detects minute electrophysiological changes in brain signals and uses them to provide a channel that does not depend on muscle movement to control computers and other devices (Wolpaw et al., 2002). In the last 15 years, research has led to the development of DBI systems to assist people with severe physical disabilities. As work in the field continues, mainstream applications for DBIs may emerge, perhaps for people in situations of imposed disability, such as jet pilots experiencing high G-forces during maneuvers, or for people in situations that require hands-free, heads-up interfaces. The DBI field is just beginning to explore the possibilities of real-world applications for brain-signal interfaces.

LOCKED-IN SYNDROME

One of the most debilitating and tragic circumstances that can befall a human being is to become "locked-in," paralyzed and unable to speak but still intact cognitively. Brainstem strokes, amyotrophic lateral sclerosis, and other progressive diseases can cause locked-in syndrome, leaving a person unable to move or communicate, literally a prisoner in his or her own body. Traditional assistive technologies, such as specialized switches, depend on small but distinct and reliable muscle movement. Therefore, until recently, people with locked-in syndrome had few options but to live in virtual isolation; observing and comprehending the world around them but powerless to interact with it.

DBIs have opened avenues of communication and control for people with severe and aphasic disabilities, and locked-in patients have been the focus of much DBI work. Experiments have shown that people can learn to control their brain signals enough to operate communication devices such as virtual keyboards, operate environmental control systems such as systems that can turn lights and TVs on and off, and even potentially restore motion to paralyzed limbs. Although DBI systems still require expert assistance to operate, they have a significant potential for providing alternate methods of communication and control of devices (Wolpaw et al., 2002).

CONTROL THROUGH DIRECT-BRAIN INTERFACES

Brain control in science fiction is typically characterized as mind reading or telekinesis, implying that thoughts can be interpreted and directly translated to affect or control objects. Most real-world DBIs depend on a person learning to control an aspect of brain signals that can be detected and measured. Some depend on the detection of invoked responses to stimuli to perform control tasks, such as selecting letters from an alphabet. Some categories of brain signals that can be used to implement a DBI are described below.

Field potentials are synchronized activity of large numbers of brain cells that can be detected by extracellular recordings, typically electrodes placed on the scalp (known as electroencephalography or EEG) (Kandel, 2000). Field potentials are characterized by their frequency of occurrence. Studies have shown that people can learn via operant-conditioning methods to increase and decrease the voltage of brain signals (in tens of microvolts) to control a computer or other device (Birbaumer et al., 2000; Wolpaw et al., 2000). DBIs based on processing field potentials have been used to implement binary spellers and even a web browser (Perelmouter and Birbaumer, 2000).

Other brain signals suitable for DBI control are related to movement or the intent to move. These typically include hand and foot movements, tongue protrusion, and vocalization. These *event-related potentials* have been recorded both from scalp EEGs to implement an asynchronous switch (Birch and Mason, 2000) and from implanted electrodes placed directly on the brain (Levine et al., 2000). Other research has focused on detecting brain signal patterns in imagined movement (Pfurtscheller et al., 2000).

Another aspect of brain signals that can be used for DBI controls is the brain's responses to stimuli. The P300 response, which occurs when a subject is presented with something familiar or surprising, has been used to implement a speller. The device works by highlighting rows and columns of an alphabet grid and averaging the P300 responses to determine which letter the subject is focusing on (Donchin et al., 2000). P300 responses have also been used to enable a subject to interact with a virtual world by concentrating on a virtual object until it is activated (Bayliss and Ballard, 2000).

Another approach to DBI control is to record from individual neural cells. A tiny hollow glass electrode was implanted in the motor cortices of three locked-

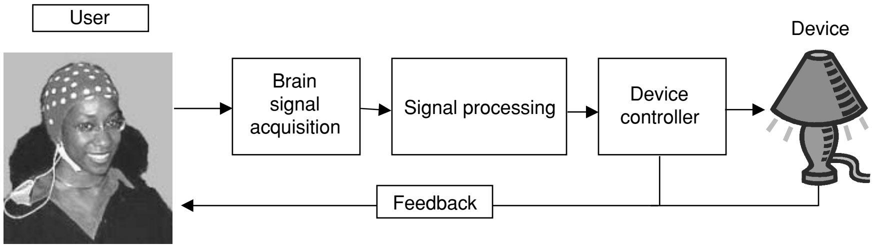

FIGURE 1 General DBI architecture. Source: Mason and Birch, in press.

in subjects enabling neural firings to be captured and recorded (Kennedy et al., 2000). Subjects control this form of DBI by increasing or decreasing the frequency of neural firings, typically by imagining motions of paralyzed limbs. This DBI has been used to control two-dimensional cursor movement, including iconic-communications programs and virtual keyboards (Moore et al., 2001).

SYSTEM ARCHITECTURE

DBI system architectures have many common functional aspects. Figure 1 shows a simplified model of a general DBI system design as proposed by Mason and Birch (in press).

Brain signals are captured from the user by an acquisition method, such as EEG scalp electrodes or implanted electrodes. The signals are then processed by a feature extractor that identifies signal changes that could signify intent. A translator then maps the extracted signals to device controls, which control a device, such as a cursor, a television, or a wheelchair.

APPLICATIONS

As the DBI field matures, considerable interest has been shown in applications of DBI techniques to real-world problems. The principal goal has been to provide a communication channel for people with severe motor disabilities, but other applications may also be possible. Researchers at the Georgia State University (GSU) BrainLab are focusing on applications for DBI technologies in several critical areas:

Restoring lost communication for a locked-in person is a critical and very difficult problem. Much of the work on DBI technology centers around communication, in the form of virtual keyboards and iconic selection systems, such as TalkAssist (Kennedy et al., 2000). Environmental control is also an important quality-of-life issue; environmental controls include turning a TV to a desired channel and turning lights on and off. The Aware Chair project at GSU is working on incorporating communication and environmental controls into a wheelchair equipped with intelligent, context-based communication that can adapt to the people in the room, the time of day, and the activity history of the user. The Aware Chair also presents context-based environmental control options (for example, if the room is getting dark, the chair provides an option for turning on the lights). Currently, the Aware Chair is being adapted for neural control using EEG scalp electrodes and a mobile DBI system.

The Internet has the potential to greatly enhance the lives of locked-in people. Access to the Internet can provide shopping, entertainment, educational, and sometimes even employment opportunities to people with severe disabilities. Efforts are under way to develop paradigms for DBI interaction with web browsers. The University of Tübingen, GSU, and University of California, Berkeley, have all developed browsers (Mankoff et al., 2002).

Another quality-of-life area lost to people with severe disabilities is the possibility of creating art and music. The GSU Neural Art Project is currently experimenting with ways to translate brain signals directly into a musical instrument device interface (MIDI) to produce sounds and visualizations of brain signals to produce graphic art.

A DBI application with significant implications is neural-prostheses or muscle stimulators controlled with brain signals. In effect, a neural prosthesis could reconnect the brain to paralyzed limbs, essentially creating an artificial nervous system. DBI controls could be used to stimulate muscles in paralyzed arms and legs to enable a subject to learn to move them again. Preliminary work on a neurally controlled virtual hand has been reported by Kennedy et al. (2000). DBI control has also been adapted to a hand-grasp neuroprosthesis (Lauer et al., 2000).

Restoring mobility to people with severe disabilities is another area of research. A wheelchair that could be controlled neurally could provide a degree of freedom and greatly improve the quality of life for locked-in people. Researchers are exploring virtual navigation tasks, such as virtual driving and a virtual apartment, as well as maze navigation (Bayliss and Ballard, 2000; Birbaumer et al., 2000).

CONCLUSION

Researchers are just beginning to explore the enormous potential of DBIs. Several areas of study could lead to significant breakthroughs in making DBI systems work. First, a much better understanding of brain signals and patterns is a difficult but critical task to making DBIs feasible. Invasive techniques, such as implanted electrodes, could provide better control through clearer, more distinct signal acquisition. Noninvasive techniques, such as scalp electrodes, can be improved by reducing noise and incorporating sophisticated filters. Although research to date has focused mainly on controlling output from the brain, future efforts will be focused on input channels (Chapin and Nicolelis, 2002). In addition to improvements in obtaining brain signals, much work remains to be done on using them to solve real-world problems.

REFERENCES

Bayliss, J.D., and D.H. Ballard. 2000. Recognizing evoked potentials in a virtual environment. Advances in Neural Information Processing Systems 12: 3-9.

Birbaumer, N., A. Kubler, N. Ghanayim, T. Hinterberger, J. Perelmouter, J. Kaiser, I. Iversen, B. Kotchoubey, N. Neumann, and H. Flor. 2000. The thought translation device (TTD) for completely paralyzed patients. IEEE Transactions on Rehabilitation Engineering 8(2): 190-193.

Birch, G.E., and S.G. Mason. 2000. Brain-computer interface research at the Neil Squire Foundation. IEEE Transactions on Rehabilitation Engineering 8(2): 193-195.

Chapin, J., and M. Nicolelis. 2002. Closed-Loop Brain-Machine Interfaces. Proceedings of Brain-Computer Interfaces for Communication and Control. Rensselaerville, New York: Wadsworth.

Donchin, E., K. Spencer, and R. Wijesinghe. 2000. The mental prosthesis: assessing the speed of a P300-based brain-computer interface. IEEE Transactions on Rehabilitation Engineering 8(2): 174-179.

Kandel, E., J. Schwartz, and T. Jessell. 2000. Principles of Neural Science, 4th ed. New York: McGraw-Hill Health Professions Division.

Kennedy, P.R., R.A.E. Bakay, M.M. Moore, K. Adams, and J. Goldwaithe. 2000. Direct control of a computer from the human central nervous system. IEEE Transactions on Rehabilitation Engineering 8(2): 198-202.

Lauer, R.T., P.H. Peckham, K.L. Kilgore, and W.J. Heetderks. 2000. Applications of cortical signals to neuroprosthetic control: a critical review. IEEE Transactions on Rehabilitation Engineering 8(2): 205-207.

Levine, S.P., J.E. Huggins, S.L. BeMent, R.K. Kushwaha, L.A. Schuh, M.M. Rohde, E.A. Passaro, P.A. Ross, K.V. Elisevish, and B.J. Smith. 2000. A direct-brain interface based on event-related potentials. IEEE Transactions on Rehabilitation Engineering 8(2): 180-185.

Mankoff, J., A. Dey, M. Moore, and U. Batra. 2002. Web Accessibility for Low Bandwidth Input. Pp. 89-96 in Proceedings of ASSETS 2002. Edinburgh: ACM Press.

Mason, S.G., and G.E. Birch. In press. A general framework for brain-computer interface design. IEEE Transactions on Neural Systems and Rehabilitation Technology.

Moore, M., J. Mankoff, E. Mynatt, and P. Kennedy. 2001. Nudge and Shove: Frequency Thresholding for Navigation in Direct Brain-Computer Interfaces. Pp. 361-362 in Proceedings of SIGCHI 2001 Conference on Human Factors in Computing Systems. New York: ACM Press.

Perelmouter, J., and N. Birbaumer. 2000. A binary spelling interface with random errors. IEEE Transactions on Rehabilitation Engineering 8(2): 227-232.

Pfurtscheller, G., C. Neuper, C. Guger, W. Harkam, H. Ramoser, A. Schlögl, B. Obermaier, and M. Pregenzer. 2000. Current trends in Graz brain-computer interface (BCI) research. IEEE Transactions on Rehabilitation Engineering 8(2): 216-218.

Wolpaw, J.R., D. J. McFarland, and T.M. Vaughan. 2000. Brain-computer interface research at the Wadsworth Center. IEEE Transactions on Rehabilitation Engineering 8(2): 222-226.

Wolpaw, J.R., N. Birbaumer, D. McFarland, G. Pfurtscheller, and T.Vaughan. 2002. Brain-computer interfaces for communication and control. Clinical Neurophysiology 113: 767-791.

The Future of Nuclear Energy

Advanced Nuclear Reactor Technologies

John F. Kotek
Argonne National Laboratory-West
Idaho Falls, Idaho

For more than a decade, when energy experts considered which technologies would be used to meet future U.S. energy needs, nuclear power was largely ignored. This has changed in the last five years, as improved performance at existing plants has shown that well run nuclear plants can be a very low-cost source of baseload electrical generation. Increasing concerns about the effects of human activity on climate, coupled with a growing desire to ensure the security of U.S. energy supplies, have led to a renewed interest in nuclear power. Other countries, including Japan, China, South Korea, Russia, and Finland are pressing ahead with plans to add new nuclear generating capacity. However, before new nuclear plants can gain a foothold in the U.S. market, the economics of constructing new plants must be improved. In addition, the standing of nuclear energy in public consciousness would be elevated if new plants offered improved nuclear-waste management strategies and were demonstrably safer than existing plants. Finally, the expansion of nuclear energy to developing countries would be greatly facilitated if stronger intrinsic barriers to proliferation were built into new nuclear energy systems.

HOW NUCLEAR REACTORS WORK

A nuclear reactor produces electricity by harnessing the energy released during the splitting, or fission, of a heavy isotope, such as uranium-235 or plutonium-239. Fission can be induced when the nucleus of one of these isotopes absorbs a free neutron. When the isotope fissions, it generally splits into two smaller isotopes (referred to as fission products) and releases two or three neutrons and about 200 MeV of energy, about 20 million times the energy released in a typical chemical reaction. The released neutrons can go on to fission another uranium or plutonium atom or can be captured in or escape from the reactor core. Power reactors are designed so the fraction of captured or escaped neutrons increases as the core temperature increases; the rate of fission reactions adjusts to maintain a nearly constant temperature, leading to a stable, self-sustaining chain reaction.

The core of a typical water-cooled reactor contains the fuel, usually uranium oxide pellets sealed in zirconium alloy cladding tubes. About 290 of these tubes are contained in a fuel assembly, and a typical core of a light-water reactor contains about 200 of these 3.5 m-long assemblies. The uranium or plutonium in the fuel is allowed to fission, and the energy released during fission is used to heat water. The heated coolant is then used, either directly or after one or more heat-transfer steps, to generate steam to drive a turbine, which generates electricity. A typical fuel assembly contains about 500 kg of uranium; if electricity sells for around 3 cents per kilowatt-hour, about $6 million of electricity would be generated over the life of the fuel assembly.

A SECOND LOOK AT NUCLEAR POWER

Several attractive features of nuclear power have aroused renewed interest in the United States. First, of course, nuclear fuel is very inexpensive compared to coal or natural gas. The nuclear fuel needed to power a 1,000 MWe plant costs about $40 million per year; the fuel for a similar-sized coal plant costs about $110 million and for a natural gas plant about $220 million (Reliant Energy, 2001). The low fuel cost more than offsets the higher operations and maintenance costs of nuclear plants. The average operations, maintenance, and fuel cost in 1999 for the 103 U.S. nuclear power plants was 1.83 cents per kilowatt-hour, lower than for coal plants (2.07 cents) and natural gas-fired plants (3.52 cents) (Utility Data Institute, 2001). Another attractive feature is that nuclear plants do not release air pollutants or carbon dioxide. Today nuclear plants provide 20 percent of U.S. electrical generation without burning fossil fuels or causing air pollution or an increase in greenhouse gases. These and other benefits, such as a small footprint per unit energy and a secure fuel supply, have put nuclear power back on the drawing board. To get beyond the drawing board, however, new nuclear power plants must overcome several hurdles.

The most challenging hurdle is the high capital cost of nuclear power plant construction. According to a 2001 report by the U.S. Department of Energy, overnight capital costs for a project must be contained at $1,500 per kWe or less (DOE, 2001). Capital costs for nuclear power plants completed in the 1980s and 1990s were in many cases several times higher, although a large fraction of the costs was interest that accumulated during construction delays. By comparison, capital costs for combined-cycle natural gas plants are currently about $500 per kWe.

Another hurdle, probably less important to the future deployment of new plants in the United States, but still significant, is the very small, but non-zero potential, for serious accidents. A third hurdle is the need for a repository to store used nuclear fuel. Finally, before nuclear power can be deployed on a wide scale, it may be necessary to reduce the potential for proliferation from civilian nuclear fuel cycles and to find better ways of managing used nuclear fuel.

NEAR-TERM PROSPECTS

No new nuclear power plants have been ordered in the United States since the 1970s, and no new plants have come on line since 1996. Despite this hiatus, work has continued on the development of more economical and safer reactor designs. Recent interest in new plants has been focused on two classes of new designs—water-cooled reactors and gas-cooled modular reactors.

Because of extensive experience in the construction and operation of water-cooled reactors and the relative maturity of the technology, advanced water reactors are, in my view, the most likely to be constructed in the United States in the near term. One promising design is the Westinghouse AP1000, an evolution of the AP600 pressurized light-water reactor plant that received U.S. Nuclear Regulatory Commission (NRC) design certification in 1999. The AP600 is a 600 MWe plant that incorporates several passive safety features (i.e., safety systems that work without requiring operator action), including passive safety injection, passive residual heat removal, and passive containment cooling. These passive systems are designed to improve safety and reduce the need for operator response in the event of an accident.

On April 2, 2002, Westinghouse submitted an application to the NRC for design certification of the AP1000 (Westinghouse hopes design certification can be achieved by the end of 2004). Westinghouse reports that more than 90 percent of the design for the plant has already been completed and that more than 80 percent of the AP600 Safety Analysis Report will remain unchanged for the AP1000. The thermal efficiency of the plant is about 32 percent, similar to existing pressurized-water reactors (Matzie, 1999).

The second class of reactors under consideration for near-term deployment is gas-cooled modular reactors. The two designs of this type that have elicited the most industry interest are the gas-turbine modular helium reactor (GT-MHR) under development by General Atomics and the pebble-bed modular reactor (PBMR), which is being designed by a team that includes British Nuclear Fuels Ltd., the parent company of Westinghouse. Instead of using the steam cycle to generate electricity, these reactors couple the reactor directly to a gas turbine. By using a direct Brayton cycle, efficiencies approaching 50 percent can be achieved. The units are small enough (300-600 MWt) to be mass produced in standardized units, which reduces capital costs while retaining safety characteristics. The small reactor and power-conversion units can be housed below ground, reducing the risk of man-made and natural hazards.

The fuel for a gas reactor uses tiny particles of uranium or plutonium oxide coated with carbon and silicon carbide. The particles create a barrier to the release of fission products and can withstand maximum attainable accident temperatures. The GT-MHR has a three-year operating fuel cycle—half of the fuel in the reactor core is replaced every 18 months while the reactor is shut down. By contrast, the PBMR has continuous refueling with the reactor in operation. Both designs use inert helium gas as a coolant.

LONG-TERM PROSPECTS

The U.S. Department of Energy is leading a 10-country effort to develop the next generation of nuclear energy systems. This program, known as Generation IV, will be guided by a technology roadmap being prepared by representatives of the 10 countries. The roadmap technical teams have evaluated many innovative concepts, including integral pressure-vessel water reactors and liquid metal-cooled fast-spectrum reactors.[1]

In integral pressure-vessel water reactors, the integral vessel houses reactor core and support structures, the core barrel, control-rod guides and drivelines, steam generators, a pressurizer, and reactor coolant pumps. This arrangement eliminates the need for separate steam generators, as well as a separate pressurizer, connecting pipes, and supports. Although the vessel is large (~18 m height and 4.4 m outside diameter), it is well within state-of-the-art fabrication capabilities. A 300-MWt reactor has 21 fuel assemblies inside a 2.6 m core barrel. Each assembly has about 440 pins in a square lattice. Integral-vessel reactors have lower power densities than light-water reactors, which allows for higher safety margins and longer core life, although they also have larger vessels and other structures. Projected system efficiencies are on the order of 36 percent.

The second class of innovative concepts is liquid metal-cooled fast-spectrum reactors ("fast-spectrum" refers to the energy of the neutrons in the reactor core). In a typical reactor, a moderator (usually water, which pulls double-duty as both neutron moderator and reactor coolant) is used to slow down neutrons because slower neutrons are more efficient at causing fission in U-235. In a fast-spectrum reactor, there is no moderator. Instead, it relies on higher energy neutrons, which are less effective at causing uranium to fission but are more effective at causing fission in plutonium and other heavy elements. For this reason, these reactors are not ideal for a uranium-based fuel cycle; but they are quite suitable for use with a fuel cycle based on plutonium and the other heavy

[1] An overview of the roadmap was released in September 2002, and can be found at *http://nuclear.gov*.

elements that accumulate in spent fuel. Therefore, fast reactors could play a vital role in a long-term nuclear energy system that recycles used fuel to minimize long-lived waste.

Most fast-spectrum reactors operated around the world use liquid sodium metal as a coolant. Future fast-spectrum reactors may use lead or a lead-bismuth alloy, or even helium, as a coolant. One of the attractive properties of metals as coolants is that they offer exceptional heat-transfer properties; in addition, some (but not all) metal coolants are much less corrosive than water. However, because sodium is reactive with air and water, fast-spectrum reactors built to date have a secondary sodium system to isolate the sodium coolant in the reactor from the water in the electricity-producing steam system. The need for a secondary system has raised capital costs for fast reactors and has limited thermal efficiencies to the range of 32 to 38 percent. Novel steam-generator designs, direct gas cycles, and different coolants are options that may eliminate the need for this secondary sodium loop and improve the economics of fast reactors (Lake et al., 2002).

The Generation IV group is looking at complete nuclear energy systems, not just reactors, which are only one part of the nuclear fuel cycle (the path of uranium from the mine through fuel fabrication and use and disposal or recycling). Because of concerns about proliferation, recycling of spent nuclear fuel was banned in the United States in April 1977; spent fuel is recycled in France, Japan, the United Kingdom, Russia, and elsewhere. The U.S. strategy for managing used nuclear fuel is to isolate it from the environment in canisters placed in a deep geologic repository; Yucca Mountain in Nevada has just been approved for this purpose. In an attempt to keep options for the long term open, Generation IV is revisiting the issue of recycling used fuel.

Nuclear fission produces a lot of heat, which can be used for more than making electricity. Several nuclear power plants around the world also provide heating for homes or for desalinating water. In the future, nuclear power plants may make a larger contribution toward meeting nonelectrical energy needs by supplying heat for industrial processes and by producing hydrogen. The advantages of hydrogen as an energy carrier have been widely publicized, but we must find ways of producing enough hydrogen to meet our needs. Currently, more than 95 percent of the hydrogen produced for refineries and chemical plants comes from the cracking of natural gas, a process that releases carbon dioxide. In the near term, nuclear-generated electricity could be used to drive electrolyzers that split hydrogen from oxygen, which would release almost no carbon dioxide. In the long term, higher temperature reactors could be used to produce heat to drive thermochemical water-cracking cycles, a process that could be nearly twice as efficient as electrolysis.

CONCLUSION

Existing nuclear power plants are a cost-effective and nonemitting contributor to the world energy system. Advanced nuclear energy systems can make a large contribution to meeting future energy needs, but the economics of these systems must be improved without compromising safety. In the long term, next-generation systems could contribute not only to our electricity needs, but also to our need for clean fuels for transportation and industry. The successful deployment of new nuclear energy systems could be a key part of a sustainable energy system.

REFERENCES

DOE (U.S. Department of Energy). 2001. A Roadmap to Deploy New Nuclear Power Plants in the United States by 2010. Washington, D.C.: U.S. Department of Energy.

Lake, James A., R.G. Bennett, and J.F. Kotek. 2002. Next generation nuclear power. Scientific American 286(1): 72-79.

Matzie, Regis. 1999. Westinghouse's advanced boiling water reactor program. Nuclear Plant Journal Editorial Archive. Available online at: <*http://npj.goinfo.com/NPJMain.nsf*> (October 30, 2001).

Reliant Energy. 2001. Reliant Energy HL&P's Nuclear Plant Has Lowest Fuel Costs of All Power Plants in the U.S. Available online at: <*http//www.reliantenergy.com/news/pressreleases/press_release_225.asp*> (July 25, 2001).

Utility Data Institute. 2001. Nuclear Energy Surpasses Coal-Fired Plants as Leader in Low-Cost Electricity Production. Available online at: <*http://www.nei.org/doc.asp?catnum =4&catid=304*> (January 9, 2001).

Licensing and Building New Nuclear Infrastructure

Peter S. Hastings
Duke Energy
Charlotte, North Carolina

Electricity demand is outpacing supply growth, and experts have calculated that the United States will need new baseload power generation (including nuclear power generation) by 2010 (NEI, 2002a). As part of the Nuclear Power 2010 Initiative, the U.S. Department of Energy (DOE) established the Near-Term Deployment Group (NTDG) to examine prospects for new nuclear plants in the United States in the next decade. According to a recent study by NTDG, a resurgence of the nuclear industry will be influenced by many factors, including economic competitiveness; deregulation of the energy industry; regulatory efficiency; existing infrastructure; the national energy strategy; safety; management of spent fuel; public acceptance; and nonproliferation (DOE, 2001). NTDG also noted that the nuclear industry is experiencing current shortfalls in several important areas (DOE, 2001):

- Qualified and experienced personnel in nuclear energy operations, engineering, radiation protection, and other professional disciplines.
- Qualified suppliers of nuclear equipment and components [including] fabrication capability and capacity for forging large components such as reactor vessels.
- Contractor and architect/engineer organizations with personnel, skills, and experience in nuclear design, engineering, and construction.

Nuclear industry infrastructure can be defined in terms of technologies, facilities, suppliers, and regulatory elements; design and operational engineering and licensing tools; and (perhaps most important) human capital to sustain the industry in the near and long terms. NTDG concluded that, although the industry has adequate industrial and human infrastructure today to build and operate a few new nuclear plants, we cannot be sure that this infrastructure can be expanded quickly enough to achieve the goal of 50 GWe in new nuclear plant installed capacity as laid out in the industry's strategy for the future, *Vision 2020* (NEI, 2002a).

The nuclear industry faces infrastructure challenges, untested implementation of new Nuclear Regulatory Commission (NRC) regulations, and uncertainties about the economic competitiveness of new nuclear plants. The near-term deployment of new nuclear-power generation will be a good litmus test of the viability of a larger expansion in nuclear production. According to NTDG, even though the level of additional capacity in the industry goals in *Vision 2020* is meager in terms of overall U.S. energy needs, achieving the goal presents major challenges (DOE, 2001).

INDUSTRY AND GOVERNMENT INITIATIVES

Various initiatives have been undertaken by the U.S. nuclear industry and DOE to encourage the domestic development of additional nuclear facilities. In 1998, DOE chartered the Nuclear Energy Research Advisory Committee (NERAC) to advise the agency on nuclear research and development (R&D) issues. NERAC released the *Long-Term Nuclear Technology R&D Plan* in June 2000 (NERAC, 2000); NERAC is also responsible for overseeing the development of plans for both NTDG and Generation IV, a project to pursue the development and demonstration of one or more "next-generation" nuclear energy systems that offer advantages in economics, safety, reliability, and sustainability and that could be deployed commercially by 2030 (DOE, 2002). In 1999, DOE initiated the Nuclear Energy Research Initiative (NERI), an R&D program to address long-term issues related to nuclear energy; in 2000, the Nuclear Energy Plant Optimization Program was initiated to focus on the performance of currently operating nuclear plants.

In 2000, NEI formed the industry-wide New Nuclear Power Plant Task Force to identify the market conditions and business structures necessary for the construction of new nuclear power plants in the United States. In April 2001, the task force published the *Integrated Plan for New Nuclear Plants*, which includes a discussion of nuclear infrastructure (NEI, 2001). More recently, NEI announced *Vision 2020*, an initiative with a goal of adding 50,000 MW of new nuclear generating capacity by 2020, along with increases in efficiency power uprates at existing plants equal to an additional 10,000 MW of generating capacity (NEI, 2002b).

DOE funding has recently been allocated to advanced reactor development, specifically for the exploration of government/industry cost sharing for the demonstration of early site permitting as part of new NRC licensing processes (which also include provisions for combined licenses and design certifications) and for national laboratory activities associated with fuel testing, code verification and validation, and materials testing associated with new reactor designs (DOE, 2001).

TECHNOLOGICAL INFRASTRUCTURE

A number of elements are required to support a new or existing nuclear plant. Suppliers and fabricators of nuclear fuel and safety-related components are clearly essential, as are suppliers of balance-of-plant equipment, construction materials, electronics and instrumentation, and countless other components.

The U.S. fuel-cycle industry has undergone significant changes in the past few years. Future fluctuations in uranium prices, the deployment of new enrichment technologies, significant consolidation of fuel-cycle supply companies, and the possible recycling of spent fuel could all affect the supply chain for nuclear fuel (i.e., mining/milling, conversion and enrichment, and fabrication into ceramic fuel pellets). The NTDG study recognized the sensitivity of new reactor deployment to these factors. NTDG solicited designs for nuclear plants that could be deployed by 2010 and attempted to identify generic issues that could impede their deployment. Proposals were received from reactor suppliers identifying eight candidate reactor designs. NTDG evaluated these designs to determine the prospects for deployment of a new nuclear plant in the United States by 2010. Candidate reactor technologies were required to demonstrate how they would operate "within credible fuel-cycle industrial structures" assuming a once-through fuel cycle using low-enriched uranium fuel and to "demonstrate the existence of, or a credible plan for, an industrial infrastructure to supply the fuel being proposed." NTDG's design-specific evaluations concluded that the candidates that would use existing fuel-cycle infrastructure could be built by 2010. However, NTDG also concluded that infrastructure expansion to achieve the industry's goal of 50 GWe of new installed capacity by 2020 was a "generic gap" that warrants government and industry action (DOE, 2001).

Siting for new reactors will be greatly influenced by how efficiently 10 CFR Part 52 (the NRC regulation for early site permits, standard design certifications, and combined nuclear plant licenses) can be implemented. NTDG concluded that the federal commitment to cost sharing via government/industry partnerships should include a demonstration of the NRC's early site permit process for a range of likely scenarios. Recently a partnership to evaluate sites for new nuclear plants was announced, and DOE selected three utilities to participate in joint government/industry projects to pursue NRC approval for sites for new nuclear power plants. These projects, the first major elements of DOE's Nuclear Power 2010 Initiative, are intended to "remove one more barrier to seeing the nuclear option fully revived" in the United States. All three companies intend to seek early site permit approvals that would enable them to locate new, advanced-technology nuclear plants at sites owned by the utilities that currently host commercial nuclear power plants (i.e., Dominion Energy's North Anna site in Virginia, Entergy's Grand Gulf site in Mississippi, and Exelon's Clinton site in Illinois). According to a press release on June 24, 2002, DOE expects applications to be submitted by late 2003.

Design concepts for the waste-management component of fuel-cycle infrastructure were also evaluated by NTDG. Each design concept addressed the on-site storage of spent nuclear fuel; but the current lack of a national system for high-level waste disposal is a programmatic gap common to all new technologies. The Nuclear Waste Policy Act (first passed in 1983, amended many times, and still the subject of heated debate) provides for the development of the Yucca Mountain site in Nevada as a mined, geologic repository for high-level waste and the development of a transportation system linking U.S. nuclear power plants, an interim storage facility, and the permanent repository (NEI, 2002c).

The recycling of spent fuel to recover fissile material and long-lived heavy elements would reduce the heat generation and volume of final waste products. When most existing U.S. nuclear plants were built, the industry—encouraged by the federal government—planned to recycle used nuclear fuel by recovering plutonium. In 1979, however, the United States deferred the reprocessing of all commercial used nuclear fuel because of concerns about the possible proliferation of nuclear weapons. Thus, the industry was forced to adopt a once-through, single-use fuel cycle. Reprocessing and recycling are not currently cost effective in the United States, and most of the discussion about recycling is now focused on increasing the capacity of waste repositories and the transmutation of actinides.

The domestic supply of certain reactor components is another long-term concern. Over time, U.S. capacity for fabricating large components with long lead times (e.g., reactor vessels and steam generators) has diminished. NTDG concluded that in most cases, other countries have the necessary capacity and could support the early expansion of nuclear plant construction in the United States, but domestic capabilities will also have to be reestablished (DOE, 2001). General Electric, for example, has indicated that many of the components and much of the hardware (including pumps, heat exchangers, nuclear fuel, control rods, and some internal reactor components, as well as most balance-of-plant equipment) for its advanced boiling water reactor (ABWR) can be produced in the United States. However, the reactor pressure vessel and the large internal components (the same components used for Japanese and Taiwanese ABWRs) will have to be fabricated by foreign suppliers that have maintained their capacity and expertise for fabricating and machining these large components. Foreign suppliers can and do meet U.S. codes and regulations; most foreign countries follow identical or similar codes, such as ASME Section III, IX, and XI, as well as U.S. NRC-imposed regulations and guidelines. NTDG concluded that this area of infrastructure will have to be reconstituted in the United States for economic reasons and that, for some components (particularly reactor vessels), the existing worldwide capacity may not be adequate to support a large expansion of reactor capacity.

REGULATORY INFRASTRUCTURE

NRC is perhaps best known for regulating reactor construction and operation. The agency also regulates nuclear fuel-cycle operations, including the mining, milling, conversion, enrichment, and fabrication of fuel and waste-management facilities. A number of proposed licensing actions and regulatory initiatives are currently under way at the NRC. The most important in terms of near-term continuity in the nuclear industry and adequate energy supply is the renewal of reactor licenses. NRC regulations limit commercial reactor licenses to an initial 40 years (based on economic and antitrust considerations) but also allow licenses to be renewed (NRC, 2002). The license renewal process involves confirmation that structures, systems, and components can continue to perform safely beyond the original term of the operating license.

License renewal is important for maintaining a significant portion of existing energy production. Currently licensed nuclear reactors generate approximately 20 percent of the electric power produced in the United States. The first plant will reach the end of its original license period in 2006; approximately 10 percent will reach 40 years by the end of 2010; and more than 40 percent will reach 40 years by 2015. As of June 2002, 10 license renewal applications had been granted, and 14 applications were under NRC review; 23 more applications are expected by 2005. License renewal also provides an important collateral benefit by keeping the industry and workforce energized and engaged, thus helping to preserve the institutional and human capital that will be necessary for the near-term deployment of new reactors.

Other high-profile licensing actions related to nuclear industry infrastructure include: an application for a license for a spent-fuel storage facility currently before the NRC and in hearings submitted by Private Fuel Storage (a group of eight electric utility companies in partnership with the Skull Valley Band of Goshute Indians); two pending applications (one from the United States Enrichment Corporation and one from Urenco) for deploying new gas-centrifuge enrichment technologies in the United States; and an application for the construction and subsequent operation of a facility to convert surplus weapons material into commercial fuel as part of an agreement between the United States and Russia to dispose of more than 60 metric tons of surplus plutonium.

A key aspect of regulatory development for near-term deployment is the efficient implementation of 10 CFR Part 52. Created in 1989, this regulation established three new licensing processes for future plants: early site permitting (i.e., NRC approval of a site before a decision has been made to build a plant); design certification (i.e., NRC approval of a standard design); and combined licenses (i.e., a combined construction permit and operating license). These processes could provide a dramatic improvement over the two-step process used for existing U.S. plants. NRC has stated that an application for a combined license that references an early site permit and a certified reactor design could result in an operating license being granted in as little as *one year* (Jeffrey S. Merrifield, NRC commissioner, remarks to American Nuclear Society Conference, June 7, 1994). In actuality, however, only the design certification process has been demonstrated thus far, and it has taken as long as 10 years.

Not surprisingly, NTDG has called the efficient implementation of 10 CFR Part 52 a high priority for short-term deployment and a matter that requires the attention of industry and government. NTDG recommends that four actions be taken: the expediting of the design certification process; the demonstration of the early site permit and combined license processes; the development of generic guidelines to ensure the efficient, safety-focused implementation of key Part 52 processes; and a demonstration of progress toward a new risk-informed, performance-based regulatory framework (DOE, 2001).

DOE's recently announced plans to work in partnership with industry to evaluate sites for new nuclear plants will test the early site permitting component of this regulation. DOE proposes that near-term investments be made as part of the Nuclear Power 2010 Initiative, in part to "demonstrat[e] this key NRC licensing process" (DOE press release, June 24, 2002).

Another evolving area of regulatory infrastructure is NRC's effort to "risk-inform" their regulatory processes. The intent is to improve the regulatory process by incorporating risk insights into regulatory decisions, thereby conserving agency resources and reducing unnecessary burdens on licensees. The risk-informed approach combines risk insights and traditional considerations to focus regulatory and licensee attention on the design and operational issues that are most important to health and safety (NRC, 2001).

In the context of the efficient implementation of 10 CFR Part 52, NTDG calls for a new regulatory framework that is fully risk-informed and performance-based and "go[es] beyond the ongoing efforts to risk-inform 10 CFR Part 50 for current plants" to improve the protection of public health and safety, eliminate regulatory burdens that do not contribute to safety, and "increase the confidence of prospective applicants in the regulatory environment for new plants and encourage business decisions to proceed with new nuclear projects" (DOE, 2001).

ENGINEERING AND LICENSING TOOLS

The skills necessary for the development of new infrastructure are largely the same as those necessary for the maintenance of existing facilities. In fact, efforts to continue improving existing facilities not only provide a performance benefit to facility stakeholders, they also help ensure that human and technological resources will be available for the development of new facilities. Thus, ongoing improvements to existing facilities are extremely important. From 1990 to 2000, for instance, improved efficiency at U.S. nuclear power plants provided the production equivalent of constructing 22 new 1,000-MW power plants—enough power to provide 22 percent of new electricity demand during that decade (NEI, 2002d). So, not only are there ample economic reasons for continuing to pursue improvements, there are also substantial collateral benefits, such as ensuring the ongoing development of technological tools and the engagement of human capital in nuclear technologies.

COMPUTER TECHNOLOGY

The most dramatic contributor to changes in design tools in the last 25 years has been computer technology. With today's computational speeds, optimization and simplification of new plant design and construction seems to be limited only by the imagination. One has only to compare modern development tools to historical methods to appreciate the change.

Core physics simulations provide a good example. In the 1970s, a typical simulation used a lattice-cell code that modeled a single fuel pin (or rod) surrounded by an infinite array of homogenized media created to look like the adjacent pins. Approximations were used to model assembly-averaged thermal-hydraulic effects and axial representations, and in-core fuel management was performed by trial-and-error shuffle schemes, using manual iterations until cycle length and power peaking requirements were met.

Modern design software typically uses a two-dimensional code that models a full fuel assembly and uses advanced ray tracing, collision probability, or Monte Carlo techniques. The core simulator is a three-dimensional advanced nodal code with pin reconstruction techniques and explicit thermal-hydraulic modeling and is capable of three-dimensional space-time calculations. Core-loading pattern development has been automated, using advanced nodal codes coupled with simulated annealing techniques to develop core-loading patterns that meet predetermined limits on pin peaking, cycle length, and other attributes to minimize fuel cost within applicable core physics limits.

Another dramatic example of the benefits of increased computer power is the advent of computer-aided design (CAD) and engineering, which not only have enabled the development of increasingly complex and sophisticated civil/structural models, but have also significantly eased the burden of physical-interference modeling and facility-configuration control. Designs in the 1970s were modeled primarily on paper and required hundreds of individual drawings. Plastic and wooden scale models were often painstakingly fabricated and maintained (see Figure 1) to help identify costly interferences between electrical, mechanical, and structural components that could be missed in two-dimensional drawings and to assist with the visualization of designs and changes to those designs.

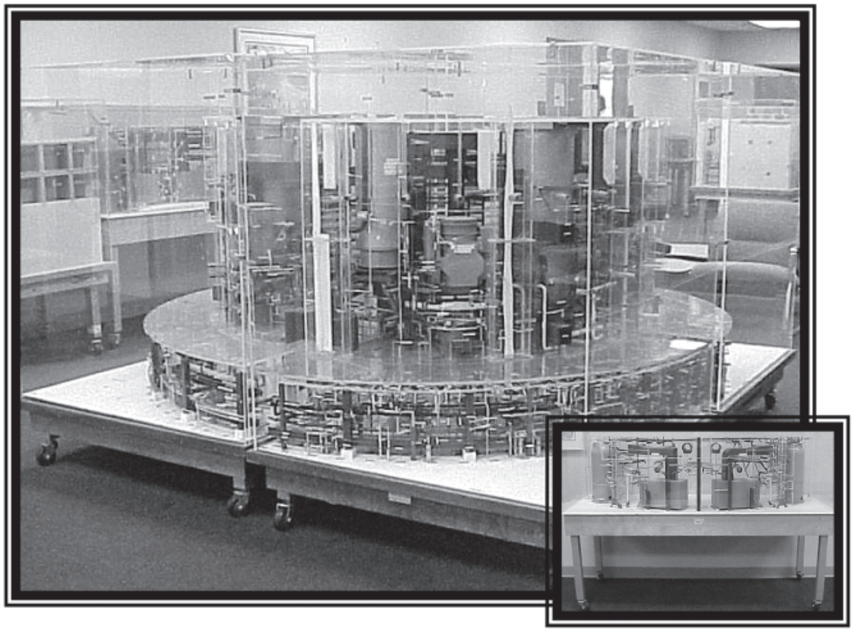

FIGURE 1 Scale model for the McGuire Nuclear Station Reactor fabricated to help identify interferences. Source: Photo courtesy of Duke Energy.

In the last decade or so, the development of more and more sophisticated models (see Figure 2) as the basis of design has improved the coordination of drawings, accelerated the communication of design alternatives, and significantly reduced the requirements for field reworking (Bernstein, 2001).

Using a central data representation results in improved integration and coordination among various design documents and between design disciplines. In addition, procurement, construction, and subsequent ongoing configuration-management of the facility—a key aspect of the design, licensing, operation, and maintenance of a nuclear facility—have been greatly facilitated (Bernstein, 2001).

HUMAN INFRASTRUCTURE

One of the most important aspects of nuclear infrastructure, particularly for a domestic resurgence of the industry, is the development of human resources. The nuclear industry faces the dual challenge of an aging workforce and a growing gap between its employment needs and the number of graduating students. Replacement of the aging workforce is essential for both existing plants and new facilities.

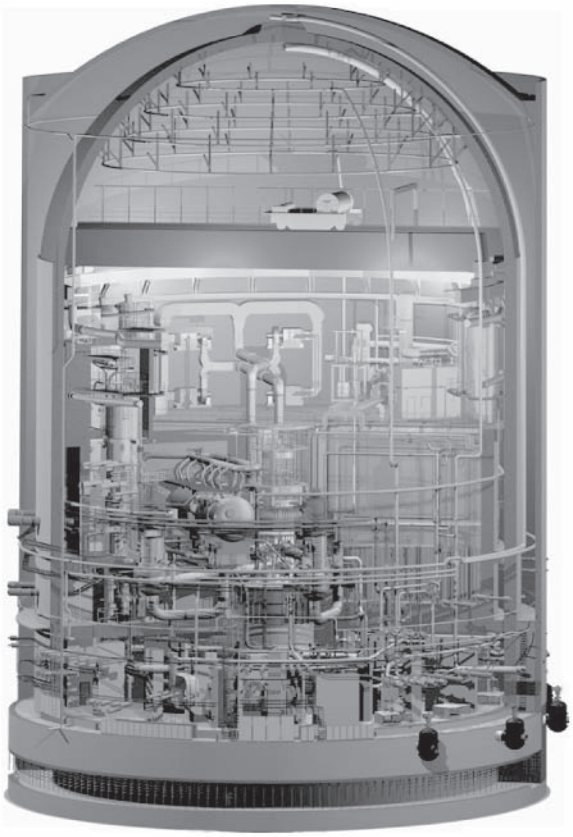

FIGURE 2 Model of 1,000-MW reactor (Lianyungang Unit 1), Elias Mayer, Fortum Engineering, Ltd., Vantaa, Finland. Source: Graphic courtesy of Intergraph Process, Power & Offshore.

NTDG cited key initiatives dealing with human resources at DOE, the Nuclear Energy Institute (NEI), and the American Nuclear Society and recommended that these initiatives be maintained and strengthened (DOE, 2001). On June 10, 2002, DOE announced the establishment of a new program, Innovations in Nuclear Infrastructure and Education, that offers several million dollars in awards to university consortia to encourage investments in programs on research reactors and nuclear engineering and in strategic partnerships with national laboratories and industry. At the same time, DOE announced that it would award more than 100 scholarships, fellowships, and grants to nuclear science/engineering institutions and students.

Other public and private efforts to bolster the educational infrastructure for the nuclear industry are under way throughout the United States. As a result, almost every university and national laboratory associated with nuclear science and engineering now offers scholarships and fellowship programs.

CONCLUSION

The potential for a resurgence of the nuclear industry in the United States is a function of many factors. New technologies continue to be developed, but key areas of infrastructure to support expansion must be maintained and (in many cases) expanded for a future nuclear option to be sustainable. Although significant challenges lie ahead, a number of government, industry, and joint public/ private efforts are under way to facilitate this expansion. The reestablishment of the industrial infrastructure that supplies materials and components, the continued improvement and demonstration of effective regulatory processes, and the development of essential human resources are all critical factors to the future of the nuclear industry in the United States.

ACKNOWLEDGMENTS

The author gratefully acknowledges the support and assistance of Dr. Per Peterson (University of California, Berkeley), Richard Clark (Duke Energy), and the Process, Power, and Offshore Division of Intergraph Corporation.

REFERENCES

Bernstein, P. 2001. 2D to 3D Challenge: Autodesk on Architectural Desktop. London: CADDesk.

DOE (U.S. Department of Energy). 2001. A Roadmap to Deploy New Nuclear Power Plants in the United States by 2010. Washington, D.C.: U.S. Department of Energy.

DOE. 2002. Generation IV website. Available online at: <*http://gen-iv.ne.doe.gov*>.

NEI (Nuclear Energy Institute). 2001. Integrated Plan for New Nuclear Plants. Washington, D.C.: Nuclear Energy Institute.

NEI. 2002a. Vision 2020 Booklet. Available online at: <*www.nei.org/documents/Vision2020 Booklet.pdf*>.

NEI. 2002b. Industry Projects to Build New Nuclear Plants. Available online at: <*www.nei.org*>.

NEI. 2002c. National Used Nuclear Fuel Management Program. Available online at: <*www.nei.org*>.

NEI. 2002d. Plant Improvement Programs. Available online at: <*www.nei.org*>.

NERAC (Nuclear Energy Research Advisory Committee). 2000. Long-Term Nuclear Technology R&D Plan. Washington, D.C.: Nuclear Energy Research Advisory Committee.

NRC (Nuclear Regulatory Commission). 2001. Risk-Informed Regulation Implementation Plan. Washington, D.C.: Nuclear Regulatory Commission.

NRC. 2002. Reactor License Renewal Overview. Available online at: <*www.nrc.gov/reactors/operating/licensing/renewal/overview.html*>.

Sustainable Energy from Nuclear Fission Power

Marvin L. Adams
Department of Nuclear Engineering
Texas A&M University
College Station, Texas

Increases in world population and per-capita energy demand in the next few centuries are expected to cause a substantial rise in world energy use. The World Energy Council predicts a doubling of consumption to 800 EJ/yr (EJ = exaJoule = 10^{18} J) by 2050 (WEC, 2002). For energy production to be sustainable, input requirements (including construction materials) must be available at reasonable financial and environmental cost, and the waste stream must have acceptably low economic and environmental impacts. No production option has yet demonstrated the ability to meet a substantial fraction of projected demand sustainably: every option needs further research and development. In this paper, I summarize options for sustainable energy production, discuss the need for a significant contribution from nuclear fission and its potential for providing such a contribution, and identify some challenges that must be met to achieve that potential.

OPTIONS

In the following discussion of the relative merits of energy technologies hundreds of years into the future, we draw seemingly reasonable conclusions based on some fundamental truths. However, the possibility of unforeseen technological developments over such a long period of time introduces considerable uncertainty. With that caveat in mind, let us boldly proceed.

The vast majority of the world's energy in the coming centuries will come from a few sources: fossil fuels, the sun, biomass, wind, geothermal sources, nuclear fission, and (potentially) nuclear fusion. Table 1 provides an overview of the general suitability of each of these sources for sustainable energy produc-

TABLE 1 Comparison of Energy Sources

Source	EJ[a]/yr today	Reserves[a]	Use of other resources	Waste	Comments
Oil / natural gas liquids[b]	140	> 6 ZJ		CO$_2$ and other emissions	Not considered sustainable
Natural gas	85	> 5 ZJ		CO$_2$ and other emissions	Not considered sustainable
Coal	90	30 ZJ	Land = moderate (mines) Construction = 0.7[c]	CO$_2$ and other emissions	Reserves greater than oil or gas, but much less than fission
Biomass	55	Up to 270 EJ/yr sustainable	Land = high	Particulates and other emissions	R&D may help somewhat
Nuclear fission	28	> 600 ZJ	Land = low Construction = 0.7	Long life Repository space Proliferation	Small waste volume Space is political issue, not technical
Hydroelectric power	9	30 EJ/yr sustainable	Land = high (lakes)		Inherently small role
Solar energy	Small		Land = high Construction = 30–80		R&D may help somewhat
Wind	Small		Land = high Construction = 5–15		Inherently diffuse source
Geothermal energy	Small	<100 EJ/yr sustainable			
Nuclear fusion	0	Enormous			Not yet demonstrated

[a] EJ = exaJoule = 10^{18} J; ZJ = zettaJoule = 10^{21} J.

[b] This does not include larger known reserves of oil shale, which is difficult and not cost effective to use with present methods.

[c] Construction number = approximate number of operation-months required to replace the energy used in construction; this is one measure of the intensity of use of other resources. Data from AWEA, 1998.

tion. Because the anticipated demand is high and because different technologies are better for different applications, it is likely that all of these sources will be tapped.

Economically recoverable oil and natural gas will probably be depleted within a century or two, and both have attractive applications besides energy production. Thus, I will not consider them as sustainable energy sources. The burning of coal produces a great deal of carbon dioxide (CO_2), a greenhouse gas. For coal to be a sustainable long-term energy option, either we must find a way to economically sequester the CO_2 (300 kg/s from each 1-GW_{el} power plant) or the world must decide that the addition of vast quantities of CO_2 to the atmosphere is environmentally acceptable. The burning of biomass also releases CO_2, but the same CO_2 is then captured by the growth of new matter. Thus, biomass does not add to atmospheric CO_2 levels. The greatest drawbacks of biomass are its large land use (according to the World Energy Council estimate, a sustainable 270 EJ/yr would require a crop area larger than the continental United States) and its high level of particulate and other emissions. Hydroelectric power is renewable and emission-free but limited to 30 EJ/yr, at most, by the availability of sites for dams. The sun seems to offer an almost limitless emission-free source of energy, and the role of solar power in the future energy supply will surely be significant, especially for small-scale local applications. However, solar energy is a diffuse source that requires large land areas, and its use of engineering materials is high per unit of energy produced. Technological advances will help, but the low energy density of solar radiation is a fundamental limitation that will be difficult to overcome for large-scale generation. Wind power is less material intensive than solar energy and will play an increasingly important role, but wind is also limited by its low energy density and relatively high land use. Other renewables, such as geothermal energy, have similar limitations. The technology for large-scale generation from nuclear fusion appears to be decades away, and its costs are uncertain. Although fusion has enormous potential, it is not yet clear how its advantages and disadvantages will compare to those of other options, including nuclear fission.

Each of these options has drawbacks, and each has positive features. Most of the drawbacks of any option can be overcome if we are willing to pay enough (in dollars, land, materials, etc.). It is reasonable to conclude that we should continue to develop all options and that all of them will be needed to some extent to meet the energy demand in the coming centuries.

SUSTAINABLE FISSION ENERGY

Potential

The known economically recoverable 3.3 million metric tons of uranium (WEC, 2001) and 4 to 6 million metric tons of thorium (UNDP, 2000) could produce 250 ZJ (zettaJoule) and 350 to 500 ZJ, respectively, if used to their full potential. Thus, more than 600 ZJ of potential nuclear fission energy—1,500 times the current total worldwide annual energy consumption—is readily available. Much more easily recoverable thorium will surely be found if a demand develops (Rubbia et al., 1995). An additional 7 million metric tons of uranium is estimated but not yet proven to be economically recoverable (WEC, 2001). Fission power uses little land and requires modest construction inputs (mainly concrete and steel) per unit of energy produced—lower than the construction inputs for wind and solar energy by factors of 10 and 100, respectively (AWEA, 1998). Thus, as far as inputs are concerned, fission power has the potential to provide a large fraction of the world's energy for many, many centuries. However, tapping the full potential energy of uranium and thorium resources will require changes from current fission-energy practice, including the use of "high-conversion" reactors and the recycling of fissionable isotopes.

The output from fission power includes modest amounts of chemical and low-level radioactive wastes, which are relatively easy to handle, as well as used ("spent") fuel, which is the main disposition challenge. Spent fuel from today's power reactors contains approximately 5 percent fission products (atoms produced by splitting another atom or by radioactive decay of another fission product), 2 percent "fissile" material (including ^{235}U, ^{239}Pu, and ^{241}Pu), and 1 percent other actinides (including ^{238}Pu and ^{241}Am), with ^{238}U comprising most of the remaining mass. ("Fissile" means the atom is likely to fission after absorbing any neutron, including a low-energy neutron.) Note that fission power produces a very small volume of spent fuel. With current technology, six years of operation of a 1-GW$_{el}$ plant yields spent fuel that could fit inside a 4-meter cube, and the vast majority of this material is recyclable. If we recycle, which is essential for the substantial use of uranium and thorium resources, then much less material will require disposal, and most of it will have a much shorter half-life.

Tapping the Energy

In the current "once-through" fission fuel cycle in the United States, uranium fuel that is enriched to 3 to 5 percent in the ^{235}U isotope remains in a power reactor for four to six years, after which it is treated as waste. In the enrichment process, each kilogram (kg) of natural uranium (which is 0.7 percent ^{235}U) is typically converted to 0.85 kg of "depleted" uranium (0.2 percent ^{235}U) and 0.15 kg of low-enriched uranium (3.5 percent ^{235}U). Thus, only 15 percent of the mined uranium reaches a reactor. During its four to six years in the reactor, approximately 5 percent of the atoms in the uranium fuel fission. Thus, current practice releases only 0.75 percent (5 percent of 15 percent) of the potential energy of the mined uranium; a great deal of the energy-storing material is treated as waste.

Essentially all natural thorium is ^{232}Th, and 99.3 percent of natural uranium is ^{238}U. These isotopes are not fissile—they do not fission when they absorb low-energy neutrons—and neither can fuel a reactor by itself. Both are fissionable (absorption of MeV-range neutron can cause fission) and fertile (capture of a neutron leads to a new fissile atom). To unlock the energy from uranium and thorium resources, the abundant fertile isotopes (^{238}U and ^{232}Th) must be converted into fissile isotopes (^{239}Pu and ^{233}U, respectively). This is simple: if either ^{238}U or ^{232}Th captures a neutron, it quickly converts via beta decay:

$$^{238}_{92}\mathrm{U} + ^{1}_{0}\mathrm{n} \rightarrow ^{239}_{92}\mathrm{U} \xrightarrow[23\,\mathrm{min}]{\beta^-} ^{239}_{93}\mathrm{Np} \xrightarrow[2.3\,\mathrm{days}]{\beta^-} ^{239}_{94}\mathrm{Pu} \tag{1}$$

$$^{232}_{90}\mathrm{Th} + ^{1}_{0}\mathrm{n} \rightarrow ^{233}_{90}\mathrm{Th} \xrightarrow[22\,\mathrm{min}]{\beta^-} ^{233}_{91}\mathrm{Pa} \xrightarrow[27\,\mathrm{days}]{\beta^-} ^{233}_{92}\mathrm{U} \tag{2}$$

The conversion ratio of a reactor is the production rate of fissile atoms (by conversion of fertile atoms) divided by the consumption rate of fissile atoms (by fission or other neutron-induced destruction). A reactor with a conversion ratio greater than unity is sometimes called a breeder reactor. Unlocking all of the potential energy in ^{238}U and ^{232}Th will require high-conversion reactors.

High-conversion reactors have been demonstrated for both ^{239}Pu/^{238}U and ^{233}U/^{238}Th cycles; however, this technology is not as mature as the technology used in current commercial reactors. There are opportunities for innovation in many aspects of the design of high-conversion reactors, including materials (for higher-temperature operation and improved resistance to corrosion and radiation damage), fuel form (e.g., oxide, metal, molten salt), and coolant (e.g., liquid metals, helium). Innovations will undoubtedly improve the technology, but even today it is clear that high-conversion reactors can be designed and operated for sustainable fission power.

To take advantage of high-conversion reactors that convert fertile material to fissile material, we must recycle the new fissile material into new fuel. In fact, to maximize the energy and minimize the waste from uranium and thorium, all actinides (not just fissile isotopes) should be recycled so that as many heavy atoms as possible can fission. This will require reprocessing spent fuel, separating out the desired elements, and using reactors that can cause most of the actinides to fission. Reprocessing is a commercial practice in England and France, as is highly efficient (> 99.8 percent) extraction of uranium and plutonium. Efficient extraction of other actinides has been demonstrated on a laboratory scale, and research and development are under way to improve these technologies and move them to commercial scale.

Because of the mix of fissile and fissionable isotopes that develops after several recycling steps, fuel from multiple recycling steps is best suited for fast-spectrum reactors. In a fast-spectrum reactor, neutrons retain relatively high energy from birth (via fission) to death (via absorption or escape). Fast-spectrum reactors can cause nonfissile actinides to fission, whereas other reactors have trouble doing so. High-conversion reactors can be fast-spectrum reactors, and thus could be both creators of fissile material and burners of nonfissile actinides.

Handling the Waste

If a substantial majority of the actinides are recycled and made to fission in fast-spectrum reactors, the remaining waste will be mainly fission products. Most fission products decay with half-lives of decades or shorter, and after 300 years the total fission-product inventory decays to a radiotoxicity level lower than that of the original ore (OECD, 1999). Fission products can be immobilized in glass; this is done commercially in England and France. The waste-containing borosilicate glass is so insoluble that if it were immersed in water, only 0.1 percent would dissolve in 10,000 years. Thus, it is not difficult to immobilize fission products for several hundred years, after which they are no more harmful than the original ore that nature placed underground. The waste-containing glass can be isolated, for example, in stable underground geologic formations.

If we wanted to reduce the radiotoxicity of fission products even further, we would need to address the fission products with long half-lives. The fission products that contribute most to long-term radiotoxicity are ^{99}Tc and ^{129}I, which have half-lives of 2.1×10^5 and 1.6×10^7 years, respectively. These isotopes could conceivably be separated from other fission products and transmuted (by absorption of neutrons from accelerators, reactors, or accelerator-driven subcritical assemblies) into shorter-lived isotopes (NRC, 1996; OECD, 2002). A recent study concludes that both of these isotopes could be transmuted so that their products would decay with 51-year half-lives (OECD, 2002). This would eliminate them as potential long-term hazards.

No separation technology is 100 percent efficient; thus, even with actinide recycling, some fraction of the actinides will remain with the fission products for disposal. The efficiency of the separations will determine the long-term (> 1,000 years) radiotoxicity of the waste from fission power. For example, if 99.9 percent of the plutonium and 99 percent of the americium (Am), curium (Cm), and neptunium (Np) were separated, with all other actinides remaining with the waste, then it would take 10,000 years for radiotoxicity to decay to ore levels—a substantial increase over the 300 years required if 100 percent of all actinides were separated out (OECD, 1999). Increasing the separation efficiency to 99.9 percent of Am, Cm, and Np would reduce this to less than 1,000 years. Thus, the efficiency of separation of actinides has a significant impact on the waste-isolation time and thus on the difficulty of the waste-isolation problem. Economical separation of > 99.8 percent of plutonium has been demonstrated on a commercial scale (Cogema's La Hague plant), and 99.9 percent separation of Am has been demonstrated on a laboratory scale. Research and development continues to improve the technology for actinide separation; thus, it seems reasonable to expect that highly efficient (≈99.9 percent) separation of key actinides will be economically achievable eventually. This raises the possibility of a relatively short waste-isolation time (a few hundred years), which would greatly simplify waste disposal.

The ^{233}U/^{232}Th cycle creates fewer transuranic atoms per unit of energy produced and, thus, in some sense creates less of a long-term waste problem (Rubbia et al., 1995). However, it also makes recycling somewhat more difficult, largely because the daughters of the recycled ^{233}U are more highly radioactive than ^{239}Pu and its daughters, which means hands-on operations must be replaced by remote-controlled operations.

In summary, if a substantial majority of actinides are recycled (which will be necessary for us to tap the majority of the potential energy of uranium and thorium resources), then the waste stream from fission power will consist of a very small volume of fission products along with a small fraction of lost actinides. This waste can be readily immobilized in an insoluble material, such as borosilicate glass. The efficiency of actinide-separation technology will determine whether this waste inventory will decay to ore-level radiotoxicity in hundreds of years or thousands of years. Thus, improvements in separation technology may have a significant impact on the waste-isolation time and on public acceptance of fission power.

One disadvantage of actinide recycling is that recycled fuel costs approximately 10 to 20 percent more than fresh fuel. Another is that recycling technology could conceivably enable countries or groups to develop nuclear weapons—the proliferation issue.

Proliferation

Nuclear weapons use highly concentrated fissile material, such as uranium that is highly enriched in ^{235}U or plutonium that is mostly ^{239}Pu. Highly enriched ^{235}U can be obtained by the same process that produces low-enriched ^{235}U for reactor fuel—the process is simply carried farther. (This is the process that North Korea recently acknowledged using in their weapons program.) ^{239}Pu is generated in every reactor that contains ^{238}U (see Eq. (1)); however, if fuel stays in a reactor for more than a few months, significant quantities of other plutonium isotopes are also created, which makes the plutonium more difficult to use for weapon design and fabrication. Nevertheless, a National Research Council report has concluded that a credible weapon could be made from plutonium of almost any isotopic composition (NRC, 1995).

The ^{232}Th/^{233}U cycle may be more proliferation-resistant than the ^{238}U/^{239}Pu cycle, because ^{233}U daughter products generate heat and radiation that could make it difficult to design, build, and maintain weapons. Nevertheless, it is technically possible to create weapons from separated ^{233}U.

Every country that has developed nuclear weapons has obtained its concentrated fissile material from dedicated military programs, not by co-opting power-reactor technology. Nevertheless, there is a concern that if recycling technology is developed and widely used in the power industry, it could be used by some nations or groups to produce plutonium for a weapons program. A challenge, therefore, is to develop an economical, practical, proliferation-resistant fission-fuel cycle. This is one goal of the Generation-IV Reactor Development Program (Kotek, 2003). One example of a proliferation-resistant technology that includes recycling of valuable actinides is the integral fast-reactor concept (Till et al., 1997). With this technology, Pu is never separated from U during reprocessing, and thus no weapons-usable material ever exists at any stage of the process.

NEAR-TERM ISSUES

Achieving sustainable fission energy will require changes in current practices, especially in the United States. Current U.S. policy is for all spent fuel to be shipped to a geologic repository (recently identified as Yucca Mountain, Nevada) with no reprocessing. This is a once-through, or "open," fuel cycle. There are several adverse consequences of not reprocessing spent fuel:

1. More than 99 percent of the potential energy in the uranium is lost.
2. Some actinides have long half-lives (24,000 years for ^{239}Pu), which leads to stringent long-term requirements for disposal technologies.
3. Some actinides generate heat, which limits repository capacity.

The first consequence is significant for the very long term. The second affects public acceptance of nuclear power and public acceptance of any given repository site (people wonder how anyone can know that the material will remain sufficiently isolated for tens of thousands of years) and poses significant technical challenges. The third affects the capacity of a given repository. The latter two effects are very important for the near term (i.e., the next few decades).

From a technical point of view, repository space is not an issue. There appear to be more than enough suitable stable geologic formations in the world to handle waste from millennia of fission reactors, especially if fissionable materials (actinides) are recycled. However, because of the political picture today, repository space is a precious commodity. After 20 years of study, more than $4 billion of expenditures, and several political battles, the Yucca Mountain site has very recently been selected by the U.S. Department of Energy and Congress as the repository location for which the DOE will attempt to obtain a license. If licensing proceeds as quickly as possible, the first spent fuel will not be delivered until 2010 or later. Given the current once-through, no-reprocessing fuel strategy, the current U.S. fleet of reactors will produce enough spent fuel by 2040 to fill the Yucca Mountain repository to its estimated technical capacity

(DOE, 2002). The addition of new reactors would of course hasten this date. Clearly, in the near term in the United States, repository capacity must be increased for fission power to achieve its potential. It seems equally clear that finding a second repository site would be challenging, especially politically.

The capacity of Yucca Mountain is limited by the thermal load it can accommodate (radioactive decay releases energy that heats the material). If the waste produced lower W/kg, more mass could be accommodated (DOE, 2002). With the current once-through fuel strategy, the main thermal load after a few decades will come from the decay of the isotopes ^{238}Pu and ^{241}Am, with half-lives of 88 and 432 years, respectively. If plutonium and americium were removed from the spent fuel, the heat load would be dominated by isotopes with 30-year half-lives; thus, the thermal load of a given mass of spent fuel would be halved every 30 years (P.F. Peterson, University of California-Berkeley, personal communication, June 2002). In other words, half of the repository capacity would effectively be regenerated every 30 years. It is easy to imagine that this might be more achievable politically than obtaining approval for another repository site.

SUMMARY AND CONCLUSIONS

Worldwide energy use is likely to increase in the foreseeable future, and sustainable energy sources are not abundant. It seems likely that many different sources will be tapped to meet energy needs. Nuclear fission could potentially provide a significant fraction of the world's energy for millennia: its inputs (fuel and construction materials) are readily available and its waste stream (fission products and lost actinides) is very small and not technically difficult to handle. Realizing the potential of fission energy will require high-conversion reactors and the recycling of fissionable atoms, which in turn will require that some technical and political challenges be met.

In the short term, especially in the United States, waste-repository capacity is a significant issue. The long-term capacity of the Yucca Mountain repository could be increased significantly by separating plutonium and americium from spent reactor fuel.

ACKNOWLEDGMENTS

I thank Per Peterson and Harold McFarlane for their helpful comments, information, and insights.

REFERENCES

AWEA (American Wind Energy Association). 1998. How Much Energy Does It Take to Build a Wind System in Relation to How Much Energy It Produces? Available online at: <*http://www.awea.org/faq/bal.html*>.

DOE (U.S. Department of Energy). 2002. Yucca Mountain Site Suitability Evaluation. DOE/RW-0549. Washington, D.C.: U.S. Department of Energy. Also available online at: <*http://www.ymp.gov/documents/sse_a/index.htm*>.

Kotek. J. 2003. Advanced Nuclear Reactor Technologies. Frontiers of Engineering: Reports on leading-Edge Engineering from the 2002 NAE Symposium on Frontiers of Engineering. Washington, D.C.: The National Academies Press.

NRC (National Research Council). 1995. Management and Disposition of Excess Weapons Plutonium: Reactor-Related Options. Washington, D.C.: National Academy Press. Also available online at: <*http://www.nap.edu/catalog/4754.html*>.

NRC. 1996. Nuclear Wastes: Technologies for Separations and Transmutation. Washington, D.C.: National Academy Press. Also available online at: <*http://www.nap.edu/ catalog/ 4912.html*>.

OECD (Organization for Economic Cooperation and Development). 1999. Status and Assessment Report on Actinide and Fission Product Partitioning and Transmutation. Nuclear Development Report No. 1507. Paris, France: Nuclear Energy Agency of the Organization for Economic Cooperation and Development. Also available online at: <*http://www.nea.fr/ html/trw/docs/neastatus99/*>.

OECD. 2002. Accelerator-Driven Systems (ADS) and Fast Reactors (FR) in Advanced Nuclear Fuel Cycles. Nuclear Development Report No. 3109 (2002). Paris, France: Nuclear Energy Agency of the Organization for Economic Cooperation and Development. Also available online at: <*http://www.nea.fr/html/ndd/reports/2002/nea3109.html*>.

Rubbia, C., J.A. Rubio, S. Buono, F. Carminati, N. Fieter, J. Galvez, C. Geles, Y. Kadi, R. Klapisch, P. Mandrillon, J.P. Revol, and C. Roche. 1995. Conceptual Design of a Fast Neutron Operated High Power Energy Amplifier. CERN-AT-95-44ET. Geneva, Switzerland: European Organization for Nuclear Research.

Till, C.E., Y.I. Chang, and W.H. Hannum. 1997. The integral fast reactor: an overview. Progress in Nuclear Energy 31: 3–11.

UNDP (United Nations Development Programme). 2000. World Energy Assessment: Energy and the Challenge of Sustainability. New York: United Nations Development Programme. Also available online at: <*http://www.undp.org/seed/eap/activities/wea/*>.

WEC (World Energy Council). 2001. Survey of Energy Resources. Part I: Uranium. Available online at: <*http://www.worldenergy.org/wec-geis/publications/reports/ ser/uranium/uranium.asp*>.

WEC. 2002. Global Energy Scenarios to 2050 and Beyond. Available online at: <*http://www.worldenergy.org/wec-geis/edc/scenario.asp*>.

Stretching the Boundaries of Nuclear Technology

JAMES P. BLANCHARD
Department of Engineering Physics
University of Wisconsin-Madison

Most people are well aware that nuclear power can be used to produce electricity, but few are aware that it can be used to provide power in many other situations. Radioisotopes have been used for decades in commercial applications, such as pacemakers and smoke detectors, and recent trends indicate that other applications are on the horizon. Two technologies being actively investigated are space nuclear power and nuclear energy for microelectromechanical systems (MEMS).

SPACE NUCLEAR POWER

In 1989, a national space policy was approved that included the goal of putting a man on Mars by 2019. By most accounts, meeting this goal will require nuclear propulsion in order to shorten the mission time, thereby reducing exposure to zero gravity conditions and cosmic rays. Hence, nuclear propulsion will play a major role in space travel beyond the moon. This year, NASA announced a five-year, $1-billion program to develop nuclear reactors to power the next generation of spacecraft.

History

Early work on nuclear propulsion was primarily focused on nuclear-thermal technologies, in which a fission reactor is used to heat a gas and accelerate it through a nozzle. Research activity in this area began in 1944 and peaked in the 1960s. A typical design uses hydrogen as a propellant and graphite-moderated carbide fuel in the reactor core. One design, called Phoebus, achieved 5,000

MW of thermal power and 1 MN of thrust, which is about half the thrust of a space-shuttle engine (Bower et al., 2002).

Another early concept for nuclear propulsion was Project Orion, which relied on a series of nuclear blasts behind the payload to create shock waves that accelerated the device (Schmidt et al., 2002). Although this technology looked promising, it was abandoned in the 1960s because of a ban on nuclear testing.

Fuel Efficiency and Mission Length

The efficiency of the fuel used for propulsion is measured by a parameter called the specific impulse, which is defined as the ratio of the thrust produced to the rate at which fuel is consumed. The units of this parameter, typically seconds, are determined by dividing force by weight of fuel consumed per unit time. Hence, a specific impulse of N seconds can be interpreted as a capability for providing a unit thrust with a unit weight of fuel for N seconds. By comparing the specific impulses for different propulsion technologies, one can assess their advantages and disadvantages. Increased fuel efficiency is manifested in several ways—shorter trips, larger payloads for a fixed total launch weight, and flexibility for scientific activities at the destination.

The specific impulses for several propulsion options are shown in Table 1. The specific impulse for nuclear fuels can be many times that of chemical fuels, while the thrust is correspondingly lower and the run time longer. Electrostatic thrusters have relatively low thrust but can run virtually continuously and, therefore, can provide short trip times and low launch weights for a given payload. In

TABLE 1 Propulsion Parameters for Several Propulsion Technologies

Technology	Specific Impulse (sec)	Thrust per Engine (N)	Run Time (duration)
Chemical	150–450	0.5–5 million	A few seconds to hundreds of minutes
Nuclear Thermal	825–925	5,000–50,000	A few minutes to several hours
Electromagnetic	2,000–5,000	10–200	A few seconds to several hundred hours
Electrostatic	3,500–10,000	1–10	A few minutes to several days or months

Source: Niehoff and Hoffman, 1996. Reprinted with permission of the American Astronautical Society.

contrast, chemical rockets tend to run at high thrust for short times, accelerating rapidly as the rocket fires and then coasting between necessary adjustments in trajectory.

Nuclear-Electric Propulsion

There are three basic types of electric propulsion systems: electrothermal, electrostatic, and electromagnetic. In electrothermal propulsion, the propellant is heated either by an electric arc or a resistance heater. The hot propellant is then exhausted through a conventional rocket nozzle to produce thrust. Electrostatic propulsion uses electric fields to accelerate charged particles through a nozzle. In electromagnetic propulsion, an ionized plasma is accelerated by magnetic fields. In all three types, electricity from a nuclear source, such as a fission reactor, is used to power the propulsion device (Allen et al., 2000; Bennett et al., 1994). The power flow for a typical nuclear-electric propulsion scheme is shown in Figure 1.

The most mature of the electric propulsion concepts is electrostatic propulsion. NASA's Deep Space 1 device (Figure 2), launched in 1998, relies on an ionized xenon gas jet for propulsion (Brophy, 2002). The xenon fuel fills a chamber ringed with magnets, which control the flow; electrons emitted from a

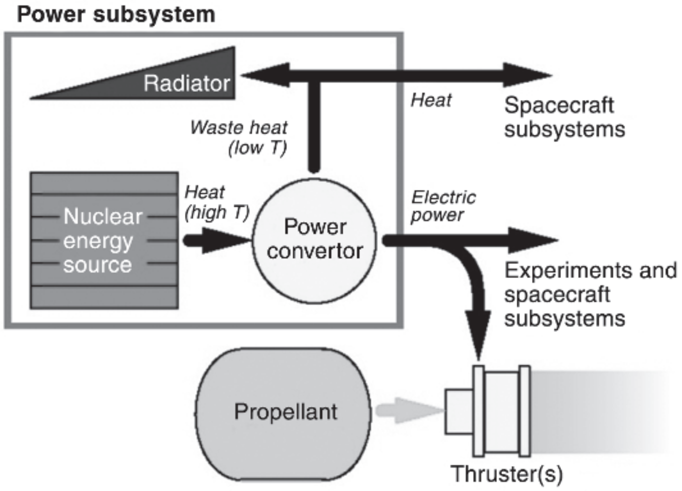

FIGURE 1 Schematic drawing of power flow for a typical nuclear-electric propulsion device. Heat is produced in the nuclear source (typically a fission reactor core) and converted to electricity in the power converter. The electricity is then used to power the thruster. The radiator dissipates the waste heat. Source: NASA, 2001.

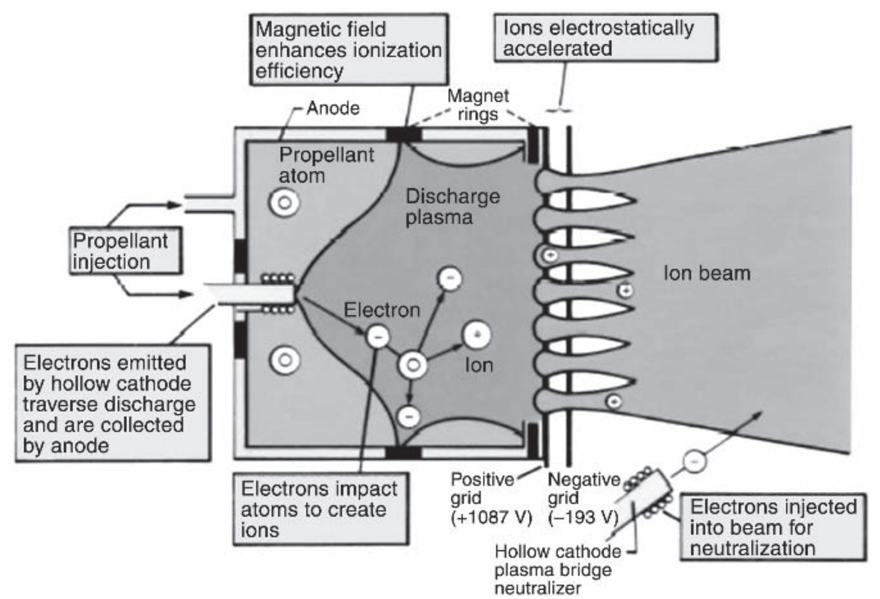
Magnetic field enhances ionization efficiency
Ions electrostatically accelerated
Anode
Magnet rings
Propellant atom
Discharge plasma
Propellant injection
Ion beam
Electron
Ion
Electrons emitted by hollow cathode traverse discharge and are collected by anode
Electrons impact atoms to create ions
Positive grid (+1087 V)
Negative grid (–193 V)
Hollow cathode plasma bridge neutralizer
Electrons injected into beam for neutralization

FIGURE 2 Schematic drawing of the Deep Space 1 ion thruster. Source: NASA, 2002.

cathode ionize the gas. The ions pass through a pair of metal grids at a potential of 1,280 volts and are thus accelerated out the back. A second electrode emits electrons to neutralize the charge on the device. The engine is capable of producing 90 mN of thrust while consuming 2,300 W of electrical power. This device is solar powered, but future designs anticipate using a fission reactor to produce the electricity.

All electric propulsion systems require supplies of electricity, and fission reactors, which have high power density, are an excellent choice for meeting this need. Numerous projects are under way to develop fission reactors with low weight, high reliability, long life without refueling, and safety during launch. A wide variety of heat-transport and energy-conversion technologies are being investigated. One example of a fission reactor is the safe, affordable fission engine (SAFE-400), a 400-kW (thermal) reactor that is expected to produce 100 kW of electric power using heat pipes for energy transport and a Brayton cycle for energy conversion (Poston et al., 2002). The core consists of 381 uranium-nitride fuel pins clad with rhenium. The uranium-nitride fuel was chosen because of its high uranium density and high thermal conductivity. Molybdenum/sodium heat pipes are used for heat transport to provide passive safety features in case of an accident.

Plasma Propulsion

An approach related to electric propulsion is the plasma rocket, exemplified by the variable specific impulse magnetoplasma rocket (VASIMR) (Diaz, 2000). Like an ion thruster, a VASIMR injects a propellant (usually hydrogen) into a cell and ionizes it. The resulting plasma is heated using radio-frequency injection and a magnetic nozzle that accelerates the gas to provide the propulsion. A second example is the gas dynamic mirror, a long, slender device in a magnetic mirror configuration. This device is powered by fusion reactions in the plasma; the thrust is produced by plasma ions exiting the end of the device. Accelerated by the mirror's magnetic-field gradients, the ions provide efficient propulsion. One concept features a 50-m long, 7-cm radius plasma and produces 50,000 N of thrust at a specific impulse of more than 100,000 seconds (Kammash et al., 1995).

SMALL-SCALE RADIOISOTOPE POWER

MEMS have the potential to revolutionize many technologies, and the number of commercial applications is increasing rapidly. Many applications, such as pumps, motors, and actuators, can be improved with onboard power supplies, and various technologies are being explored to provide such power. Obvious choices, such as chemical batteries, fuel cells, and fossil fuels, show some promise, but none of them can match radioisotope power for long, unattended operation (Blanchard et al., 2001). This is because of the larger energy density available with nuclear sources.

Radioisotopes can be used to produce power in a variety of ways. Thermoelectric and thermionic technologies convert the heat generated by the decay to electricity; other approaches make more direct use of the released energy. Thermoelectric conversion uses a thermal gradient between two different materials to create a current via the Seebeck effect. Thermionic conversion creates a current by boiling electrons off a cathode (at high temperature) and catching them at an anode. Techniques for more direct methods include simple collection of the emitted charged particles, ionization near a P-N or P-I-N junction in a semiconductor, and conversion of the decay energy to light and subsequent conversion to electricity in a photovoltaic.

History

Radioisotopes have been used as power sources for decades. Early pacemakers were powered by approximately 0.2 grams (3 Ci) of ^{238}Pu, producing about 0.2 mW and delivering about 0.05 mW to the heart muscle (Parsonnet, 1972). Whereas pacemakers powered by chemical batteries have lives of less than 10 years and thus require replacement in most patients, the half-life of ^{238}Pu (approximately 86 years) permits radioisotope-powered devices to last the life of the patient.

Although a smoke detector is not strictly a power source, many smoke detectors contain radioisotopes (usually 1 to 5 microcuries of ^{241}Am). The source ionizes air between a pair of parallel plates, and a chemical battery (or house current) is used to collect these charges and thus measure the degree of ionization in the gap. When smoke enters the gap, the increased ionization trips the sensor.

Radioisotope thermoelectric generators (RTGs) are used in many applications, including underwater power and lighting in remote locations, such as the Arctic (Lange and Mastal, 1994). RTGs were also used to provide power for the Cassini and Voyager missions. Much like the pacemakers mentioned above, RTGs create power by thermoelectric conversion. Most RTGs are modular, with each module containing approximately 2.7 kg of Pu (133 kCi) and measuring approximately 42 cm in length and 114 cm in diameter. The modules produce 276 W of electric power at the beginning of life and, despite decay of the isotope, will produce approximately 216 W after 11 years of unattended operation.

Current research is focused mostly on the miniaturization of RTGs for many applications, such as MEMS; in addition, efforts to improve the efficiency of existing RTGs are ongoing.

Nuclear Microbatteries

Thermal devices, such as RTGs, are difficult to reduce to the microscale because, as the size is decreased, the surface-to-volume ratio increases, thus increasing the relative heat losses and decreasing the efficiency of the device. Hence, microbattery designs have tended to focus on direct methods of energy conversion. For example, one can construct a diode from silicon using a layer of P-type silicon adjacent to a layer of N-type silicon and a radioactive source placed on the top of the device. As the source decays, the energetic particles penetrate the surface and create electron-hole pairs in the vicinity of the P-N junction. This creates a potential across the junction, thus forming a battery. Figure 3a is a schematic drawing of such a device, and Figure 3b is a photograph of one concept created at the University of Wisconsin. The device shown in Figure 3b is fairly large, measuring approximately 0.5 cm on each side, but one can easily imagine using a single pit from the device as a power source. This would provide a microbattery measuring approximately 400 microns by 400 microns by 50 microns; using a beta emitter (^{63}Ni), it could produce approximately 0.2 μW of electrical power. An early prototype of the device pictured in Figure 3b, loaded with a weak source (64 microcuries of ^{63}Ni), produced approximately 0.07 nW of power. Given that the thermal energy of 64 microcuries of ^{63}Ni is 6.4 nW, this device is about 1 percent efficient. Placing a second diode on top of the source would nearly double the efficiency (because the decay

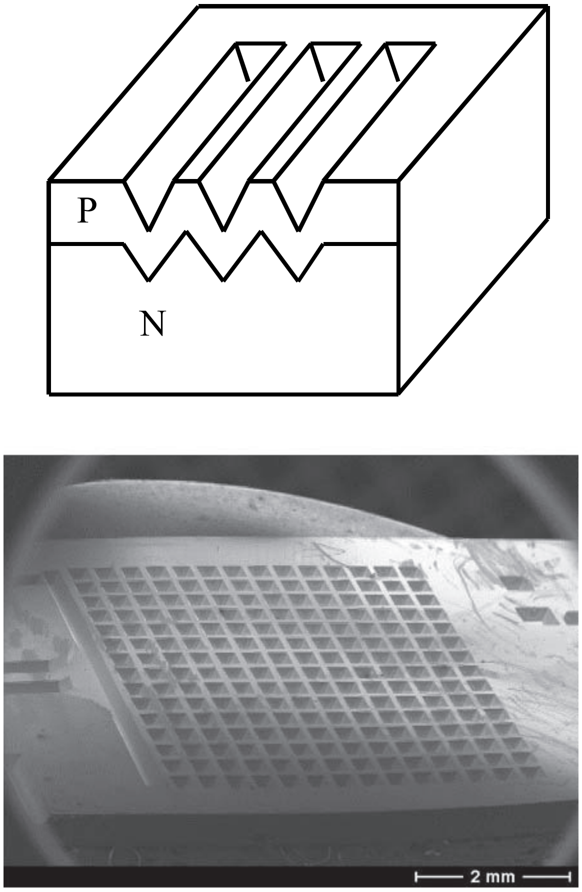

FIGURE 3(a) A schematic drawing of a simple device using a P-N junction and radioactive source to produce the potential. **(b)** Photograph of a device using pits rather than trenches to hold the source.

products are produced isotropically). Work is also under way to improve the efficiency by optimizing the design.

When an alpha source is used in such a battery, the available energy for a fixed activity level is increased by several orders of magnitude. Unfortunately, the high-energy alpha particles damage the silicon lattice as they pass through and quickly degrade the power. Attempts are being made to overcome this limitation by using materials that are resistant to damage. Materials being considered include wide-band-gap semiconductors, such as gallium nitride, which might improve radiation stability and device efficiency (Bower et al., 2002).

Thermoelectric devices are another approach to using alpha sources. These devices use the heat from the source to produce a temperature gradient across the thermoelectric device to produce power. Thus, there is no risk of radiation damage, and alpha particles could be used. Hi-Z Technology, Inc. (San Diego), developed a 40 mW device using a radioisotope heater unit (RHU) that was produced by NASA several years ago. The RHU uses about 2 grams of ^{238}Pu to produce 1 W of thermal power. Using this as a heat source, Hi-Z produced a thermoelectric device that established a temperature difference of approximately 225°C throughout the device and provided 40 mW of power. The efficiency of the device was approximately 4 percent. Some improvement can be gained through improved insulation and thermoelectrics.

A third approach to creating a micropower device uses radioisotopes to excite phosphors that emit photons, which can then be collected in a standard or modified solar cell. This protects the photovoltaic from damage but increases losses in the system. In addition, the phosphor may be damaged. Typical organic scintillators have energy-conversion efficiencies of 1 percent, whereas inorganic crystals can achieve efficiencies of up to 30 percent. TRACE Photonics, Inc. (Charleston, Illinois), has built a scintillation glass using sol-gel processes with high light-conversion efficiency under radiation exposure (Bower et al., 2002). Current overall efficiencies are approximately 1 percent, but device integration can probably be improved because of the low weight and direct conversion.

Applications of Micropower Sources

All current MEMS devices sold commercially are passive devices. Hence, there is no existing market for micropower sources. Nevertheless, one can envision many future applications of MEMS devices with onboard micropower sources, such as small drug dispensers placed directly into the bloodstream and laboratories-on-a-chip that can carry out real-time blood assays. Researchers at UCLA and UC Berkeley have been investigating so-called “smart-dust” concepts for using wireless communications to create large-scale sensor networks (Kahn et al., 1999). This approach involves distributed sensors that can communicate with each other through a network and thus “provide a new monitoring and control capability for transportation, manufacturing, health care, environmental monitoring, and safety and security” (Asada et al., 1998). These devices will require power for data collection and storage, as well as for the delivery of information between neighboring devices.

A new application of nuclear power is the self-powered cantilever beam produced at the University of Wisconsin (Li et al., 2002). This device, shown in Figure 4, places a conducting cantilever beam in the vicinity of a radioisotope, in this case ^{63}Ni. As the beam collects the electrons emitted from the source, it becomes negatively charged, and the source becomes positively charged. The

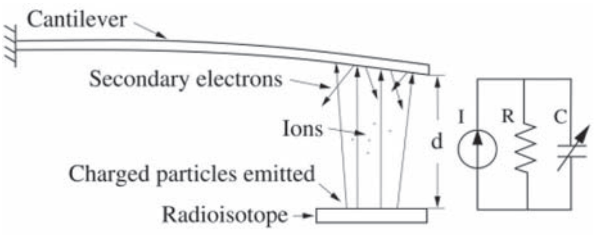

FIGURE 4 Schematic drawing of a self-oscillating cantilever beam. Devices can be modeled as capacitors in parallel with leakage resistors, with most of the leakage resulting from ionization in the gap.

beam is thus attracted to the source until contact is made and the device discharges. This causes the beam to be released and return to its original position. The process then repeats itself. Hence, the beam undergoes a repetitive bending and unbending; the period of the oscillation is determined by the strength of the source, the beam stiffness, and the initial separation between the beam and the source. Work is ongoing to produce wireless communication devices based on this design.

CONCLUSIONS

Nuclear power is the best, perhaps the only, realistic power source for both long-distance space travel and long-lived, unattended operation of MEMS devices. Much more research will have to be done to optimize the currently available technologies for future applications, but nuclear technologies will clearly provide viable, economic solutions, and they should be given continued attention and support as they approach commercialization

REFERENCES

Allen, D.T., J. Bass, N. Elsner, S. Ghamaty, and C. Morris. 2000. Milliwatt Thermoelectric Generator for Space Applications. Pp. 1476-1481 in Proceedings of Space Technology and Applications International Forum-2000. New York. American Institute of Physics Press.

Asada, G., T. Dong, F. Lin, G. Pottie, W. Kaiser, and H. Marcy. 1998. Wireless Integrated Network Sensors: Low Power Systems on a Chip. Pp. 9-16 in Proceedings of the 1998 European Solid State Circuits Conference. Paris: Seguir Atlantica.

Bennett, G., H. Finger, T. Miller, W. Robbins, and M. Klein. 1994. Prelude to the Future: A Brief History of Nuclear Thermal Propulsion in the United States. Pp. 221-267 in A Critical Review of Space Nuclear Power and Propulsion, 1984–1993, edited by M. El-Genk. New York: American Institute of Physics Press.

Blanchard, J., R.M. Bilboa y Leon, D.L. Henderson, and A. Lai. 2001. Radioisotope Power Sources for MEMS Devices. Pp. 87-88 in Proceedings of 2001 ANS Annual Meeting. Washington, D.C.: American Nuclear Society.

Bower, K., X. Barbanel, Y. Shreter, and G. Bohnert. 2002. Polymers, Phosphors, and Voltaics for Radioisotope Microbatteries. Boca Raton, Fla.: CRC Press.

Brophy, J. 2002. NASA's Deep Space 1 ion engine. Revue of Scientific Instruments 73(2): 1071-1078.

Diaz, F. 2000. The VASIMR rocket. Scientific American 283(5): 90-97.

Kahn, J.M., R.H. Katz, and K.S.J. Pister. 1999. Mobile Networking for Smart Dust. Pp. 271-278 in Proceedings of the ACM/IEEE International Conference on Mobile Computing and Networking. New York: ACM Press.

Kammash, T., M. Lee, and D. Galbraith. 1995. High-Performance Fusion Rocket for Manned Space Missions. Pp. 47-74 in Fusion Energy in Space Propulsion, edited by T. Kammash. Progress in Astrophysics and Aeronautics 167.

Lange, R., and E. Mastal. 1994. A Tutorial Review of Radioisotope Power Systems. Pp. 1-20 in A Critical Review of Space Nuclear Power and Propulsion 1984–1993, edited by M. El-Genk. New York: American Institute of Physics Press.

Li, H., A. Lal, J. Blanchard, and D. Henderson. 2002. Self-reciprocating radioisotope-powered cantilever. Journal of Applied Physics 92(2): 1122-1127.

NASA (National Aeronautics and Space Administration). 2001. The Safe Affordable Fission Engine (SAFE) Test Series. Available online at*: <http://www.spacetransportation.com/ast/presentations/7b_vandy.pdf>.*

NASA. 2002. DS1: How the Ion Engine Works. Available online at: *<http://www.grc.nasa.gov/WWW/PAO/html/ipsworks.htm>.*

Niehoff, J., and S. Hoffman. 1996. Pathways to Mars: An Overview of Flight Profiles and Staging Options for Mars Missions. Pp. 99-125 in Strategies for Mars: A Guide for Human Exploration, edited by C.R. Stoker and C. Emmart. Paper no. AAS 95-478. Science and Technology Series Vol. 86. San Diego, Calif.: Univelt. (Copyright © 1996 by American Astronautical Society Publications Office, P.O. Box 28130, San Diego, CA 92198; Website: *<http://www.univelt.com>*. All Rights Reserved. This material reprinted with permission of the AAS.)

Parsonnet, V. 1972. Power sources for implantable cardiac pacemakers. Chest 61: 165-173.

Poston, D., R. Kapernick, and R. Guffee. 2002. Design and Analysis of the SAFE-400 Space Fission Reactor. Pp. 578-588 in Space Technology and Applications International Forum. New York: American Institute of Physics Press.

Schmidt, G., J. Bonometti, and C. Irvine. 2002. Project Orion and future prospects for nuclear propulsion. Journal of Propulsion Power 18(3): 497-504.

Engineering Challenges for Quantum Information Technology

Quantum Cryptography

Steven J. van Enk
Bell Laboratories, Lucent Technologies
Murray Hill, New Jersey

Cryptography has a very long history (Singh, 2000). The ancient civilizations of the Chinese, Egyptians, Greeks, Romans, and Arabs all developed methods of keeping messages secret. Early on, often all that was necessary was hiding the *existence* of a message (steganography). A method developed in China, for instance, involved shaving the head of a messenger, writing the message there, and waiting for the hair to grow back. This method was obviously both inconvenient and time consuming. But the main drawback of steganography is that once the hiding place of a message has been discovered, all of the information in the message is revealed at once.

Thus, cryptography, the art of hiding the *meaning* of a message, was born. There are two basic methods of encrypting a message—transposition and substitution. In transposition, the order of the letters is changed; in substitution, each character (or sometimes a whole word) is replaced by another character (or word), according to certain procedures known to both sender and receiver. Over the course of thousands of years, increasingly complicated versions of encryption protocols have been designed, but in the end, almost every one has been broken. Every language contains structure—particularly certain letters or letter combinations that appear more often than others—and, if one is not careful, encrypted texts reveal the same structure. Nowadays we know that the only secure way of encrypting a text is by shifting each letter by a random amount. In modern terminology, each bit must be XOR-ed with a random bit. Only then does the encrypted text itself contain no information.

The remaining problem is key distribution, how to get the sender and receiver to agree on the random sequence of bits to be used. Modern methods, the ones used for encryption of Internet communications for instance, all rely on so-called one-way functions—functions that are easy to calculate (roughly speaking, with resources that scale polynomially with the number of bits of the numbers involved) but very hard (scaling exponentially) to invert. For example, public key cryptography (particularly, the well known RSA encryption scheme) is based on the difficulties of factoring large numbers and of taking the discrete logarithm. This type of protocol has two weaknesses. First, neither of these two tasks has been proven to be exponentially hard. Indeed, a quantum computer can solve both problems in polynomial time (Ekert and Jozsa, 1996; Shor, 1997). Second, with increasingly powerful classical computers and algorithms, or with a powerful quantum computer, a code that cannot be cracked now may be cracked in the future. Therefore, if the encrypted message is stored, it could eventually be deciphered. For most messages this may not be important, but for certain military messages, it may be essential, especially if the time span turns out to be short. Perhaps surprisingly, both problems can be solved with quantum mechanics, which provides an unconditionally secure protocol that relies only on the laws of physics and not on unproven mathematical assumptions (Bennett and Brassard, 1984; Bennett et al., 1992).

Here is how quantum key distribution (QKD) works. Alice (the sender) and Bob (the receiver) wish to create a shared random key. Alice sends Bob a series of polarized photons. First she tosses her coin many times to produce two random sequences of 0's and 1's. One sequence determines which bit she is going to send, the other which polarization basis she will use, either horizontal/vertical polarization, or left-hand/right-hand circular polarization. An eavesdropper cannot find out with certainty which bit was sent. For example, a polarization filter set to block vertical polarization will make no distinction at all between the two circular polarizations because both will be blocked 50 percent of the time. Moreover, if Alice uses such a filter, even a circularly polarized photon that passes through will have vertical polarization. The disturbance of the quantum state caused by measurements is what Alice and Bob will notice. Let's say Bob chooses a basis too randomly for his polarization measurement. In the absence of errors and eavesdroppers, Bob would find the correct polarization if he happened to choose the same basis as Alice (which occurs with 50 percent probability). To detect the presence of an eavesdropper, they can check a small subset of the bits that should be the same. The fraction of errors they find is the upper bound on how much information an eavesdropper may have gathered. It is only an upper bound, however, because errors may also be caused by other, innocent effects. Once they have an upper bound, Alice and Bob can distill, by purely classical privacy amplification techniques, a shorter random key that is secure to an arbitrary degree. In a simplified example, let's suppose Alice and Bob share two random bits, but they know an eavesdropper knows at most one. If they simply take the sum (XOR) of their two bits, they will have a perfectly secure new random bit.

I have just sketched the ideal protocol. In practice, single photons are very hard to generate. Instead, faint laser pulses are used (the number of photons in a pulse is not fixed but distributed according to a Poisson distribution). The laser light is attenuated so strongly that the average number of photons is only about 0.1 to 0.01. Despite these differences, and despite all kinds of other inevitable imperfections (e.g., losses in fibers, misalignments of optical elements, etc.), such protocols have been proven secure (Inamori, 2002). Because the total number of errors is limited (otherwise the upper bound on the information the eavesdropper has would indicate she knows everything!), and because the main errors are caused by losses inside optical fibers, QKD is possible only over a limited distance.

The state of the art in QKD (Gisin et al., 2002; Townsend, 1998) is as follows: the world record in distance for free-space QKD is 23 km (Kurtsiefer et al., 2002). To avoid turbulence, the communication line was between two mountaintops in Germany. The secure bit rate was about 300 to 400 bits/sec. For QKD over standard optical telecom fibers, the record is 67 km with a secure bit rate of about 60 bits/sec (Stucki et al., 2002). It is important to note that these numbers refer to a protocol with conditional, rather than unconditional, security. The security is based on relaxed, but realistic, assumptions—the eavesdropper cannot store photons, cannot measure the number of photons in a pulse without destroying them, and cannot replace the channel by a channel with no losses. Secure distances for the unconditional security promised by the original protocols are limited to about 15 to 20 km for fibers, less for free-space QKD. Nevertheless, because no one knows at this moment how to store quantum data for a substantial amount of time, QKD does provide a new type of protection against eavesdropping. In effect, today's QKD messages are safe against tomorrow's technology, which cannot be said of RSA-encrypted data.

As the state-of-the-art experiments have shown, QKD has been advanced beyond proof-of-principle demonstrations in physics laboratories. In fact, the devices developed for QKD are available as commercial products (Stucki et al., 2002), and a whole QKD setup, including the software to perform the classical calculations for privacy amplification, may soon be available. Does this mean a quantum computer is around the corner? Unfortunately, the answer is no, because two requirements for quantum computing are very hard to achieve, although they are not necessary for quantum communication. First, in a quantum computer the quantum states of many qubits, the quantum counterpart of the classical bit, must be stored for roughly as long as the computation runs; for QKD, qubits may be measured and destroyed as soon as they have reached the receiver. Second, the many qubits of a quantum computer must interact with each other in a carefully controlled way; in QKD, the qubits can be sent separately and never have to interact with each other.

Future research will focus on two aspects of QKD—miniaturization of the devices and improved secure bit rates and longer distances for secure key distribution. The latter will require the development of better single-photon detectors and true single-photon generators. In addition, until "lossless" fibers are developed, the only technique for overcoming fiber losses is quantum error correction, which is a challenging task, comparable in difficulty to building a small quantum computer (van Enk et al., 1998). Despite the challenges that lie ahead, quantum mechanics will someday make communications more secure.

REFERENCES

Bennett, C.H., and G. Brassard. 1984. Quantum Cryptography: Public Key Distribution and Coin Tossing. Pp. 175-179 in Proceedings of the IEEE International Conference on Computers, Systems, and Signal Processing. New York: IEEE.

Bennett, C.H., G. Brassard, and A. Ekert. 1992. Quantum cryptography. Scientific American 269(4): 26-33.

Ekert, A., and R. Jozsa. 1996. Quantum computation and Shor's factoring algorithm. Review of Modern Physics 68(3): 733-753.

Gisin, N., G. Ribordy, W. Tittel, and H. Zbinden. 2002. Quantum cryptography. Review of Modern Physics 74(1): 145-195.

Inamori, H., N. Lutkenhaus, and D. Meyers. 2002. Unconditional Security of Practical Quantum Key Distribution. Available online at: <*http://xxx.lanl.gov/abs/quant-ph/0107017*>.

Kurtsiefer, C., P. Zarda, M. Halder, H. Weinfurter, P.M. Gordon, P.R. Tapster, and J.G. Rarity. 2002. A step towards global key distribution. Nature 419(6906): 450.

Shor, P.W. 1997. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM Journal on Computing 26(5): 1484-1509.

Singh, S. 2000. The Code Book. New York: Anchor Books.

Stucki, D., N. Gisin, O. Guinnard, G. Ribordy, and H. Zbinden. 2002. Quantum key distribution over 67km with a plug&play system. New Journal of Physics 4(41): 1-8. Available online at: <*http://www.idquantique.com/qkd.html*>.

Townsend, P.D. 1998. Quantum cryptography on optical fiber networks. Optical Fiber Technology 4(4): 345-370.

van Enk, S.J., J.I. Cirac, and P. Zoller. 1998. Photonic channels for quantum communication. Science 279(5348): 205-208.

Ion-Trap Quantum Computation

Dietrich Leibfried
National Institute of Standards and Technology
University of Colorado
Boulder, Colorado

In the early 1980s, Feynman, Benioff, Bennett, and Deutsch pointed out that a quantum computer could have advantages in speed and overhead for solving certain problems; and because of their reversibility, quantum computers might dissipate much less heat than irreversible classical computers. In 1994, Shor presented a quantum algorithm for efficiently finding the prime factors of large numbers that gave the field practical meaning. For example, because most modern encryption schemes work on the assumption that it is inefficient to find prime factors of large numbers, a quantum computer would put the safety of all of these schemes in jeopardy. Shor's algorithm also raised hopes that other quantum algorithms could outperform their classical counterparts. Moreover, if one takes Moore's law seriously, in a few years (around 2020) the size of a memory unit in a classical computer should be as small as a single atom. At atomic dimensions, the unit's function will be determined by quantum mechanics, no matter whether it is part of a classical or quantum computer.

Why is quantum computing more powerful than its classical counterpart? A classical bit can have the logical values 0 or 1. The qubit, its quantum counterpart, can have the values 0 and 1 at the same time. This is directly related to other seemingly paradoxical statements often heard in connection with quantum mechanics, such as, "Schrödinger's cat is both dead and alive." This example can be easily associated with a qubit by assigning "dead" to 0 and "alive" to 1.

In the trapped-ion system, the qubit levels are two electronic states of the trapped charged atom. Using superpositions, states in which the qubit is in 0 and 1, one can process all possible input states at once with a given algorithm (quantum parallelism). The only catch is that once the algorithm terminates, the answer must be read out. Similar to Schrödinger's cat, which, upon looking, is found either dead or alive, reading the output register collapses the superposition of all possible answers to one particular answer. Knowing this, one can infer the structure of problems for which a quantum computer could be useful. The input should contain a great number of possibilities, for example, for factoring large numbers, all integer numbers up to the given large number. The output should contain a much smaller number of meaningful answers, in our example all of the prime factors of that number. Even if the output collapses to only one of the factors upon readout, we can divide our large number by that factor and thus simplify the original problem.

ION TRAPS, A SUCCESS STORY

In 1989, Wolfgang Paul and Hans Dehmelt won the Nobel Prize for their work on ion traps. The most precise atomic clock built so far (at the National Institute of Science and Technology [NIST]) has a single mercury ion in a Paul trap (Diddams et al., 2001). The clock will lose about one second in the estimated age of the universe (14 billion years). Atomic clocks have a lot in common with qubits. Both rely on a two-level system that is extremely well isolated from its environment but can still be manipulated in a precisely controlled way, with lasers, for example.

ARCHITECTURE OF AN ION-TRAP-BASED QUANTUM COMPUTER

Ignacio Cirac and Peter Zoller first proposed using a string of ions confined in an ion trap for a quantum computer in 1995 when they realized that ions could be exceptionally good qubits and that their motion in the trap could provide a means of performing logic gates between them. Ions are charged, so they experience a strong repelling interaction. When lined up like a string of pearls, a kick to the ion on one end of the string propagates to the other end and back, similar to the way a collection of balls connected by springs responds. Because of this coupling, the motion of these ions must be considered a collective motion that can be described with normal modes. Ions in a trap can be laser cooled to the quantum mechanical state with the lowest energy (ground state), very close to absolute zero temperature.

Cirac and Zoller realized that the collectively shared motion could transfer information between different ion-qubits. A gate between qubit-ion A and qubit-ion B amounts to the following steps: (1) the state of the ion-qubit A is "written" onto the state of the collective motion with a laser pulse; (2) depending on the state of the motion, the state of qubit-ion B is either flipped or not; (3) the state of the motion is then transferred back onto qubit-ion A.

The NIST group was able to demonstrate step 2 that same year (Monroe et al., 1995). Since then, several extensions and refinements have been added to the original proposal. In 2000, the NIST group was able to demonstrate quantum

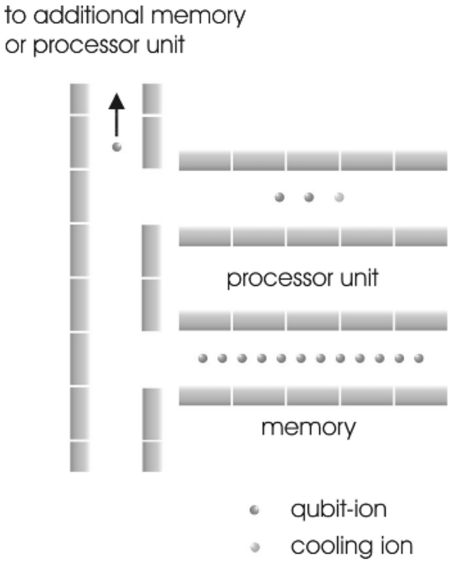

FIGURE 1 Architecture of a multi-trap quantum computer.

gates on two and four ions (Sackett et al., 2000). In the same year, David DiVincenco published five criteria for quantum computation that have been widely adopted as a test for determining if a physical system is a serious candidate for quantum computing (DiVincenco, 2001). Four-and-one-half of the criteria have been experimentally demonstrated in the ion-trap system; the only undemonstrated point is the scalability to a large qubit number in DiVincenco's first criterion, "a scalable physical system with well characterized qubits."

To fulfill this last criterion, it seemed appropriate to modify the original proposal slightly. Instead of one long string of ions in one trap, the computer could consist of several interconnected traps (see Figure 1) (Kielpinski et al., 2002; Wineland et al., 1998). The ion-qubits could be moved in and out of these traps by an appropriate sequence of voltages applied to their electrodes. Similar to a classical computer, some of the trap units would serve as "memory" (the qubit-ions would be just stored there, but no gates performed). Other units would be the "processors" in which the qubit-ions would be brought together; a third ion of a different species could be added as a cooling agent. The collective motion in the trap would couple the three ions so they could be cooled by applying laser pulses to the third ion without ever touching the internal states of the two qubit-ions. After cooling, one could perform quantum gates on the two qubit-ions. They would then be transferred back into the memory unit, and the algorithm could proceed by applying gates to different qubits. Once the algorithm had terminated, the final state of the qubits could be read out with high efficiency (>99 percent) by laser-induced fluorescence.

FROM QUANTUM MECHANICS TO QUANTUM ENGINEERS

In 2001, we built a two-trap setup to demonstrate the basic features of our envisioned architecture. To implement our ideas, we had to resort to manufacturing techniques developed for microengineering and integrated circuits. The double trap (Figures 2a and 2b) consists of two alumina wafers into which the basic electrode structure is cut with lasers. A thin film of gold (ca. 0.5 µm) is then deposited through a shadow mask onto the wafers where conducting surfaces are desired. Subsequently, an additional layer of gold 3 µm thick is electroplated onto these surfaces to ensure uniform, low-resistance electrodes. The trap areas 2 and 4 have a distance of 1.1 mm. In this trap array, we moved an ion from area 2 to area 4 within 28 µs, while preserving the qubit information, and we separated two ions trapped in area 3 into areas 2 and 4, thereby proving the feasibility of two main ingredients of the multitrap scheme (Rowe et al., 2002). We have just verified the final ingredient, cooling of the qubit-ion species with a different species without disturbing the qubit information. Simultaneously we are working on two traps, based on different technologies (Figures 3 and 4).

In the boldest approach so far, we are trying to build an all-silicon trap (Figure 4). The conductive electrodes are made of silicon doped with boron;

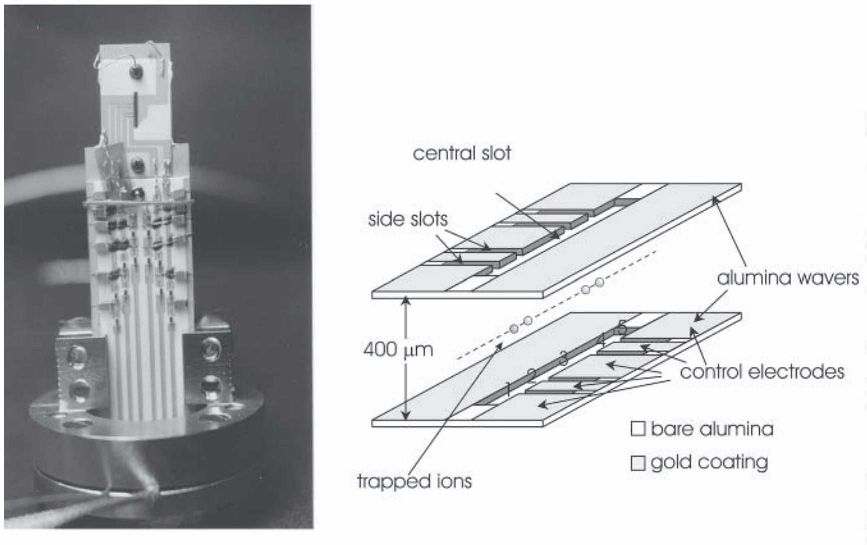

FIGURE 2(a) Photograph of the complete trap structure, including SMD filter electronics. **(b)** Schematic drawing of the trap. The numbers over the lower wafer indicate the trap units.

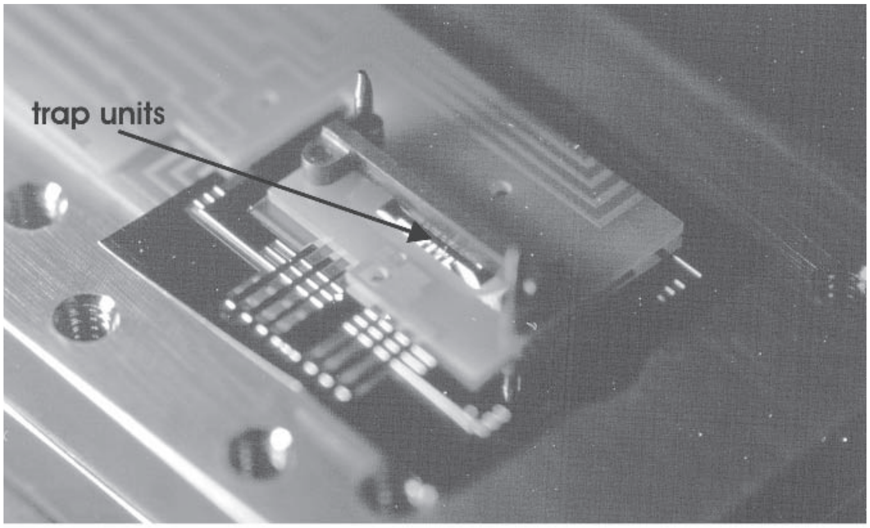

FIGURE 3 Gold-leaf trap. The electrodes are made of a thin gold foil (200 μm) clamped by two alumina wafers.

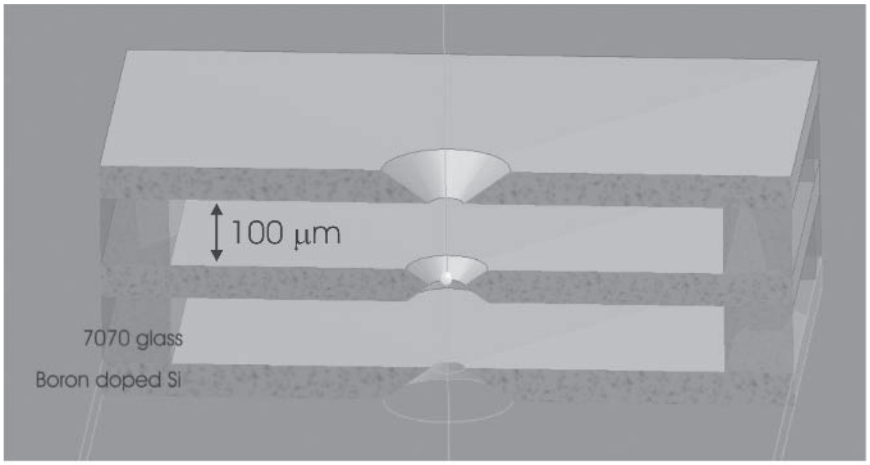

FIGURE 4 Schematic drawing of the prototype silicon trap for one ion (shown as a dot in the center hole). The silicon is doped with boron to make it conductive, and all substructures are fabricated by etching. The three boards are fixed to 7070 glass spacers by anodic bonding.

isolating spacers are made of Corning 7070 glass. The monocrystalline silicon electrodes should provide an atomically smooth, grain-boundary-free surface that might allow us to scale the trap size down from the 400 μm range currently used to about 40 μm with a greatly improved packing density and tight confinement of the trapped ions.

As we close in on our goal, we will have to tackle problems familiar to engineers. We will have to route smaller and more complex electrode structures and switch their voltages quickly with minimum cross talk. We will have to further miniaturize and integrate our approach using cutting-edge technologies. We would also like to integrate the laser optics on the chip using optical fibers and microoptics components. The ultimate ion-trap computer would only have to be hooked up to power supplies and laser light sources. Even those might someday be integrated using on-chip laser diodes.

REFERENCES

Diddams, S.A., Th. Udem, J.C. Bergquist, E.A. Curtis, R.E. Drullinger, L. Hollberg, W.M. Itano, W.D. Lee, C.W. Oates, K.R. Vogel, and D.J. Wineland. 2001. An optical clock based on a single trapped 199Hg$^+$ ion. Science 293: 825-828.

DiVincenco, D.P. 2001. The Physical Implementation of Quantum Computation. Pp. 7-13 in Scalable Quantum Computers: Paving the Way to Realization, S.L. Braunstein and H.K. Lo, eds. Berlin: Wiley-VCH.

Kielpinski, D., C.R. Monroe, and D.J. Wineland. 2002. Architecture for a large-scale ion-trap quantum computer. Nature 417: 709-711.

Monroe, C., D.M. Meekhof, B. E. King, W.M. Itano, and D.J. Wineland. 1995. Demonstration of a fundamental quantum logic gate. Physics Review Letters 75(25): 4714-4717.

Rowe, M.A., A. Ben-Kish, B. DeMarco, D. Leibfried, V. Meyer, J. Beall, J. Britton, J. Hughes, W.M. Itano, B. Jelenkovic, C. Langer, T. Rosenband, and D.J. Wineland. 2002. Transport of quantum states and separation of ions in a dual RF ion trap. Quantum Information and Computation 2(4): 257-271.

Sackett, C.A., D. Kielpinski, B.E. King, C. Langer, V. Meyer, C.J. Myatt, M. Rowe, Q.A. Turchette, W.M. Itano, D.J. Wineland, and C. Monroe. 2000. Experimental entanglement of four particles. Nature 404: 256-258.

Wineland, D.J., C. Monroe, W.M. Itano, D. Leibfried, B.E. King, and D.M. Meekhof. 1998. Experimental issues in coherent quantum-state manipulation of trapped atomic ions. Journal of Research at the National Institute of Standards and Technology 103(3): 259-328.

ADDITIONAL READING

Nielsen, M.A., and I.L. Chuang. 2000. Quantum Computation and Quantum Information. Cambridge, U.K.: Cambridge University Press.

Scalable Quantum Computing Using Solid-State Devices

BRUCE KANE
Laboratory for Physical Sciences
University of Maryland
College Park, Maryland

The field of solid-state quantum computing is in its infancy. Coherent operations on single qubits (the simplest type of quantum logical operation) have only recently been demonstrated (Nakamura et al., 1999; Vion et al., 2002). Nevertheless, there is a great deal of optimism that solid-state implementations of quantum computers will ultimately lead to scalable architectures in the same way that the invention of the transistor and integrated circuit presaged the development of large-scale and networked conventional computers. The optimism is based on the tremendous amount of research being done over a broad front that is heading steadily toward the development of devices for conventional computation built on nearly the atomic scale. This limit is the end of Moore's Law scaling, but only the threshold of the quantum realm. Thus, the nascent field of solid-state quantum computing can capitalize on research intended for the development of smaller and faster conventional logical devices.

Many difficulties will have to be overcome before qubits can be integrated into solid-state systems. Because there are on the order of 10^{23} atoms in a solid-state device, it is very difficult to attain the decoupling from extraneous degrees of freedom that is necessary for large-scale quantum computing. Hence, much of the early research on solid-state quantum computing has been focused on identifying potential qubits inherently isolated from their surroundings. Most current research is focused either on superconducting qubits (quantum information is stored on flux states in a SQUID or on charge states of a small "Cooper pair box") or on electron-spin or nuclear-spin qubits in semiconductors. To have the necessary long decoherence times, these devices must invariably operate at low temperatures (< 1 K). Even if research is successful, it is still not obvious that these technologies will be scalable (it is noteworthy that many conventional solid-state devices, such as bipolar transistors and tunnel diodes, proved to be unsuitable for very large-scale integrated circuits for reasons that only became apparent years after they were developed). Figure 1 shows the range of research being done on quantum computers.

A major surprise in the early days of quantum computing theory was that quantum error correction was possible at all; it has been shown that if a qubit of quantum information is redundantly coded into several qubits, errors in quantum computation can be reduced just as they can be corrected in classical communications channels (Nielsen and Chuang, 2000). One certainty is that the operation of scalable quantum computers will rely heavily on error correction. There is a "threshold for error corrected continuous quantum computation." When errors at the single-qubit-level quantum operations are reduced below this threshold, quantum computation becomes possible.

The error threshold is still very stringent. At most, one error can occur in

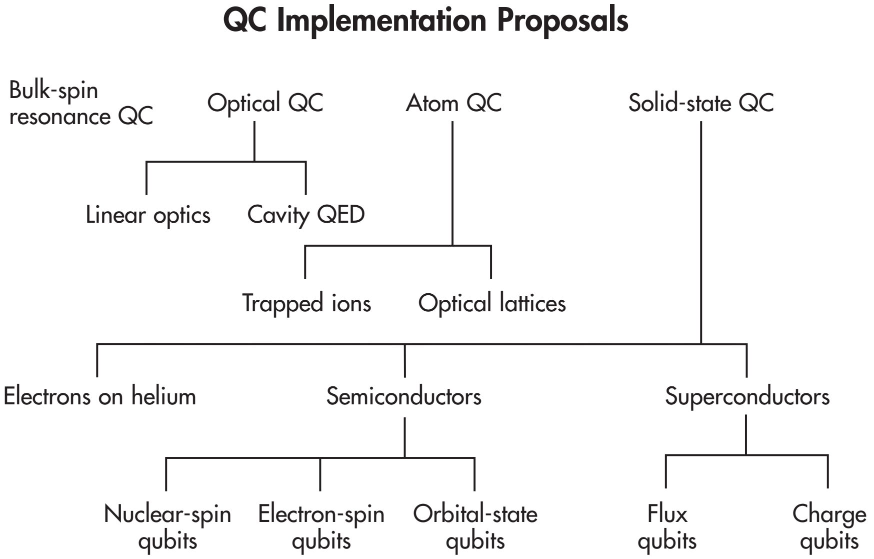

FIGURE 1 Diagram of the many possible realizations of quantum computers currently being explored experimentally. Bulk-spin resonance, optical, and atom-based qubits have all demonstrated elementary quantum logical operations, but solid-state implementations will particularly benefit from fabrication technologies being developed for conventional computer manufacturing. This synergy means that solid-state technologies are likely to be scalable and to benefit from continued advances in solid-state technology.

every ~10^4 operations. This level of accuracy is beyond the level of device variability of most, if not all, solid-state devices currently manufactured. Meeting the accuracy threshold will undoubtedly be one of the major challenges advocates for solid-state quantum computation will have to overcome. Even if the accuracy threshold can be reached, error correction will place a substantial overhead on the resources of a quantum computer because many physical qubits encode the same logical qubit. Also, most logical operations in the computer will simply be implementing error correction protocols (Steane, 2002).

KEY ISSUES

Scaling of quantum computing will undoubtedly be a formidable task that justifies the skepticism expressed by many people (Keyes, 2001). Nevertheless, an outline of the leading issues can be helpful for a preliminary assessment of current approaches to large-scale quantum computing. Many of these issues relate to information *flow* rather than to individual quantum operations on which most current research is focused.

Efficient on-chip quantum communication will be essential for the development of large-scale solid-state quantum computing. Communication between devices is also important in conventional computers, but the need for quantum error correction necessitates the continuous transfer of redundant qubits throughout the computer. Rapid flow of quantum information will thus be essential for scalable quantum computing.

Cross talk must be minimized. As the number of qubits increases, the potential for errors due to unwanted coupling between qubits also increases. Spatially localized qubits could minimize this type of error.

Parallel logical and measurement operations must be possible. We usually distinguish between the coherent logical operations of a quantum computer and the measurements of 0's and 1's that must ultimately produce the results of an algorithm. Error correction, however, requires that both types of operations go on together; in large-scale quantum computers they should occur simultaneously.

Logic, measurement, and communication must be able to be performed at high speed. Because of the exponential acceleration of quantum algorithms, a quantum computer operating at any clock speed will outperform its classical counterpart. However, given a choice between two otherwise equivalent quantum computer architectures, the faster architecture will be the most desirable.

Individual quantum devices must be precisely engineered. The accuracy threshold for quantum computing suggests that the variability between quantum logical devices must differ by no more than 1 part in 10^4.

MAJOR LIMITATIONS

As yet, no proposed implementation for large-scale quantum computers has addressed all of these criteria. However, it is possible to outline the major limitations of the approaches currently receiving the most attention.

Trapped-ion quantum computers. Although ion traps are not traditionally thought of as solid-state devices, recent proposals include complex metallizations on substrates to control and move ions between sites and are similar in many important ways to solid-state implementations. Kielpinski et al. (2002) have addressed many scaling issues, but moving ions to move quantum information is an intrinsically slow process because ion masses are typically 10^5 that of electrons. Therefore, it would be desirable if quantum information could be carried on lighter particles (like electrons).

Superconducting-qubit quantum computers. The primary advantage of superconducting qubits is that they are macroscopic. Consequently, devices are being fabricated with relatively simple micron-scale lithography. A drawback is that quantum information must be moved in the computer by electromagnetic excitations on metal traces. Large-scale implementations of superconducting quantum computers may thus be vulnerable to cross talk, a problem also encountered in conventional computers that have complex metal interconnects.

Electron-spin and nuclear-spin quantum computers. Quantum computer architectures that use electron spins or nuclear spins as qubits in a solid-state quantum computer are perhaps the most technologically challenging (Kane, 2000). These architectures also have some compelling advantages: spin qubits are known to be extremely well isolated from their environments in some materials and have longer quantum lifetimes (measured using electron-spin and nuclear-spin resonance techniques) than any other qubit under investigation in solids. Interestingly, silicon—the material in which almost all conventional computers are implemented—is also characterized by extremely long-lived spin states that can potentially be used for quantum computing. Spin coupling is extremely local, so cross talk is minimal if spins can be well isolated. Spin can be conveniently transported rapidly on electrons driven by electric fields (Skinner et al., 2002). Finally, parallel, rapid quantum logic and measurement appear to be possible using spin qubits (Figure 2).

The major difficulty with spin-based implementations of quantum computers is that they require measurement and control of single electron spins or nuclear spins, a task that is only now on the threshold of realization. Scalable spin-based quantum computing will probably require the development of a new technology in which single atoms can be accurately positioned to create devices with the precision necessary for quantum computation. Fortunately, several promising approaches (Figure 3) for fabricating these single-atom devices are currently being explored (O'Brien et al., 2001).

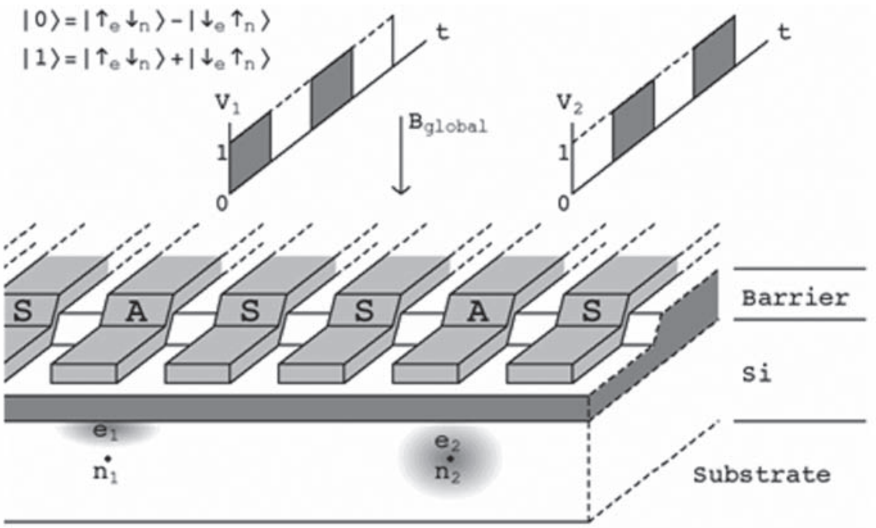

FIGURE 2 Diagram of a silicon-based quantum computer architecture. Single nuclear spins are the qubits; mobile electrons are used to manipulate the nuclear-spin states and move quantum information within the computer. Voltage pulses applied to metal gates on the top of the structure control the positions of electrons inside the silicon.

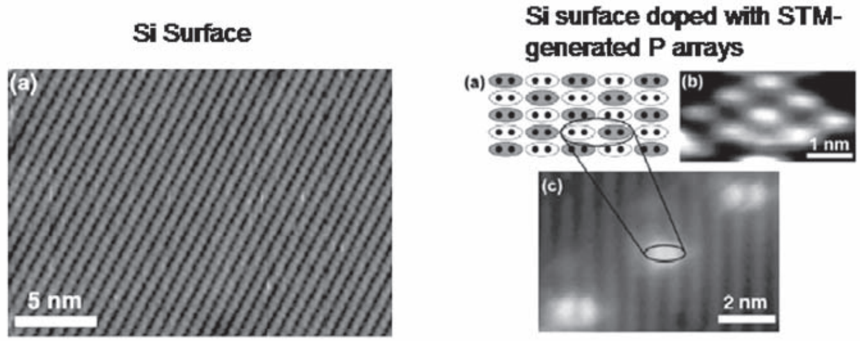

FIGURE 3 One possible route to the realization of single-spin quantum logical devices in silicon (Si) is via scanning tunneling microscope (STM) lithography. A single monolayer of hydrogen is formed on a pure Si surface (left). Using an STM, holes are punctured in the hydrogen; phosphorus is then introduced through the holes onto arrays in the Si crystal (right). Source: O'Brien et al., 2001. Reprinted with permission.

CONCLUSION

The development of large-scale quantum computers will require an unprecedented combination of precision and complexity. Extremely complex conventional information processors are only possible because digital logic is tolerant to device variation. Quantum logic is much less forgiving. Thus, realizing the dream of a large-scale quantum computer will require that engineers overcome the daunting challenge of combining the precision of an atomic clock with the complexity of a modern microprocessor.

REFERENCES

Kane, B.E. 2000. Silicon-based quantum computation. Fortschritte der Physik 48(9/11): 1023-1042.

Keyes, R.W. 2001. The cloudy crystal ball: electronic devices for logic. Philosophical Magazine B 81(9): 1315-1330.

Kielpinski, D., C. Monroe, and D.J. Wineland. 2002. Architecture for a large-scale ion-trap quantum computer. Nature 417(6890): 709-711.

Nakamura, Y., Y.A. Pashkin, and J.S. Tsai. 1999. Coherent control of macroscopic quantum states in a single-Cooper-pair box. Nature 398(6730): 786.

Nielsen, M.A., and I.L. Chuang. 2000. Quantum Computation and Quantum Information. Cambridge, U.K.: Cambridge University Press.

O'Brien, J.L., S.R. Schofield, M.Y. Simmons, R.G. Clark, A.S. Dzurak, N.J. Curson, B.E. Kane, N.S. McAlpine, M.E. Hawley, and G.W. Brown. 2001. Towards the fabrication of phosphorus qubits for a silicon quantum computer. Physical Review 64(16): 161401(R).

Skinner, A.J., M.E. Davenport, and B.E. Kane. 2002. Hydrogenic spin quantum computing in silicon: a digital approach. Available online at: <*http://arxiv.org/abs/quant-ph/0206159*>.

Steane, A.M. 2002. Quantum computer architecture for fast entropy extraction. Quantum Information and Computation 2: 297-306.

Vion, D., A. Aassime, A. Cottet, P. Joyez, H. Pothier, C. Urbina, D. Esteve, and M.H. Devoret. 2002. Manipulating the quantum state of an electrical circuit. Science 296(5569): 886-888.

Dinner Speech

The Science, Technology, and Business of Digital Communication

ANDREW J. VITERBI
Viterbi Group, LLC
San Diego, California

ABSTRACT

Wire-line telegraphy systems in the nineteenth century and wireless in the early twentieth century used rudimentary digital communication. Modern digital communication technology originated in the middle of the twentieth century and blossomed late in the century. The impetus was two-fold—solid-state integration that grew exponentially according to Moore's law and the development of system theories of information and communication. Together they made possible the sophisticated algorithms that enable efficient digital communications both via satellite and terrestrially. Advanced communication systems, which were first used for military and government satellites, became economically viable and universally available only in the 1990s. Digital satellite broadcasting and wireless cellular voice and data transmissions are the beneficiaries of this half-century of remarkable progress.

INTRODUCTION

A crude form of digital communication began at the turn of the twentieth century with Guglielmo Marconi's experiments. These early radio components generated pulses of energy of varying lengths, but not continuous waveforms. Analog communication really began with Lee De Forest's triode amplifier. But modern digital communication encompasses more than the transmission of waveforms representing 1's and 0's. It includes elaborate processing of information to maximize the efficiency and accuracy of the message, whether it is audio, visual, or textual. Processing goes well beyond the capabilities of simple analog modulation. This phase of the development of digital communication dates from the late 1940s, when two groundbreaking events took place within months of each other in the same complex of Bell Telephone Laboratory buildings in Murray Hill, New Jersey. The first was the development of the transistor and the birth of solid-state electronics. The second was the founding of the field of information theory with its remarkable implications for communication.

SCIENCE

The scientific basis for modern digital communication derives in equal measure from physical principles and mathematical concepts. Electromagnetic theory provides the foundation for all forms of electrical communications, wired or wireless. This scientific field originated in the eighteenth century and flourished in the nineteenth, culminating in James Clerk Maxwell's equations and Heinrich Hertz's propagation experiments. But modern digital communication was enabled by another field of physical research, solid-state electronics, which was developed in the middle of the twentieth century. The development of the transistor in 1947 led two decades later to the beginning of solid-state circuit integration with the ensuing exponential growth in digital processing and memory capabilities. The mathematical origins of digital communication theory are as remote as the Gauss, Euler, Fourier, and Laplace papers of the eighteenth and early nineteenth centuries and as recent as Shannon's theory of information in the mid-twentieth century.

Following the development of the transistor in 1947 at Bell Laboratories by Bardeen, Brattain, and Shockley, for which they earned a Nobel Prize, numerous government and commercial efforts were undertaken to replace bulky, power-hungry, and vulnerable vacuum tubes. William Shockley left Bell Laboratories to found Shockley Laboratories, which lost its key researchers to Fairchild Corporation, from which emerged the founders of Intel Corporation, the prime mover in the creation of the personal computer through the development of the microprocessor. In fact, in 1965, Intel's cofounder Gordon Moore foresaw the exponential growth in solid-state integration and predicted that device density (the number of devices per silicon integrated circuit) would double every 18 months. Although Moore's law is based on qualitative socioeconomic arguments rather than quantitative physical theories, it has proven to be amazingly accurate. In fact, from 1965 to 2002, the growth rate was slightly ahead of the 2^{25} growth rate Moore predicted. The decrease in cost and power consumption has been proportional to the increase in device density. Indeed, the concept of "system-on-a-chip" has become commonplace in large-volume electronic manufacturing in the last few years. The increase in device speeds has also been exponential, rising from kilo-operations per second (K ops) in the 1960s to M ops in the 1980s to G ops today.

This physical capability would be in search of an application were it not for concurrent advances in system concepts rooted in applied mathematics. Although this field is the product of great European mathematicians of previous centuries, the impetus for the great strides in communication theory in the twentieth century was largely the work of midcentury American mathematicians, mostly the results of war-related research. The three prophets of modern communication theory were Norbert Wiener, Stephen Rice, and Claude Shannon. Wiener, a professor at MIT is often cited as the father of cybernetics. In 1949, he published a monograph entitled *Extrapolation, Interpolation and Smoothing of Stationary Time Series*, known to students as the Yellow Peril for the color of its binding. This work, which was grounded in both harmonic analysis and mathematical statistics, influenced the design of early radar and analog communication systems. Rice's work at Bell Laboratories, which was published in 1944 in a paper in the *Bell System Technical Journal* entitled "Mathematical Theory of Noise," applied random-process theory to the communication problem. In 1948, his colleague Claude Shannon, who had spent the war years theorizing on cryptographic concepts, published papers in two issues of the same journal entitled "A Mathematical Theory of Communication," which introduced startling new concepts that went well beyond any previously well established theory. (In fact, initially it was underestimated or misunderstood by both physicists and mathematicians).

Lest we appear nationalistic in heralding this purely American school, we should note that there was a nearly parallel Russian school led by Khinchine, Kotelnikov, and Kolmogorov that produced approximately parallel results. The practical applications of their work, however, never achieved wide acceptance or had the same impact. In the ensuing decades, schools of communication theory emerged in Hungary, Israel, Canada, Sweden, Italy, Germany, France, and Switzerland. The Hungarians followed the Russian model; all the rest followed the American model.

Shannon's theories, although difficult to master, are very easy to describe in terms of the less than obvious answers to two basic questions about the limits of digital communication. The first question, "compression" or "source coding," asks how few bits per second are required for the faithful reproduction of a source (voice, video, or text). The second, "transmission" or "channel coding," asks how many bits per second can be accurately transmitted over a noisy and otherwise impaired medium. The two papers published by Shannon in 1948 fully answered both questions! It took nearly half a century, however, to develop communication systems that have almost reached the performance predicted by Shannon.

TECHNOLOGY

Solid-state integrated circuitry, the technology that physically enabled advanced digital communication, would not have been possible without the evolution of computing capabilities. Integration dramatically lowered the price of computation to the level of consumer appliances. This enormous expansion of the market resulted in economies of scale that further reduced prices. But the requirements of digital communication processors go beyond the processing speed and memory requirements of a computation processor. Best described as real-time digital-signal processing, it includes the basic operations of Fourier transforms and statistical operations, such as likelihood function computations involved in demodulation and decoding.

The initial impetus for the development of these technologies came from the U.S. government. Starting in the 1950s, military research and development (R&D) by government agencies explored digital communications for their added security, covertness, and suppression of intentional interference (jamming). Similar requirements and techniques were simultaneously evolving for radar, which reinforced the communication research. In the 1960s and 1970s, NASA funded efforts to improve the efficiency of communication to and from space vehicles and satellites. The twin drivers were the minimizing of weight in satellite orbit and the maximizing of the communication range of space vehicles for a given earth antenna size (or conversely minimizing the antenna size for a given range). Government-sponsored R&D was performed by a variety of organizations, ranging from Defense Department and NASA laboratories to federally contracted research centers to universities and private contractors large and small. Later, through the 1990s, commercially motivated R&D led to great advances in wire-line modems, data-transmitting and broadcasting satellite systems operating with very small-aperture antennas (VSATs); in the diffusion of Internet connectivity; and in the explosive growth of wireless telephony.

BUSINESS

Large-scale commercialization of digital communication in the 1980s and 1990s spawned at least five interrelated industries in approximately the following order.

In the 1970s, **wire-line modems**, the first personal modems, transmitted in the low K bits per second. Since then, the combination of improved lines and sophisticated signal-processing algorithms has increased transmission rates to M bits per second for digital subscriber loop (DSL) service. At the same time, coaxial cable systems installed to provide analog television now carry data at a rate of several M bits per second or several channels of digital television in the same bandwidth originally occupied by one analog channel.

Starting in the 1960s, **communication satellites** began transmitting analog programming to cable TV head-ends. Each program occupied one satellite transponder and required antennas several meters in diameter. With digital compression and signal processing, as well as high-power satellites, the number of programs per channel increased more than four-fold, and antenna diameters were reduced to fractions of a meter, opening up a large consumer market.

The virtually unlimited bandwidth of **fiber-optic networks** developed in the last quarter of the twentieth century created an ideal foundation for networks carrying all forms of communication. **Packet switching**, which was developed through military research in the 1950s, provides tremendous flexibility and efficiency for digital transmission.

The greatest beneficiary of all has been the **Internet**. Conceived as the ARPANET in the late 1960s for sharing programs, processing capabilities, and databanks among a few dozen laboratories and universities working on defense research, the Internet now connects countless nodes and services hundreds of millions of users through the Worldwide Web.

The most recent and most widespread digital communication application, cellular **wireless telephony,** now serves more than one billion subscribers, one-sixth of the world population. Launched in the 1980s as an analog technology (basically two-way FM radio), in the 1990s, second-generation digital technology increased the market a hundred-fold. Some of the most sophisticated compression and transmission technologies have reduced the size, lowered the cost, and reduced the power consumption (and hence increased battery life) of cellular phones. In addition to cellular networks, also known as wireless wide-area networks (WANs), wireless local-area networks (LANs) are now used in homes, business enterprises, and urban "hot spots." Finally, wireless personal-area networks (PANs), which transmit over a distance of a few meters, avoid cables and wires within a home or workplace.

The two major wireless technologies provide a study in contrasts. Communication satellites in geosynchronous orbit must transmit a wide bandwidth signal over a range of 40,000 kilometers resulting in a very low signal-to-noise ratio. By contrast, terrestrial wireless transmission has a range of a few kilometers at most, but each user's signal must contend with interference from a multitude of other users' signals arriving simultaneously at the same base station or access point. A particularly successful wireless cellular technology, known as spread-spectrum or code-division multiple access (CDMA), reduces the interfering signal to appear as wideband noise, similar to the satellite receiver's noise, which is of thermal origin. Thus, the receiver processing technologies of these two widely disparate digital communication technologies are surprisingly similar.

The Wireless Cellular Industry

I will conclude with a brief history of the biggest proliferation of digital communication devices. In less than a decade, the worldwide wireless cellular industry has put into service more than a billion digital cellular phones and data terminals. The evolution of this industry has been defined in terms of product

TABLE 1 Wireless Generations Defined

1G	Analog Voice	1980s	
2G	Digital Voice and 10 Kbps Circuit Switched Data	1990s	TDMA, GSM, CDMA
3G	Digital Voice and Packet Data: 384 Kbps to 2 Mbps	2000+	CDMA 2000, WCDMA

generation by standards bodies and the media (Table 1). The first generation of cellular handsets and infrastructure involved purely analog transmission and processing. Service began in the early 1980s and was well established in North America by 1990, when digital handsets and the infrastructure of the second generation (2G) were introduced. In Europe, where analog wireless technology had not been widely adopted, the new digital system, designated Global System for Mobiles (GSM), was highly successful largely because it was standardized by the European Telecommunication Standards Institute (controlled by major wireless carriers and manufacturers) so it could be adopted throughout the European Union. European subscribers could maintain service while roaming over a large geographical area. The time-division multiple-access (TDMA) transmission technology of GSM was emulated in North America (IS-54) and Japan (PDC), although with much narrower bandwidths. All three systems failed to achieve the industry's goal of increasing bandwidth efficiency over analog by a factor of ten. PDC was adopted only in Japan; the North American TDMA system attracted limited support, mainly in Latin America, but did not displace analog service, even in the United States.

Code-division multiple-access (CDMA), the only digital technology that fulfilled the goal of bandwidth efficiency, was proposed in the late 1980s. After initial industry resistance, CDMA was accepted in 1993 as an alternate North American standard (IS-95 or CDMA One) and entered service in the mid-1990s. South Korea adopted CDMA as its sole digital standard, began service in 1996, and was remarkably successful, not only in penetrating more than half of the population but also in converting the nation from an importer of technology to a major exporter of technology. Ultimately, CDMA service was offered by one or more 2G wireless service providers in most nations following the American and Korean example, usually in competition with providers of GSM or other TDMA services. The notable exception was Western Europe, where competition was excluded by the EU's regulatory adherence to GSM.

When the International Telecommunications Union (ITU) set about selecting an access technology for the so-called third generation (3G) to provide for higher speed data and further efficiencies in voice telephony, it approved two enhanced forms of CDMA. The fundamental reason for this choice is shown in

TABLE 2 2G Voice Efficiencies

Relative number of users in given bandwith	
Analog	1
GSM	3 to 4
TDMA	4 to 5
CDMA	10 to 12
Beyond 2G:	
CDMA 2000-1x	21
WCDMA	?

Source: Seybold, 2001

Table 2, which compares various 2G technologies (based on research by a noted wireless analyst). The basis of comparison is the number of digital voice users that can be supported in the same bandwidth required to support one analog voice conversation.

The principal difference between the two 3G versions, CDMA 2000 and WCDMA, is that the former is a direct evolution of the 2G version CDMA One and therefore requires synchronization of the (base station) infrastructure; the latter, WCDMA, puts the burden on subscriber handsets through increased processing and consequently higher power consumption. The CDMA 2000 subscriber count is already more than twenty million and has demonstrated a near doubling in capacity over that of the 2G CDMA System. WCDMA is off to a slow start, partly because of technical difficulties, but largely because of the poor financial health of European carriers caused by the huge debt they incurred as a result of excessive payments for the 3G spectrum. This cost was mostly avoided by North American and Asian service providers who operate 3G CDMA 2000 in the same spectrum as 2G CDMA One, from which it evolved.

The ultimate success of 3G, however, will depend on the benefits it provides to subscribers, in the form of new and enhanced applications, and to service providers, in increased revenue and better returns on infrastructure investments. Four benefits are already evident:

- The use of spectrum (which has become a costly resource) for voice and data is much more efficient.
- Through greatly increased data transmission speed and, consequently, reduced latency, the sharing of video clips and still photos has become more appealing (as has already been demonstrated in some Asian countries). Even real-time game playing over the wireless Internet is being proposed.

- For the same reason, the downloading of Web pages to PCs or PDAs away from one's office or home is also more attractive.
- Enterprises with wireless LANs can extend them seamlessly to remote locations, even with mobility.

With these advantages, the upgraded digital wireless industry is certain to rebound from its current recession, although more slowly than proponents have envisioned.

REFERENCE

Seybold, A. 2001. Silicon Insights: Spectral Efficiency. Available online at: <*http://abcnews.go.com/sections/business/DailyNews/silicon_insights_seybold_010716.html*>.

Appendixes

Contributors

Marvin Lee Adams is a professor of nuclear engineering at Texas A&M University. Prior to joining the faculty there, he was a nuclear engineer with the Tennessee Valley Authority and a code physicist at Lawrence Livermore National Laboratory. He serves on numerous advisory committees and panels with the U.S. Department of Energy, and its laboratories, and has served on the United States-Russia Joint Technical Working Group on options for the disposition of weapons plutonium. Dr. Adams is a reviewer for several technical journals and serves on the editorial board of *Transport Theory and Statistical Physics*. He is a fellow of the American Nuclear Society and the recipient of numerous awards, including Texas A&M University Faculty Fellow (2001) and Montague Center for Teaching Excellence Scholar (1995). He received a B.S. from Mississippi State University and an M.S.E. and Ph.D. from the University of Michigan, all in nuclear engineering. (*mladams@tamu.edu*)

James P. Blanchard is a professor in the Engineering Physics Department at the University of Wisconsin-Madison (UW). His fields of interest are radiation damage, fusion technology, inertial fusion, reactor component lifetime, solid mechanics, and nuclear microbatteries. Dr. Blanchard is the recipient of the National Science Foundation Presidential Young Investigator Award and the UW Chancellor's Distinguished Teaching Award. He received a Ph.D. in nuclear engineering from the University of California, Los Angeles, in 1988. (*blanchard@ engr.wisc.edu*)

Mary Czerwinski is a research manager in the Large Display User Experience Group at Microsoft Research. Her group is responsible for studying and designing advanced technology that leverages human capabilities across a wide variety of input and output channels. Dr. Czerwinski's primary research areas include spatial cognition, multimodal user-interface design, and the intelligent design of notifications. She has been an affiliate assistant professor at the Department of Psychology, University of Washington since 1996. She has also held positions at Compaq Computer Corporation, Rice University, Lockheed Engineering and Sciences Corporation, and Bell Communications Research. Dr. Czerwinski received a Ph.D. in cognitive psychology from Indiana University in Bloomington. She is active in the field of Human-Computer Interaction, publishing and participating in a wide number of conferences, professional venues, and journals. (*marycz@microsoft.com*)

David Lee Davidson passed away suddenly on October 27, 2002. He was a fellow at Solutia, Inc., in Pensacola, Florida, where he was responsible for product and process development for fibers, polymers, and chemicals. He joined Monsanto/Solutia in 1992; prior to that, he held positions at Westinghouse Idaho Nuclear Company, Hercules, Inc., and Air Products and Chemicals, Inc. Dr. Davidson received a B.S. from the University of Delaware, an M.S. from the University of Massachusetts, and a Ph.D. from Princeton University, all in chemical engineering.

Juan J. de Pablo is Howard Curler Distinguished Professor in the Chemical Engineering Department at the University of Wisconsin-Madison. His fields of interest are molecular thermodynamics and statistical mechanics. Dr. de Pablo was recipient of a National Science Foundation (NSF) Presidential Young Investigator Award, an NSF Presidential Early Career Award for Science and Engineering, and the Camille Dreyfus Teacher-Scholar Award, among others. He received a B.S. from the Universidad Nacional Autónoma de México and a Ph.D. from the University of California at Berkeley. (*depablo@ engr.wisc.edu*)

Thomas A. Dingus is director, Virginia Tech Transportation Institute, and Newport News Shipbuilding/Tenneco Professor, Charles E. Via, Jr. Department of Civil and Environmental Engineering at Virginia Polytechnic Institute and State University. Prior to joining Virginia Tech, he was an associate director, University of Iowa Center for Computer-Aided Design, where he was responsible for the administration of the human factors research program associated with the Iowa Driving Simulator. He was also the founding director of the National Center for Advanced Transportation Technology at the University of Idaho. Dr. Dingus received a Ph.D. in industrial engineering and operations research from Virginia Tech (1987). Prior to attending graduate school, he spent one year as a research scientist at the Air Force Human Resources Laboratory Advanced Simulation Techniques Branch and four years as a human factors engineer and senior human factors engineer for Martin-Marietta Aerospace. He is the author of more than 120 scientific articles and technical reports. His research on intelligent-vehicle highway systems, driver attention demand, driver workload, advanced information display design, human factors, and safety has been supported by numerous companies and government agencies. Dr. Dingus has also been involved in the specification, design, and construction of numerous instrumented research vehicles, has conducted 12 major on-road instrumented vehicle studies, and has been involved in nine major driving-simulation studies. He is a member of the Intelligent Transportation Society of America, the American Society of Safety Engineers, and the Human Factors and Ergonomics Society. (*tdingus@vtti.vt.edu*)

Peter S. Hastings is an engineering manager with Duke Energy in Charlotte, North Carolina. He has 20 years' experience in engineering and management in the commercial nuclear industry, both in industry and with the U.S. Department of Energy (DOE). As manager of licensing and safety analysis for the Mixed Oxide Fuel Fabrication Facility, he oversees implementation for the construction authorization and possession-and-use license for the facility and the Integrated Safety Analysis (ISA), and participated in the rule-making and development of regulatory guidance for the recent change to 10 CFR Part 70. Prior to his current assignment, he established and oversaw processes required for licensing the nation's first high-level radioactive-waste repository and managed nuclear safety analyses and long-term performance assessment for that project. He established the National Regulatory Commission (NRC) licensing and programmatic basis for the repository's preclosure nuclear safety and accident analysis program and established DOE's program for assessing long-term performance impacts during site characterization to meet NRC requirements. In addition, he has several years of experience in nuclear station operations, start-up testing, surveillance, and design engineering. Mr. Hastings received a B.S. in nuclear engineering from North Carolina State University. *(pshastings@dukeengineering.com*)

Bruce E. Kane is a member of the research staff in the Department of Physics and Laboratory for Physical Sciences at the University of Maryland, College Park. He has a B.A. in physics from the University of California, Berkeley, and a Ph.D. in physics from Princeton University. Dr. Kane, developer of the silicon-based quantum computer concept, began his research career as an experimentalist in 1987 in the Princeton laboratory of D. C. Tsui, winner of the 1998 Nobel Prize in physics. Since his arrival at the University of Maryland in 1999, Dr. Kane has devoted his attention to developing experimental methods for single spin detection using single electron transistors and scanned probe techniques on doped silicon devices. He hopes to demonstrate single spin detection and rudimentary quantum logic in these devices in the next few years. (*bekane@umd.edu*)

John F. Kotek is affiliated with the Advanced Reactor Programs at Argonne National Laboratory-West. In December 2002, he completed a one-year congressional fellowship in the office of Senator Jeff Bingaman (D-NM), chairman, Senate Energy and Natural Resources Committee. Mr. Kotek assisted Senator Bingaman on energy and national security-related issues and helped prepare the Senate's comprehensive energy legislation. At Argonne, he managed Argonne-West's participation in the Generation IV program, which focuses on the development of next-generation nuclear energy systems. Prior to joining Argonne, he spent nine years at the U.S. Department of Energy Office of Nuclear Energy, Science and Technology, where he held numerous positions, including, manager, university support programs; assistant manager, medical isotopes production project; chief of staff; associate director for management and administration; and associate director for technology. He received a B.S. in nuclear engineering from the University of Illinois at Urbana-Champaign and an M.B.A. from the University of Maryland. (*john.kotek@anlw.anl.gov*)

Dietrich Leibfried is a research associate at the National Institute of Standards and Technology (NIST) and the University of Colorado in Boulder, where he is coleader of a research project on quantum information processing with trapped Be^+ ions. Prior to this, he was a postdoctoral fellow at the University of Innsbruck, a guest researcher at NIST, and a staff scientist at the Max Planck Institute in Garching, Germany. Dr. Leibfried received an M.S. and Ph.D. in physics from the Ludwig-Maximilians-Universität in Munich, Germany. He has received numerous awards, including the Helmholtz Award from the Federal Technical Institute (PTB) in Braunschweig, Germany, for a new determination of the Rydberg constant, and a START award from the Austrian Fonds zur Förderung der Wissenschaften, the highest award in Austria for junior researchers, for work in the field of quantum information. Dr. Leibfried is a member of the German Physical Society and the Optical Society of America, a reviewer for several journals, and the author of numerous articles published in *Physics Today* and *Physikalische*, the corresponding journal of the German Physical Society. (*dil@boulder.nist.gov*)

Melody M. Moore is an assistant professor in the Computer Information Systems Department, College of Business Administration, Georgia State University (GSU). Her research interests are in the areas of brain-computer interfaces, software evolution, and user interface reengineering. For nine years prior to joining the GSU faculty, Dr. Moore was on the faculty of the College of Computing, Georgia Institute of Technology, where she directed the Open Systems Laboratory and taught software engineering. Before entering academia, she worked for nine years in industry at Texas Instruments, Sperry, and National Semiconductor as a professional software engineer developing real-time embedded systems, secure operating systems, networking, and compilers. She received a B.A. in computer science from the University of Texas at Austin (1980) and an M.S. in information and computer science and a Ph.D. in computer science from Georgia Institute of Technology. *(melody@gsu.edu*)

Steven J. van Enk is a member of the technical staff at Bell Laboratories/Lucent Technologies in Murray Hill, New Jersey. Prior to assuming this position, he had postdoctoral fellowships at the Max-Planck Institute in Germany, the Theoretical Institute at the University of Innsbruck, and the California Institute of Technology. His research interests are in quantum communication and quantum information (physical implementation, quantum cryptography, teleportation) and quantum optics and QED (mechanical effects of light, angular momentum of light, and the Casimir effect). Dr. van Enk received an M.S. from the University of Utrecht and a Ph.D. from the University of Leiden in the Netherlands, both in physics. (*vanenk@research.bell-labs.com*)

Kim J. Vicente is professor in the Department of Mechanical and Industrial Engineering and founding director of the Cognitive Engineering Laboratory, University of Toronto. He also holds appointments in the Institute of Biomaterials and Biomedical Engineering and the Department of Computer Science, University of Toronto. In 2002-2003, he is Jerome Clarke Hunsaker Distinguished Visiting Professor of Aerospace Information Engineering at the Massachusetts Institute of Technology. Dr. Vicente received a B.A.Sc. in industrial engineering from the University of Toronto (1985), an M.S. in industrial engineering and operations research from the Virginia Polytechnic Institute and State University (1987), and a Ph.D. in mechanical engineering from the University of Illinois at Urbana-Champaign (1991). He has held positions at the Section for Informatics and Cognitive Science of the Risø National Laboratory in Roskilde, Denmark, and Georgia Institute of Technology. His interests include the design of interfaces for complex sociotechnical systems, the study of expertise, and the analysis and design of complex work environments. He is on the editorial boards of several journals and a member of the Committee for Human Factors of the National Research Council/The National Academies. He has applied his extensive research on cognitive work analysis and human-computer interface design for complex sociotechnical systems to a number of diverse domains, including animation, aviation, engineering design, medicine, network management, nuclear power, and petrochemical processes. Dr. Vicente is the author of *Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-based Work* (Lawrence Erlbaum Associates, 1999), the first textbook in the area of cognitive work analysis. In 1999, he was one of the 25 Canadians under the age of 40 chosen by *Time Magazine* as “Leaders for the 21st century who will shape Canada’s future.” (*vicente@mie.utoronto.ca*)

Andrew J. Viterbi is president of the Viterbi Group, LLC, which advises and invests in start-up companies, predominantly in the wireless communications and network infrastructure fields. In July 1985, Dr. Viterbi cofounded QUALCOMM, Inc., a developer and manufacturer of mobile satellite communications and digital wireless telephony; he was chief technical officer until 1996, and vice chairman of the company until 2000. Previously, in 1968, Dr. Viterbi cofounded LINKABIT Corporation, a digital communications company; he was executive vice president and president in the early 1980s. From 1963 to 1973, Dr. Viterbi was professor, School of Engineering and Applied Science, University of California, Los Angeles, where he conducted fundamental work in digital communication theory and wrote numerous research papers and two books. He continued teaching on a part-time basis at the University of California, San Diego, until 1994, and is currently professor emeritus. From 1957 to 1963, Dr. Viterbi was a member of the Communications Research Section of the California Institute of Technology Jet Propulsion Laboratory. Dr. Viterbi received a B.S. and M.S. from the Massachusetts Institute of Technology (MIT) in 1957, and a Ph.D. from the University of Southern California in 1962. He has received numerous awards and international recognition for his leadership and substantial contributions to communications theory and its industrial applications over the years. He is a Fellow of the IEEE, a Marconi Fellow, and a member of the National Academy of Engineering, the National Academy of Sciences, and the American Academy of Arts and Sciences. All four international standards for digital cellular telephony use the Viterbi algorithm for interference suppression, as do most digital satellite communication systems, both for business applications and for direct satellite broadcast to residences. (*andrew.viterbi@ viterbigroup.com*)

John M. Vohs is the Carl V.S. Patterson Professor and chair of the Department of Chemical and Biomolecular Engineering at the University of Pennsylvania. He joined the faculty there after receiving a B.S. degree from the University of Illinois and a Ph.D. from the University of Delaware. Dr. Vohs' research interest is in the field of surface and interfacial science, particularly the relationships between the local atomic structure of surfaces and their chemical reactivity. His work on structure-activity relationships for metal-oxide catalysts, especially those used for selective oxidation reactions and automotive emissions control systems, is widely known. In recent years, he has collaborated in the development of solid-oxide fuel cells that run on readily available hydrocarbon fuels, such as natural gas and diesel. Dr. Vohs has received numerous honors, including an NSF Presidential Young Investigator Award and two Union Carbide Research Innovation Awards. (*vohs@seas.upenn.edu*)

Program

NATIONAL ACADEMY OF ENGINEERING

Eighth Annual Symposium on
Frontiers of Engineering
September 19–21, 2002

CHEMICAL AND MOLECULAR ENGINEERING IN THE 21ST CENTURY
Organizers: Pablo Debenedetti and Brigette Rosendall

Fuel Cells That Run on Common Fuels
John M. Vohs, University of Pennsylvania

Dimension-Dependent Properties of Macromolecules in Nanoscopic Structures
Juan J. de Pablo, University of Wisconsin-Madison

The Role of Computational Fluid Dynamics in Process Industries
David Lee Davidson, Solutia, Inc.

* * *

TECHNOLOGY FOR HUMAN BEINGS
Organizers: Ann Bisantz and Rick Kjeldsen

The Human Factor
Kim J. Vicente, University of Toronto

Human Factors Applications in Surface Transportation
Thomas A. Dingus, Virginia Polytechnic Institute and State University
(talk given by Vicki Neale)

Implications of Human Factors Engineering for Novel Software User-Interface Design
Mary Czerwinski, Microsoft

Frontiers of Human-Computer Interaction: Direct-Brain Interfaces
Melody M. Moore, Georgia State University

* * *

DINNER SPEAKER

The Science, Technology, and Business of Digital Communication
Andrew J. Viterbi, President, Viterbi Group, LLC

* * *

THE FUTURE OF NUCLEAR ENERGY

Organizers: Kathryn McCarthy and Per Peterson

Advanced Nuclear Reactor Technologies
John F. Kotek, Argonne National Laboratory-West

Licensing and Building New Nuclear Infrastructure
Peter S. Hastings, Duke Energy

Sustainable Energy from Nuclear Fission Power
Marvin L. Adams, Texas A&M University

Stretching the Boundaries of Nuclear Technology
James P. Blanchard, University of Wisconsin-Madison

* * *

ENGINEERING CHALLENGES FOR QUANTUM INFORMATION TECHNOLOGY

Organizers: Ike Chuang and Hideo Mabuchi

Quantum Cryptography
Steven J. van Enk, Bell Labs, Lucent Technologies

Ion-Trap Quantum Computation
Dietrich Leibfried, National Institute of Standards and Technology

Scalable Quantum Computing Using Solid-State Devices
Bruce Kane, University of Maryland

Participants

NATIONAL ACADEMY OF ENGINEERING

Eighth Annual Symposium on
Frontiers of Engineering
September 19-21, 2002

Gregory D. Abowd
Associate Professor
College of Computing
Georgia Institute of Technology

Marvin L. Adams
Professor
Department of Nuclear Engineering
Texas A&M University

Charles H. Ahn
Assistant Professor
Department of Applied Physics
Yale University

Wole C. Akinyemi
Technical Specialist
Cummins, Inc.

Phillip A. Armstrong
Project Manager
Air Products and Chemicals, Inc.

Zachi Baharav
Researcher
Research and Development
Agilent Laboratories

Jane Bare
Chemical Engineer
Office of Research and Development
U.S. Environmental Protection Agency

William C. Beavin
Associate Technical Fellow
Boeing Company

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

Gary H. Blackwood
StarLight Project Manager
Jet Propulsion Laboratory

James P. Blanchard
Professor
Department of Engineering Physics
University of Wisconsin-Madison

Matt Blaze
Research Scientist
AT&T Laboratories

Eric T. Boder
Assistant Professor
Department of Chemical Engineering
University of Pennsylvania

Christopher N. Bowman
Professor and Gillespie Faculty Fellow
Department of Chemical Engineering
University of Colorado

Walter F. Buell
Research Scientist
The Aerospace Corporation

Paul R. Bunch
Manager, Discovery Operations
Lilly Research Laboratories
Eli Lilly and Company

Timothy J. Bunning
Senior Materials Research Engineer
Materials and Manufacturing Directorate
Air Force Research Laboratory

Peter J. Burke
Assistant Professor
Department of Electrical and Computer Engineering
University of California, Irvine

James P. Calamita
Member of Research and Technology Staff
Xerox Corporation

Elizabeth D. Carey
Executive Director, Engineering
Cummins, Inc.

Scott B. Carlson
Vice President, Division Manager
SAIC

Jeffrey J. Chalmers
Professor
Department of Chemical Engineering
Ohio State University

Gang Chen
Associate Professor
Department of Mechanical Engineering
Massachusetts Institute of Technology

Yen-Kuang Chen
Staff Researcher, Distributed Signal Processing and Exploratory Architecture
Microprocessor Research Labs
Intel Corporation

Naomi C. Chesler
Assistant Professor
Department of Biomedical Engineering
University of Wisconsin-Madison

Isaac Chuang
Associate Professor
Media Laboratory
Massachusetts Institute of Technology

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

Kevin D. Costa
Assistant Professor
Department of Biomedical Engineering
Columbia University

Victoria L. Coverstone
Associate Professor
Department of Aeronautical and Astronautical Engineering
University of Illinois at Urbana-Champaign

Mary Czerwinski
Research Manager
Large Display User Experience Group
Microsoft Research

Phillippe A. Daniel
Vice President
CDM

David Lee Davidson
Fellow
Solutia, Inc.

Juan J. de Pablo
Howard Curler Distinguished Professor
Department of Chemical Engineering
University of Wisconsin-Madison

Pablo G. Debenedetti
Class of 1950 Professor and Chair
Department of Chemical Engineering
Princeton University

Casimer M. DeCusatis
Senior Engineer
IBM Corporation

Rachelle Delaney
Safety Engineer
Ford Motor Company

David W. DePaoli
Group Leader
Oak Ridge National Laboratory

Thomas A. Dingus *(unable to attend)*
Director, Virginia Tech Transportation Institute
Virginia Polytechnic Institute and State University

Joseph A. Donndelinger
Senior Research Engineer
General Motors R&D Center

Rhonda Franklin Drayton
Assistant Professor
Department of Electrical and Computer Engineering
University of Minnesota

Stefan M. Duma
Assistant Professor, Mechanical Engineering
Director, Impact Biomechanics Laboratory
Virginia Polytechnic Institute and State University

James A. Dyer
Consultant
DuPont Engineering Technology

Matthew Franklin
Associate Professor
Department of Computer Science
University of California, Davis

Roger H. French
Senior Research Associate
DuPont Co. Central Research

Michael R. Furst
Research Manager
Xerox Innovation Group/Research & Technology
Xerox Corporation

Paul A. Gillis
Research Scientist
Texas Operations
Dow Chemical Company

Stuart Goose
Research Project Manager
Siemens Corporate Research

Anand K. Gramopadhye
Professor
Department of Industrial Engineering
Clemson University

Augusto L. Gutierrez-Aitken
Assistant Manager, Technology Development
TRW

Robert T. Hanlon
Senior Development Engineer
Engineering Division
Rohm and Haas Company

Peter S. Hastings
Engineering Manager
Duke Energy

Kayleen L.E. Helms
Senior Packaging Engineer
Advanced Technology Development
Intel Corporation

Nicola A. Hill
Associate Professor
Materials Department
University of California, Santa Barbara

Klaus-Dieter Hilliges
Research Project Manager
Agilent Laboratories

Joseph B. Hughes
Professor and Chair
Department of Civil and Environmental Engineering
Rice University

Andrew T. Hunt
CEO, CTO, and Founder
MicroCoating Technologies, Inc.

Glen C. Irvin, Jr.
Senior Engineer
Eastman Kodak Company

Jingyue Ju
Associate Professor and Head of DNA Sequencing and Chemical Biology
Columbia Genome Center and Department of Chemical Engineering
Columbia University

Bruce E. Kane
Researcher
Department of Physics and Laboratory for Physical Sciences
University of Maryland

Thomas E. Kazior
Senior Principal Engineer
Raytheon RF Components

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

Joseph C. Klewicki
Professor and Chair
Department of Mechanical Engineering
University of Utah

John F. Kotek
Advanced Reactor Programs
Argonne National Lab-West

Glen R. Kowach
Distinguished Member of Technical Staff
Bell Labs, Lucent Technologies

Paul E. Krajewski
Laboratory Group Manager
Materials & Processes Laboratory
GM Research & Development & Planning
General Motors Corporation

Kevin E. Lansey
Professor
Department of Civil Engineering and Engineering Mechanics
University of Arizona

Daniel D. Lee
Assistant Professor
Department of Electrical and Systems Engineering
University of Pennsylvania

Stephen J. Lee
Chemist
U.S. Army Research Office

Dietrich Leibfried
Research Associate
Time and Frequency Division
National Institute of Standards and Technology

Eric K. Lin
Chemical Engineer
Polymers Division
National Institute of Standards and Technology

Hideo Mabuchi
Associate Professor
Department of Physics
California Institute of Technology

Ivan Marusic
Associate Professor
Department of Aerospace Engineering and Mechanics
University of Minnesota

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

Hendrik J. Meerman
Senior Scientist
Genencor International, Inc.

Michael E. Miller
Senior Human Factors Engineer
Eastman Kodak Company

Michele H. Miller
Associate Professor
Department of Mechanical Engineering-Engineering Mechanics
Michigan Technological University

Melody M. Moore
Assistant Professor
Computer and Information Systems Department
Georgia State University

Vicki L. Neale
Group Leader
Safety and Human Factors Division
Virginia Tech Transportation Institute
Virginia Polytechnic Institute and State University

Mitsunori Ogihara
Professor
Department of Computer Science
University of Rochester

Daniel W. Pack
Assistant Professor
Department of Chemical Engineering
University of Illinois at Urbana-Champaign

Scott L. Painter
Senior Research Scientist
Center for Nuclear Waste Regulatory Analyses
Southwest Research Institute

Ioannis Ch. Paschalidis
Associate Professor
Department of Manufacturing Engineering
Center for Information and Systems Engineering
Boston University

David W. Payton
Principal Research Scientist & Department Manager
Information Sciences Lab
HRL Laboratories

Sury Peddireddi
Advanced Manufacturing Engineer
Delphi Energy and Chassis

Tonya L. Peeples
Associate Professor
Chemical and Biochemical Engineering
University of Iowa

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

Claudio S. Pinhanez
Research Staff Member
TJ Watson Research Center
IBM Software Group

Lili Qiu
Researcher
MSR Systems and Networking Group
Microsoft Research

James Reuther
Aerospace Engineer
NASA Ames Research Center

Brigette Rosendall
Engineering Specialist
Bechtel Corporation

Anna L. Schauer
Technical Manager
Surety, Electronics, and Software Department
Sandia National Laboratories

Robert J. Schoelkopf
Assistant Professor
Applied Physics
Yale University

Fotis Sotiropoulos
Associate Professor
School of Civil and Environmental Engineering
Georgia Institute of Technology

Vivek Subramanian
Assistant Professor
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley

Michael P. Tarka
Senior Staff Systems Engineer
Lockheed Martin Air Traffic Management

Vahid Tarokh
Gordon McKay Professor and Vinton Hayes Fellow
Division of Engineering and Applied Science
Harvard University

Claire J. Tomlin
Assistant Professor
Department of Aeronautics and Astronautics
Stanford University

Steven J. van Enk
Member of Technical Staff
Computing Concepts Research Department
Bell Labs, Lucent Technologies

Kim J. Vicente
Professor and Director, Cognitive Engineering Laboratory
Department of Mechanical and Industrial Engineering
University of Toronto

John M. Vohs
Carl V.S. Patterson Professor and Chair of Chemical Engineering
Department of Chemical and Biomolecular Engineering
University of Pennsylvania

Lijun Wang
Research Scientist
NEC Research Institute, Inc.

John W. Weatherly
Research Scientist
Cold Regions Research and Engineering Lab
U.S. Army Corps of Engineers

Robert G. Welch
Associate Director
Fabric and Home Care
Procter and Gamble Company

Richard Wesel
Associate Professor
Electrical Engineering Department
University of California, Los Angeles

Shu Yang
Member of Technical Staff
Bell Labs, Lucent Technologies

Dinner Speaker

Andrew J. Viterbi
President
Viterbi Group, LLC

IBM Graduate Fellows

Sung K. Chang
Graduate Research Assistant
Department of Biomedical Engineering
University of Texas at Austin

Pei-Yu Chiou
Graduate Student
Department of Electrical Engineering
University of California, Los Angeles

Gregory M. Gratson
Graduate Student
Department of Materials Science and Engineering
University of Illinois at Urbana-Champaign

Abhijit S. Joshi
Graduate Research Assistant
Department of Mechanical Engineering
Clemson University

Saharon Rosset
Ph.D. Student
Statistics Department
Stanford University

Krystyn J. Van Vliet
Graduate Student
Department of Materials Science and Engineering
Massachusetts Institute of Technology

Science Academies and Councils of the Middle East

Abeer Ribhi Zakaria Arafat
Mechanical Design and Technology Center
Royal Scientific Society, Jordan

Michael Greene
Associate Director
Department of Development, Security, and Cooperation
Policy and Global Affairs Division
The National Academies

Joachim Meyer
Senior Lecturer
Department of Industrial Engineering and Management
Ben Gurion University of the Negev

Hashem Shahin
Department of Human Genetics and Molecular Medicine
Sackler School of Medicine
Tel Aviv University

Guests

Scott Andresen
Features Editor
IEEE Computer Society

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

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

Leila Belkora
Freelance Writer

Anita K. Jones
Lawrence R. Quarles Professor
School of Engineering and Applied Science
University of Virginia

Emir Macari
Program Director
Centers of Research Excellence in Science and Technology
National Science Foundation

Sharon Nunes
Director
Life Sciences Solution Development
Corporate Technology Group
IBM 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

Katharine Gramling
Research Associate

Lance Telander
Senior Project Assistant