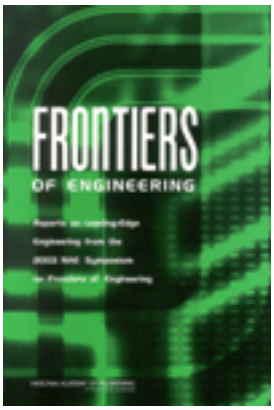

NINTH ANNUAL SYMPOSIUM ON FRONTIERS OF ENGINEERING

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

Funding for the activity that led to this publication was provided by the Air Force Office of Scientific Research, Defense Advanced Research Projects Agency, Department of Defense–DDR&E-Research, National Aeronautics and Space Administration, Eastman Kodak, Microsoft Corporation, ATOFINA Chemicals, Inc., Cummins, Inc., Science Applications International Corporation, Air Products and Chemicals, Inc., Millipore Corporation, GE Foundation, Dr. John A. Armstrong, and other individual donors.

International Standard Book Number 0-309-09139-X (Book)
International Standard Book Number 0-309-52992-1 (PDF)

Additional copies of this report are available from The National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, D.C. 20001; (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

PABLO G. DEBENEDETTI (Chair), Class of 1950 Professor and Chair, Department of Chemical Engineering, Princeton University
JANE BARE, Chemical Engineer, Office of Research and Development, U.S. Environmental Protection Agency
MATT BLAZE, Research Scientist, AT&T Laboratories
GANG CHEN, Associate Professor, Department of Mechanical Engineering, Massachusetts Institute of Technology
JOSEPH B. HUGHES, Professor and Chair, School of Civil and Environmental Engineering, Georgia Institute of Technology
LILA KARI, Associate Professor and Canada Research Chair in Biocomputing, Department of Computer Science, University of Western Ontario
STEPHEN J. LEE, Chemist, U.S. Army Research Office
MITSUNORI OGIHARA, Professor, Department of Computer Science, University of Rochester
ROBERT J. SCHOELKOPF, Assistant Professor, Applied Physics, Yale University

Staff

JANET R. HUNZIKER, Program Officer
JENNIFER M. HARDESTY, Senior Project Assistant
MARY W. L. KUTRUFF, Administrative Assistant

Preface

This volume highlights the papers presented at the Ninth 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 2003 symposium was held September 18–20 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 2003 SYMPOSIUM

The presentations this year covered four broad areas: environmental engineering, the fundamental limits of nanotechnology, counterterrorism technologies and infrastructure protection, and biomolecular computing. The talks in the Environmental Engineering session illustrated the interface between engineering and the natural sciences, resource economics, systems analysis, and risk management. Specifically, the presentations delved into how an understanding of microbial mineral respiration can inform the remediation of contaminated environments, the interface of water resource engineering with economics and public policy, and the use of life cycle analysis to address sustainability issues. While nanotechnology has been a topic at several past Frontiers meetings, the talks at this year's meeting focused on the fundamental limits of nanotechnology and sought to answer the question: How far down is the bottom? The speakers addressed the prospects in top-down approaches as applied to silicon microelectronics, the limits in the bottom-up approach to molecular electronics, the limits of storage in magnetic materials, and limits imposed on nanotechnology by the behavior of thermal systems at very small length scales. Presentations in the Counterterrorism Technologies and Infrastructure Protection session addressed chemical and biological threats, as well as threats to infrastructure systems such as power distribution, telecommunications, and transportation systems that are dependent on computing systems for their control and operation. Here presentations covered biocatalytic decontamination/demilitarization, the use of engineer-

ing problem-solving approaches for dealing with biological terrorism, and software and network security issues. The symposium concluded with a session on Biomolecular Computing, also known as DNA computing, biocomputing, and molecular computing. The primary goal of this field is to investigate the potential of molecules such as DNA for massively parallel computation. Here the talks covered some of the most promising areas in the field: computation based on DNA self-assembly, molecular breeding of DNA sequences by DNA shuffling, and programmable biological cells. (See Appendix C 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's dinner speaker was William F. Ballhaus, Jr., president and CEO of The Aerospace Corporation, whose talk was titled, *"The Most Important Lessons You Didn't Learn in Engineering School."* The full text of Dr. Ballhaus's remarks are included in this volume.

NAE is deeply grateful to the following organizations for their support of the Ninth 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, Eastman Kodak, Microsoft Corporation, ATOFINA Chemicals, Inc., Cummins, Inc., Science Applications International Corporation, Air Products and Chemicals, Inc., Millipore Corporation, GE Foundation, and Dr. John A. Armstrong and other individual donors. NAE would also like to thank the members of the Symposium Organizing Committee (see p. *iv*), chaired by Pablo Debenedetti, for planning and organizing the event.

Contents

COUNTERTERRORISM TECHNOLOGIES AND INFRASTRUCTURE PROTECTION

Biological Counterterrorism Technologies

Infrastructure Protection

BIOMOLECULAR COMPUTING

DINNER SPEECH

APPENDIXES

NINTH ANNUAL SYMPOSIUM ON FRONTIERS OF ENGINEERING

ENVIRONMENTAL ENGINEERING

Microbial Mineral Respiration

DIANNE K. NEWMAN
Divisions of Geological and Planetary Sciences
and Engineering and Applied Sciences
California Institute of Technology
Pasadena, California

Contributions from anthropogenic sources generally dominate the discussion on global change. And yet, although human activities have unquestionably left their mark on the environment, when averaged over geologic time, their importance pales in comparison with changes that have been effected by the activities of unicellular microorganisms (e.g., Bacteria, Archaea, and single-celled Eucarya). Not only does the number of microorganisms on Earth significantly exceed the number of humans (Whitman et al., 1998), microorganisms, unlike humans, are also found everywhere, are remarkably efficient at catalyzing a wide range of chemical reactions through their metabolisms, and have been around for billions of years (Figure 1).

Microbial metabolisms have brought about many changes in the Earth's environment. Microorganisms have altered the chemistry of the atmosphere via oxygenic photosynthesis, nitrogen fixation, and carbon sequestration. They have modified the composition of oceans, rivers, and pore fluids by controlling mineral weathering rates or by inducing mineral precipitation. They have changed the speciation of metals and metalloids in water, soils, and sediments by releasing complexing agents and/or by enzymatically catalyzing redox reactions. And they have shaped the physical world by binding sediments, precipitating ore deposits, and weathering rocks (Newman and Banfield, 2002).

MINERAL RESPIRATION

Respiratory metabolisms have impacted mineral formation and/or dissolution. The thought of respiring a mineral may seem suffocating, but bacteria have been doing it for billions of years (Figure 2). Respiration is fundamentally the

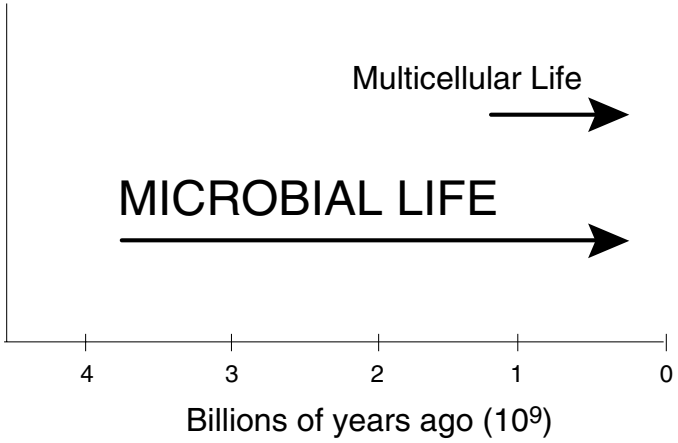

FIGURE 1 Microorganisms have dominated the history of life on Earth and have been major agents of global change. Although we will never know the exact date when microbial life first developed, a variety of geochemical and geobiological indicators suggest that this happened several billion years ago.

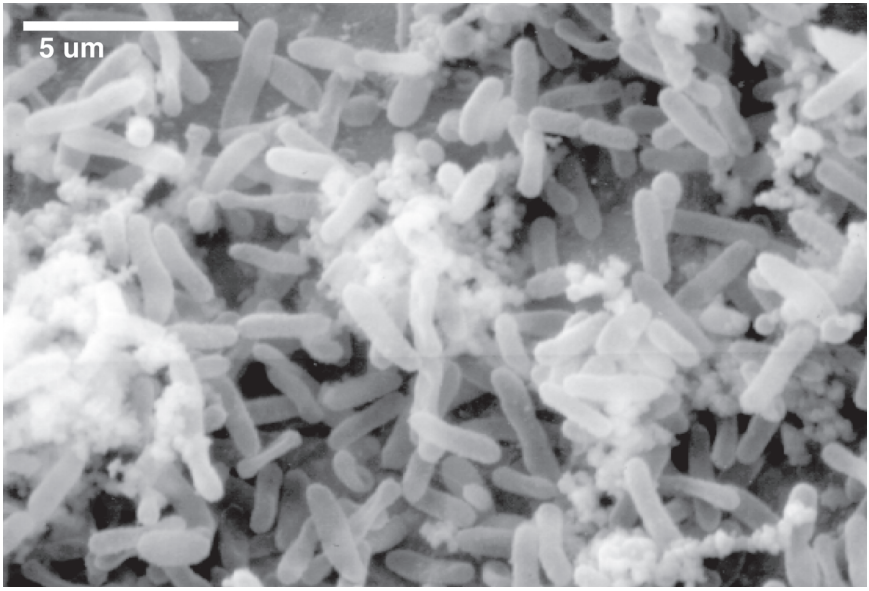

FIGURE 2 A scanning electron micrograph of the metal-reducing bacterium, *Shewanella oneidensis* strain MR-1, attached to a steel surface. As steel corrodes, ferric iron oxides are produced. MR-1 can reduce these minerals during respiration.

process of making energy available by transferring electrons from an electron donor to an electron acceptor. Typically, the transfer occurs down a respiratory chain embedded in the cell membrane; specific molecules hand off electrons from one end to the other. In the process, they generate a potential across the membrane that can be harnessed to do work (i.e., to store chemical energy in the form of ATP) (Mitchell, 1961). For respiration to succeed, a terminal electron acceptor, such as oxygen, must be available to receive the electrons. Before the evolution of oxygen in the atmosphere, microorganisms had to respire with alternative electron acceptors.

Most of the terminal electron acceptors used by bacteria for respiration, such as oxygen, nitrate, and sulfate, are soluble. This means they can make their way to the cell to receive electrons from the membrane-bound molecules of the respiratory chain. The real question is how bacteria transfer electrons to solids like hematite (α·Fe_2O_3) and goethite (α·FeOOH) (Figure 3). Because these minerals are effectively insoluble under environmentally relevant conditions, simple dissolution and diffusion of ferric iron to the cell cannot be the answer (ferric iron is the constituent of the mineral that receives electrons). Therefore, bacteria must have other strategies for transferring electrons to minerals during respiration. The question is, what are they?

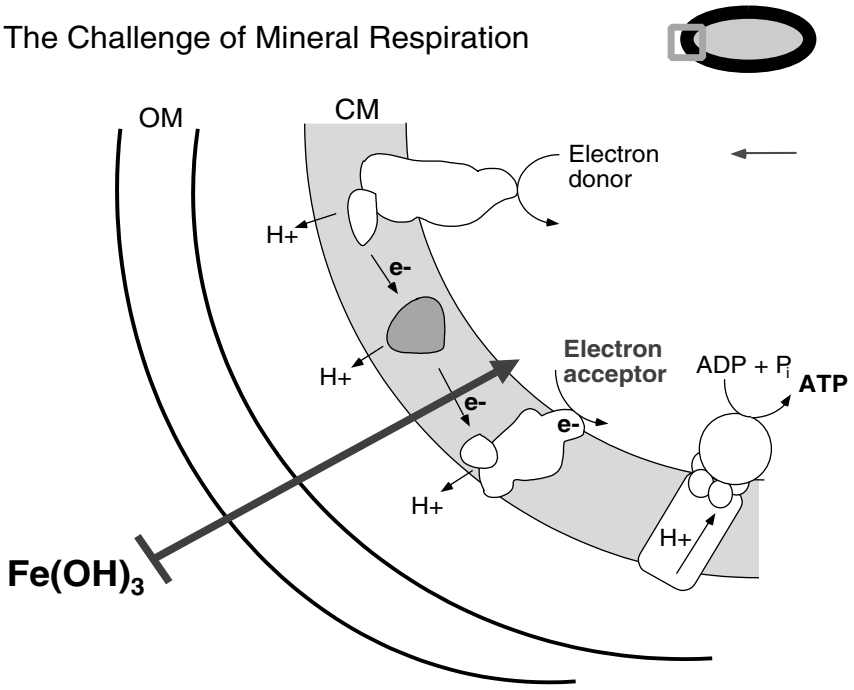

FIGURE 3 A schematic drawing illustrating the process of mineral respiration.

Several mechanisms have been proposed (Figure 4). Some have suggested that bacteria solubilize the minerals by producing chelators. Although the addition of synthetic chelators has been shown to stimulate microbial electron transfer to iron minerals, no evidence has been found that bacteria use this mechanism in respiration. Another suggestion is that they may use soluble shuttles, such as organic compounds with quinone moieties, to transfer electrons from the cell to the mineral. These shuttles may be exogenous substances, or they may be substances produced by the organisms themselves (Lovley et al, 1998; Newman and Kolter, 2000).

Another mechanism, possibly the dominant one, is that bacteria transfer electrons directly from the cell surface to the mineral after a regulated search and attachment process. A variety of biomolecules (including cytochromes, quinones, and dehydrogenases) have been identified as part of this electron-transfer pathway (Schröder et al., 2003). Several of these biomolecules are located on the outer membrane of the cell and presumably make contact with the mineral directly (Lower et al., 2001). Given that the initial rate and long-term extent of electron transfer is correlated with their surface area and the concentration of reactive sites, this seems like a reasonable explanation (Zachara et al., 1998).

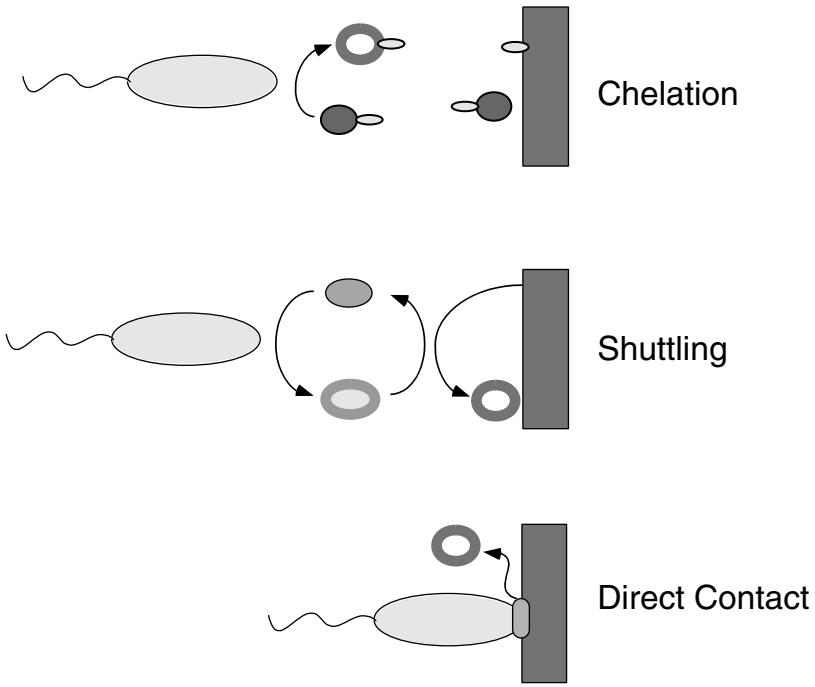

FIGURE 4 Three models that explain how bacteria respire minerals.

Yet the nature of the electron-transfer event remains obscure and is a subject of active research.

PUTTING MINERAL RESPIRATION TO WORK

Despite uncertainties about the molecular mechanisms of mineral respiration, environmental microbiologists and engineers have been putting it to work for more than two decades. The best example is bioremediation; microbial metabolisms based on mineral respiration have been used to clean up organic and/or inorganic contaminants in groundwater (Figure 5).

In the late 1980s, researchers at the U.S. Geological Survey observed that the oxidation of aromatic hydrocarbons (e.g., benzene, xylenes, and toluene) in contaminated shallow aquifers was associated with the depletion of Fe(III) oxides from contaminated sediments and the accumulation of dissolved Fe(II) over time (Lovley et al., 1989). Hypothesizing that Fe(III)-respiring microorganisms may have been responsible for this phenomenon, Derek Lovley and his coworkers showed that the oxidation of added toluene to CO_2 was dependent on active microbial metabolism. They went on to demonstrate that *Geobacter metallireducens* strain GS-15, an Fe(III)-respiring bacterium, could oxidize a variety of

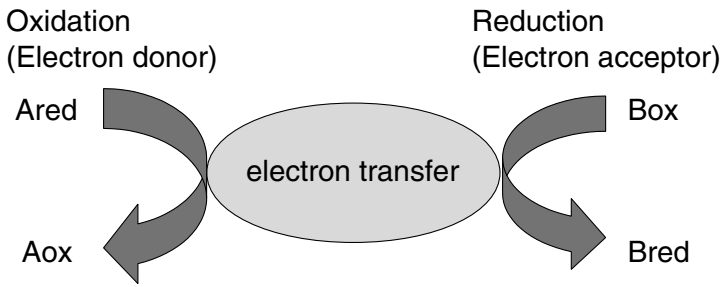

FIGURE 5 Catabolic electron-transfer metabolisms (e.g., fermentation, respiration, and photosynthesis) generate energy. Microorganisms are extraordinarily versatile and can use a variety of compounds as electron donors or electron acceptors in these reactions. For example, Ared can be toxic organic compounds, such as benzene and toluene. Microbes can oxidize these to CO_2 under both aerobic and anaerobic conditions; in anaerobic groundwater systems, Fe(III) is the primary oxidant (i.e., electron acceptor). In addition to Fe(III), bacteria can use other metal(loids) as electron acceptors. These include As(V), Cr(VI), U(IV), Tc(VII), Se(V), and graphite electrodes. When these inorganic compounds are used as electron acceptors, their reduction may lead to their mobilization or precipitation, depending on the metal(loid). Such reactions may be important for the bioremediation of contaminated sites and/or the generation of electricity from marine sediments. Note: In a chemical reaction, Ared + Box = Aox + Bred. Ared = the electron donor for the reaction. Box = the electron acceptor for the reaction. Aox = the oxidized product of the reaction. Bred = the reduced product of the reaction.

aromatic compounds. Two decades later, members of the *Geobacteraceae* family have been observed to account for a significant portion of the microbial population in contaminated sediments (Snoeyenbos-West et al., 2000), and the U.S. Department of Energy (DOE) is actively investing in research to get a better understanding of and to stimulate Fe(III) respiration for bioremediation (DOE, 2003a).

In addition to coupling the oxidation of organic contaminants to the reduction of Fe(III), microbial activity can be of value for the bioremediation of inorganic contaminants, uranium (U) and technetium (Tc), for example, two abundant radioactive metals that contaminate the subsurface environments at some DOE sites. Fe(III)-respiring bacteria have been shown to enzymatically reduce highly soluble U(VI) carbonate complexes and $Tc(VII)O_4^-$ to the insoluble tetravalent phases, UO_2 and TcO_2 (Lloyd, 2003). The theory is that if organisms with this capacity are stimulated *in situ*, toxic inorganic compounds may be precipitated from groundwater and immobilized in the subsurface (Barkay and Schaefer, 2001).

Although the long-term efficacy of this approach is still a subject of debate, encouraging preliminary field studies show that a significant percentage of soluble U can be removed rapidly from groundwater by stimulating the indigenous microbial population (Finneran et al., 2002). Similarly, hexavalent chrominum [Cr(VI)] is a strong, highly mobile carcinogen that forms an insoluble Cr(III) precipitate when reduced. Efforts are currently under way at the Hanford DOE site to determine whether stimulation of the indigenous microbial population with lactate (a carbon source) can enhance Cr immobilization (DOE, 2003b).

Besides removing toxic inorganic compounds from groundwater, microbial respiratory metabolisms also have the potential to generate electricity. In a recently reported example, members of the family *Geobacteraceae* were shown to grow by oxidizing organics with a graphite electrode as the sole electron acceptor (Bond et al., 2002). When fuel cells consisting of anodes embedded in the sediment connected to cathodes positioned in the overlying seawater were deployed in two coastal marine environments (a salt marsh near Tuckerton, New Jersey, and the Yaquina Bay Estuary near Newport, Oregon), oxidation of both organic and inorganic electron donors in the sediment supported power generation (Tender et al., 2002). Although much work remains to be done before we can understand how microbial communities catalyze electron-transfer reactions to the anode, this process has the potential to sustain long-term power generation from marine sediments.

Lest microbial respiratory metabolisms be considered uniformly beneficial, it is important to point out that mineral respiration does not always improve water quality. A tragic example of this can be seen today in Bangladesh, where thousands of people are dying from drinking arsenic-contaminated well water. It is now widely believed that microbial respiratory activities are contributing to this problem by mobilizing arsenic in the groundwater (Harvey et al., 2002).

Although the details have not yet been determined, evidence points to the reductive dissolution of iron arsenate minerals as a likely mechanism (Oremland and Stolz, 2003).

CONCLUSIONS

Microbial respiratory metabolisms based on minerals are fascinating from a purely scientific standpoint because of what they can teach us about electron-transfer reactions. They are also of great interest to environmental engineers seeking novel ways to remediate contaminated environments and/or to generate electricity. How organisms evolved the capacity to transfer electrons to mineral surfaces is not well understood but merits further investigation. It is possible that horizontal gene transfer has accelerated the distribution of this capability in both time and space; if so, this mechanism could be exploited in the future to deliver useful genetic material to targeted microbial populations.

ACKNOWLEDGMENTS

I would like to thank the Clare Boothe Luce Foundation, the Packard Foundation, and the Office of Naval Research for providing generous research support.

REFERENCES

Barkay, T., and J. Schaefer. 2001. Metal and radionuclide bioremediation: issues, considerations and potentials. Current Opinion in Microbiology 4(3): 318–323.

Bond, D.R., D.E. Holmes, L.M. Tender, and D.R. Lovley. 2002. Electrode-reducing microorganisms that harvest energy from marine sediments. Science 295(5554): 483–485.

DOE (U.S. Department of Energy). 2003a. Natural and Accelerated Bioremediation (NABIR) Program. Available online at: *<http://www.lbl.gov/NABIR/>*.

DOE. 2003b. NABIR Research Program. Available online at: *<http://www.lbl.gov/NABIR/researchprogram/awards/em_projects02.html#hazen>*.

Finneran, K.T., R.T. Anderson, K.P. Nevin, and D.R. Lovley. 2002. Potential for bioremediation of uranium-contaminated aquifers with microbial U(VI) reduction. Soil and Sediment Contamination 11(3): 339–357.

Harvey, C.F., C.H. Swartz, A.B.M. Badruzzaman, N. Keon-Blute, W. Yu, M.A. Ali, J. Jay, R. Beckie, V. Niedan, D. Brabander, P.M. Oates, K.N. Ashfaque, S. Islam, H.F. Hemond, and M.F. Ahmed. 2002. Arsenic mobility and groundwater extraction in Bangladesh. Science 298(5598): 1602–1606.

Lloyd, J.R. 2003. Microbial reduction of metals and radionuclides. FEMS Microbiology Reviews 27(2-3): 411–425.

Lovley, D.R., M.J. Baedecker, D.J. Lonergan, I.M. Cozzarelli, E.J.P. Phillips, and D.I. Siegel. 1989. Oxidation of aromatic contaminants coupled to microbial iron reduction. Nature 339(6222): 297–299.

Lovley, D.R., J.D. Coates, E.L. Blunt-Harris, E.J.P. Phillips, and J.C. Woodward. 1998. Humic substances as electron acceptors for microbial respiration. Nature 382(6564): 445–448.

Lower, S.K., M.F. Hochella, Jr., and T.J. Beveridge. 2001. Bacterial recognition of mineral surfaces: nanoscale interactions between *Shewanella* and α-FeOOH. Science 292(5520): 1360–1363.

Mitchell, P. 1961. Coupling of phosphorylation to electron and hydrogen transfer by a chemiosmotic type of mechanism. Nature 191: 144–148.

Newman, D.K., and J.F. Banfield. 2002. Geomicrobiology: how molecular-scale interactions underpin biogeochemical systems. Science 296(5570): 1071–1077.

Newman, D.K., and R. Kolter. 2000. A role for excreted quinines in extracellular electron transfer. Nature 405(6782): 94–97.

Oremland, R.S., and J.F. Stolz. 2003. The ecology of arsenic. Science 300(5621): 939–944.

Schröder, I., E. Johnson, and S. de Vries. 2003. Microbial ferric iron reductases. FEMS Microbiology Reviews 27(2-3): 427–447.

Snoeyenbos-West, O.L., K.P. Nevin, R.T. Anderson, and D.R. Lovley. 2000. Enrichment of Geobacter species in response to stimulation of Fe(III) reduction in sandy aquifer sediments. Microbial Ecology 39(2): 153–167.

Tender, L.M., C.E. Reimers, H.A. Stecher III, D.E. Holmes, D.R. Bond, D.A. Lowy, K. Pilobello, S.J. Fertig, and D.R. Lovley. 2002. Harnessing microbially generated power on the seafloor. Nature Biotechnology 20(8): 821–825.

Whitman, W.B., D.C. Coleman, and W.J. Wiebe. 1998. Prokaryotes: the unseen majority. Proceedings of the National Academy of Sciences 95(12): 6578–6583.

Zachara, J.M., J.K. Fredrickson, S.-M. Li, D.W. Kennedy, S.C. Smith, and P.L. Gassman. 1998. Bacterial reduction of crystalline Fe^{3+} oxides in single phase suspensions and subsurface materials. American Mineralogist 83(11-12, Part 2): 1426–1443.

Water-Resource Engineering, Economics, and Public Policy

GREGORY W. CHARACKLIS
Department of Environmental Sciences and Engineering
University of North Carolina at Chapel Hill

INTRODUCTION

Water-resource engineering is a branch of environmental engineering that involves the analysis and manipulation of hydrologic systems, particularly in areas related to water quality and water availability (e.g., water supply and flooding). Since the mid-1970s, concerns about improving water quality and meeting future demands have multiplied in scope and magnitude, prompting significant changes in public policy at the federal, state, and local levels (Cech, 2003). The strong influence of public policy on the research agenda in water-resource engineering (and most other environmental fields) differentiates it from other engineering disciplines in some important respects.

In non-environmental disciplines, the primary motivation for developing a solution (e.g., a product or process) is economic—doing something "faster, better, cheaper." The solution is then implemented by consumers and organizations based on their determination of how well the new product or process fits their needs. By contrast, the motivation for developing solutions to water-resource challenges is often related to meeting regulatory objectives (e.g., public health or environmental quality) that cannot be easily quantified in economic terms. Although attempts are made to estimate costs and benefits when evaluating and setting regulatory policy, implementation of a policy is generally dictated by centrally controlled directives that often entail high compliance costs.

A number of factors can justify this command-and-control approach, but the growing number of environmental regulations, and their increasingly restrictive nature, have stimulated a growing interest in supplementing the traditional approach with market principles. The development of market-based schemes for

regulating water resources can be greatly facilitated by interdisciplinary evaluations that integrate analytical tools from engineering and economics. An interdisciplinary analysis can not only help in the identification of potential methods of regulatory implementation, but can also improve assessments of the consequences of policy choices.

WATER-RESOURCE POLICY

Regulations of water-resource systems usually involve limits on access and generally come in two forms. The first, restrictions on the physical removal of water from a system, is based on concerns about causing adverse environmental impacts. The second involves limiting access to a hydrologic system's capacity to assimilate waste by restricting the kind and amount of waste that may be released into it (e.g., wastewater effluent standards). In both cases, limits are generally based on some form of scientific analysis. Once a target range for limiting access has been defined, engineering analyses are performed to generate information on the effectiveness and cost of available technical solutions. These might involve the development of new water sources or the installation of a new technology that reduces emissions. The cost associated with each alternative is then calculated, and policy alternatives based on these technical solutions are assembled.

Although this approach does provide useful results, it often leaves out important considerations. Water-resource policy making generally involves a combination of moral, technical, and economic factors, and fully informed decisions require that these disparate kinds of information be integrated. The moral aspects are usually judged by elected officials (or their surrogates at the regulatory level) on the basis of personal values or political motivations. Technical and economic information, however, can be more difficult to evaluate and generally requires some level of synthesis before it can be readily "digested" by policy makers. Furthermore, even though the technical and economic implications of water-resource policy decisions are often inextricably linked, they are frequently analyzed independently, using tools specific to each discipline.

Translating technical and economic factors into a common context involves more than just estimating the costs and benefits of a potential regulatory scheme (although these are important). It also entails an analysis of trade-offs and an exploration of alternative means of implementation (NRC, 1996). Interdisciplinary analyses can integrate engineering and economic considerations into terms that policy makers can more readily incorporate into their decision-making process.

The following case study involves a region in which restrictions were placed on aquifer access in response to environmental concerns about declining aquifer levels. In this case, studies have been done to characterize the environmental impact of current pumping rates, with study results used as the basis for setting limits on withdrawals that are designed to mitigate adverse effects. Policy deci-

sions must now be made on the method of implementing and managing these regulatory limits.

CASE STUDY: EDWARDS AQUIFER

The Edwards Aquifer is a porous, fractured-limestone aquifer in south-central Texas that has been the sole source of water for most of the region's agricultural, municipal (e.g., San Antonio), and industrial users. Regional population growth and economic development have increased pumping rates to the point that the water height in the aquifer frequently drops to levels that substantially reduce the flow of water to several natural springs. These springs provide habitat for several endangered species, and a federal court has ruled that allowing spring flow to drop below specified levels constitutes a "taking" under the Endangered Species Act (Votteler, 1998). As a result, the state of Texas has been instructed to take whatever actions are necessary to ensure that acceptable flow rates are maintained.

This ruling led to a number of biological and hydrological studies to determine the effects of various pumping rates on the aquifer level and water flow in the springs. Based on the results of these studies, the amount of water allowed to be pumped from the aquifer has been limited to 450,000 acre-feet per year, a level that will be insufficient for any party to continue using water at the present rate. In addition to the annual limits, more severe temporary restrictions will be imposed when aquifer levels fall below specified levels. In response to these limits on aquifer access, regional municipalities are considering the development of new water supplies, but the few alternative sources available in this semi-arid region would be very expensive to exploit (SCTRWPG, 2000). Economic analyses of agricultural water use suggest that the value of water in irrigation is quite low relative to the cost of accessing new sources (Keplinger and McCarl, 2000), making transfers of water from agriculture to municipal and industrial users an economically attractive solution (Figure 1).

Consequently, a property-rights system has been put in place to allocate the reduced pumping capacity of the Edwards Aquifer among regional consumers, and institutions have been established to facilitate the trading of these rights. Because of concerns about the impact of this regulatory framework on the rural economy, authorities were generous in their initial allocation of rights to irrigators, but left municipal users with a deficit of approximately 43,000 acre-feet per year relative to their current average use (and a considerably larger deficit during dry years). Municipalities also bear the burden of "emergency" measures, which call for urban users to reduce monthly pumping by 5 percent, 10 percent, or 15 percent during months in which the aquifer level drops below 650 feet (above mean sea level), 640 feet, or 630 feet, respectively.

To augment their supplies, municipalities have resorted to purchasing agricultural water rights, the most inexpensive and immediately available source of

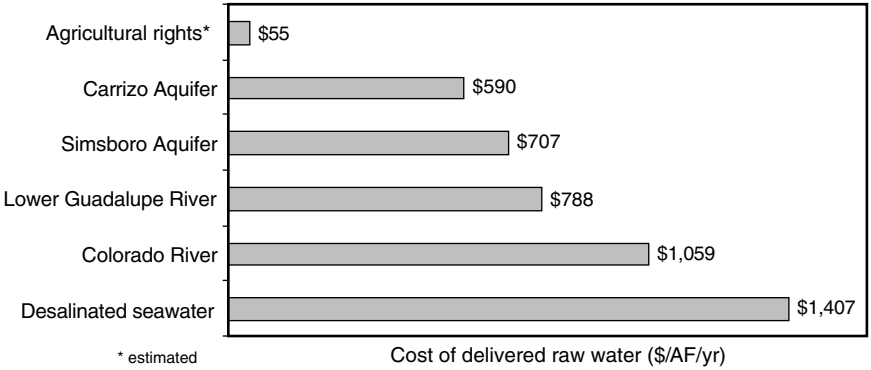

FIGURE 1 Cost of alternative water supplies in the Edwards Aquifer region. Source: Adapted from SCTRWPG, 2000.

water. Although a general policy of market-based trading has been used to allocate the newly scarce pumping capacity, critical issues are still unresolved. Policy makers would like to find a solution that minimizes the amount of water transferred out of agriculture. But if municipalities cannot achieve their supply-reliability goals under the proposed framework, then the market-based solution becomes much less valuable (McCarl et al., 1999).

An analysis that incorporates both hydrologic modeling and market simulation was undertaken to evaluate approaches that would maximize the average volume of water available to agriculture and still meet municipal objectives related to supply reliability (high) and cost (low). Although water transfers from agricultural to urban use are inevitable, if the only transfer mechanism available is a permanent sale, municipalities will be forced to buy rights and maintain a volume well in excess of average usage to ensure supply reliability during periodic droughts. These purchased rights are unlikely to be transferred back to agriculture once they are acquired. Thus, in many normal and wet years a significant fraction of the aquifer's total capacity may not be used. If more flexible types of transfer were available, such as transfers that would allow water to be acquired on an "as needed" basis, the capacity maintained by municipalities could be reduced and the average volume of water available to agriculture increased (Characklis et al., 1999).

Purchase decisions regarding these more flexible transfer instruments could be greatly facilitated by improved information about future supplies. In the case of the Edwards Aquifer, future shortfalls will be largely the result of the temporary pumping restrictions imposed when aquifer levels drop below specified levels, conditions that can be accurately predicted through hydrologic modeling (Durbin, 2002).

Flexible transfers allow for more responsive purchase decisions, particularly when supply conditions are known in advance. Annual leases provide for the temporary transfer of pumping rights over a one-year period (beginning January 1), and probabilistic information on future aquifer levels at the time of purchase could be used to estimate the volume necessary to meet supply-reliability goals. Another potential transfer instrument is the option, a transaction in which the buyer pays a small fee at the beginning of the year for the right to purchase water, or exercise the option, at a later date (June 1) at a prearranged price. The accuracy of predictions of aquifer level would have a significant bearing on decisions to purchase and exercise options. Finally, spot-market leases would provide ultimate flexibility, allowing municipalities to acquire water at any time, but also subjecting them to price volatility (i.e., dry weather = high prices).

An assessment of the benefits of including all or some of these transaction types could indicate whether the costs of developing the marketing and monitoring framework necessary to regulate these more complex transactions is justified. Because the Edwards Aquifer market is still in its infancy, little information is available on market behavior. As a result, a market simulation was developed to evaluate monthly supply and demand throughout a given year. The information was then used to calculate spot-market prices as a basis for calculating the prices of the other transaction types.

The monthly pumping restrictions imposed as a result of declining aquifer levels can considerably disrupt municipal water supplies, so the ability to assess the likelihood of restrictions can be valuable when formulating planning strategies. Therefore, a hydrologic model was developed to predict the aquifer level in the "J-17" well (the regulatory reference point) as a function of the previous month's well level, aquifer inflows, agricultural pumping, and municipal pumping. A comparison of model results with empirical data measured over a 25-year period suggests that the model provides a very good estimate of the aquifer level (Figure 2); therefore, the model was embedded in the market simulation.

Market simulations are run many times for a range of initial aquifer levels, with input obtained via Monte Carlo sampling from a joint, multivariate distribution created from inflow and withdrawal data. Simulated supply and demand conditions are translated into market prices for each transfer type, and the expected cost and reliability of various combinations, or "portfolios," of transfer types can be computed. The transfer types are specified for each scenario, and a sequential search method is then used to identify minimum cost portfolios that meet designated supply-reliability constraints (Figure 3). Differences in the cost of the respective portfolios indicate the value of including each transaction type in the market, as well as how the cost of market-based approaches compares to the development of the least expensive new water source (Carrizo Aquifer).

Although the model shows that spot-market leasing would have considerable economic advantages, because of the risk averse nature of municipal utilities, a strategy that involves exposure to spot-market price volatility, even if the

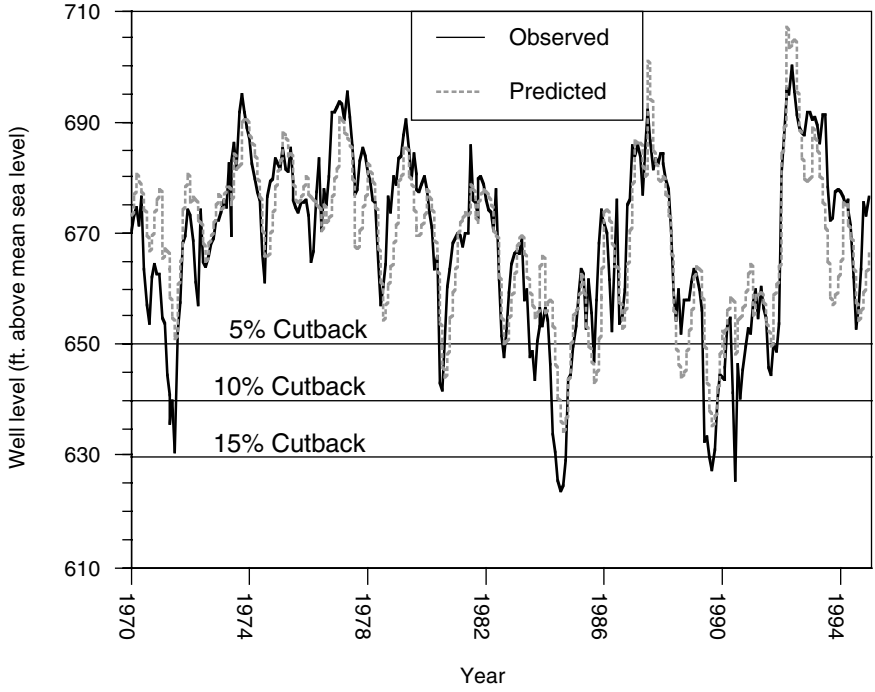

FIGURE 2 Predicted vs. observed aquifer level in the J-17 reference well. Source: Durbin, 2002.

expected costs are lower, is not likely to appeal to municipal decision makers. They may, however, feel comfortable entering into annual leasing and option contracts, which provide much greater certainty about water availability and cost. The advantages of using these two instruments include savings of close to $1 million dollars per year relative to the cost of depending on permanent transfers alone. More flexible transfers would also increase the average amount of water available to agriculture by 5,000 to 25,000 acre-feet per year, a feature that is important to policy makers. Consequently, the dual objectives of maintaining municipal supply reliability and reducing the economic impacts on agriculture would appear to be furthered by the introduction of more flexible market instruments. Thus, an excellent argument can be made to justify an investment in institutions to monitor and regulate these types of transfers.

CONCLUSIONS

Integrating technical and economic analytical techniques provides a means of identifying and evaluating a range of solutions to water-resource challenges.

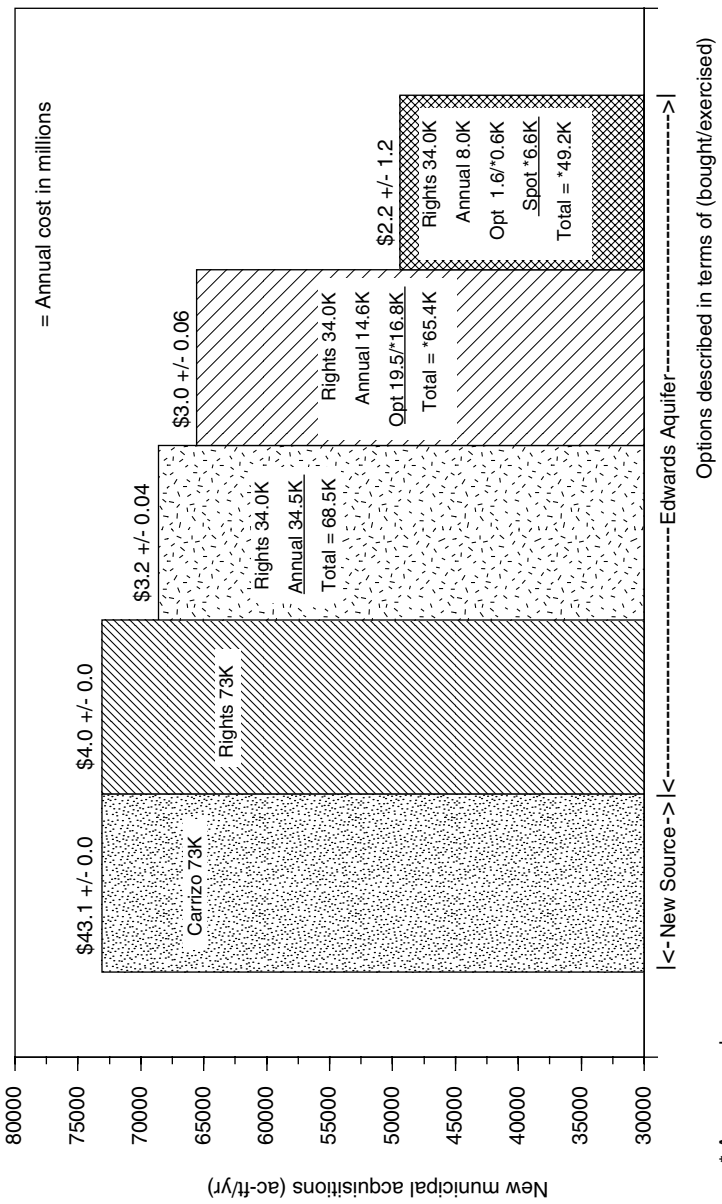

FIGURE 3 Comparison of approaches to meeting municipal demand with at least 99 percent monthly reliability.

The example above describes the application of these methods to the development of strategies for managing regulatory limits on resource acquisition; similar approaches could be used for regulating emissions. As demands continue to increase for both water resources and the assimilative capacity these resources provide, the need for creative, interdisciplinary solutions will also increase.

REFERENCES

Cech, T.V. 2003. Principles of Water Resources: History, Development, Management, and Policy. New York: John Wiley and Sons.

Characklis, G.W., R.C. Griffin, and P.B. Bedient. 1999. Improving the ability of a water market to efficiently manage drought. Water Resources Research 35(3): 823–832.

Durbin, S.W. 2002. Utilization of a Water Market to Optimize Water Acquisition in the Edwards Aquifer Region. Unpublished M.S. thesis, University of North Carolina at Chapel Hill.

Keplinger, K.O., and B.A. McCarl. 2000. An evaluation of the 1997 Edwards Aquifer irrigation suspension. Journal of the American Water Resources Association 36(4): 889–901.

McCarl, B.A., C.R. Dillon, K.O. Keplinger, and R.L. Williams. 1999. Limiting pumping from the Edwards Aquifer: an economic investigation of proposals, water markets, and spring flow guarantees. Water Resources Research 35(4): 1257–1268.

NRC (National Research Council). 1996. Linking Science and Technology to Society's Environmental Goals. Washington, D.C.: National Academy Press.

SCTRWPG (South Central Texas Regional Water Planning Group). 2000. Water Supply Options for South Central Texas. Austin, Texas: Texas Water Development Board.

Votteler, T.H. 1998. The little fish that roared: the Endangered Species Act, state groundwater law, and private property rights collide in the Texas Edwards Aquifer. Environmental Law 28(4): 845–904.

Life Cycle Development: Expanding the Life Cycle Framework to Address Issues of Sustainable Development

GREGORY A. NORRIS
Sylvatica
North Berwick, Maine
Harvard School of Public Health
Boston, Massachusetts

The environmental movement of the late twentieth century began three decades ago amid concerns about energy depletion, water and air pollution, and mounting flows of postconsumer packaging wastes. Independently in the United States, continental Europe, and the United Kingdom, small teams of engineers and physicists were called upon by policy makers to provide a comprehensive account of the environmental and resource implications of increasingly popular disposable, plastic packaging. Each of these teams invented a methodology that came eventually (in the 1990s) to be called environmental life-cycle analysis (LCA).

With the increasing importance and visibility of "product policy" and "extended product (or producer) responsibility" during the 1990s, LCA changed from a little-known cottage industry to an internationally standardized analytical tool in support of environmental management. LCA is now used by thousands of companies, many governments, consumer and environmental groups, even the United Nations Environment Program, to shed light on the cradle-to-grave environmental consequences of product-related decisions.

In 2002, the leaders of many national governments converged, along with representatives from industry and civil society, in Johannesburg at the World Summit on Sustainable Development (WSSD). Participants took stock of the successes and failures of the past 30 years and looked ahead to the promise and perils facing humanity in relation to sustainable development—development that meets the needs of the present without compromising the ability of future generations to meet their needs (WCED, 1987). The WSSD issued the *Plan of Implementation for Changing Unsustainable Patterns of Consumption and Production,* which called on producers to "Improve the products and services pro-

vided, while reducing environmental and health impacts, using where appropriate, science-based approaches, such as life-cycle analysis" (WSSD, 2002). Thus, LCA, which was originally developed to inform environmental policies, is now being called upon to assist in the search for sustainable patterns of consumption and production. The interdisciplinary frontiers of this engineering-based methodology are constantly evolving in response to demands in the policy arena for sustainable development.

ORIGINS

The field of environmental LCA originated during a period of rising concern about disposable packaging. In 1969, the packaging manager at the Coca-Cola Company was considering whether the company should begin to manufacture its own beverage containers (Hunt and Franklin, 1996). At the same time, the previously stable world of beverage container materials and designs was shifting to new materials, such as plastic, and from returnable containers to disposable, or "one way," packaging systems.

To help inform this major business decision in the context of emerging concerns about resources and the environment, the Coca-Cola packaging manager began to visualize a new type of study "that would attempt to quantify the energy, material, and environmental consequences of the entire life cycle of a package from the extraction of raw materials to disposal" (Hunt and Franklin, 1996). He commissioned a study that eventually led to the development of LCA. Personnel involved in the original project migrated to the post of deputy assistant administrator for solid waste at the newly formed Environmental Protection Agency (EPA), just when EPA was considering using taxes, subsidies, or direct regulation to stem the momentum of change from returnable to disposable beverage containers. EPA commissioned an LCA of the issue from the same people who had done the original study for Coke (Hunt and Franklin, 1974).

Contemporaneously, the German government commissioned a study of packaging materials, including an assessment of the potential environmental benefits of biodegradable polymers (Oberbacher et al., 1996), and the Glass Manufacturers Federation commissioned an LCA of returnable and nonreturnable bottles in the United Kingdom (Boustead, 1996). Both the German and U.K. studies drew upon the newly developed methods of Hunt and Franklin.

LCAs attempt to provide comprehensive analyses in two dimensions. First, the "full" set of manufacturing, transportation, and solid-waste management processes required to support the manufacture, use, and disposal of a product is considered to be potentially within the scope of the study. In practice, however, boundary rules must be adopted to keep the data requirements within feasible limits (ISO, 1998). Second, the "full" set of relevant flows to and from the environment is potentially within the scope of the study. Rules are adopted to make the task manageable.

The whole LCA enterprise can only be feasible over the long term with real-world budget constraints if studies contribute to ongoing, cumulative, reusable, modular, and "generic" databases about average or representative unit processes. These radically comprehensive LCA studies provided counterintuitive, nondefinitive results. First, the studies often showed that product properties, such as lightness or longevity-of-use, had life-cycle energy advantages that compensated for the environmental impacts of nonrenewable materials. Second, by including a broad array of pollutants and environmental impacts in LCAs, manufacturers were sometimes able to argue that the problem was much more complicated than the narrow issue that may have motivated the study or policy debate; LCAs revealed trade-offs among environmental impacts. In general, LCAs supported (and continue to support) two major results:

- LCAs look beyond narrow, suboptimization that shifts the environmental burden from one life-cycle stage to another or from one environmental impact category to another.
- LCAs can reveal leverage points or hot spots, the life cycle of a product system that provides the greatest leverage to improve its environmental profile.

ATTRIBUTIONAL AND CONSEQUENTIAL PERSPECTIVES

For nearly 30 years, LCA models were built by analysts responding to the (often implicit) question of how a specific product is made and disposed of. The static nature of this question, the engineering background of most LCA analysts, and the practical requirement for generic and reusable modular databases, led to static, scale-independent models of the flows of materials and energy among linear, engineering-unit processes. As LCA evolved, models describing how current inflows and outflows link these processes were built according to accounting conventions (such as those laid out in detail by the ISO standards for LCA). These models could allocate or attribute total economy-wide pollution and resource consumption burdens among the total economy-wide output of products (Heijungs, 1997).

Until quite recently, this engineering-based, "attributional" perspective was unquestioningly considered to lead to results valuable for identifying environmentally preferable decisions. But a decision-making perspective is a what-if perspective that requires a considerably different model than the model based on the "how are things made" perspective.

For example, consider a product *P* whose manufacture is electricity intensive. An attributional LCA model is built by asking where the electricity comes from. The model then characterizes the fuel shares for electricity generation in the region of interest and attributes the burdens among the power plant types based on their shares of energy contribution to the regional mix (e.g., during a recent year). By contrast, a what-if perspective might ask what the environmen-

tal consequences of making more or less of product *P* are. In this case, when the demand for product *P* goes up or down, not all power-generating plant types respond in equal proportion to their current output, either over the short term or long term. For example, if hydro power contributes 50 percent of currently generated kWh in the region of interest, increasing the demand for product *P* will not lead to more rain (more hydropower output) in the short or medium term. And over the long term, the plant types expected to be the most cost effective will be added to the generating mix.

Thus, attributional (or pollution-cause-oriented) and consequential (effect-oriented) perspectives can lead to very different models that generate significantly different results (Ekvall, 1999). This realization, and the fact that the usual purpose of LCAs is to help identify decisions or courses of action that will lead to beneficial environmental effects or consequences, has generated lively debate in the LCA community about fundamental criteria for model selection and database development.

At the same time, LCAs are increasingly integrating information from databases and perspectives from beyond the engineering-unit process world. For example, to make products ranging from computers to buildings requires not only energy and materials, but also product designers, architects, engineers, marketers, managers, and so on. These people tend to travel as part of their work. It turns out that the environmental impacts of business travel contribute significantly to the total environmental burdens of a wide range of products (Känzig, 2003). By drawing upon models and databases of economic input/output analysis, these burdens can be addressed.

SUSTAINABLE DEVELOPMENT

The present application context of LCA is sustainable development, which is commonly considered to have three pillars: economic growth, ecological balance, and social progress (WBCSD, 2003). Traditional LCA addressed only the environmental pillar. However, when economic modeling and a what-if perspective are combined, LCA can potentially address more of the sustainability agenda, thereby avoiding burden shifting among the social, environmental, and economic objectives and potentially helping to build a broader base of support for policy proposals by addressing the concerns of a wide range of stakeholders.

This larger perspective is made possible in two ways. First, by integrating economic models and databases, LCA can address impacts and performance measures of groups that are routinely tracked at the level of economic sectors rather than at the level of engineering-unit processes. An example of an impact group is occupational health and safety. A recent investigation concluded that the health effects of occupational health and safety issues and incidents in product supply chains appear to be on the same order of magnitude as the expected

near-term human health consequences of supply chain pollution releases (Hofstetter and Norris, 2003).

Second, by integrating the economic and consequential modeling approaches and databases, LCA can show that product supply chain activities bring benefits as well as burdens for the agenda of sustainable development. Sustained increases in economic output among developing countries is linked to major improvements in human health, through the mechanisms of income-poverty reduction, increased investment in and access to education, and increased public investment in the public health infrastructure (World Bank, 2001). Traditional LCA, which was focused strictly on pollution impacts and was blind to the benefits of development, is considered biased against the primary concerns of developing countries by some sustainability analysts. By addressing the benefits of economic development as well as the costs of pollution and resource degradation, LCA has the potential to address these concerns and suggest truly sustainable solutions to the challenges of consumption and production in the twenty-first century.

REFERENCES

Boustead, I. 1996. LCA—How it came about: the beginning in U.K. International Journal of Life Cycle Assessment 1(3): 147–150.

Ekvall, T. 1999. System Expansion and Allocation in Life Cycle Assessment. AFR Report 245. Goteborg, Sweden: Chalmers University of Technology.

Heijungs, R. 1997. Economic Drama and the Environmental Stage. Leiden, Netherlands: University of Leiden.

Hofstetter, P., and G. Norris. 2003. Why and how should we assess occupational health impacts in integrated product policy? Environmental Science and Technology 37(10): 2025–2035.

Hunt, R., and W. Franklin. 1996. Personal reflections on the origin and the development of LCA in the USA. International Journal of Life Cycle Assessment 1(1): 4–7.

Hunt, R., and W. Franklin. 1974. Resource and Environmental Profile Analysis of Nine Beverage Container Alternatives. EPA 530-SW-91c. Washington, D.C.: Environmental Protection Agency.

ISO (International Standards Organization). 1998. Environmental Management–Life Cycle Assessment: Goal and Scope Definition and Inventory Analysis. ISO Standard 14041. Geneva: International Standards Organization.

Känzig J. 2003. Input/Output Life Cycle Assessment of Air Transportation. Masters thesis, Life Cycle Systems, Environmental Science and Engineering. Lausanne, Switzerland: Ecole Polytechnique Fédérale de Lausanne (EPFL).

Oberbacher, B., H. Nikodem, W. Klopffer. 1996. LCA—How it came about: an early systems analysis of packaging for liquids which would be called an LCA today. International Journal of Life Cycle Assessment 1(2): 62–65.

WBCSD (World Business Council for Sustainable Development). 2003. Information available online at: *<http://www.wbcsd.ch/templates/TemplateWBCSD1/layout.asp?type=p&MenuId=NjA&doOpen=1&ClickMenu=LeftMenu>*.

WCED (World Commission on Environment and Development). 1987. Our Common Future. Oxford, U.K.: Oxford University Press.

World Bank. 2001. Attacking Poverty. World Development Report 2000/2001. Washington, D.C.: World Bank.

WSSD (World Summit on Sustainable Development). 2002. Plan of Implementation for Changing Unsustainable Patterns of Consumption and Production. Johannesburg, South Africa: World Summit on Sustainable Development. Available online at: <*http://www.johannesburgsummit.org/html/documents/summit_docs/plan_final1009.doc*>.

Fundamental Limits of Nanotechnology: How Far Down is the Bottom?

Status, Challenges, and Frontiers of Silicon CMOS Technology

JACK HERGENROTHER
IBM T.J. Watson Research Center
Yorktown Heights, New York

For more than three decades, continued improvements in silicon (Si) transistor density, integration, switching speed and energy, and cost per electronic function have driven the $160-billion semiconductor industry, one of the most dynamic industries in the world. These faster and cheaper technologies have led to fundamental changes in the economies of the United States and other countries around the world. The exponential increase in transistor count that has occurred over the past few decades was accurately predicted by Gordon Moore in 1965. To continue to power the information technology economy, however, the Si industry must remain on the Moore's Law trajectory.

Because Moore's Law has accurately predicted the progress of Si technology over the past 38 years, it is considered a reliable method of predicting future trends. These extrapolated trends set the pace of innovation and define the nature of competition in the Si industry. Progress is now formalized each year in an annual update of the International Technology Roadmap for Semiconductors (ITRS), also known as "The Roadmap" (ITRS, 2003). Table 1 shows the status and key parameters in the 2002 update. ITRS extends for 15 years, but there are no guarantees that the problems confronting Si technology will be solved over this period. ITRS is simply an assessment of the requirements and technological challenges that will have to be addressed to maintain the current rate of exponential miniaturization. At the current pace, it is widely believed that the industry will reach the end of conventional scaling before the end of 2016, perhaps as early as 2010.

The device that supports Moore's Law, known as the metal-oxide semiconductor field-effect transistor (MOSFET), comes in both n-channel and p-channel flavors depending on whether the primary current is carried by electrons or

TABLE 1 2001 Status and Summary of Key Parameters from the 2002 ITRS Update

Year	Technology Node	MPU Gate Length (nm)	ASIC Gate Length (nm)	Logic Density (Mtransistors/cm^2)	SRAM Density (Mtransistors/cm^2)	DRAM Density (Gbit/cm^2)
2001	"130 nm"	65	90	39	184	0.55
2004	"90 nm"	37	53	77	393	1.49
2007	"65 nm"	25	32	154	827	3.03
2010	"45 nm"	18	22	309	1718	6.10
2013	"32 nm"	13	16	617	3532	18.4
2016	"22 nm"	9	11	1235	7208	37.0

Source: ITRS, 2003.

holes, respectively. The types and placement of dopants in a MOSFET determine whether it is n-channel or p-channel. The basic structure of the nMOSFET (Figure 1) consists of a moderately p-doped channel formed near the top surface of a polished Si wafer between two heavily n-doped source and drain regions. On top of the channel is a thin insulating layer of silicon dioxide or oxynitride, which separates the heavily n-doped polysilicon gate from the channel. For gate voltages above the threshold voltage V_T (typically 0.3 to 0.5 *V*), electrons are attracted to the gate but remain separated from it by the insulating gate oxide. These electrons form an "inversion layer" that allows a significant electron current to flow from the source to the drain. The magnitude of this "drive current" I_D is a key MOSFET performance metric.

In digital circuits, MOSFETs essentially behave like switches, changing the current by as much as 12 orders of magnitude via the gate voltage. To enable low-power chips, MOSFETs must be excellent switches and have very small

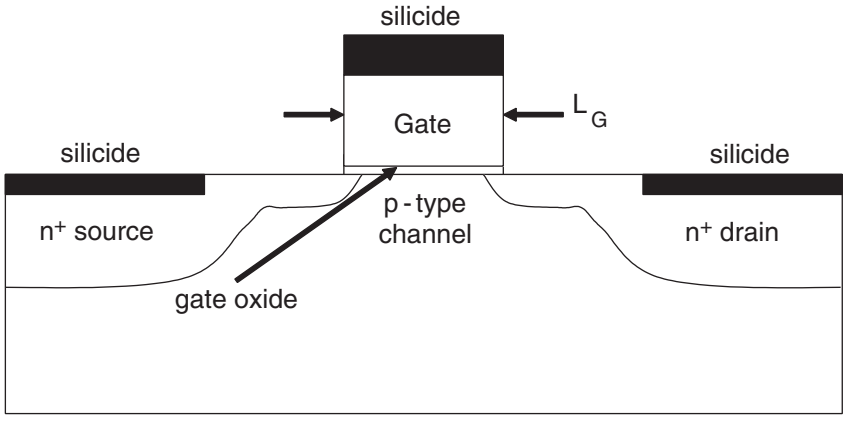

FIGURE 1 Basic structure of a traditional planar nMOSFET.

subthreshold leakage currents (I_{off}). This is the undesirable current that leaks from the source to the drain when 0 V is applied to the gate. In today's state-of-the-art MOSFETs, metal silicides, such as cobalt silicide ($CoSi_2$) and nickel silicide (NiSi), are used to increase the electrical conductivity of the source, drain, and gate. Individual MOSFETs are electrically separated by silicon dioxide deposited in shallow trenches and connected by many levels of miniature copper wire interconnects.

nMOSFETs and pMOSFETs are combined in circuits to form complementary MOS (CMOS) technology. The general principle of traditional CMOS scaling is the reduction of the horizontal and vertical MOSFET dimensions as well as the operating voltage by the same factor (typically 0.7× per technology generation every two to three years) to provide simultaneous improvements of 2× in areal transistor density, 0.7× in switching time, and 0.5× in switching energy. Figure 2 shows the intrinsic switching time of nMOSFETs plotted against their physical gate lengths L_G. The switching time is essentially the time it takes a first transistor carrying a current I_D to charge the input gate capacitance C_G of an identical transistor to the supply voltage V_{DD}.

Note that, although the fastest microprocessors being produced today have clock frequencies in the range of 2 GHz = 1/(500 ps), the intrinsic switching time

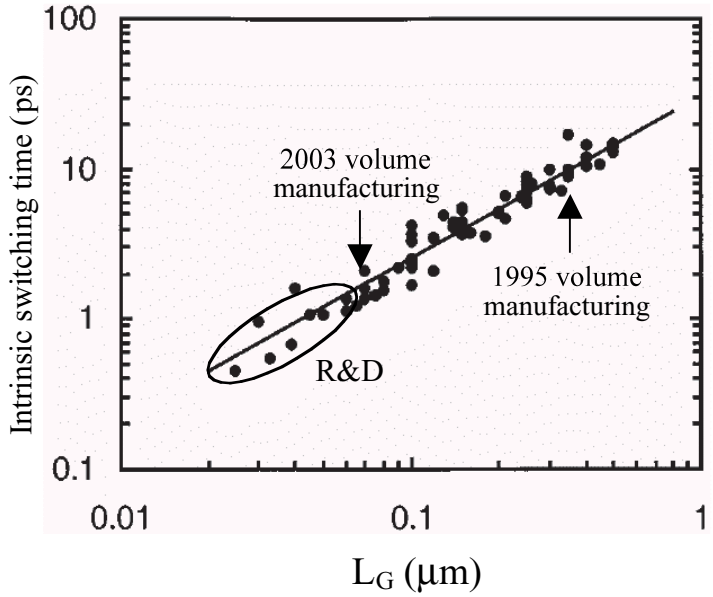

FIGURE 2 The trend of intrinsic switching time $C_G V_{DD}/I_D$ versus physical gate length. The fastest Si MOSFETs produced to date have intrinsic switching times well under 500 fs. Source: Bohr, 2002. Reprinted with permission.

of nMOSFETs is only about 1.5 ps, about 300 times shorter than the inverse clock time. The difference is a consequence of several basic elements of CMOS design architecture: (1) transistors must often drive more than one input gate capacitance; (2) interconnect capacitances must also be charged; (3) typically, about 20 logic stages are cascaded and have to be switched within one clock cycle; and (4) the clock frequency must accommodate the slowest of these cascaded paths. The intrinsic switching time is the critical metric (as opposed to the inverse clock frequency) to keep in mind when comparing the performance of Si MOSFETs with alternative nanodevices intended for logic applications. It is likely that today's research devices with intrinsic switching times under 500 fs will be produced in volume before the end of the decade.

Figure 3 shows the switching energy $C_G V_{DD}^2$, a simple metric that leaves out second-order effects. Note that the fundamental limit on switching-energy transfer during a binary switching transition is $k_B T ln2 \approx 3 \times 10^{-21}$ *J* (Meindl and Davis, 2000). This is roughly five orders of magnitude smaller than the switching energy of Si devices currently being manufactured, indicating that this truly fundamental limit is not an imminent concern.

When discussing state-of-the-art CMOS devices, one must be careful to distinguish research devices from the devices at the cutting edge of volume

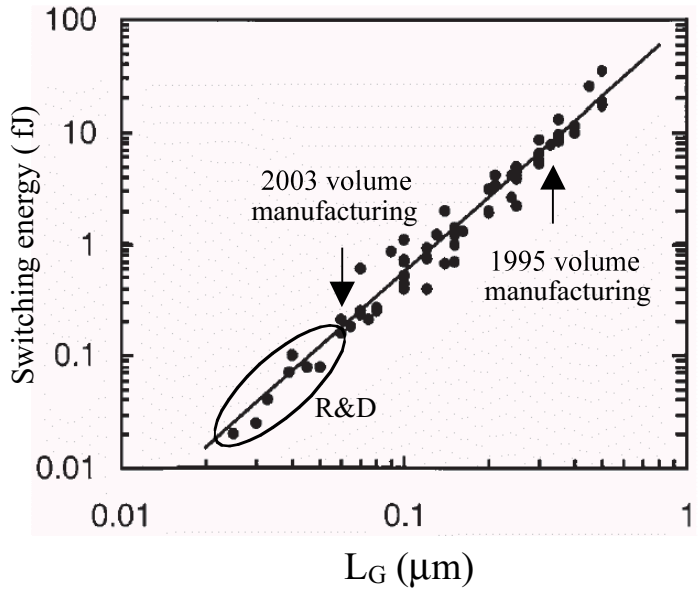

FIGURE 3 The trend of switching energy $C_G V_{DD}^2$ versus physical gate length. Si MOSFETs with less than 30 aJ switching energies have been fabricated. Source: Bohr, 2002. Reprinted with permission.

manufacturing. Manufactured devices must be so finely tuned that they are able to pass a long list of stringent performance, yield, and reliability tests. Depending on the structural, material, or process innovation in question, it may take anywhere from 5 to 15 years of significant, and in some cases worldwide, investment to prepare a new research concept for volume manufacturing. State-of-the-art, bulk Si (Thompson et al., 2002) and Si-on-insulator (SOI) technologies (Khare et al., 2002), which are very near manufacturing, feature gate lengths of 50 nm or smaller and physical gate oxide thicknesses of 12 Å (just a few Si-Si atomic bond lengths) and can pack a six-transistor static RAM (SRAM) cell within a 1.0 μm^2 area. These devices are interconnected with as many as 10 layers of Cu wiring. The critical features are patterned in engineering marvels known as scanners that use deep ultraviolet light with a wavelength of 193 nm. The entire process is carried out on 300-mm diameter wafers in fabrication lines that cost about $3 billion each to build. Note that in recent years physical gate lengths have been shrunk faster than the technology node would suggest. This "hyperscaling" of the gate length, achieved via photolithographic enhancements that are suitable for the reduction of isolated feature sizes, as well as controllable linewidth reduction ("trimming") techniques, has accelerated performance improvements.

Super-scaled bulk MOSFET research devices (in which there is not as high a premium placed on line-width control and optimized performance) have been demonstrated with gate lengths as small as 15 nm (Hokazono et al., 2002; Yu et al., 2001). Although these devices have record-breaking intrinsic switching times and switching energies, they must still undergo significant optimization over the next five to seven years to have a shot at meeting the ITRS performance targets.

Showstoppers to continued scaling have been predicted for three decades, but a history of innovation has sustained Moore's Law in spite of these challenges. However, there are growing signs today that MOS transistors are beginning to reach their traditional scaling limits (ITRS, 2003; Meindl, 2001). The most notable sign is the ever increasing subthreshold leakage current I_{off}, which results from thermionic emission over the source barrier. The subthreshold leakage dictates that the exponential decay of the subthreshold current with gate voltage can be no faster than k_BT/q, or about 60 mV/decade at room temperature.

In practice, MOSFETs do not turn off as abruptly as suggested by the fundamental 60 mV/decade limit because of capacitive divider effects and, more important, various short-channel effects that become significant for sub-100 nm channel lengths. As gate and channel lengths continue to shrink, it is becoming increasingly difficult to design devices with proper electrostatic scalability, that is, reasonable immunity to these short-channel effects. The short-channel effects that impact I_{off} fall into three categories: (1) threshold voltage roll-off, the decrease of threshold voltage with decreasing channel length; (2) drain-induced barrier lowering (DIBL), the decrease of threshold voltage with increasing drain voltage; and (3) degradation of the subthreshold swing with decreasing channel

length. All three effects are related in that they are the simple result of two-dimensional electrostatic effects, and they can dramatically increase the subthreshold leakage current. Because there is a fundamental limit to subthreshold swing at room temperature, short-channel effects put a lower limit on the threshold voltage of the target device. Until recently, V_T has been scaled downward in proportion to the supply voltage V_{DD}.

The floor on V_T has a significant impact of the MOSFET drive current, which depends on the gate overdrive V_{DD}-V_T. To maintain the drive current performance dictated by The Roadmap, subthreshold leakage targets have been significantly relaxed. For example, in the 0.25 μm generation that entered manufacturing circa 1997, I_{off} was maintained at approximately 1 nÅ/μm. In the 90 nm generation soon to be ready for manufacturing, I_{off} for high-performance devices are typically about 40 nÅ/μm. If the leakage goes much higher, it is unlikely that the power generated by chips with hundreds of millions of these transistors can be tolerated. In fact, difficulty in controlling the subthreshold leakage current has already led to a scaling bifurcation between high-performance and low-power transistors that has been formally adopted in the ITRS.

Another significant challenge to continued scaling is the rapid increase of the quantum mechanical tunneling current that passes through the gate oxide as it is progressively thinned. This gate current is rapidly approaching the size of the subthreshold leakage. Current 90 nm generation high-performance devices soon to be in manufacturing have 12 Å SiO_2-based gate oxides that are within at most 2 Å of the limit imposed by gate leakage. Although this gate leakage does not destroy the functionality of digital circuits, in the absence of a solution, chip power-dissipation levels will be unacceptably high.

Lithography, which has been one of the key enablers of scaling, has recently emerged as one of the key challenges to scaling. The Rayleigh equation defines the half-pitch resolution as $R=k_1(\lambda/N_A)$ where k_1 is a coefficient that accounts for the effectiveness of resolution enhancement techniques, λ is the wavelength of the light used, and N_A is the numerical aperture of the lens. In the past, the evolution of lithography occurred primarily through the adoption of successively shorter exposure wavelengths. More recently, because of increasing difficulty of moving to shorter wavelengths, progress in lithography has depended somewhat more heavily on resolution enhancement techniques (such as the use of phase-shift masks and customized illumination techniques), and the development of higher N_A lenses. At the 90 nm node, 193 nm wavelength lithography is used for critical mask levels, whereas more mature 248 nm lithography is widely used for other mask levels.

Although 157 nm lithography was originally targeted for introduction at the 65 nm node (mainstream volume manufacturing in 2007), significant unforeseen problems have arisen that have delayed its availability. As a result, lithographers are facing the challenge of extending 193 nm lithography to the

65 nm node. The lithography required for the 45 nm node (2010) will have to be a very mature 157 nm lithography or a next-generation lithography (NGL) technique such as extreme ultraviolet (EUV) lithography (Attwood et al., 2001) or electron-projection lithography (EPL) (Harriott, 1999). However, the step from 157 nm to 13.5 nm (EUV) is huge, because virtually all materials absorb rather than transmit light at 13.5 nm wavelengths; thus, the only option is reflection optics. With such a short wavelength, EUV allows the use of conservative k_1 and N_A values, but it faces significant challenges: (1) source power and photoresist sensitivity; (2) the production of defect-free reflection masks; and (3) contamination control in reflection optics.

Even if advanced lithographic techniques can be developed to pattern the required ultrafine features in photoresist, these features must be transferred reliably and with high fidelity into the materials that make up the transistors and interconnect. This will require significant advances in Si process technology, in particular in etching and film deposition. A wide range of processes are under development to meet these future needs (Agnello, 2002). Among them are self-limited growth and etch processes, such as atomic-layer deposition, that have the potential to provide atomic-level control (Ahonen et al., 1980; Suntola and Antson, 1977). In MOSFETs at the limit of scaling, it is also critical to provide precise dopant-diffusion control while also enabling the high levels of dopant activation required for low parasitic resistances.

The frontiers of research on Si technology feature a wide spectrum of work aimed at enhancing CMOS performance, solving (or at least postponing) some of the major issues, such as gate leakage, and providing options in case the traditional scaling path falters. One of the most active areas of research is the use of strain to enhance carrier-transport properties (Hwang et al., 2003; Rim et al., 2002a, 2002b; Thompson et al., 2002; Xiang et al., 2003). Si under biaxial tensile strain is well known to exhibit higher electron and hole mobilities than unstrained Si. Devices built with strained-Si channels have shown drive currents as much as 20 percent higher than in unstrained-Si channels (Hwang et al., 2003; Rim et al., 2002a; Thompson et al., 2002; Xiang et al., 2003). A variety of approaches have been used to provide strained-Si channels. The dominant approaches include: (1) the formation of a thin, strained-Si layer on top of a carefully engineered SiGe, relaxed graded buffer layer (Fitzgerald et al., 1992); and (2) the formation of a strained-Si layer on top of a relaxed SiGe on insulator (SGOI) (Huang et al., 2001; Mizuno et al., 2001). In both cases, the larger lattice constant of SiGe constrains the lattice of Si, essentially pulling it into tensile strain. The strained-Si layer must be quite thin to prevent strain relief from occuring through dislocations. Currently intense efforts are being made in several areas: (1) studying how much mobility enhancement translates to an improvement in drive in short-channel devices; (2) reducing the defect densities of strained-Si starting materials; and (3) solving integration issues (e.g., rapid arsenic diffusion) that result from the presence of SiGe in the active area of the

device. In spite of these problems, strained Si is one of the most promising nontraditional technology options.

A much more radical method to improve carrier-transport properties involves the use of germanium as the channel material itself (Chui et al., 2002; Shang et al., 2002). This is primarily driven by its significantly higher bulk mobilities for electrons (2×) and holes (4×). One of the primary challenges is the lack of a stable native Ge oxide, which makes it difficult to passivate the surface of Ge. Renewed interest in Ge MOSFET is due in large part to recent advances in the understanding of oxynitrides and high-k dielectrics (initially aimed at Si-based transistors), suggesting that these materials might be used effectively to passivate the surface of Ge. Work on Ge MOSFETs is still in a very early phase. At the time of this writing, only Ge pMOSFETs have been demonstrated, with several groups attempting to build the first Ge nMOSFETs.

Returning to the imminent challenges facing Moore's Law, one possible solution to the difficult gate-leakage problem is the introduction of high-k (high dielectric constant) gate dielectrics (Wallace and Wilk, 2002). Many different high-k materials, such as Si_3N_4, HfO_2, ZrO_2, TiO_2, Ta_2O_5, La_2O_3, and Y_2O_3, have been studied for use as a MOSFET gate dielectric. The materials that have received the most attention recently are hafnium oxide (HfO_2) (Gusev et al., 2001; Hobbs et al., 2003), hafnium silicate ($HfSi_xO_y$) (Wilk et al., 2000), and nitrided hafnium silicate ($HfSi_xO_yN_z$) (Inumiya et al., 2003; Rotondaro et al., 2002). Because the main objective of scaling the gate oxide is to increase the capacitance density of the gate insulator, high-k dielectrics are attractive because the same capacitance density can be achieved with a physically thicker (and if the new material's barrier height is sufficient, a lower gate-leakage) layer. However, after 40 years, the Si/SiO_2 system has almost ideal properties, many of which have yet to be satisfied by a new high-k dielectric. Replacing SiO_2 is not as easy as it appears. Among these requirements is material-bulk and interface properties comparable to those of SiO_2. The new material must also exhibit thermal stability during the temperature cycles required in Si processing, low dopant-diffusion coefficients, and sufficient reliability under constant voltage stress. Although rapid progress has been made, no group to date has demonstrated a sufficiently thin high-k dielectric that preserves the transport properties (i.e., carrier mobilities) of the Si/SiO_2 system. An interesting fact about high-k gate dielectrics is that they will be required for moderate performance, low-power applications (e.g., hand-held communication devices) before they are needed for high-performance applications (e.g., microprocessors).

Another approach to increasing the capacitance density of the doped polySi/gate dielectric/Si stack is to replace the doped polySi in the gate with a metal (Kedzierski et al., 2002; Lee et al., 2002; Ranade et al., 2002; Samavedam et al., 2002; Xiang et al., 2003). This would eliminate undesirable polysilicon-depletion effects, which typically add an effective thickness of 4 to 5 Å to the thickness of the gate oxide. With physical gate-oxide thicknesses currently in the 12 Å range,

there is a significant degradation of capacitance density. One of the challenges to metal-gate technology is that the work function of the gate material is used to set the threshold voltage of the MOSFET. For bulk MOSFETs, a suitable process must be devised that allows the formation of tunable or dual work function metal gates (Lee et al., 2002; Ranade et al., 2002).

In recent years, double-gate MOSFETs have received a good deal of attention, and a variety of approaches to building them have been investigated (Guarini et al., 2001; Hisamoto et al., 1998; Wong et al., 1997) (Figure 4). These devices provide several key advantages over traditional planar MOSFETs (Wong et al., 2002): (1) enhanced electrostatic scalability because two gates are in close proximity to the current-carrying channel; (2) reduced-surface electric fields that can significantly improve the carrier mobilities and the $C_G V_{DD}/I_D$ performance; and (3) because the electrostatics are controlled with geometry (rather than with doping as in the planar MOSFET), they can be built with undoped channels. This eliminates the problems caused by statistical fluctuations in the number of dopants in the device channel, which is becoming an increasing concern for planar MOSFETs.

Double-gate MOSFETs present difficult fabrication challenges. The most significant of these is that the Si channel must be made sufficiently thin (no thicker than about 2/3 L_G) and must be precisely controlled. In addition, the two gates must be aligned to each other as well as to the source-drain doping to minimize the parasitic capacitance that could otherwise negate the advantage derived from enhanced electrostatic scalability.

The most thoroughly studied double-gate MOSFET is known as the FinFET (Hisamoto et al., 1998; Kedzierski et al., 2001; Yu et al., 2002) and in slightly modified forms as the tri-gate transistor (Doyle et al., 2003) or omega-FET

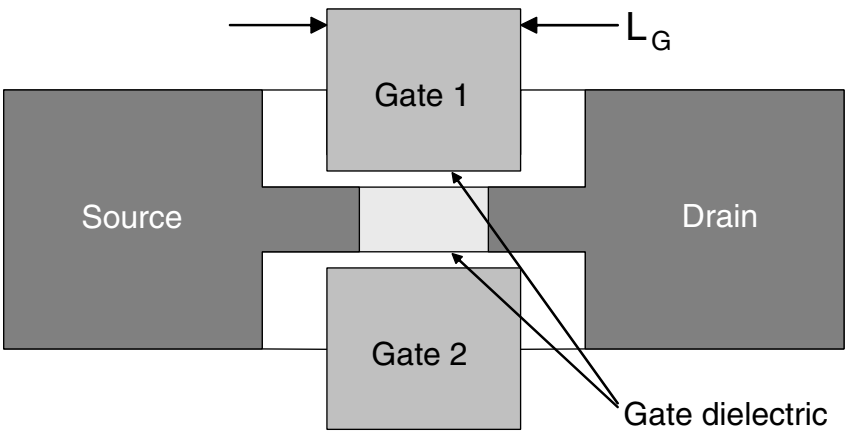

FIGURE 4 Structure of a representative double-gate MOSFET.

(Park et al., 2003; Yang et al., 2002) (Figure 5). Of all double-gate MOSFETs that have been fabricated, the FinFET is the most similar to the traditional planar MOSFET. This is a huge advantage when one considers the possibility of altering the device structure that supports a $160-billion world market and that chips must be manufactured in which every single one of a billion of these devices all work, satisfies tight electrical performance tolerance requirements, and continues to operate reliably for 10 to 30 years. In the FinFET, a thin (10 to 30 nm) vertical fin of Si is created on top of an oxide layer. The gate of the device wraps around both sides and the top of the fin to provide excellent control of the channel potential and resistance to short-channel effects. Within the past year, aggressively-scaled FinFETs with 10 nm gate lengths and intrinsic switching times ($C_G V_{DD}/I_D$) of 300 fs have been demonstrated (Yu et al., 2002), and a functional six-transistor SRAM cell using FinFETs was realized (Nowak et al., 2002).

In conclusion, it can be argued that one of the most significant challenges to the continuation of traditional scaling is the need for continued improvements in performance (e.g., in $C_G V_{DD}/I_D$) amidst other scaling constraints, which are primarily manifested as limits on static-power dissipation, power that is dissipated even when transistors are not switching. In light of this fact, it is very unlikely that Moore's Law will ever hit a brick wall. More likely, it will first enter a regime in which CMOS scaling provides diminishing marginal performance benefits and, ultimately, a negative marginal performance benefit. Some engineers in the field refer to this scenario as Moore's Summit rather than as the end of Moore's Law. Clearly, there will be no single, well-defined "summit," but

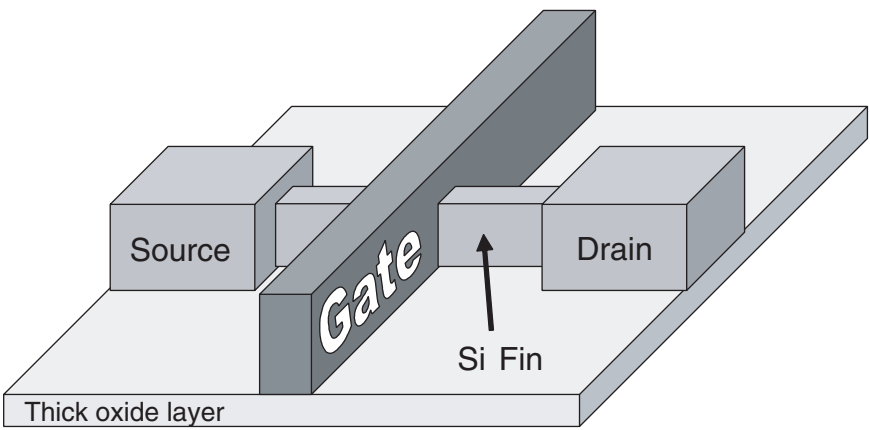

FIGURE 5 Basic structure of the FinFET, the most thoroughly studied double-gate MOSFET.

many summits depending on the application (Frank, 2002; Frank, et al., 2001). Device designs for different applications are likely to continue to diverge and may possibly use different device structures and materials. Rather than a single end point for scaling, there will be a variety of end points optimized for particular applications.

REFERENCES

Agnello, P.D. 2002. Process requirements for continued scaling of CMOS: the need and prospects for atomic-level manipulation. IBM Journal of Research and Development 46 (2/3): 317–338.

Ahonen, M., M. Pessa, T. Suntola. 1980. A study of ZnTe films grown on glass substrates using an atomic layer evaporation method. Thin Solid Films 65: 301–307.

Attwood, D., G. Kubiak, D. Sweeney, S. Hector, C. Gwyn. 2001. Progress and future directions of the U.S. EUV lithography program. International Microprocesses and Nanotechnology Conference, Digest of Papers 78.

Bohr, M.T. 2002. Nanotechology goals and challenges for electronic applications. IEEE Transactions on Nanotechnology 1(1): 56–62.

Chui, C.O., K. Hyoungsuh, D. Chi, B.B. Triplett, P.C. McIntyre, K.C. Saraswat. 2002. A sub-400/spl deg/C germanium MOSFET technology with high-/spl kappa/dielectric and metal gate. International Electron Devices Meeting Technical Digest 437–440.

Doyle, B., B. Boyanov, S. Datta, M. Doczy, S. Hareland, B. Jin, J. Kavalieros, T. Linton, R. Rios, R. Chau. 2003. Tri-gate fully-depleted cmos transistors: fabrication, design and layout. Symposium on VLSI Technology, Digest of Technical Papers 133–134.

Fitzgerald, E.A., Y.-H. Xie, D. Monroe, P.J. Silverman, J.M. Kuo, A.R. Kortan, F.A. Thiel, and B.E. Weir. 1992. Relaxed GexSi1–x structures for III–V integration with Si and high mobility two-dimensional electron gases in Si. Journal of Vacuum Science and Technology B 10(4): 1807–1819.

Frank, D.J. 2002. Power-constrained CMOS scaling limits. IBM Journal of Research and Development 46(2/3): 235–344.

Frank, D.J., R.H. Dennard, E. Nowak, P.M. Solomon, Y. Taur, H.-S.P. Wong. 2001. Device scaling limits of Si MOSFETs and their application dependencies. Proceedings of the IEEE 89(3): 259–288.

Guarini, K., P.M. Solomon, Y. Zhang, K.K. Chan, E.C. Jones, G.M. Cohen, A. Krasnoperova, M. Ronay, O. Dokumaci, J.J. Bucchignano, C. Cabra. Jr., C. Lavoie, V. Ku, D.C. Boyd, K.S. Petrarca, I.V. Babich, J. Treichler, P.M. Kozlowski, J.S. Newbury, C.P.D Emic, R.M. Sicina, and H.-S.P. Wong. 2001. Triple-self-aligned, planar double-gate MOSFETs: devices and circuits. International Electron Devices Meeting Technical Digest 425–428.

Gusev, E., D. Buchanan, E. Cartier, A. Kumar, D. DiMaria, S. Guha, A. Callegari, S. Zafar, P. Jamison, D. Neumayer, M. Copel, M. Gribelyuk, H. Okorn-Schmidt, C. D'Emic, P. Kozlowski, K. Chan, N. Bojarczuk, L.-A. Rannarsson, P. Ronsheim, K. Rim, R. Fleming, A. Mocuta, and A. Ajmera. 2001. Ultrathin high-*k* gate stacks for advanced CMOS devices. International Electron Devices Meeting Technical Digest 451–454.

Harriott, L.R. 1999. A new role for e-beam: electron projection. IEEE Spectrum 36(7): 41.

Hisamoto, D., W.-C. Lee, J. Kedzierski, E. Anderson, H. Takeuchi, K. Asano, K, T.-J. King, J. Bokor, C. Hu. 1998. A folded-channel MOSFET for deep-sub-tenth micron era. International Electron Devices Meeting Technical Digest 1032–1034.

Hobbs, C., L. Fonseca, V. Dhandapam, S. Sarnavedarn, B. Taylor, J. Grant, L. Dip, D. Triyoso, R. Hegde, D. Gilmer, R. Garcia, D. Roan, L. Lovejoy, R. Rai, L. Hebert, H. Tseng, B. White, P. Tobin. 2003. Fermi level pinning at the polySi/metal oxide interface. Symposium on VLSI Technology, Digest of Technical Papers 9–10.

Hokazono, A., K. Ohuchi, M. Takayanagi, Y. Watanabe, S. Magoshi, Y. Kato, T. Shimizu, S. Mori, H. Oguma, T. Sasaki, H. Yoshimura, K. Miyano, N. Yasutake, H. Suto, K. Adachi, H. Fukui, T. Watanabe, N. Tamaoki, Y. Toyoshima, H. Ishiuchi. 2002. 14 nm gate length CMOSFETs utilizing low thermal budget process with poly-SiGe and Ni salicide. International Electron Devices Meeting Technical Digest 639–642.

Huang, L.J., J.O. Chu, D.F. Canaperi, C.P. D'Emic, R.M. Anderson, S.J. Koester, and H.-S. Philip Wong. 2001. SiGe-on-insulator prepared by wafer bonding and layer transfer for high-performance field-effect transistors. Applied Physics Letter 78(9): 1267–1269.

Hwang, J.R., J.H. Ho, S.M. Ting, T.P. Chen, Y.S. Hsieh, C.C. Huang, Y.Y. Chiang, H.K. Lee, A. Liu, T.M. Shen, G. Braithwaite, M. Currie, N. Gerrish, R. Hammond, A. Lochtefeld, F. Singaporewala, M. Bulsara, Q. Xiang, M.R. Lin, W.T. Shiau, Y.T. Loh, J.K. Chen, S.C. Chien, F. Wen. 2003. Performance of 70mn strained-Si CMOS devices. Symposium on VLSI Technology, Digest of Technical Papers 103–104.

Inumiya, S., K. Sekine, S. Niwa, A. Kaneko, M. Sato, T. Watanabe, H. Fukui, Y. Kamata, M. Koyama, A. Nishiyama, M. Takayanagi; K. Eguchi, Y. Tsunashima. 2003. Fabrication of HfSiON gate dielectrics by plasma oxidation and nitridation, optimized for 65nm node low power CMOS applications. Symposium on VLSI Technology, Digest of Technical Papers 17–18.

ITRS (International Technology Roadmap for Semiconductors). 2003. Available online at: *<http://public.itrs.net/Home.htm>*.

Kedzierski, J., D.M. Fried, E.J. Nowak, T. Kanarsky, J. Rankin, H. Hanafi, W. Natzle, D. Boyd, Y. Zhang, R. Roy, J. Newbury, C. Yu, Q. Yang, P. Saunders, C. Willets, A. Johnson, S. Cole, H. Young, N. Carpenter, D. Rakowski, B.A. Rainey, P. Cottrell, M. Ieong, and P. Wong. 2001. High-performance symmetric gate and CMOS-compatible Vt asymmetric gate FinFET devices. International Electron Devices Meeting Technical Digest 437–440.

Kedzierski, J., E. Nowak, T. Kanarsky, Y. Zhang, D. Boyd, R. Carruthers, C. Cabral, R. Amos, C. Lavoie, R. Roy, J. Newbury, E. Sullivan, J. Benedict, P. Saunders, K. Wong, D. Canaperi, M. Krishnan, K.-L. Lee, B.A. Rainey, D. Fried, P. Cottrell, H.-S.P. Wong, M. Ieong, W. Haensch. 2002. Metal-gate FinFET and fully-depleted SOI devices using total gate silicidation. International Electron Devices Meeting Technical Digest 247–250.

Khare, M., S.H. Ku, R.A. Donaton, S. Greco, C. Brodsky, X. Chen, A. Chou, R. DellaGuardia, S. Deshpande, B. Doris, S.K.H. Fung, A. Gabor, M. Gribelyuk, S. Holmes, F.F. Jamin, W.L. Lai, W.H. Lee, Y. Li, P. McFarland, R. Mo, S. Mittl, S. Narasimha, D. Nielsen, R. Purtell, W. Rausch, S. Sankaran, J. Snare, L. Tsou, A. Vayshenker, T. Wagner, D. Wehella-Gamage, E. Wu, S. Wu, W. Yan, E. Barth, R. Ferguson, P. Gilbert, D. Schepis, A. Sekiguchi, R. Goldblatt, J. Welser, K.P. Muller, P. Agnello. 2002. A high performance 90nm SOI technology with 0.992 /spl mu/m/sup 2/6T-SRAM cell. International Electron Devices Meeting Technical Digest 407–410.

Lee, J.-H., H. Zhong, Y.-S. Suh, G. Heuss, J. Gurganus, B. Chen, V. Misra. 2002. Tunable work function dual metal gate technology for bulk and non-bulk CMOS. International Electron Devices Meeting Technical Digest 359–362.

Meindl, J. 2001. Special issue on limits of semiconductor technology. Proceedings of the IEEE 89(3): 223–226.

Meindl, J.D., and J.A. Davis. 2000. The fundamental limit on binary switching energy for terascale integration (TSI). IEEE Journal of Solid-State Circuits 35(10): 1515–1516.

Mizuno, T., N. Sugiyama, A. Kurobe, S.-i. Takagi. 2001. Advanced SOI p-MOSFETs with strained-Si channel on SiGe-on-insulator substrate fabricated by SIMOX technology. IEEE Transactions on Electron Devices 48(8): 1612–1618.

Nowak, E.J., B.A. Rainey, D.M. Fried, J. Kedzierski, M. Ieong, W. Leipold, J. Wright, M. Breitwisch. 2002. A functional FinFET-DGCMOS SRAM cell. International Electron Devices Meeting Technical Digest 411–414.

Park, T., S. Choi, D.H. Lee, J.R.Yoo, B.C. Lee, J.Y. Kim, C.Q. Lee, K.K. Chi, S.H. Hong, S.J. Hyun, Y.G. Shin, J.N. Han, I.S. Park, S.H. Chung, J.T. Moon, E. Yoon, J.H. Lee. 2003. Fabrication of body-tied FinFETS (omega MOSFETS) using bulk Si wafers. Symposium on VLSI Technology, Digest of Technical Papers 135–136.

Ranade, P., Y.-K. Choi, D. Ha, A. Agarwal, M. Arneen, T.-J. King. 2002. Tunable work function molybdenum gate technology for FDSOI-CMOS. International Electron Devices Meeting Technical Digest 363–366.

Rim, K., J. Chu, H. Chen, K.A. Jenkins, T. Kanarsky, K. Lee, A. Mocuta, H. Zhu, R. Roy, J. Newbury, J. Ott, K. Petrarca, P. Mooney, D. Lacey, S. Koester, K. Chan, D. Boyd, M. Ieong, H.-S. Wong. 2002a. Characteristics and device design of sub-100 nm strained Si N- and PMOSFETs. Symposium on VLSI Technology, Digest of Technical Papers 98–99.

Rim, K., S. Narasimha, M. Longstreet, A. Mocuta, J. Cai. 2002b. Low field mobility characteristics of sub-100 nm unstrained and strained Si MOSFETs. International Electron Devices Meeting Technical Digest 43–46.

Rotondaro, A.L.P., M.R. Visokay, J.J. Chambers, A. Shanware, R. Khamankar, H. Bu, R.T. Laaksonen, L. Tsung, M. Douglas, R. Kuan, M.J. Bevan, T. Grider, J. McPherson, L. Colombo. 2002. Advanced CMOS transistors with a novel HfSiON gate dielectric. Symposium on VLSI Technology, Digest of Technical Papers 148–149.

Samavedam, S.B., L.B. La, J. Smith, S. Dakshina-Murthy, E. Luckowski, J. Schaeffer, M. Zavala, R. Martin, V. Dhandapani, D. Triyoso, H.H. Tseng, P.J. Tobin, D.C. Gilmer, C. Hobbs, W.J. Taylor, J.M. Grant, R.I. Hegde, J. Mogab, C. Thomas, P. Abramowitz, M. Moosa, J. Conner, J. Jiang, V. Arunachalarn, M. Sadd, B.-Y. Nguyen, B. White. 2002. Dual-metal gate CMOS with HfO/sub 2/ gate dielectric. International Electron Devices Meeting Technical Digest 433–436.

Shang, H., Okorn-Schmidt, K.K. Chan, M. Copel, J.A. Ott, P.M. Kozlowski, S.E. Steen, S.A. Cordes, H.-S.P. Wong, E.C. Jones, W.E. Haensch. 2002. High mobility p-channel germanium MOSFETs with a thin Ge oxynitride gate dielectric. International Electron Devices Meeting Technical Digest 441–444.

Suntola, T., and J. Antson. A Method for Producing Compound Thin Films. U.S. Patent 4,058,430. Filed 1975, issued 1977.

Thompson, S., N. Anand, M. Armstrong, C. Auth, B. Arcot, M. Alavi, P. Bai, J. Bielefeld, R. Bigwood, J. Brandenburg, M. Buehler, S. Cea, V. Chikarmane, C. Choi, R. Frankovic, T. Ghani, G. Glass, W. Han, T. Hoffmann, M. Hussein, P. Jacob, A. Jain, C. Jan, S. Joshi, C. Kenyon, J. Klaus, S. Klopcic, J. Luce, Z. Ma, B. Mcintyre, K. Mistry, A. Murthy, P. Nguyen, H. Pearson, T. Sandford, R. Schweinfurth, R. Shaheed, S. Sivakumar, M. Taylor, B. Tufts, C. Wallace, P. Wang, C. Weber, M. Bohr. 2002. A 90 nm logic technology featuring 50 nm strained Si channel transistors, 7 layers of Cu interconnects, low k ILD, and 1 /spl mu/m/sup 2/ SRAM cell. International Electron Devices Meeting Technical Digest 61–64.

Wallace, R.M., and G. Wilk. 2002. High-k gate dielectric materials. Materials Research Society Bulletin 27(3): 192–197.

Wilk, G.D., R.M. Wallace, and J.M. Anthony. 2000. Hafnium and zirconium silicates for advanced gate dielectrics. Journal of Applied Physics 87(1): 484–492.

Wong, H.-S.P., 2002. Beyond the conventional transistor. IBM Journal of Research and Development 46(2/3): 133–168.

Wong, H.-S.P., K. Chan, and Y. Taur. 1997. Self-aligned (top and bottom) double-gate MOSFET with a 25nm thick Si channel. International Electron Devices Meeting Technical Digest 427–430.

Xiang, Q., J.-S. Goo, J. Pan, B. Yu, S. Ahmed, J. Zhang, M.-R. Lin. 2003. Strained Si nmos with nickel-silicide metal gate. Symposium on VLSI Technology, Digest of Technical Papers 101–102.

Yang, F.-L., H.-Y. Chen, F.-C. Chen, C.-C. Huang, C.-Y. Chang, H.-K. Chiu, C.-C. Lee, C.-C. Chen, H.-T. Huang, C.-J. Chen, H.-J. Tao, Y.-C. Yeo, M.-S. Liang, C. Hu. 2002. 25 nm CMOS omega FETs. International Electron Devices Meeting Technical Digest 255–258.

Yu, B., L. Chang, S. Ahmed, H. Wang, S. Bell, C.-Y. Yang, C. Tabery, C. Ho, Q. Xiang, T.-J. King, J. Bokore, C. Hu, M.-R. Lin, D. Kyser. 2002. FinFET scaling to 10 nm gate length. International Electron Devices Meeting Technical Digest 251–254.

Yu, B., H. Wang, A. Joshi, Q. Xiang, E. Ibok, and M.-R. Lin. 2001. 15 nm gate length planar CMOS transistor. International Electron Devices Meeting Technical Digest 937–939.

Molecular Electronics

JAMES R. HEATH
Department of Chemistry
California Institute of Technology
Pasadena, California

In molecular-electronics research, molecules are used to yield the active and passive components (switches, sensors, diodes, resistors, LEDs, etc.) of electronic circuits or integrated circuits. For certain applications, such as molecular-based memory or logic circuitry, the devices are simply molecular-based analogues of devices with more conventional silicon-based circuitry. In those cases, molecular-electronics components may have the advantages of less manufacturing complexity, lower power consumption, and easier scaling. The field of molecular electronics is evolving rapidly, and even though there are no commercial applications as yet, the science coming out of this research is spectacular.

Consider the circuits of nanowires shown in the electron micrograph in Figure 1 (Melosh et al., 2003). The smallest (100-element) crossbar in this image is patterned at a density approaching $10^{12}/cm^2$, and the wire diameter is approximately 8 nm. With species like boron or arsenic, at a doping level of $10^{18}/cm^3$, a similar 8-nm diameter micrometer-long segment of silicon wires would have 20 to 30 dopant atoms; a junction of two crossed wires would contain approximately 0.1 to 0.2 dopant atoms. Thus, conventional field-effect transistors fabricated at these wiring densities might exhibit nonstatistical and perhaps unpredictable behavior. In fact, the patterning method (called superlattice nanowire pattern transfer [SNAP]) that produced the generation of patterns shown in Figure 1 can be used to prepare ultradense arrays of silicon nanowires. Thus, for the first time, researchers can interrogate the statistics of doping and other materials-type fluctuations that are expected to become important at the nanoscale.

Doping fluctuations, however, are relatively trivial compared to other problems encountered by engineers trying to scale conventional, silicon-based inte-

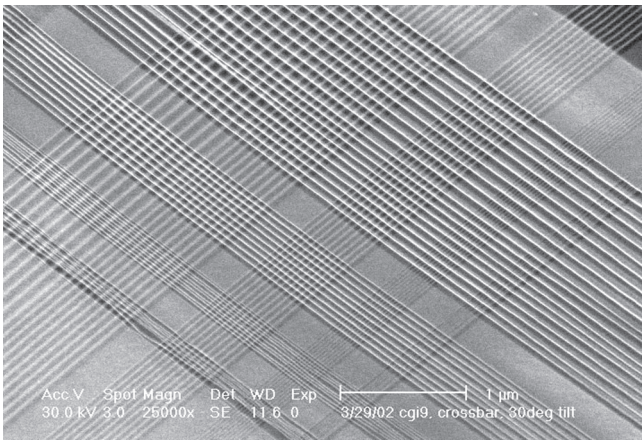

FIGURE 1 A series of 100-element crossbar circuits, the prominent circuits of molecular electronics. The molecule of interest is typically sandwiched between the intersection of two crossed wires. This very simple circuit can be used even in the presence of manufacturing defects and can be fabricated at dimensions that far exceed the best lithographic methods. The device densities in these circuits approach $10^{12}/cm^2$ for the smallest crossbars.

grated circuits to significantly higher densities than are produced now. Power consumption (just from leakage currents through the gate oxide) is perhaps the most serious issue, but the lack of patterning techniques and high fabrication costs are also important.

In fact, no one is seriously contemplating scaling standard electronics device concepts to molecular dimensions (Packan, 1999); alternative strategies are being pursued, although they are all in their early stages. These alternatives include molecular electronics, spintronics, quantum computing, and neural networks, all of which have so-called "killer applications." For quantum computing, it is the reduced scaling of various classes of NP-hard problems. For spintronics, it is a memory density that scales exponentially with numbers of coupled spin transistors. For molecular electronics, it is vastly improved energy efficiency per bit operation, as well as continued device scaling to true molecular dimensions. For true neural networks, it is greatly increased connectivity and, therefore, a greatly increased rate of information flow through a circuit.

Because molecular electronics borrows most heavily from current technologies, in the past few years it has advanced to the point that many major semiconductor-manufacturing companies, including IBM, Hewlett Packard, and LG (Korea), are launching their own research programs in molecular electronics.

At device areas of a few tens of square nanometers, molecules have a certain fundamental attractiveness because of their size, because they represent the ulti-

mate in terms of atomic control over physical properties, and because of the diverse properties (e.g., switching, dynamic organization, and recognition) that can be achieved through such control.

In the crossed-wire circuit shown in Figure 1 (called a crossbar circuit), the molecular component is typically sandwiched between the intersection of two crossing wires. Molecular-electronics circuits based on crossbar architectures can be used for logic, sensing, signal routing, and memory applications (Luo et al., 2002). To realize such applications, many things must be considered simultaneously: the design of the molecule; the molecule/electrode interface; electronically configurable and defect-tolerant circuit architectures; methods of bridging the nanometer-scale densities to the submicrometer densities achievable with lithography; and others (Heath and Ratner, 2003). Using a systems approach in which all of these issues are dealt with consistently and simultaneously, we have been able to fabricate and demonstrate simple molecular-electronics-based logic, memory, and sensing circuitry.

The active device elements in these circuits are molecular-mechanical complexes (Figure 2) organized at each junction within the crossbar, as shown in

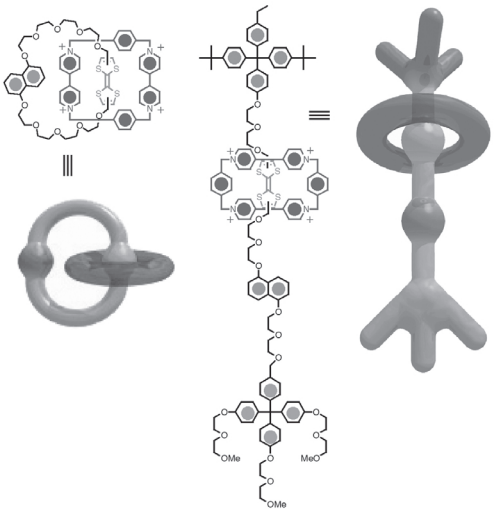

FIGURE 2 Two types of molecular-mechanical complexes that have been demonstrated to work as molecular-electronic switches. At left is a [2]catenane complex, and at right is a [2]rotaxane complex. For both structures, the tetracationic cyclophane (TCP^{4+}) ring encircles the tetrathiafulvalene (TTF) recognition unit. The lowest energy-oxidation state of either complex corresponds to removal of an electron from the TTF group. This leads to a coulombic repulsion of the TCP^{4+} ring so that it encircles the dioxynapthyl group. Molecular switches based on this concept have been demonstrated to work in solution, in solid polymer matrices, immobilized on solid surfaces, and sandwiched between two electrodes. Source: Collier et al, 2002.

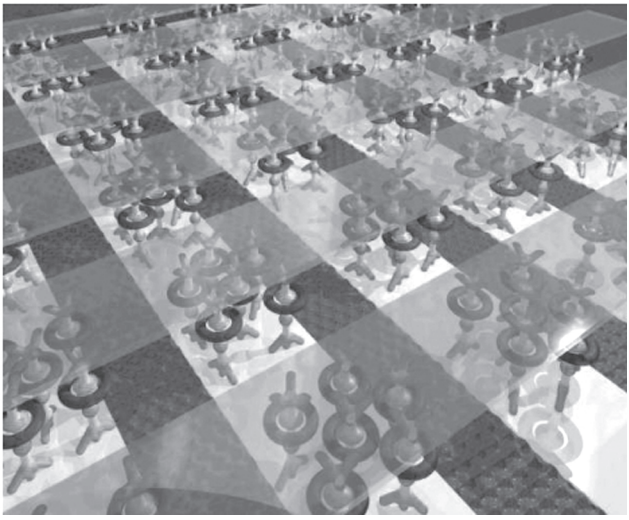

FIGURE 3 A computer graphic showing the molecular components in a crossbar circuit. The [2]rotaxanes molecules shown here are bistable molecular-mechanical compounds. The switching mechanism involves the translation of the ring between two binding sites along the backbone of the molecule. Molecular-electronic circuits have been demonstrated to work as both random-access memory circuits and simple logic circuits.

Figure 3 (Heath and Ratner, 2003; Luo et al., 2002). The molecular structure in Figure 2 (left) is a [2]catenane that consists of two mechanically interlocked rings. One ring is a tetracationic cyclophane (TCP^{4+}); the other is a crown-ether-type ring with two chemical recognition sites, a dioxynapthyl (DN) group and a tetrathiafulvalene group (TTF). The structure on the right side of Figure 2 is a [2]rotaxane that consists of similar chemical motifs. Here the TCP^{4+} ring encircles a dumbbell-shaped structure that has both DN and TTF recognition groups. The molecules are switched via a one- or two-electron process that results in a molecular-mechanical transformation (and a significant change in the electronic structure) of the molecule. This type of molecular actuation, which can be rationally optimized through molecular design and synthesis, provides the basis for information storage or for defining the "open" and "closed" states of a switch.

Typical data from one of our molecular switches is shown in Figure 4 (we have incorporated additional data from various control molecules in this figure). For the structures shown in Figure 2, the controls include [2]catenanes with identical recognition sites (i.e., two DN groups), the dumbbell component of the [2]rotaxane structure, the TCP^{4+} ring, and others. One always does control experiments, of course, but controls are critical here, because these devices are difficult to characterize fully, and one of the few experimental variables is mo-

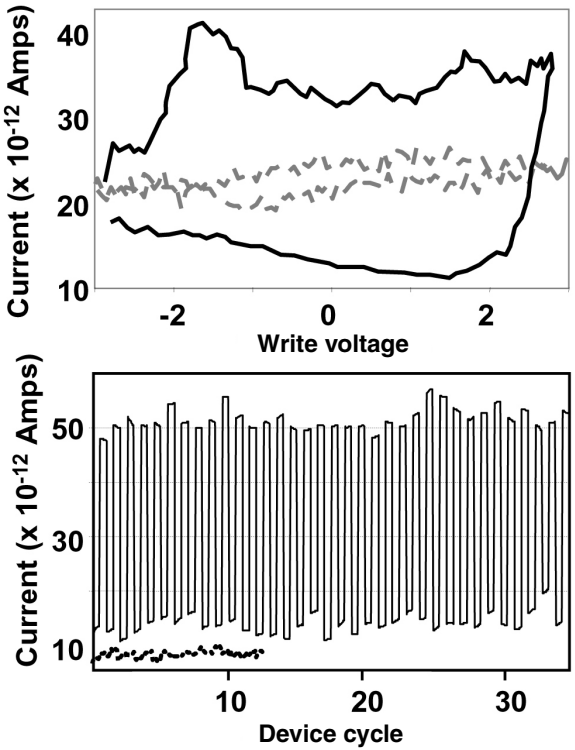

FIGURE 4 The hysteretic response (top) and switch cycling (bottom) of a molecular-electronic switch consisting of a bistable [2]rotaxane sandwiched between a polysilicon bottom electrode and a Ti/Al top electrode. The device dimensions are approximately 50 × 50 nm. Control data from the dumbbell component (the [2]rotaxane without the TCP^{4+} ring) are included.

lecular structure. Perhaps the most fundamental challenge facing scientists constructing solid-state molecular-electronic devices and circuits is developing an intuition for guiding the design of the molecular components.

Charge transport through molecules has been known and studied for a long time, but it has traditionally been a solution-phase science (Joachim et al., 2000; Kwok and Ellenbogen, 2002; Mujica and Ratner, 2002; Nitzan, 2001; Ratner, 2002). For example, a critical component of electron-transfer theory in molecules is the solvent-reorganization coordinate. Molecular synthesis is also a solution-phase endeavor, and all of the analytical techniques (e.g., mass spectrometry, nuclear magnetic resonance (NMR), optical spectroscopy, etc.) are primarily designed to investigate the structure and dynamics of molecules in solution. When molecules are sandwiched between two electrodes, none of these theories and analytical techniques is of any use, and new concepts must be developed.

This situation is exacerbated because certain critical aspects of basic solid-state-device physics cannot be translated directly into the world of molecular electronics. For example, consider the following two fundamental tenets of solid-state materials. First, when two different materials are brought together, their Fermi levels align with each other. Second, if a conducting wire of length *L* has a measured resistance of *R*, then a wire of the same material and diameter, but with a length *2L*, will exhibit a resistance of *2R*. This is called ohmic conductance. Neither of these rules holds true for molecules. This is because molecular orbitals are spatially localized, whereas in solids the atomic orbitals form extended energy bands. When a molecule is adsorbed or otherwise attached to the surface of a solid, there is no straightforward way to think about the position of the molecular orbitals with respect to the Fermi level of the solid. Furthermore, if a molecule of length *L* yields a resistance value of *R*, then a similarly structured, but longer molecule could actually be a better conductor or, perhaps, a far worse conductor. There is no real ohmic conductance in molecules.

Thus, the fundamental challenge is to develop an intuition of how molecules behave in solid-state settings and to use that intuition as feedback to molecular synthesis. In fact, we need a method of characterizing molecular-electronic devices that yields the type of rich information generated by modern NMR spectroscopic methods. Perhaps the most promising approach is single-molecule, three-terminal devices in which a single molecule bridges a very narrow gap (called a break junction) between a source and a drain electrode. A third electrode (called the gate) provides an electric field for tuning the molecular-electronic energy levels into and out of resonance with the Fermi energies of the source and drain electrodes. These devices were originally developed by Park and McEuen and were then further explored by their two groups separately (Liang et al., 2002; Park et al., 2002). Figure 5 shows data from a single-molecule device containing the dumbbell component of the [2]rotaxane molecule shown in Figure 2 (Yu et al., 2003). These data show clearly that these types of device measurements represent a very high information-content analytical method. In an analogy to optical methods, the energy levels resolved in such a device span a very broad range of the spectrum, from ultraviolet to far infrared. This means that molecular orbital energies, and even low-frequency molecular vibrations, might be observable. How molecular orbitals align with the Fermi levels of electrodes, the coupling of molecular vibrations with charge transport through the molecule, and the nature of the molecule/electrode interface states are questions being addressed in various laboratories around the world.

Some of the critical length scales of circuits that can now be fabricated, such as the diameter of the wires and the interwire separation distance (or pitch), are more commonly associated with biological macromolecules, such as proteins, mRNA oligonucleotides, and so on, than with electronics circuitry. In fact, one unique application of nanoscale molecule-electronics circuitry, and perhaps the

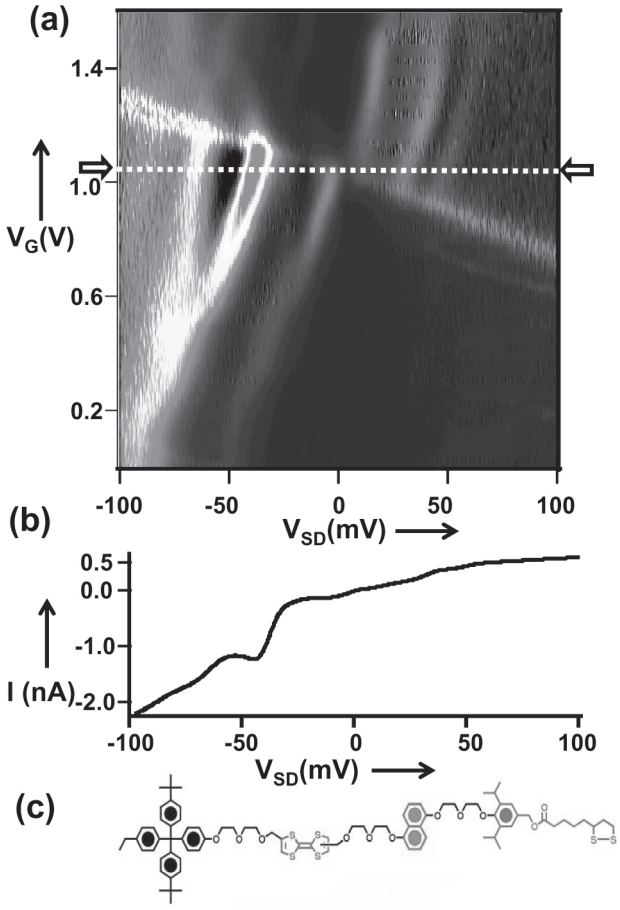

FIGURE 5 Data from a single-molecule, three-terminal device in which the dumbbell component of a [2]rotaxane molecular switch bridges a 4-nm gap separating a source and drain electrode. Voltage on a gate electrode is used to tune the molecular-energy levels into resonance with the electrode-energy levels. **5a.** The conductance (dI/dV) of the molecular junction is plotted as a function of source-drain voltage (x-axis) and gate voltage (y-axis). Light colors indicate high conductance values. **5b.** The current-voltage trace, drawn through the point indicated by the arrows and the dashed white line on the conductance plot. **5c.** The structure of the molecule being measured in this junction. All measurements are carried out at 2 Kelvin.

application that sets this field apart from traditional electronics, is the potential construction of an electrical interface to a single biological cell (Cui et al., 2001). One of our ongoing projects is constructing such an interface designed to measure, simultaneously and in real time, thousands of molecular signatures of gene and protein expression (Heath et al., 2003; Zandonella, 2003). This type of

circuitry may eventually provide a direct connection between the worlds of molecular biology and medicine and the worlds of electrical engineering and integrated circuitry.

ACKNOWLEDGMENTS

The work described in this paper has been supported by the Defense Advanced Research Projects Agency, the MARCO Center, the Semiconductor Research Corporation, the Office of Naval Research, the U.S. Department of Energy, and the National Science Foundation. The results presented here were collected by an outstanding group of students and postdoctoral fellows at UCLA and Caltech.

REFERENCES

Collier, C.P., G. Mattersteig, E.W. Wong, Y. Luo, K. Beverly, J. Sampaio, F.M. Raymo, J.F. Stoddart, and J.R. Heath. 2002. A [2]-catenane based solid-state electronically reconfigurable switch. Science 289(5482): 1172–1175.

Cui, Y., Q. Wei, H. Park, and C.M. Lieber. 2001. Nanowire nanosensors for highly sensitive and selective detection of biological and chemical species. Science 293(5533): 1289–1292.

Heath, J.R., M.E. Phelps, and L. Hood. 2003. Nanosystems biology. Molecular Imaging and Biology 5(5): 312–325.

Heath, J.R., and M.A. Ratner. 2003. Molecular electronics. Physics Today 56(5): 43–49.

Joachim, C., J.K. Gimzewski, and A. Aviram. 2000. Electronics using hybrid-molecular and monomolecular devices. Nature 408(6812): 541–548.

Kwok, K.S., and J.C. Ellenbogen. 2002. Moletronics: future electronics. Materials Today 5(2): 28–37.

Liang, W., M.P. Shores, M. Bockrath, J.R. Long, and H. Park. 2002. Kondo resonance in a single-molecular transistor. Nature 417(6890): 725–729.

Luo, Y., C.P. Collier, K. Nielsen, J. Jeppesen, J. Perkins, E. DeIonno, A. Pease, J.F. Stoddart, and J.R. Heath. 2002. Two-dimensional molecular electronics circuits. ChemPhysChem 3(6): 519–525.

Melosh, N.A., A. Boukai, F. Diana, B. Geradot, A. Badolato, P.M. Petroff, and J.R. Heath. 2003. Ultrahigh-density nanowire lattices and circuits. Science 300(5616): 112–115.

Mujica, V., and M.A. Ratner. 2002. Molecular Conductance Junctions: A Theory and Modeling Progress Report. Pp. 10-1–10-27 in Handbook of Nanoscience, Engineering, and Technology, D. Brenner, S. Lyshevski, G. Iafrate, W.A. Goddard III, eds. Boca Raton, Fla.: CRC Press.

Nitzan, A. 2001. Electron transmission through molecules and molecular interfaces. Annual Review of Physical Chemistry 2(1): 681–750.

Packan, P. 1999. Pushing the limits. Science 24(5436): 2079–2081.

Park, J., A.N. Pasupathy, J.I. Goldsmith, C. Chang, Y. Yaish, J.R. Petta, M. Rinkoski, J.P. Sethna, H.D. Abruña, P.L. McEuen, and D.C. Ralph. 2002. Coulomb blockade and the Kondo effect in single-atom transistors. Nature 417(6890): 722–725.

Ratner, M.A. 2002. Introducing molecular electronics. Materials Today 5(2): 20–27.

Yu, H., Y. Luo, H.R. Tseng, K. Beverly, J.F. Stoddart, and J.R. Heath. 2003. The molecule-electrode interface in single-molecule transistors. Angewandte Chemie 42: 5706–5711.

Zandonella, C. 2003. Cell nanotechnology: the tiny toolkit. Nature 423(6935): 10–12.

Limits of Storage in Magnetic Materials

THOMAS J. SILVA
Magnetic Technology Division
National Institute of Standards and Technology
Boulder, Colorado

The modern hard disk drive is a marvel of multidisciplinary technology. The basic concept behind the disk drive is simple enough: A magnetic solenoid (the write head) applies a localized magnetic field to the surface of a disk coated with a thin film of magnetic material (the media). When the magnetic field from the write head exceeds a threshold (the coercivity), the magnetic polarity of the media is reversed, resulting in a "bit" of recorded data. To read the data back from the media, a field sensor is passed over the same disk surface. Stray magnetic fields from the recorded "bits" alter the resistivity of the field sensor, which is measured with a current source and broadband voltage preamplifier. The disk rotates under the impetus of a precision direct-drive motor, with rotation rates of 15,000 rpm in high-performance drives. Because the head applies localized magnetic fields to write data, an actuator and armature swivels the head to access different tracks of data recorded at specific radii on the disk surface.

Localization of both the write fields and read head sensitivity is achieved by positioning both the write head and read head sensor in very close proximity of the disk surface. The heads are bonded to the rear end of a "slider" that skims across the disk surface on a thin layer of air entrained by the rotating disk. Thus, the operation of a disk drive requires a delicate balance of mechanical, electrical, and magnetic properties, with all three optimized for both economy and performance.

Although the technology is conceptually trivial, in its execution, it strains the limits of human imagination. Magnetic recording capacities have been increasing at an astonishing rate of 100 percent compounded annually for the last

seven years. Areal densities for commercial drives are approaching 1.6×10^{10} bits/cm^2 (100 Gbits/in^2), with a magnetic bit area of only 10,000 square nanometers. While the actuator is reading and writing data to the disk surface, it must keep the head over tracks of data 250 nm wide with nanometer precision. The head used to read and write data from the magnetic platter flies over the disk surface with an average separation of only 10 nm at the breathtaking speed of 215 km/hr (134 mph).

At the same time, the cost per gigabyte for disk storage is unrivaled in the on-line data storage sector. 160 GB drives can now be purchased for a street price of $230, with an equivalent cost per megabyte of only 14 cents! (By comparison, the street price for flash RAM is approaching $250 for 1 GB, or 25 cents per MB—more than *two orders of magnitude* more expensive than the cost per MB for hard-disk-drive storage.) The economic advantage of the magnetic recording process has sustained the disk-drive industry throughout the entire period of computer commercialization. Replacing the disk drive with purely solid-state memory in the computer architecture has never been economically feasible despite the disadvantages associated with a mechanical device.

The economic advantage of disk-drive architecture rests on the recording of data on a featureless surface. The media in a disk drive does not require *any* lithographic processing, which can be heavily leveraged to provide high-density data storage at a highly competitive price. The fabrication of the recording head, on the other hand, relies heavily on precise nanofabrication techniques, including lithography and chemomechanical polishing. Yield demands are predicated on the capability of a single functioning head to access as much as 40 gigabytes of data on one side of a disk three inches in diameter.

One of the astounding dimensions associated with mechanical disk drives is the extremely thin air bearing that separates the media from the head. The reason for such small head/media spacing is quickly apparent when you consider the frequency components for magnetic fields in free space generated by an arbitrary magnetization distribution in a plane at $z = 0$. Because the fields in free space ($z > 0$) must satisfy the Laplace equation, the Fourier components of the magnetic field at spatial frequency k decay as

$$\tilde{H}(k) \propto e^{-kz} \tag{1}$$

This results in a 55 dB signal loss for every increase in spacing equal to a multiple of the recorded bit pattern wavelength. The exponential decay of the stray fields from the recorded pattern in media mandates very small head/media spacing and very small media thickness.

Magnetic data storage depends upon a hysteresis, a fundamental physical property of all ferromagnets. When a magnet field is applied to a magnet and then turned off, the final state of the magnet is a function of the applied field amplitude and direction. Thus, the magnet "remembers" the applied field, permitting the recording of information in a nonvolatile and erasable format. The

fundamental fact that ferromagnets exhibit hysteresis is a manifestation of the spontaneous symmetry breaking that is characteristic of all phase transitions (Pathria, 1996). In the ferromagnetic phase, the collective alignment of the uncompensated electron spins lowers the free energy of the electronic system. However, the collective behavior of the spins requires they all point in some direction. A prerequisite for the ferromagnetic order parameter is the existence of a preferred direction for the ferromagnetic order parameter to align with. However, the existence of a preferred direction in an inversion symmetric material also implies that there must be more than one direction the magnetization tends to align with. By such symmetry arguments, one can show that the free energy of the magnetization will have an angular dependence that exhibits even-order symmetry, with respect to the anisotropy axis, a symmetry axis of the magnetic material. This anisotropy energy is generally written as an energy density (Morrish, 2001):

$$U_K(\theta) = K_1 \sin^2\theta + K_2 \sin^4\theta + \cdots \tag{2}$$

The coefficients K_n in the anisotropy energy expansion are the energies required to rotate the magnetization out of the easy axis direction. This perturbation energy, acting on the ferromagnetic order parameter, imbues the ferromagnet with its hysteretic properties (Stoner and Wohlfarth, 1948). To understand the role of anisotropy in the magnetic recording process, we need only consider the lowest order term in the anisotropy expansion of Eq.2, also referred to as uniaxial anisotropy energy density, K_u.

SUPERPARAMAGNETISM

Let us consider a system in equilibrium that is in contact with a thermal reservoir. From Boltzmann statistics, we know that the *relative* probability that the system occupies a state of energy density ΔU above the ground state energy density U_H is given by

$$\begin{aligned} \Pr(\Delta U) &= \frac{e^{-\frac{(\Delta U + U_H)V}{k_B T}}}{e^{-\frac{(U_H)V}{k_B T}}} \\ &= e^{-\frac{(\Delta U)V}{k_B T}} \end{aligned} \tag{3}$$

where V is the total volume of the magnetic system under consideration. Now, let us assume that the system hops between different energy states with a characteristic time scale of τ_0. If we consider that the probabilities associated with the occupation of any given energy state is given by Eq.3, it quickly becomes apparent that the average frequency $\bar{\nu}$ with which the system occupies any given energy level is given by

$$\bar{\nu} = \frac{1}{\tau_0} e^{-\frac{\Delta U}{k_B T}} \tag{4}$$

So far, we have considered thermal equilibrium processes that determine the rate at which thermal fluctuations perturb the energy of a system. Now, let us consider a magnetic system consisting of the uniform magnetization state, such as a single-domain grain in a polycrystalline magnetic film. The magnetic domain is in thermal contact with several thermal baths, including both vibrations of the crystal lattice (the phonon bath) and nonuniform spin fluctuations in the magnetization itself (the magnon bath.) (We can isolate the uniform magnetization state from all the other nonuniform modes by the same considerations that lead to the Planck distribution for an ensemble of harmonic oscillators, where we consider magnetic excitations to obey just such a distribution.)

The fluctuations in the uniform state manifest themselves as random variations in the magnetization orientation θ relative to the anisotropy axis. If, however, the thermal fluctuations cause the magnetic orientation to move to a position where the total magnetic energy is maximized, we assume that the magnetic state will change its ground state configuration, so that the probability distribution for fluctuations will now be referenced to a new ground state oriented 180 degrees relative to the original magnetization direction with energy density U_L; the system "hops" over the energy barrier imposed by the rotational anisotropy of the magnetic energy. The rate at which a sudden reorientation of the energy ground state (from U_H to U_L) can occur is given by

$$\nu_S = \frac{1}{\tau_0} e^{-\frac{K_u V}{k_B T}} \tag{5}$$

We note that a system with only anisotropy energy will hop between different ground state configurations with equal probability, so that an ensemble that starts in a uniformly magnetized state (such as a collection of grains that form a single magnetic bit in the disk media) will eventually demagnetize (i.e., $|\vec{M}| \to 0$) as a result of thermal hopping over the energy barrier. Under the assumption that the magnetization fluctuations are uncorrelated, the relaxation process should follow the form for a Poisson point process (Helstrom, 1984), which has the solution

$$\begin{aligned} \left|\vec{M}\right| &= M_0 e^{-\nu t} \\ &= M_0 e^{-t/\tau}; \tau = \tau_0 e^{\frac{K_u V}{k_B T}} \end{aligned} \tag{6}$$

This is the Arrhenius-Neel Law for thermal relaxation of magnetic systems (Néel, 1949). It is important to consider that the solution is only appropriate if

all single-domain articles in the system have exactly the same anisotropy energy. A very elegant experimental confirmation of Eq.6 was recently obtained using lithographically patterned arrays of identical magnetic particles, individually measured using spin-dependent tunneling methods (Rizzo et al., 2002).

A note concerning the energy fluctuation time τ_0 (sometimes referred to as the "attempt" time) is in order. This parameter is one of the least understood in the context of superparamagnetic behavior, in part because direct experimental determination of this time scale has only recently been achieved (Rizzo et al., 2002). However, a rather simple theoretical understanding of nonequilibrium statistical mechanics can shed a great deal of light on the nature of the attempt time. The Onsager regression hypothesis states that the correlation time of thermal fluctuations for a system in thermal equilibrium is identical to the time scale at which the same system relaxes back into equilibrium after perturbation by an external stimulus (Chandler, 1987). For example, suppose that a sudden magnetic field pulse is applied to a magnetic system and that the ground state orientation is rotated by an angle $\Delta\theta$. The magnetic system energy will relax into the reoriented ground state in an exponential fashion,

$$U(t) - U_L = \Delta U e^{\frac{t}{\tau_0}} \tag{7}$$

The time scale τ_0 in Eq.7 can be thought of as the time for inelastic scattering events that decrease the net energy of the spins that comprise the magnetic system. According to Onsager's regression hypothesis, the τ_0 in Eq.7 is identical to the attempt time τ_0 in Eq.5.

The relaxation time τ_0 in Eq.7 is also called the "damping" time, empirically derived values that are obtained by fitting data to phenomenological models for magnetization dynamics. The most widely cited of these models is that of Landau and Lifshitz (1935),

$$\frac{d\vec{M}}{dt} = -|\gamma|\mu_0(\vec{M}\times\vec{H}) + \frac{\alpha\gamma\mu_0}{M_S}(\vec{M}\times(\vec{M}\times\vec{H})) \tag{8}$$

The first term describes the purely elastic process of spin precession, by which the spins in the ferromagnet coherently orbit along a path of constant energy. The second term is a phenomenological description of the damping process. By construction, the torque represented by the second term always points in a direction orthogonal to the spin precession and orthogonal to the magnetization. Both of these constraints ensure that the damping describes an inelastic process that conserves the net moment of the magnetic object under consideration. The step-response solution of Eq.8 for a thin-film geometry in the limit of small angle motion about the *y*-axis yields (Silva et al., 1999)

$$M_x(t) \approx \Delta M_x\left(1-\cos(\omega_0 t)e^{-\frac{t}{\tau_0}}\right)$$
$$M_z(t) \approx \Delta M_z \sin(\omega_0 t)e^{-\frac{t}{\tau_0}} \quad (9)$$

where

$$\tau_0 = \frac{2}{\alpha\gamma\mu_0 M_S}; \omega_0 = \gamma\mu_0\sqrt{(M_S+H_K)(H_K)} \quad (10)$$

and the z-axis is normal to the film plane. For most magnetic materials, damping values are experimentally measured to lie in the range $0.1 > \alpha > 0.01$, and the magnetic moment for data storage media is typically on the order of $\mu_0 M_S \sim 0.5$ T. Thus, damping times for media are expected to range over $0.2 < \tau_0 < 2$ ns. Experimental determination of damping is more easily performed with soft magnetic materials that exhibit an effective anisotropy field of less than 8 kA/m (100 Oe). Figure 1 shows free induction decay data obtained from the alloy $Ni_{0.8}Fe_{0.2}$. The data are well fitted to Eq.9, with the measured signal proportional to the in-plane transverse magnetization M_x.

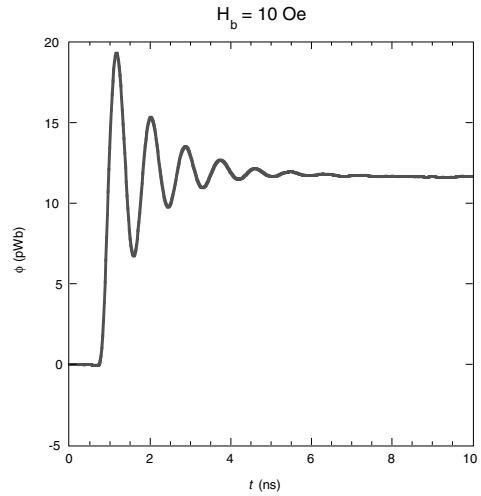

FIGURE 1 Free induction decay data for a 50 nm $Ni_{0.8}Fe_{0.2}$ film. The magnetization is responding to a 50 ps rise time magnetic field pulse applied transverse to the initial magnetization direction. Data were obtained with a pulsed inductive microwave magnetometer, described in detail in Silva et al. (1999). The damping time is 1.2 ns, with an equivalent damping parameter of alpha = 0.008.

Upon inclusion of a statistical distribution for the energy barriers, one finds that the relaxation process is no longer exponential, but is instead logarithmic, with a logarithmic relaxation rate proportional to the width of the energy barrier distribution (Street and Wooley, 1949). To understand why this is the case, a graphical analysis is a convenient pedagogical tool. In Figure 2, we see a plot for the exponential relaxation process at room temperature for a ferromagnetic system with various energy barrier heights, ranging over $3.6 < K_u V/k_B T < 37$, assuming an energy fluctuation rate of 10 GHz. The time axis in Figure 2 is logarithmic. Only an order of magnitude variation in the height of the energy barrier results in a 12 order of magnitude variation in the relaxation time. At the same time, exponential relaxation looks like a step function when plotted on a logarithmic time axis. If one assumes that distribution of energy barriers is flat, integration of the temporal relaxation process is quasi-logarithmic (i.e., the average response for a distribution of step functions in log time will be a linear function in log time). In Figure 3, we show the experimental data from *The Physical Principles of Magnetism* where the thermal switching times τ were continuously varied by applying a longitudinal bias field antiparallel to the original direction of the magnetization in a patterned magnetic element (Morrish, 2001). The data show the clear hallmarks of thermally driven reversal—exponential dependence of switching probability on applied field duration, and a continuously decreasing time constant that depends on the height of the energy barrier that prevents switching.

The transition from exponential relaxation to logarithmic relaxation in an inhomogeneous system will occur under conditions when the energy barrier is negligible. This was observed in Rizzo et al. (1999) in an experiment where thin-film media were subjected to magnetic field pulses of varying intensity and duration. In Figure 4, the data from this experiment are reproduced. For pulses of less than 10 ns duration, the response is clearly exponential, with a time constant of several nanoseconds. For pulses of longer duration, the response is exponential, reflecting thermal activation with a broad distribution of energy barriers.

The dependence of hysteresis on temperature is a fundamental limit that impedes the perpetual increase in the areal density of disk drives (Lu and Charap, 1994). To understand why this is the case, we must consider the various noise sources in the magnetic recording process. There are three principle contributions to the noise in disk drives: (1) Johnson noise, (2) media noise, and (3) magnetic fluctuation noise. Johnson noise is associated with voltage fluctuations in the read sensor and preamp used to read back the magnetic data. As long as the electrical resistance of the read head is kept constant by proper use of dimension scaling rules, this noise source can be kept under control. Media second noise source stems from the spatial variation of the magnetization distribution in the media. Although the media is a thin metallic polycrystalline film of magnetic material, growth conditions for the sputter deposition are chosen to

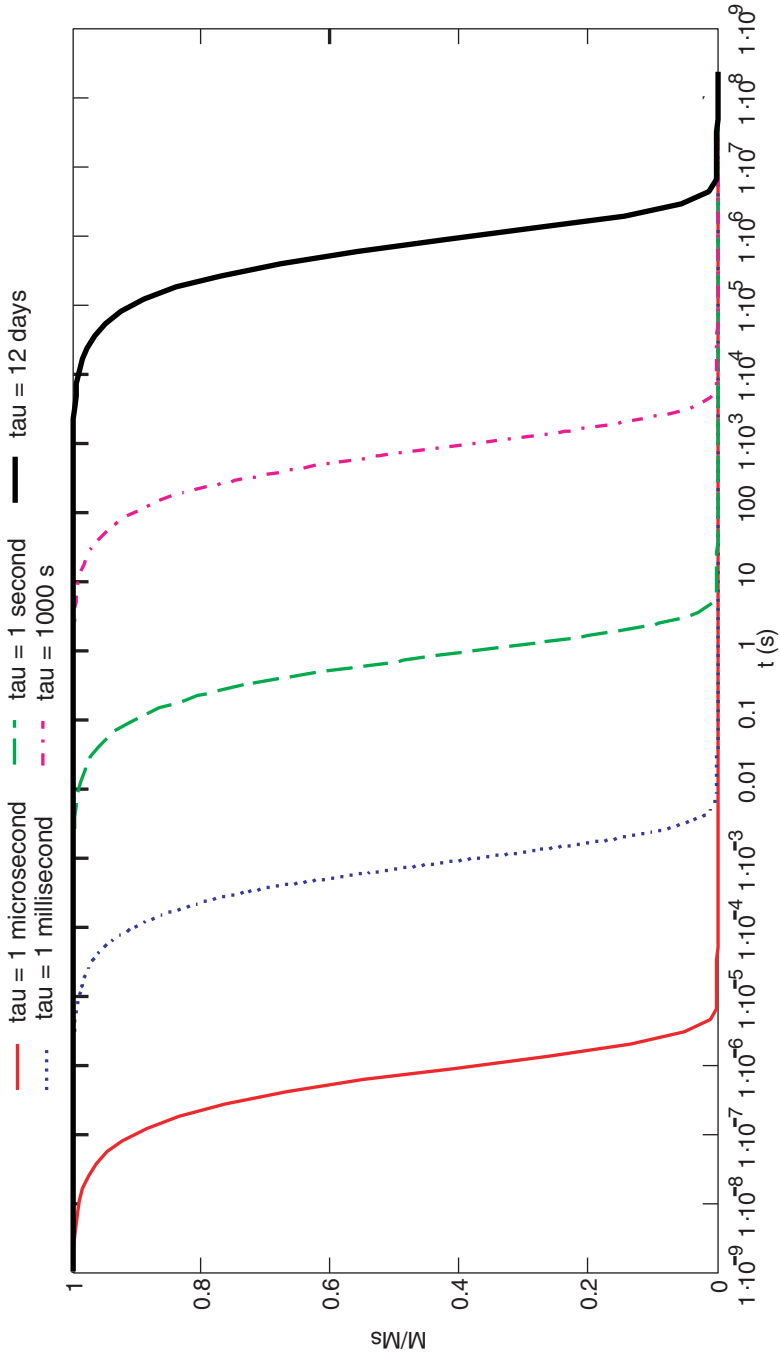

FIGURE 2 Plot of Eq.5 for various value of decay time tau.

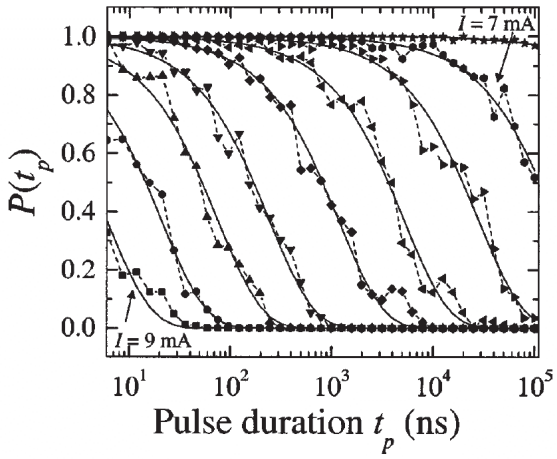

FIGURE 3 Probability for reversal of a lithographically patterned magnetic element, as a function of pulse duration. Different curves are for varying pulse amplitudes. Data are well fitted to exponentials, with a decreasing time constant for increasing pulse amplitude. Source: Rizzo et al., 2002. Reprinted with permission.

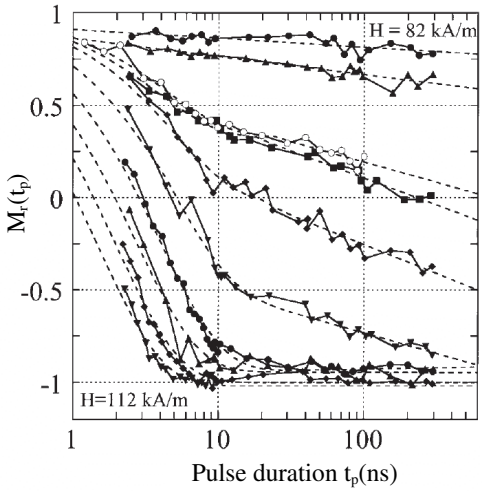

FIGURE 4 Magnetization reversal process for thin-film recording media. The data indictate that the reversal process is logarithmic for field pulses longer than 10 ns, and exponential for shorter duration pulses. Source: Rizzo et al., 1999. Reprinted with permission.

minimize the exchange coupling between individual grains of the film. The exchange coupling between the spins within a ferromagnet is what leads to ferromagnetic order below the Curie temperature. Nevertheless, frustration of the exchange interaction between crystallites in a continuous magnetic film does not preclude ferromagnetic order as long as the anisotropy energy for the grains is sufficient to prevent thermal erasure, as per Eq.6. By reducing the intergranular exchange coupling, the characteristic length scale for magnetic variations *between* bits is minimized at the expense of magnetic uniformity *within* the bit. As a first approximation, media noise scales inversely as the square root of the number of grains in a bit. Thus, to keep media noise in check, grain size must also scale with the reduced bit dimension as areal density is increased. As a rule of thumb, each bit must contain ~ 100 grains for a tolerable level of media noise.

By Eq.6, to prevent thermal erasure, any decrease in grain volume must be accompanied by an increase in the anisotropy energy K_u. An optimal energy barrier size that maintains data longevity for the 10- to 20-year lifetime of the disk drive is approximately $K_u V/k_B T \sim 50$. The bit volume at an areal density of 100 Gbit/in^2 is 5×10^4 nm^3 (assuming 10 nm film thickness), with an average grain volume of only 500 nm^3. To be thermally stable, the anisotropy energy density must be greater than 70 μeV/nm^3. This is equivalent to an anisotropy "field" H_k of $\mu_0 H_k = 2K_u/M_S \approx 2.7$ T. The fabrication of thin magnetic films with such large crystalline anisotropies is not in itself a limitation. The Co-based alloys used for data storage exploit the large uniaxial anisotropy associated with hexagonal close-packed (hcp) structure of Co. (Single crystal hcp Co has a uniaxial anisotropy of 0.6 T. Alloying Co with various noble and refractory metals can enhance the anisotropy by a factor of 2 to 4 [Lu et al., 2003].)

A much more serious issue is the writing process. The coercive field required to switch the magnetization orientation in the media scales roughly as half the anisotropy field. Thus, media for 100 Gbit/in^2 recording densities mandates write fields of 1.4 T. To generate such large fields, the recording head itself must have saturation magnetization levels of approximately twice the coercivity, or $\mu_0/M_S \approx 2.7$ T. (The write head is essentially a ferrous-core electromagnet that uses the soft magnetic properties of high-moment alloys to convert weak currents of a few hundred mA into large magnetic fields.) At this point, we are now limited by the excess spin that can exist in a ferromagnet. FeCo alloys have the highest saturation magnetization levels of 2.4 T, with 2.5 Bohr magnetons per atom. Thus, exceeding densities of 100 Gbit/in^2 may be possible, but fundamental limitations associated with the generation of large recording fields will severely impede further progression of simple scaling laws.

MAGNETIC FLUCTUATION NOISE

Recently, it has been found that thermal fluctuations in the read head can also impede further improvements in areal density (Smith and Arnett, 2001;

Stutzke et al., 2003). These "soft" fluctuations also increase in magnitude as the magnetic read sensor is scaled down in size due to the fixed thermal energy per excitation mode. In this case, however, the thermal fluctuations are not detected as irreversible jumps in the ground state configuration of the sensor, but simply as voltage fluctuations in the sensor output.

The read head in a modern disk drive uses the giant magnetoresistance (GMR) effect for magnetic multilayers. In the late 1980s it was discovered that the resistance in an artificial superlattice of magnetic and nonmagnetic metallic films was a strong function of relative magnetization orientation in the superlattice structure. Resistance changes on the order of several tens of percent were realized for such structures. Implementation of the new technology was very quick; today, all commercial disk drives use GMR sensors for the read-back process. For such sensors, two magnetic layers are separated by a 2 to 5-nm thick nonmagnetic spacer of Cu. One of the magnetic layers is kept fixed through exchange biasing techniques; the other sensor is free to rotate under the influence of stray magnetic fields from the recording media. The voltage produced when a current is driven through the multilayer device is given by

$$V = I_0\left(R_0 + \frac{\Delta R}{2}(1-\cos\theta)\right) \tag{11}$$

where I_0 is the bias current, R_0 is the nonmagnetic contribution to the device resistance, ΔR is the GMR resistance magnitude for antiparallel orientation of the two layers, and θ is the relative orientation of the two layers. The GMR sensor is biased so the magnetization of the free layer is oriented perpendicular to the magnetization of the fixed layer in order to linearize the response of the device to fields from the written bit pattern:

$$\delta V \approx I_0\left(\frac{\Delta R}{2}\delta\theta\right) \tag{12}$$

where $\theta = \frac{\pi}{2} + \delta\theta$. The rotation of the magnetization by external fields is in proportion to the strength of the external field H_0 and the effective anisotropy of the free magnetic layer:

$$\delta\theta \approx \frac{H_0}{H_k} \tag{13}$$

However, the free layer can also respond to fluctuations associated with the thermal bath. To understand how strong these fluctuations can be, we must understand the nature of thermal fluctuations in ferromagnets far from the Curie temperature. In this case, fluctuations are quantized much the same as for the problem of black body radiation. Thus, the distribution of energy among the

different magnetic eigenmodes (also referred to as magnons) is given by a Planck distribution. In the limit of large temperatures relative to the energy of the eigenmodes, the distribution becomes classical and the energy per eigenmodes is simply $k_B T$. The number of magnons $N(\omega)$ for a given mode of frequency ω is simply

$$N(\omega) \approx \frac{k_B T}{\hbar \omega} \tag{14}$$

For a ferromagnetic sensor with an anisotropy of 1.6 kA/m (20 Oe) and a magnetization of 1 *T*, the eigenfrequency for the lowest order mode is simply that of Eq.10, with a characteristic frequency of 1.3 GHz. Thus, the number of excitations occupying the lowest order mode at an operational temperature of T = 400 K is $N(\omega) \sim 7000$. The number of unpaired spins in a free layer is proportional to the volume of the magnetic film (typically 0.18 μm × 0.06 μm = 5 nm = 5.4×10^4 nm^3) and the magnetization density of the material. The relative fraction $n(\omega)$ of spin excitations is therefore

$$\begin{aligned} n(\omega) &= \frac{N(\omega)}{\left(\frac{M_S V}{g\mu_B}\right)} \\ &= \frac{\gamma k_B T}{\omega M_S V} \end{aligned} \tag{15}$$

From linear magnon theory (Sparks, 1965), the rms amplitude $\delta\theta_{rms}$ for these thermally driven magnetic excitations is related to the magnon density by

$$n_0 = \frac{1 - \cos\left(\frac{\delta\theta_{rms}}{\varepsilon}\right)}{2} \tag{16}$$

where ε is the ellipticity for gyromagnetic precession in a thin-film geometry,

$$\varepsilon = \sqrt{\frac{M_S}{H_k}} \tag{17}$$

Solving for $\delta\theta_{rms}$, in the limit of small angle fluctuations, we end up with

$$\delta\theta_{rms} = 2\varepsilon\sqrt{\frac{\gamma k_B T}{\omega M_S V}} \tag{18}$$

Assuming a magnetic volume of 5.4×10^4 nm^3, the fluctuations have an amplitude of $\delta\theta_{rms} \sim 70$ degrees: A significant fraction of the sensor dynamic

range! Of course, these fluctuations are largest at the resonance frequency ω. Away from the resonance, we use the linearized solution to Eq.7 to determine the amplitude of this magnetic noise. The solution to Eq.7 in the limit of $M_S \gg H_k$ and small angle motion is that of a harmonic oscillator, with a susceptibility spectrum given by

$$\chi(\omega) = \chi_0 \left(\frac{\omega_0^2}{\omega_0^2 - \omega^2 - 2i\left(\frac{\omega}{\tau_0}\right)} \right) \tag{19}$$

The noise power spectrum is proportional to $|\chi(\omega)|^2$. Experimental measurements have verified that the magnetic fluctuations take the spectral shape of the ferromagnetic resonance, as shown in Figure 5.

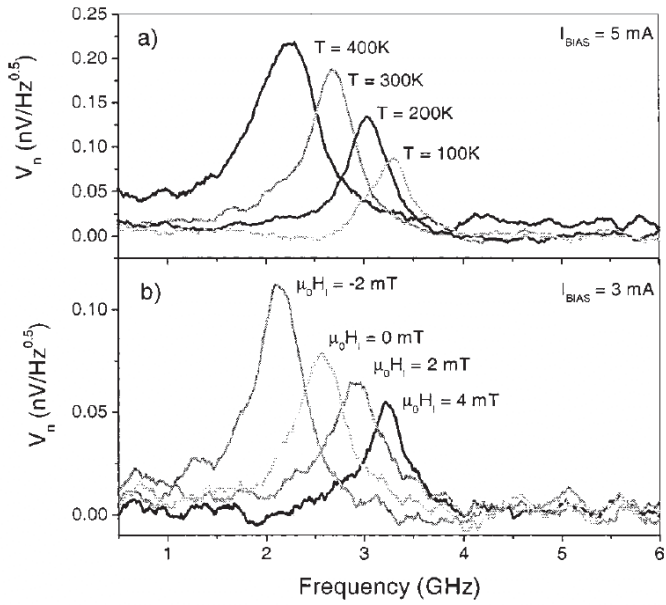

FIGURE 5 Magnetic fluctuation noise in a GMR sensor. The resonant peak is at the ferromagnetic resonance frequency for the given temperature (upper figure) and given applied longitudinal field (lower figure). Increasing the longitudinal bias field increases the resonance frequency, in accordance with Eq.10. Source: Stutzke et al., 2003. Reprinted with permission.

The integral of the power spectrum is equal to the total noise power:

$$C\int_0^\infty |\chi(\omega)|^2 \, d\omega = (\delta\theta_{rms})^2 \tag{20}$$

Solving for the normalization constant, we obtain

$$C = \frac{1}{\frac{\pi}{4}(\tau_0\omega_0^2)}\left(\frac{\delta\theta_{rms}}{\chi_0}\right)^2 \tag{21}$$

The noise amplitude at low frequencies $\delta\theta_{dc}$ is approximately white in the limit $\tau\omega_0 \gg 1$, with an amplitude of

$$\begin{aligned}\delta\theta_{dc} &= \\ &= \sqrt{C\chi(0)} \\ &= 2\sqrt{\frac{1}{(\pi\tau_0\omega_0^2)}\left(\frac{\delta\theta_{rms}}{\chi_0}\right)}\chi_0 \\ &= 2\sqrt{\frac{1}{(\pi\tau_0\omega_0^2)}}\delta\theta_{rms}\end{aligned} \tag{22}$$

in units of radians per $\sqrt{Hz}$. Using Eq.10 in the limit of $M_S \gg H_k$, we end up with

$$\delta\theta_{dc} = \frac{4}{\sqrt{2\pi}}\sqrt{\frac{\alpha k_B T}{\gamma\mu_0^2 M_S^{1/2} H_k^{5/2} V}} \tag{23}$$

The rms fluctuations in the GMR signal are proportional to the integrated noise spectrum $\Delta\theta$, which we approximate as the product of Eq.23 and the square root of the electronic bandwidth Δf:

$$\begin{aligned}\Delta\theta &= \frac{4}{\sqrt{2\pi}}\sqrt{\frac{\alpha k_B T \Delta f}{\gamma\mu_0^2 M_S^{1/2} H_k^{5/2} V}} \\ &= \frac{2}{\sqrt{\pi}}\sqrt{\alpha\varepsilon\left(\frac{\Delta f}{\omega_k}\right)\left(\frac{k_B T}{K_u V}\right)}\end{aligned} \tag{24}$$

where $\omega_k = \gamma\mu_0 H_k$ is the effective gyromagnetic frequency of the anisotropy. Notice that the noise power is proportional to the ratio of the thermal energy to

the anisotropy energy. As was the case for the thermal erasure of recorded bits in thin-film media, the ratio of the anisotropy to thermal energy determines the stability of the magnetization in the presence of thermal fluctuations.

Using typical parameter values ($\alpha = 0.01$, $H_k = 1.6$kA/m (20 Oe), $M_S = 800$ kA/m (800 emu/cm^3), $T = 400$ K, $\Delta f = 400$ MHz, and $V = 5.4 \times 10^4$ nm^3), the integrated noise amplitude is $\Delta\theta \sim 12$ degrees, reducing the maximum dynamic range of the GMR sensor to 90/12 = 7.5. This is the bare minimum necessary for proper operation of the disk drive. Any further reductions in the sensor volume will result in higher noise levels because the total thermal noise power of the magnetic fluctuations is independent of sensor size, whereas the magnetostatic energy associated with the rotation of the magnetization is directly proportional to device volume. The noise level can be reduced by increasing the anisotropy H_k, but only at the expense of the read head sensitivity, thus mandating improvements in the GMR resistance ΔR. Some progress has been made along these lines, but present values are already $\Delta R/R \sim 0.2$. Even at $\Delta R/R$ of 100 percent, H_k can only be increased by a factor of 5. To maintain the same level of magnetic fluctuation noise, the sensor volume can be reduced by the same factor, with an effective increase in areal density by a factor of 3. However, once the technological hurdles have been surpassed to obtain large GMR signals, this improvement in density no longer scales. The maximum possible recording densities will have been achieved.

An alternative strategy is to reduce the damping parameter α for the GMR sensor material. By reducing the relaxation rate, thermal fluctuations are effectively isolated at the FMR frequency, far above the operational frequency of the drive. However, the nature of damping in ferromagnetic metals remains phenomenological; there has been little success in developing a theory for the damping process that can be used to engineer low damped materials. Instead, measurements of a wide variety of alloys suggest that damping parameter values of less than 0.01 are not in the offing as a result of a simple search through material parameter space. Instead, we need to understand the fundamental mechanisms that give rise to damping. Numerous mechanisms have been suggested (Abraham and Kittel, 1952; Korenman and Prange, 1972; Suhl, 1998), but to date none has demonstrated the predictive power necessary to qualify as a true physical theory for damping. This remains a serious deficiency in the theory of ferromagnetic materials that deserves more attention in the physics community.

FUTURE PROSPECTS

It appears that the use of scaling in disk-drive technology is no longer a viable means of improving drive capacity. The problem is the fundamental nature of thermal fluctuations in magnetic systems at small volumes, which limits both the bit capacity of the recording media and the read sensors used to read

back the bits from the media. Alternative technical strategies have been proposed (perpendicular magnetic recording, patterned media, heat assisted magnetic recording), but all of these approaches offer only marginal improvements over current areal densities based on incremental advances in recording efficiency. For all practical purposes, the most cost-effective means of increasing value for the data storage consumer is the same as for other commodity items—manufacturing finesse, improved reliability, and marketing.

Nevertheless, it should be appreciated that magnetic recording technology has reached nanoscale dimensions in spite of the mechanical components that have always been considered a hindrance; bit dimensions are almost at 35 nm (length) × 180 nm (width) × 10 nm (thickness). By any measure, such numbers are nothing short of astonishing. However, moving to even smaller bit dimensions will require more than improvements in engineering. It will require a greater understanding of the fundamental physics underlying the phenomena of hysteresis and the thermodynamics of the ferromagnetic phase. Such understanding might enable the discovery of new physics that would allow for new modalities of the recording process.

It is worth the time to consider this last point in more detail. Hysteresis in a spin system, at its most elementary level, is an astonishing effect. It should be kept in mind that the switching of magnetic bits involves transitions between nearly degenerate energy levels. Recording a bit does not involve the storage of energy in the archival medium, but rather the expenditure of work to cause a thermodynamically irreversible alternation in the magnetization state of the medium. Thus, the magnetic bit is essentially an indelible record that some fraction of the electromagnetic work provided by the recording head has been converted to heat via the spin system of the ferromagnet, resulting in a net increase in entropy, $\Delta S = K_u V/T$; the more entropy generated in the writing of a bit, the more stable the recorded mark.

Such a process stands in marked contrast to our basic understanding of spin dynamics at the quantum level. One need only consider that the moment of a ferromagnetic volume is composed of electron spins, each of which would have a very short memory time if isolated from all the other spins in the ferromagnetic system, even at low temperatures. Electron spin resonance measurements typically yield decoherence times ranging from nanoseconds to microseconds in most conductors. The ferromagnetic phase, enabled by the exchange interaction between the uncompensated electron spins, is stabilized against such relaxation processes, but at a cost. Switching the magnetization in a stable fashion requires the irreversible conversion of work into heat.

Thus, the effort to increase recording densities is essentially a struggle to find increasingly efficient means of converting the same amount of work into heat per bit, even though the size of the bits is shrinking. Current recording technology relies upon intrinsic relaxation processes induced by large applied magnetic fields. Borrowing terminology from nuclear magnetic resonance, large

Zeeman splittings induced by the recording heads drives spin reversal via longitudinal relaxation processes. The rephrasing of the recording process in the context of spin resonance techniques suggests one possible strategy for increasing recording densities. It is well known that resonant techniques (e.g., adiabatic fast passage) are far more efficient means of manipulating spin states in paramagnetic systems. Perhaps similar methods could be developed for the improved focus of electromagnetic energy to yet smaller spin volumes in a ferromagnetic medium.

REFERENCES

Abraham, E. and C. Kittel. 1952. Spin lattice relaxation in ferromagnets. Physical Review 88(5): 1200.

Chandler, D. 1987. Introduction to Modern Statistical Mechanics. New York: Oxford University Press.

Helstrom, C.W. 1984. Probability and Stochastic Processes for Engineers. New York: Macmillan.

Korenman, V. and R.E. Prange. 1972. Anomalous damping of spin waves in magnetic metals. Physical Review B 6(7): 2769–2777.

Landau, L., and E. Lifshitz. 1935. On the theory of the dispersion of magnetic permeability in ferromagnetic bodies. Physik. Z. Sowjetunion 8: 153–169.

Lu, P.-L., and S.H. Charap. 1994. Magnetic viscosity in high-density recording. Journal of Applied Physics 75(10): 5768–5770.

Lu, B., D. Weller, A. Sunder, G. Ju, X. Wu, R. Brockie, T. Nolan, C. Brucker, and R. Ranjan. 2003. High anisotropy CoCrPt(B) media for perpendicular magnetic recording. Journal of Applied Physics 93(10): 6751–6753.

Morrish, A.H. 2001. The Physical Principles of Magnetism. New York: IEEE Press.

Néel, L. 1949. Theorie du trainage magnetique des ferromagnetiques en grains fins avec applications aux terres cuites. Annual Geophysics 5: 99–136.

Pathria, R.K. 1996. Statistical Mechanics, 2nd Ed. New York: Butterworth-Heinemann.

Rizzo, N.D., M. DeHerrera, J. Janesky, B. Engel, J. Slaughter, and S. Tehrani. 2002. Thermally activated magnetization reversal in submicron magnetic tunnel junctions for magnetoresistive random access memory. Applied Physics Letters 80(13): 2335–2337.

Rizzo, N.D., T.J. Silva, and A.B. Kos. 1999. Relaxation times for magnetization reversal in a high coercivity magnetic thin film. Physical Review Letters 83(23): 4876–4879.

Silva, T.J., C.S. Lee, T.M. Crawford, and C.T. Rogers. 1999. Inductive measurement of ultrafast magnetization dynamics in thin-film permalloy. Journal of Applied Physics 85(11): 7849–7862.

Smith, N., and P. Arnett. 2001. White-noise magnetization fluctuations in magnetoresistive heads. Applied Physics Letters 78(10): 1448–1450.

Sparks, M. 1965. Ferromagnetic Relaxation Theory. New York: McGraw-Hill.

Stoner, E.C., and E.P. Wohlfarth. 1948. A mechanism of magnetic hysteresis in heterogeneous alloys. Philosophical Transactions of the Royal Society of London, Series A 240(826): 599–642.

Street, R.C., and J.C. Wooley. 1949. A study of magnetic viscosity. Proceedings of the Physical Society, Section A 62(9): 562–572.

Stutzke, N., S.L. Burkett, and S.E. Russek. 2003. Temperature and field dependence of high-frequency magnetic noise in spin valve devices. Applied Physics Letters 82(1): 91–93.

Suhl, H. 1998. Theory of the magnetic damping constant. IEEE Transactions on Magnetics 34(4): 1834–1838.

Thermodynamics of Nanosystems

CHRISTOPHER JARZYNSKI
Los Alamos National Laboratory
Los Alamos, New Mexico

The field of thermodynamics was born of an engineering problem. *Reflections on the Motive Power of Fire*, an analysis of the efficiency of steam engines published by the French engineer Sadi Carnot in 1824 led ultimately to the elucidation of the second law of thermodynamics, the unifying principle that underlies the performance of all modern engines, from diesel to turbine. Today engineers are excited about nanosystems, including machines that operate at molecular length scales. If history is any guide to the present, a grasp of thermodynamics at the nanoscale is essential for the development of this field. With this in mind, the focus of this brief talk is on the following question: *If we were to construct an engine the size of a large molecule, what fundamental principles would govern its operation?* To put it another way, what does thermodynamics look like at the nanometer length scale?

It is perhaps not immediately obvious that thermodynamics should "look different" at the microscopic scale of large molecules than at the macroscopic scale of car engines. Of course, at the microscopic level, quantum effects might play a significant role, but in this talk we will assume the effects are negligible. For present purposes, the difference between the macro- and the microworld boils down to this. On the scale of centimeters, it is safe to imagine that matter is composed of continuous substances (e.g., fluid or solid, rigid or elastic, etc.) with specific properties (e.g., conductivity, specific heat, etc.). On the scale of nanometers, by contrast, the essential granularity of matter (i.e., that it is made up of individual molecules and atoms) becomes impossible to deny.

Once we consider systems small enough that we can distinguish individual molecules or atoms, *thermal fluctuations* become important. If we take a large system, such as a rubber band, and stretch it, its response can accurately be

predicted from knowledge of the properties of the elastic material of which it is made. But, imagine that we take a single RNA molecule, immerse it in water, and stretch it using laser tweezers. As a result of continual bombardment by the surrounding water molecules, the RNA molecule jiggles about in a way that is essentially random. Moreover, *because the RNA is itself a very small system*, these thermal motions are not negligible; the noise-to-signal ratio, so to speak, is substantial. Thus, the response of the RNA strand has a considerable element of randomness in it, unlike the predictable response of an ordinary rubber band. This naturally leads us to suspect that the laws of thermodynamics familiar in the context of large systems might have to be restated for nanosystems in a way that accounts for the possibility of sizeable thermal fluctuations.

Admittedly, machines the size of single molecules have not yet been constructed, but the issues I have raised are not merely speculative. The RNA-pulling scenario outlined above was carried out by an experimental group at University of California, Berkeley that measured the resulting fluctuations stretching the RNA strand (Liphardt et al., 2002). Independently, researchers at the Australian National University in Canberra and Griffith University in Brisbane have used laser tweezers to drag microscopic beads through water, with the explicit aim of observing microscopic "violations" of the second law of thermodynamics caused by thermal fluctuations (Wang et al., 2002). In a sense, the research is happening in reverse order to the research that gave rise to nineteenth-century thermodynamics. We do not yet have a nanomachine (a twenty-first century analogue of a steam engine), but we have begun to play with its potential components, with the aim of teasing out the fundamental principles that will govern the device when (or if) we ultimately construct one.

The study of the thermal behavior of tiny systems is by no means a new field. A century ago, Einstein's and Smoluchowski's quantitative explanations of *Brownian motion*—the spastic movement of pollen particles first observed by the British botanist Robert Brown in 1827—not only helped clinch the atomic hypothesis, but also led to the *fluctuation-dissipation theorem*, a remarkably simple relationship between friction and thermal noise. In the context of the RNA experiments mentioned above, the fluctuation-dissipation theorem predicts that:

$$\langle W_{\text{diss}} \rangle = \frac{\sigma_W^2}{2k_B T} \tag{1}$$

where $\langle W_{\text{diss}} \rangle$ is the average amount of work that is dissipated when we stretch the RNA; $\sigma_W^2 = \langle W^2 \rangle - \langle W \rangle^2$ is the variance in work values; T is the temperature of the surrounding water; and $k_B = 1.38 \times 10^{-23}\ J\ K^{-1}$ is Boltzmann's constant. Here the words *average* and *variance* refer to many repetitions of the pulling experiment; because of thermal noise, the exact amount of work performed in stretching the RNA molecule differs from one repetition to the next, as illustrated in Figure 1. By *dissipated work*, we mean the amount by which the total work performed (during a single realization) exceeds the reversible work,

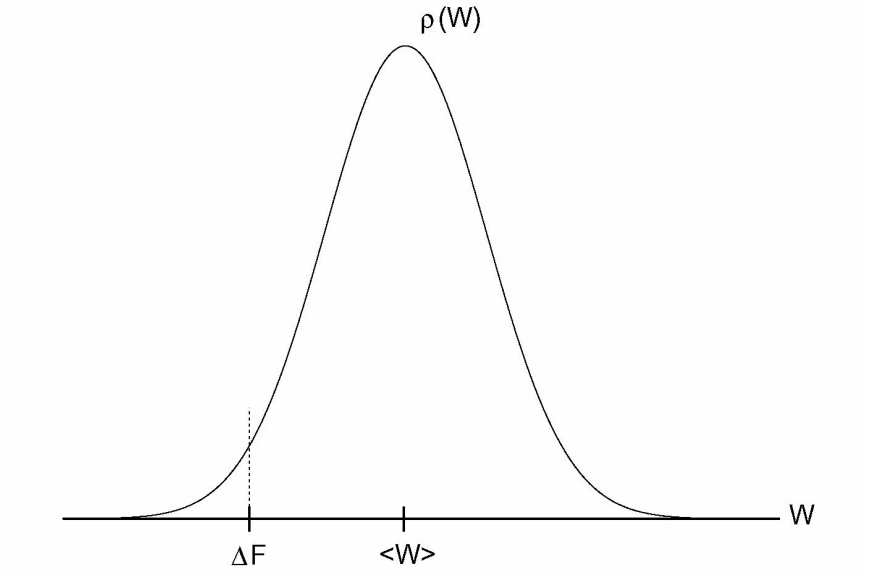

FIGURE 1 Distribution of work values for many repetitions (realizations) of a thermodynamic process involving a nanoscale system. The process might involve the stretching of a single RNA molecule or, perhaps, the compression of a tiny quantity of gas by a microscopic piston. The tail to the left of ΔF represents apparent violations of the second law of thermodynamics. The distribution ρ(*W*) satisfies the nonequilibrium work theorem (Eq.6), which reduces to the fluctuation-dissipation theorem (Eq.1) in the special case of near-equilibrium processes.

$$W_{\text{diss}} \equiv W - W_{\text{rev}} \tag{2}$$

(As discussed below, this is equivalent to $W - \Delta F$, where ΔF is the free energy difference between the initial and final states of the system.)

The fluctuation-dissipation theorem, along with most subsequent work in this field, pertains to systems near thermal equilibrium. (Eq.1 will not be satisfied if we stretch the RNA too rapidly!) In the past decade or so, however, a number of predictions have emerged that claim validity even for systems *far* from equilibrium. These have been derived by various researchers using a variety of theoretical approaches and more recently have been verified experimentally. It is impossible to do justice to a decade of work in such a short space, but I will attempt to convey the flavor of this progress by illustrating one result on a toy model.

Consider that familiar workhorse of thermodynamic pedagogy, a container filled with a macroscopic quantity of gas closed off at one end by a piston. Suppose we place the container in thermal contact with a heat bath at a temperature *T*, and we hold the piston fixed in place, allowing the gas to relax to a state

of thermal equilibrium. Then we vary the position of the piston so as to compress the gas, from an initial volume V_A to a final volume V_B. In doing so we perform external *work*, W, on the gas. If we carry out this process *reversibly* (slowly and gently), then the gas stays in equilibrium at all times, and the work performed on it is equal to the net change in its free energy:[1]

$$W_{\text{rev}} = F_B - F_A \equiv \Delta\text{F} \tag{3}$$

If instead we carry out the process *irreversibly* (perhaps even rapidly and violently), the work, W, will satisfy the inequality

$$W > \Delta F \tag{4}$$

with ΔF defined exactly as above. Eq.4 is simply a statement of the second law of thermodynamics, when we have a single heat bath.

To understand what Eq.3 and 4 look like for microscopic systems, imagine that we scale our system down drastically, to the point that our container is measured in tens of nanometers and contains only a handful of gas molecules. Again we place it in contact with a heat bath, and we consider a process whereby we move a tiny piston so as to compress the volume of the container from an initial volume V_A to a final volume V_B. (I do not mean to suggest that a realistic nanomachine will be composed of nanopistons; this familiar example is used only for illustration). Suppose we repeat this process very many times, always manipulating the volume in precisely the same way. During each repetition, or *realization*, we perform work as the gas molecules bounce off the moving piston. However, because the precise motion of the molecules differs each time we carry out the process, so does the precise value of W. In effect, the *thermal* fluctuations of the gas molecules give rise to *statistical* fluctuations in the value of W from one realization to the next. After very many repetitions, we have a distribution of work values, $\rho(W)$, where $\rho(W)dW$ is the probability that the work during a single realization will fall within a narrow window dW around the value W. What is the relationship between this distribution and the free-energy difference $\Delta F = F_B - F_A$ between the equilibrium states corresponding to the initial and final volumes? Based on our knowledge of macroscopic thermodynamics (Eq.3 and 4), we might guess that the *average* work value is no less than ΔF, i.e.,

$$\langle W \rangle \equiv \int dW\, \rho(W) W \geq \Delta F \tag{5}$$

[1]Recall that the free energy associated with a state of thermal equilibrium, at temperature T, is given by $F = E - ST$, where E and S are, respectively, the internal energy and entropy of that state. In Eq.3, F_A and F_B are associated with the equilibrium states corresponding to container volumes V_A and V_B.

with the equality holding only if the process is carried out reversibly. However, this inequality does not rule out the possibility that $\rho(W)$ has a small "tail" extending into the region $W < \Delta F$; see Figure 1.

If we were to observe a value of W less than ΔF for a process carried out with a *macroscopic* piston, this would be an extraordinary event—a true violation of the second law of thermodynamics. But in a system with relatively few degrees of freedom, this would not be such a shocking occurrence. We will refer to such an event ($W < \Delta F$ in a microscopic system) as an *apparent* violation of the second law.

Eq.5 is correct, but we can do better. Instead of considering the average of W over many repetitions of our thermodynamic process, let us consider the average of $\exp(-W/k_BT)$ over these realizations. It turns out that this average obeys a simple and somewhat surprising equality: $\int dW\ \rho(W)\exp(-W/k_BT) = \exp(-\Delta F/k_BT)$, or more succinctly

$$\left\langle e^{-W/k_BT} \right\rangle = e^{-\Delta F/k_BT} \tag{6}$$

This is the *nonequilibrium work theorem*, which remains valid no matter how gently or violently we force the piston in changing the volume from V_A to V_B. This result has been derived using various theoretical approaches (Crooks, 1998, 1999; Hummer and Szabo, 2001; Jarzynski, 1997a,b) and confirmed by the Berkeley RNA-pulling experiment discussed above (Liphardt et al., 2002).

Eq.5 and 6 both make predictions about the probability distribution of work values, $\rho(W)$, but the latter is a considerably stronger statement than the former. For instance, Eq.6 immediately implies that $\rho(W)$ cannot be located entirely in the region $W > \Delta F$. (If that were the case, then e^{-W/k_BT} would be less than $e^{-\Delta F/k_BT}$ for every realization of the process, and there would be no way for Eq.6 to be satisfied.) Thus, if the typical value of W is greater than ΔF, as will be the case for an irreversible process, then *there must also be some realizations for which* $W < \Delta F$.

Eq.6 not only tells us that $\rho(W)$ must have a tail extending into the region $W < \Delta F$, but it also establishes a tight bound on such apparent violations of the second law, namely,

$$\int_{-\infty}^{\Delta F - nk_BT} dW\ \rho(W) \le e^{-n} \tag{7}$$

for any $n > 0$. This inequality can be stated in words: the probability that we will observe a work value that falls below ΔF by at least n units of k_BT is bounded from above by e^{-n}. This means that substantial violations of the second law ($n \gg 1$) are extremely rare, as expected. (The derivation of Eq.7 from Eq.6 is left as an exercise for the reader.)

If perchance we change the volume of the container relatively slowly so that the gas remains close to thermal equilibrium at all times, then $\rho(W)$ ought

to be a normal (Gaussian) distribution. (I will not try to justify this expectation here beyond saying that it follows from the Central Limit Theorem.) It takes a few lines of algebra to show that if $\rho(W)$ is a Gaussian distribution that satisfies Eq.6, then its mean and variance must be related by Eq.1. Thus, although the nonequilibrium work theorem (Eq.6) is valid regardless of whether we move the piston slowly or quickly, in the former (near-equilibrium) case it reduces to the well-known fluctuation-dissipation theorem (Eq.1).

Eq.6 is closely related to another result, which can be stated as follows. Suppose we perform many repetitions of the process described above, in which we change the volume of our tiny container of gas *from* V_A *to* V_B, according to some schedule for moving the piston. Then suppose we perform many repetitions of the reverse process, in which we move the piston according to the reverse schedule, thus changing the container volume *from* V_B *to* V_A. For both sets of experiments, we observe the work performed during each realization of the process, and in the end we construct the associated distributions of work values, $\rho_{A\to B}(W)$ and $\rho_{B\to A}(W)$. Then these two distributions satisfy

$$\frac{\rho_{A\to B}(W)}{\rho_{B\to A}(-W)} = e^{(W-\Delta F)/k_B T} \tag{8}$$

where $\Delta F = F_B - F_A$ as before (Crooks, 1998, 1999). This is remarkable because the response of the gas during one process, $V_A \to V_B$, is physically quite different from the response during the reverse process, $V_B \to V_A$. For example, when we push the piston into the gas, a modest shock wave (a region of high density) forms ahead of the piston. For the reverse case, a region of *low* density trails the piston as it is pulled away from the gas. Nevertheless, the two distributions are related by a simple equality, Eq.8 above.

Eq.6 and 8 are two examples of recent results pertaining to microscopic systems away from thermal equilibrium. They make very specific predictions about the fluctuations of an observed quantity, in this case the *work* we perform on our system. The *fluctuation theorem* is a collective term for another set of such results (Evans and Searles, 2002), originally derived in the context of systems in nonequilibrium steady states and recently confirmed by the Australian bead-dragging experiment mentioned earlier (Wang et al., 2002). All of these predictions have potentially important design implications for nanomachines, for which thermal fluctuations are bound to play an important role. For instance, if we design a microscopic motor fueled by the transfer of molecules from high to low chemical potential, then it is of interest to estimate how often random fluctuations might cause this motor to run backward, by moving a molecule or two from low to high chemical potential, in flagrant (if brief) violation of the second law. Eq.7 addresses this sort of issue. I don't want to overstate the case here, but it seems reasonable to assume that the better we understand the laws of thermodynamics at the nanoscale, the more control we will have over such systems.

CONCLUSIONS

When considering the thermodynamics of systems at the scale of nanometers, *fluctuations* are important. If we carefully repeat an experiment many times, we will obtain noticeably different results from one repetition to the next. The take-home message of this brief talk is that although these fluctuations originate in thermal randomness, they obey some surprisingly stringent and general laws (such as Eq.6 and 8), which are valid even far from equilibrium. Moreover, these laws are fundamentally microscopic; they do not follow by educated guess from our knowledge of macroscopic thermodynamics. However, the story is not yet complete. So far, we have a jumble of related predictions, as well as experimental validation, but nothing like the beautifully coherent structure of macroscopic thermodynamics. Perhaps as nanotechnology develops and the need for a better understanding of such phenomena becomes more pressing, a more complete and satisfying picture will emerge.

REFERENCES

Crooks, G.E. 1999. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. Physical Review E 60(3): 2721–2726.

Crooks, G.E. 1998. Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems. Journal of Statistical Physics 90(5/6): 1481–1487.

Evans, D.J., and D.J. Searles. 2002. The fluctuation theorem. Advances in Physics 51(7): 1529–1585, and many references therein.

Hummer, G., and A. Szabo. 2001. From the cover: free energy reconstruction from nonequilibrium single-molecule pulling experiments. Proceedings of the National Academy of Sciences 98 (7): 3658–3661.

Jarzynski, C. 1997a. Nonequilibrium equality for free energy differences. Physical Review Letters 78(14): 5018–5035.

Jarzynski, C. 1997b. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. Physical Review E 56(5): 5018–5035.

Liphardt, J., S. Dumont, S.B. Smith, I. Tinoco, and C. Bustamante. 2002. Equilibrium information from nonequilibrium measurements in an experimental test of Jarzynski's Equality. Science 296(5574): 1832–1835.

Wang, G.M., E.M. Sevick, E. Mittag, D.J. Searles, and D.J. Evans. 2002. Experimental demonstration of violations of the Second Law of Thermodynamics for small systems and short time scales. Physical Review Letters 89(5): 050601-1–050601-4.

Counterterrorism Technologies and Infrastructure Protection

Using Biotechnology to Detect and Counteract Chemical Weapons

ALAN J. RUSSELL, JOEL L. KAAR, AND JASON A. BERBERICH
Departments of Chemical Engineering and Bioengineering
University of Pittsburgh
Agentase, LLC
Pittsburgh, Pennsylvania

The tragedy of September 11, 2001, and the ensuing anthrax attacks heightened public and governmental awareness of the need for reliable, cost-efficient, and deployable diagnostic and treatment systems for chemical weapons. Many of the present methods require cumbersome equipment and complex analytical techniques. In addition, many decontamination solutions are toxic, corrosive, and flammable, and therefore not appropriate for use over large areas or with personnel. Biocatalytic methods of detection and decontamination/demilitarization offer several advantages over existing methods.

Enzymes are environmentally benign, highly efficient biological catalysts on appropriate reactants, with hydrolysis rate enhancements exceeding a million fold over the uncatalyzed reactions. They may also be used in conjunction with other processes to limit the environmental impact of existing systems. Perhaps the most attractive feature of enzymes is that they can function effectively under ambient conditions, thus decreasing energy costs and increasing safety (especially when agents are stored in proximity to explosives). To date, enzymes are the most effective known catalysts for degrading nerve agents (organophosphates [OP] are the class of compounds commonly referred to as nerve agents). Enzyme concentrations in the micromolar range are sufficient to degrade nerve agents on contact, and just 10 milligrams of enzyme can degrade as much nerve agent as 1 kilogram of concentrated bleach (a conventional method of degrading nerve agent).

The study of OP biocatalysis dates back to the mid-1940s, when Abraham Mazur found that mammalian tissue hydrolyzed OP esters (Mazur, 1946). Douglas Munnecke (1977) used cellular extract from a mixed bacterial culture as a catalyst to detoxify organophosphorous pesticide. The rate of enzymatic hy-

drolysis of parathion, an OP ester commonly used in pesticides, was found to be nearly 2,500 times the rate of hydrolysis in 0.1 normal sodium hydroxide. Munnecke (1979) later attempted to immobilize the extract on porous glass and porous silica beads. It was reported that 2 to 5 percent of activity present in a cellular extract could be retained by the beads, and 50 percent of the immobilized activity could be maintained for a full day under ambient conditions.

The enzyme organophosphorus hydrolase (OPH) was later isolated and shown to be an effective catalyst for the degradation of a range of OP esters. Frank Raushel and colleagues have studied extensively the activity and stability of OPH from *Pseudomonas diminuta* (Dumas et al., 1989). At Texas A&M University, James Wild's laboratory has even produced the enzyme in corn (personal communication). Squid diisopropylfluorophosphotase (DFPase), an enzyme capable of degrading nerve agents, such as soman, is now in commercial production. In expanding the library of such enzymes, Joseph DeFrank and colleagues from the U.S. Army have discovered and analyzed another nerve agent-degrading enzyme, organophosphorus acid anhydrolase (OPAA) (DeFrank and Cheng, 1991).

Many species produce enzymes that efficiently degrade nerve agents, although the natural function of these enzymes remains unknown. The class of OP compounds has only been exposed to nature for a few decades, certainly not long enough for natural systems to have evolved enzymes to degrade them. Thus, the ability of enzymes to degrade these compounds is still an unexplained quirk of fate. The only nerve agent-degrading enzyme with a known natural function is OPAA, which is a peptidase enzyme (an enzyme that catalyzes the hydrolysis of peptides into amino acids). Interestingly, this enzyme is far less active with its natural substrate than with soman.

Cells must constantly control their environments and, therefore, must be in a position to rapidly manipulate the concentrations of biocatalysts. Thus, many proteins, such as enzymes, last for only minutes or hours under ambient conditions, and enzymes have generally evolved to be unstable molecules. A number of research groups around the world have been working for many years to stabilize enzymes for numerous applications, including catalysts for the decontamination of nerve agents. These mostly military researchers are also part of a NATO project group (PG31) working to formulate "green" enzyme-based decontamination systems.

Because the preparation and purification of enzymes can be costly, enzymes must be immobilized to increase their reusability and stability. Effective immobilization requires that the enzyme-containing material be prepared so that the enzyme maintains most of its native activity, maintains a high operational stability in its working environment, and maintains a high storage stability. In conventional immobilization methods, either covalent or ionic interactions link the protein to a support material. Enzymes have been immobilized on a wide variety of support materials, including alumina pellets, trityl agarose, and glass/silica

beads. Polymers also make excellent enzyme-support materials because of their structural flexibility and solvent resiliency. Numerous polymers, including nylons, acrylates, and several copolymer blends, have been used as effective supports.

Combining the power of biology with the sophistication of polymers presents considerable opportunities. Synthesis of enzyme-containing polymers involves the covalent immobilization of the enzyme directly into the polymer network via reactive functionalities on the enzyme surface, thus ensuring retention of the enzyme in the polymeric material. Bioplastics prepared in this way exhibit remarkable stability under normally deactivating conditions and, thus, are ideal for the development of the next generation of responsive, smart biomaterials. One of the most pressing needs at the beginning of the twenty-first century is for active, smart materials for chemical defense.

The threat of chemical weapons is now ubiquitous, and these poor man's nuclear weapons can kill on contact. Existing chemical defense polymers are designed to provide an impenetrable barrier between us and our environment. By contrast, layered materials use polymers to bind toxic chemical agents irreversibly to the material. Degrading toxic chemicals requires catalysts that can be incorporated into the polymer in a stable and active form.

In protective clothing, a layer of polyurethane, foam-entrapped, activated charcoal is embedded between several layers of polyester fabric. Polyurethanes are effective substrates for OP adsorption, and reports have documented that polyurethane foam particles can be used as adsorbent materials for pesticide vapors in farming fields. If proteins could be incorporated into polyurethanes, some interesting materials might emerge.

Enzyme-containing polyurethane materials are ideal matrices because of their ease of preparation, the large range of polymer properties that can be prepared, and multipoint, covalent attachment of the enzyme to the polymer. Bioplastics are prepared by reacting a polyurethane prepolymer that contains multiple isocyanate functionalities with an aqueous solution containing the enzyme. The solubility of enzymes in this aqueous phase can be significant, enabling loadings up to 15 percent. Foams, gels, and coatings can be prepared depending on the reactivity of the isocyanate. Any enzyme that is present in the aqueous solution can participate in the polymer synthesis, effectively creating an enzyme-containing-polymer network with multipoint attachment.

The key question is how well such biopolyurethanes function. The answers are striking. Enzymes can be stabilized from days to years, and the enzymes in no way alter the physical properties of the polyurethane. The derived materials are so active that just 1 kilogram of enzyme immobilized in a multipoint covalent fashion is enough to degrade up to 30,000 tons of chemical agent in one year.

Detoxifying a chemical agent is only part of the technology we will need to combat the real and present danger posed by these materials. Once a toxic agent

binds to a surface, it must be degraded and identified. Indeed, every time a chemical or biological sensor gives a false-positive result, tens of thousands of dollars are invested in responses that are not needed. In the Middle East and Africa, false positives could lead to war; and spurious, sensor-triggering events could change the course and nature of a war.

Chemical sensors in particular are generally unreliable, and engineering solutions to the problem will require overcoming some formidable technical challenges. We must first understand how chemical weapons work and use that knowledge to design biologically based sensing devices. The common feature of biosensors to date has been a lack of stability. An effective chemical-weapon sensor must not only be able to operate under ambient conditions, but must also be able to operate in extreme desert and arctic conditions.

A material that is inherently catalytic and can therefore be used to monitor continuously and detoxify detected molecules in real time represents the "holy grail" of this emerging discipline. Indeed, a surface that can both self-decontaminate and sense the progress of decontamination would represent a quantum improvement in protection for war fighters. The rapid detection of chemical agents is critical to inspections performed under the Chemical Weapons Convention, because inspectors are rarely able to control the environment at the inspection site. It is therefore essential that analytical tools vary in sensitivity, specificity, and simplicity.

The selectivity of enzymes that are active on, or inhibited by, chemical agents of interest, makes them attractive components for biosensor technology. Nerve agents exert their biological effect by inhibiting the hydrolytic enzyme, acetylcholinesterase. Thus, the inhibition of this enzyme makes it ideal for use in nerve agent-sensing systems. The most common kit available today for agent detection, the M256A1, uses enzymes and colored substrates to detect nerve agents. Unfortunately, current continuous enzyme-based sensing is limited by the instability of most enzymes and their sensitivity to changes in the environment. The M256A1 functions by reacting phenylacetate with eel acetylcholinesterase to produce a colored material. If an agent or other inhibitor is present, the color does not appear. The test takes 15 minutes to perform and is subject to interference from a number of sources (Table 1). Another drawback is the inability of the M256A1 to distinguish between different nerve agents. All of these techniques rely on the absence of a reaction to signal the presence of an agent. These sensors would be much more informative if they provided positive responses (e.g., undergoing a color change when the surface is either clean or contaminated). Nevertheless, enzyme-inhibition assays are commonly used to detect nerve agents.

In 2001, an enzyme-based sensor was developed and fielded that combines all of the desirable features and is resistant to almost all types of interference. The sensor is based on coupling the hydrolytic activity of acetylcholinesterase (an acid-producing reaction) with the biocatalytic hydrolysis of urea (a base-producing

TABLE 1 Sensitivity of Nerve-Agent Sensors to Various Kinds of Interference

	Sensor				
Interference	M8	M9	M256	M272	CAM
High temperature	Yes	Yes	Yes	Yes	No
Cleaning solvents	Yes	Yes	No	No	No
Petroleum products	Yes	Yes	Yes	Yes	Yes
Antifreeze	No	Yes	No	No	No
Insect repellant	Yes	Yes	Yes	No	Yes
Dilute bleach	Yes	Yes	Yes	Yes	No
Water	No	Yes	No	No	No

Source: Longworth et al., 1999.
Note: "Yes" indicates the sensor is sensitive to the interference. "No" indicates the sensor is not sensitive to the interference.

reaction). In an unbuffered solution containing substrate, the formation of acid by acetylcholinesterase-catalyzed hydrolysis would lead to a rapid drop in pH. Because changes in pH affect the ionization state of amino-acid residues in a protein's structure, every enzyme has an optimal pH at which its catalytic activity is greatest (Figure 1). As the pH deviates from the optimal level, enzyme activity decreases. Therefore, the pH of the unbuffered solution ultimately reaches a point at which acetylcholinesterase is completely inactivated.

Now consider the idealized situation in which the unbuffered solution containing the acid-producing enzyme is supplemented with a second enzyme that catalyzes the formation of base (Figure 1). In this system, the base-producing enzyme (light line) has a pH optimum significantly lower than that of the acid-producing enzyme (dark line). Thus, the formation of base will counteract the production of acid, thereby maintaining pH by creating a dynamic pH equilibrium between the competing enzyme reactions (pH 7.5 in Figure 1). Because the generation of base is biocatalytic, base is produced only in response to a decrease in pH. If the catalytic activity of one of the enzymes is altered via inhibition, the dynamic equilibrium of the system is destroyed; the subsequent shift in pH caused by the inactivation of the dynamic equilibrium can then be used to induce a signal indicating the presence of a target enzyme inhibitor.

This is the basis for biocatalytic, dynamic-reaction equilibrium sensing. The advantages over traditional enzyme sensors include: rapid signal development; strong and intuitive responses; and high resistance to interference by temperature. An interferant-resistant nerve-agent sensor manufactured by Agentase, LLC, couples acetylcholine, acetylcholinesterase, urea, and urease in a polymer with a pH-sensitive dye.

In the rare event a weapon of mass destruction is used, such as another release of sarin in the Tokyo subway, no effective, environmentally benign decontamination systems are available. Even though enzymes provide an efficient,

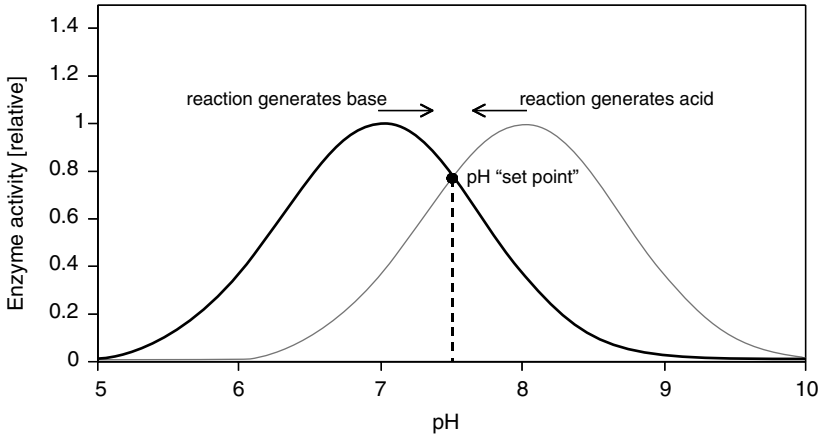

FIGURE 1 The pH dependence curves of two enzyme-catalyzed reactions. The base-producing enzyme (light line) has a lower optimal pH than the acid-producing enzyme (dark line). When both enzymes are present, the pH instantaneously stabilizes at the intersection point, which we call the pH "set point."

safe, and environmentally benign method of decontaminating nerve agents, several major hurdles still hinder their use. To decontaminate significant quantities of nerve agent, the system will require a good buffering system to handle the large amounts of acid produced from the enzyme-catalyzed hydrolysis. This problem is amplified in low-water environments, where the localized concentration of agent can easily reach 0.3 molar.

Currently, the preferred buffer for enzymatic decontamination systems under development by the U.S. Army is ammonium carbonate. The addition of solid ammonium carbonate to water results in a pH of 8.5 to 9.0 with no adjustment needed, and the ammonium ions are known to stimulate the activity of OPAA (Cheng and Calomiris, 1996). Ammonium carbonate, however, does not have a high buffering capacity, thus making it impractical for use in large-scale decontamination.

Recent studies at Porton Down, United Kingdom, demonstrated that in detergent and microemulsion systems containing 50 millimolar (mM) ammonium carbonate buffer, enzymatic degradation of high concentrations of soman (1.3 to 1.5 percent) resulted in a drop in pH from 9 to 6 in less than five minutes, at which point the enzyme was inactive. Complete decontamination of the agent was not achieved. Increasing the concentration of buffer two-fold decreased the rate of drop in pH; the pH fell to 6 in less than eight minutes. However, complete decontamination was still not achieved. The use of conventional biological buffers at concentrations sufficient to maintain pH in an optimum range for

enzyme activity introduces additional obstacles, including potential enzyme inhibition and chelation of essential metals in the enzyme structure.

The biocatalytic, dynamic-reaction equilibrium approach previously described can overcome such obstacles in the large-scale decontamination of nerve agents. Base can be generated on demand in response to the decrease in pH resulting from agent degradation. Consider an unbuffered aqueous system that includes a nerve agent-degrading enzyme, urea, and urease. In the presence of a nerve agent, biocatalytic hydrolysis would cause a rapid decrease in pH. In a method analogous to biocatalytic, dynamic-reaction, equilibrium sensing, this decrease activates urease, which generates base via the conversion of urea. Our laboratory has successfully demonstrated using biocatalytic, dynamic-reaction, equilibrium sensing for the complete decontamination of paraoxon by OPH, using urease-catalyzed urea as a buffering agent (Russell et al., 2002). Such defense systems present a significant advancement in chemical weapons defense, giving military personnel and emergency response teams the ability to achieve complete decontamination using minimal amounts of buffering material.

In summary, chemical weapons exert their terrifying effects on human physiology by interrupting biological processes. The chemicals bind tightly to enzymes, but they can also be degraded by enzymes through fortuitous side reactions. The inhibition and activity of these enzymes can be part of our defensive arsenal. Thus, biotechnology is both the target and a potential weapon in the war on terrorism.

ACKNOWLEDGEMENTS

This work was partially funded by a research grant from the Army Research Office (DAAD19-02-1-0072) and by the U.S. Department of Defense Multidisciplinary University Research Initiative (MURI) Program administered by the Army Research Office (DAAD19-01-1-0619). The principal author has an equity stake in Agentase, LLC.

REFERENCES

Cheng T.-c., and J.J. Calomiris. 1996. A cloned bacterial enzyme for nerve agent decontamination. Enzyme and Microbial Technology 18(8): 597–601.

DeFrank, J.J., and T.-c. Cheng. 1991. Purification and properties of an organophosphorus acid anhydrase from a halophilic bacterial isolate. Journal of Bacteriology 173(6): 1938–1943.

Dumas, D.P., S.R. Caldwell, J.R. Wild, and F.M. Raushel. 1989. Purification and properties of the phosphotriesterase from *Pseudomonas diminuta*. Journal of Biological Chemistry 264(33): 19659–19665.

Longworth, T.L., J.C. Cajigas, J.L. Barnhouse, K.Y. Ong, and S.A. Procell. 1999. Testing of Commercially Available Detectors against Chemical Warfare Agents: Summary Report. Aberdeen Proving Ground, Md.: Soldier and Biological Chemical Command, AMSSB-REN.

Mazur, A. 1946. An enzyme in animal tissue capable of hydrolyzing the phosphorus-fluorine bond of alkyl fluorophosphates. Journal of Biological Chemistry 164: 271–289.

Munnecke, D.M. 1979. Hydrolysis of organophosphate insecticides by an immobilized-enzyme system. Biotechnology and Bioengineering 21(12): 2247–2261.

Munnecke, D.M. 1977. Properties of an immobilized pesticide-hydrolyzing enzyme. Applied and Environmental Microbology 33: 503–507.

Russell, A.J., M. Erbeldinger, J.J. DeFrank, J. Kaar, and G. Drevon. 2002. Catalytic buffers enable positive-response inhibition-based sensing of nerve agents. Biotechnology and Bioengineering 77(3): 352–357.

An Engineering Problem-Solving Approach to Biological Terrorism

MOHAMED ATHHER MUGHAL
Homeland Defense Business Unit
U.S. Army Soldier and Biological Chemical Command (SBCCOM)
Aberdeen Proving Ground, Maryland

In the fall of 2001, anthrax-laced letters resulted in 22 confirmed infections and five deaths. Since then, the Federal Bureau of Investigation and local law-enforcement agencies have responded to thousands of reports of the use or threatened use of biological agents. Clearly, biological terrorism is a real and growing threat in the United States. But what is biological terrorism? And what approaches can we use to develop strategies to protect ourselves from this emerging threat?

BIOLOGICAL TERRORISM

The primary consequence of a large-scale biological terrorist attack would be a catastrophically large number of patients, and response systems must be capable of providing the appropriate types and amounts of medical treatments and services. However, the full spectrum of potential consequences goes far beyond medical casualties. A well planned biological terrorist attack would strain the public health surveillance systems; it would require that responders make quick, accurate disease identifications and medical diagnoses. By definition, biological terrorism is a criminal act, so an attack would entail a comprehensive criminal investigation. Depending on the biological agent used in an attack, residual environmental hazards may result. A significant portion of the population in the target area might have to be medically managed and physically controlled.

These diverse responses will require an integrated command-and-control system that includes federal, state, and local jurisdictions. In short, managing

the consequences would involve multiple professional disciplines that span three levels of government.

ASSEMBLING A TEAM

The U.S. Army Soldier and Biological Chemical Command (SBCCOM) is a unique national repository of expertise in chemical and biological defense. SBCCOM's Biological Weapons Improved Response Program (BWIRP) is designed to leverage this expertise to make sure the emergency services community is fully prepared to respond to and mitigate the consequences of a domestic bioterrorist attack. BWIRP maintains a partnership between military bioexperts and emergency managers and responders at the federal, state, and local levels. Civilian participants represent functional specialties, including emergency management, law enforcement, firefighting, emergency medical services, the handling and control of hazardous materials, and public health.

From its inception, BWIRP maintained an analytical focus on the real-world needs of emergency response professionals. The final team, a nationally representative group of civilian emergency managers and responders, included more than 60 federal, state, and local responders from nine states, eight federal agencies, six U.S. Department of Energy (DOE) national laboratories, and 11 U.S. Department of Defense agencies.

FOCUSING ON PROCESS

Once a strong team had been assembled, we defined the broad parameters of the overall process for BWIRP. The process first had to provide a forum for educating and informing the entire interdisciplinary, multiagency team on the offensive and defensive aspects of defending against bioterrorism. Second, the process had to yield initial integrated response activities for managing and mitigating the full spectrum of consequences of a large-scale, domestic, bioterrorist event.

The BWIRP process was designed to be carried out through five three-day technical workshops conducted over a six-month period. Each day of each of the five workshops was similar in structure but different in content. Day one of each workshop consisted of one-hour tutorials on preselected topics, such as the physics of aerosol dispersion, the pathogenic microbiology of biological agents, biodetection, medical treatments and intervention, and the decontamination of and physical protection against biological agents. Although the topics remained the same for each workshop, the tutorials became increasingly complex and comprehensive as the team progressed from one workshop to the next.

Day two of each workshop began with the presentation of a hypothetical bioterrorist attack scenario. From Workshop 1 through Workshop 5, the attack scenarios increased in scale from an attack on a single building to an attack on

an entire metropolitan area. After reviewing each scenario, workshop participants identified specific emergency response activities to mitigate the emerging consequences.

On Day three of each workshop, the team reviewed and integrated the entire set of response activities. The team also analyzed the integrated activities to identify shortfalls and opportunities for improvement. Throughout the reviews, the team took a "bottom-up" approach, that is, *let the problem drive the solution*.

ORGANIZING THE SOLUTION

The BWIRP team identified a myriad of response activities that spanned numerous functional areas. The next step was to organize them into a logical, integrated system. The team formulated a generic bioresponse template (Figure 1) that included the emergency response functions a city would need to respond to a bioterrorist event. This template can (and has) provided a starting point for cities and states preparing local emergency response plans for bioterrorism.

Public health surveillance, the first component of the template, should operate continuously to improve the chances for the quick detection of unusual medical events in the local population. As part of the medical surveillance systems, several communities are currently monitoring hospital admissions, 911 calls, and unexplained deaths. Once an anomaly is detected, a medical diagnosis must be made to identify and confirm its cause. If a specific disease is confirmed, the public health community will most likely undertake an epidemiological investigation to determine the distribution of cases and the source or sources of the outbreak.

Concurrent with public health activities, the law-enforcement community initiates a criminal investigation to assess the ongoing threat, safeguard evidentiary materials, and identify and apprehend suspects. Local communities should develop sampling protocols for law-enforcement personnel investigating potential biological terrorist events. The protocols should not only protect personnel and evidentiary materials, but they should also be coordinated with recipient laboratories to ensure specimens are appropriately collected and handled.

While the criminal investigation is in process, and pending the identification of a specific disease agent, local officials may begin a mass prophylaxis campaign to prevent disease and death in exposed victims. This involves the large-scale distribution and application of antibiotics, vaccines, or other medications. The speed at which medical prophylaxis is implemented is a key to success. For instance, the administration of antibiotics shortly after exposure to anthrax can significantly reduce the occurrence of disease and death; delayed administration could be ineffective.

Early, coordinated decision making on prophylactic treatment among agencies—especially public health, law-enforcement, and emergency-management agencies—is essential for a successful mass prophylaxis campaign. Because of

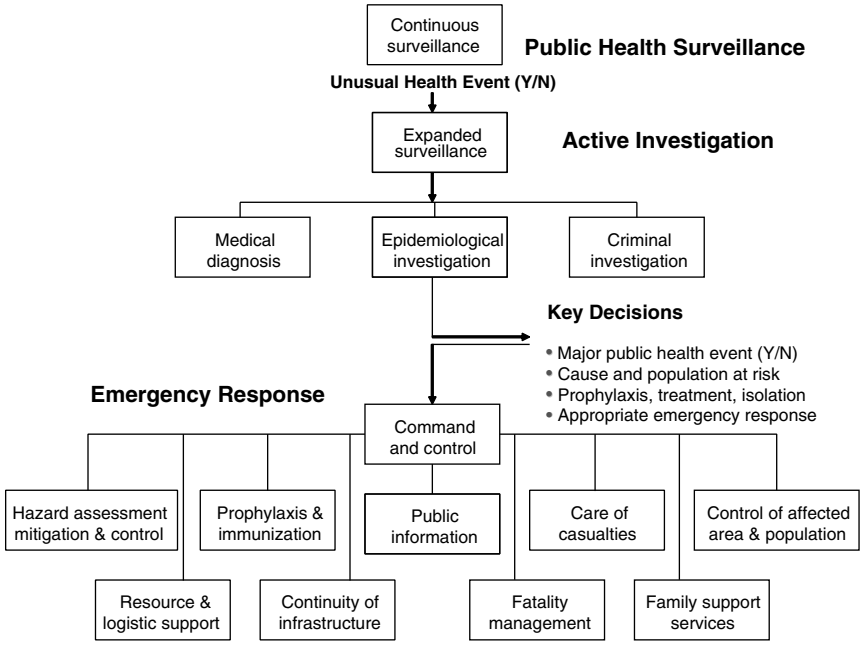

FIGURE 1 Generic bioresponse template. Source: SBCCOM.

the extreme, even global, mobility of the population, the exchange and coordination of information on the state and federal levels will most likely be necessary to support local planning for mass prophylaxis.

Depending on the biological agent, an assessment of residual hazards may be necessary to evaluate risks and mitigation measures to protect the population from further exposure to environmental hazards. Assessment and mitigation may include environmental sampling of air, water, and soil, as well as screening of insects and animals.

In the case of a contagious disease, physical control of the affected population may be necessary to minimize secondary infections. Strategies could include controlling ingress and egress points, such as bridges and tunnels, and providing security for vital sites, such as airports, hospitals, pharmaceutical distribution points, and utilities.

For an effective response, the emergency activities must take place at a rapid pace and in high volume, while ensuring continuous operation of critical infrastructure, such as communications, power generation, and water and sanitation services. The local emergency operations center (EOC) and, most likely,

state and federal EOCs—perhaps through the unified command system—will have to lead and manage the multitude of participants and resources involved.

FINE TUNING THE TEMPLATE

The template was derived through intensive analysis of five credible biological threat scenarios. By design, these scenarios were confined to infectious but noncontagious agents. Once a practical, comprehensive strategy for responding to a noncontagious agent had been developed, the strategy had to be modified to accommodate the more complex scenario of a contagious agent. To analyze and develop solutions to this problem, BWIRP and the Centers for Disease Control and Prevention (CDC) conducted a technical workshop to refine the CDC's Smallpox Control Plan/Strategy to respond to a credible, contagious bioterrorist attack scenario. The workshop panel was comprised of experts in biological warfare, medical/public health practitioners, law-enforcement officials, and emergency responders and managers. The workshop focused on vaccination, quarantine/isolation, and medical surveillance.

The panel found that the response template, with certain modifications, provides a practical strategy for minimizing the consequences of a bioterrorist attack with a contagious agent. The modifications include: adding contact-tracing to the epidemiological investigation; implementing personal protective measures for criminal investigators; establishing community outreach teams to implement mass immunizations at private homes rather than bringing potentially contagious persons together at public facilities; limiting public gatherings and mass transportation; implementing geographic isolation/quarantining; and establishing more stringent requirements for handling, burying, and disposing of fatalities.

To validate and demonstrate the applicability of the template to communities of different sizes in different regions of the country, the BWIRP team assembled and conducted investigational workshops with local first responders and emergency-management teams in three communities: Wichita, Kansas; Pinellas County, Florida; and Dover, Delaware. In all three, the template proved to be a valuable starting point for the development of customized emergency response plans.

CONCLUSIONS

BWIRP used an engineering problem-solving approach to provide a logical analytical framework for emergency responders and planners as a starting point for improving their overall readiness for a biological terrorist attack. To address this very complex, multidimensional problem, the framework was developed by participants from many different professional disciplines. By working directly with emergency responders, BWIRP was able to identify actual problems and design solutions that work in the real world. Follow-on analyses

using the same problem-solving approach both validated and improved these solutions.

In addition to providing tangible benefits to emergency responders, BWIRP's success shows that research and development engineers are a valuable national resource who can provide benefits far beyond classical engineering scenarios. Through BWIRP, mechanical, electrical, chemical, and industrial engineers have worked effectively with federal agencies as diverse as the Federal Bureau of Investigation, the Federal Emergency Management Agency, the U.S. Department of Health and Human Services, the Environmental Protection Agency, and the U.S. Department of Agriculture. Indeed, engineers worked side by side with state and local representatives in functional specialties from law enforcement, hazardous spill management, firefighting, and emergency medical services. Engineers in BWIRP feel privileged to be able to provide practical benefits and to be able to continue working on critical homeland security issues.

SUGGESTED READING

CDC (Centers for Disease Control and Prevention). 2001. The Public Health Response to Biological and Chemical Terrorism: Interim Planning Guidelines for State Public Health Officials. Washington, D.C.: Centers for Disease Control and Prevention.

CDHS (California Department of Health Services). 2001. California Hospital Bioterrorism Response Planning Guide. Available online at: *<http://www.dhs.cahwnet.gov/BioTerrorism%20Headline/Revised%20BT%20Response%20master%20document.pdf>* (October 5, 2001).

FEMA (Federal Emergency Management Agency). 1996. Guide for All-Hazard Emergency Operations Planning: State and Local Guide (SLG) 101. Available online at: *<http://www.fema.gov/rrr/gaheop.shtm>* (September, 1996).

Hutchinson, R.W., G. Christopher, M.A. Mughal, and R. Gougelet. 2003. Mass casualty prediction: by the numbers. Military Medical Technology 7(3): 12–16. Also available online at: *<http://www.mmt-kmi.com/archive_article.cfm?DocID=59>*.

Hutchinson, R.W., S. English, and M.A. Mughal. 2002. A general problem-solving approach for wicked problems: theory and application to chemical weapons verification and biological terrorism. Group Decision and Negotiation 11(4): 257–279.

Hutchinson, R.W., and M.A. Mughal. 1999. Executive Summary, BW Response Template and Response Improvements. Available online at: *<http://hld.sbccom.army.mil>* (March 1999).

Mughal, M.A. 2002. Biological terrorism: practical response strategies. Homeland Defense Journal 1(21): 11–14.

Mughal, M.A. 2002. Mitigating the unthinkable. Military Medical Technology 6(6): 30–34.

Mughal, M.A. 2002. Responding to bio-terrorism requires a concerted effort. National Defense 86(583): 34–35.

Mughal, M.A., and P. Fedele. 2001. Helping the civilian community: the Improved Response Program. Army Chemical Review PB-3-01-1: 12–16.

Mughal, M.A., and R. Fitton. 2002. Are we ready? A strategy for responding to biological terrorism. Armed Forces Journal International 140: 42–44+.

Mughal, M.A., and J. Mughal. 2003. In case of emergency. Advance for Nurses 5(7): 35.

Mughal, M.A., K. Quinn-Doggett, N. Yura, and G. Mrozinski. 2002. Homeland defense: a new mission for the Army's research, development and engineering centers. AL&T Magazine September-October: 35–37.

Mughal, M.A., and R. Vigus. 2003. Involving the Army's research, development and engineering centers in homeland defense. Homeland Defense Journal 2(4): 1–5.

SBCCOM (U.S. Army Soldier and Biological Chemical Command). 1999. Improving Local and State Agency Response to Terrorist Incidents Involving Biological Weapons: Interim Planning Guide. Available online at: <*http://hld.sbccom.army.mil*> (August 1999).

Stiner, R.S., and M.A. Mughal. 2002. Biological terrorism: new challenges for law enforcement. Sheriff Magazine 54(5): 24–27, 54–55.

Internet Security

WILLIAM R. CHESWICK
Lumeta Corporation
Somerset, New Jersey

One of the design principles of the Internet was to push the network intelligence to the "edges," to the computers that use the network rather than the network itself (Saltzer et al., 1984). Any given edge computer could send packets to any other edge host, leaving the details of packet delivery to the routers in the center of the network. This principle greatly simplified the design of the routers, which simply had to hot-potato packets toward the appropriate edge host as efficiently as possible. Routers could drop packets if there was congestion, leaving the edge hosts to provide reliable data delivery. Under this scheme, routers could relay packets with few complications, and they could be implemented in state-of-the-art hardware. Routers at the core of the Internet have benefited from Moore's Law and operated at its limits since the late 1970s.

This approach is in direct contrast to the standard telephone system, in which the intelligence resides in centrally controlled phone switches, and the edges have dumb terminals (i.e., standard telephones). In the phone system, the phone company invents and implements the technology. In the decentralized Internet approach, edge computers can install arbitrary new protocols and capabilities not envisioned by the designers of the Internet; the World Wide Web is the most obvious example.

As a result of this design, most current newspaper articles covering the Internet refer to things that happen at the edges. Viruses infect PC clients, worms attack network servers, and the Internet dutifully delivers denial-of-service attack packets to the edges. (For the security concerns of edge hosts, see Wagner, 2003.)

Most Internet edge hosts play the role of "clients" or "servers," despite the popular use of the term "peer-to-peer" networking. A client requests a connec-

tion, and a server provides the service. Servers usually take all comers and must be accessible to a large number of hosts, most of which are clients and are privately owned. Hosts must connect to servers but aren't necessarily servers, which makes them harder to attack.

A large majority of edge computers on the Internet are probably susceptible to attack and subversion at any given time. Because it is not economically feasible to harden all of these hosts, they are isolated from the rest of the Internet and many potential attacks by breaking the end-to-end model. It is impossible to mount a direct attack on a computer that cannot be reached. The trade-off between security and functionality does decrease the envisioned power and functionality of the Internet somewhat. Often, there is no satisfactory trade-off: the choice is between an unsafe service and no service at all.

Enclaves of edge computers are isolated topologically, through firewalls, routing tricks, and cryptography. At the same time, trusted communities of hosts, called intranets, allow businesses to operate with a reduced probability of successful attacks. This approach tends to create an organization with a hard outside and a soft, chewy center. Insider attacks can (and do) occur fairly often, and perimeter defenses often, even usually, contain holes.

FIREWALLS

Although network services can (and should) be hardened when necessary, the easiest defense is to get out of the game—to turn off the service or render it unreachable by attacking hosts. This can done by completely disconnecting the network from the Internet (as some high-security government networks have done) or by connecting it through a device that blocks or filters incoming traffic, called a firewall. Complete disconnection offers the best security, but it is very unpopular with most users, who may then seek to use *sub rosa* connections.

Nearly all corporate networks, and portions of many university networks (which have traditionally been open and unfiltered), use firewalls of varying strictness to cleanse incoming packet flows. Most consumers also have firewalls in their cable modems, DSL routers, wireless base stations, or client computers.

Firewalls provide a central site for enforcing security policies. Many companies provide a few firewalls as gateways between the corporate intranet and the Internet, equipped with the rules and filters to enforce the company's security policies. The Internet security expertise, which is scarce and expensive, is situated at a few centralized choke points. This lowers the costs of configuring perhaps hundreds of thousands of corporate hosts to resist software attacks from random strangers and malicious software ("malware").

Firewalls, a form of the classic perimeter defense, are supposed to offer the only connections between an intranet and the Internet, but long perimeter defenses can be difficult to monitor. A firewall may amount to circling the state of Wyoming, but modern corporate intranets can span the world. Therefore, it can

be easy to add an unauthorized connection that breaches the perimeter without the knowledge of the centralized authority. Company business units and rogue employees may believe they have a compelling need—often a business need—to access the Internet in a way not permitted by corporate policy. For example, it may take time to get approvals for new services, and the new services may require filters that take time to install or may simply be impractical—overconstrained by a poorly designed protocol that doesn't admit to easy filtering. An employee who decides it is better to apologize later than to get permission now might make his or her own connection.

Unauthorized connections that pierce the corporate perimeter may come from business partners, misconfigured gateways, or newly acquired companies. Even if unauthorized connections have been installed for good business reasons, the reason, and the person who understood the reason, may be long gone by the time a problem is identified. A new network administrator may have to choose between (1) closing down a suspect Internet connection and risking a vital business service and (2) leaving the connection up and allowing invaders into the company.

Rogue connections are often mistakes. It is easy to misconfigure virtual private networks (VPNs), Internet tunnels linking disconnected islands of trust (e.g., linking remote office networks or home computers with headquarters). Company employees often connect to a corporate network using encryption to conceal their data, thus creating a VPN that can interconnect home and office. An employee can also create a hole in the corporate perimeter. (My company's principal business is to find these holes.)

If you are in charge of such a network, how can you deal with these problems? The most effective network administrators aggressively manage their networks. They control new and existing connections carefully. They use mapping tools to detect new routes and connections. They are equipped with, and use, aggressive management controls—violators of policy can be and are fired and even prosecuted.

Perimeter defenses like firewalls can be effective, and there are periodic pop-quizzes from the Internet itself that prove this. For example, if a widespread worm does not appear on an intranet, you may be reasonably sure that the perimeter does not have overt holes. Firewalls have limitations, of course. Attacks can come from inside an intranet, and many attacks or weakness are caused by stupidity rather than malice. (One government network administrator said he had more problems with morons than with moles.) An intranet must be hardened from internal attacks, which brings us back to host-based security.

ROUTING LIMITATIONS

You cannot directly attack a host you cannot reach. Besides firewalls, which can block access to a community of hosts, the community can also use network

addresses that are not known to the Internet. If the Internet routers don't know how to reach a particular network, an Internet edge host probably cannot send packets to that network. A community with an official set of private addresses that are guaranteed never to be announced on the Internet can use these addresses to communicate locally, but the addresses are not reachable from the Internet. The community can still be attacked through intermediate hosts that have access to both private and public address spaces, and many viruses and worms spread this way.

Some attacks on the Internet come from host communities that are only announced on the Internet intermittently. The networks connect, the attacks occur, and then the networks disconnect before a pursuit can be mounted. To give a community using private address space access to the Internet, one uses network address translation.

NETWORK ADDRESS TRANSLATION

The Internet has grown, with very few tweaks, by more than nine orders of magnitude—a testament to its designers. But we are running uncomfortably close to one design limit of IP version 4, the technical description of the Internet and intranets, and their basic protocols. We have only about three billion network addresses to assign to edge hosts, and about half of them are currently accessible on the Internet, and more have been assigned. Thus, address space has gotten tight; by one common account, a class B network (roughly 65,000 IP addresses) has a street value of $1 million.

A solution that emerged from this shortage was a crude hack called "network address translation" (NAT). Many hosts hide on a private network, using any network address space they wish. To communicate with the real Internet, they forward their packets through a device that runs NAT, which translates the packet and sends it to the Internet using its own address as the return address. Returning packets are translated to the internal address space and forwarded to the internal host. A home network full of game-playing children probably looks like one busy host from the Internet.

Instead of the end-to-end model, in this model many hosts route their packets through a single host, sharing that host's network address. Only one, or one small set of network addresses is used, conserving official Internet address space.

This model has had a useful implication for security. Hosts on the home network cannot be easily scanned and exploited from the outside. Indeed, it takes some work to detect NAT at all (Bellovin, 2002). Internal hosts enjoy considerable protection in this arrangement, and it is likely that some form of NAT will be used often, even when (if?) IP version 6 (a new protocol description with support for a much larger address space) is deployed. A special configura-

tion is required on the NAT device to provide services to the Internet. The default configuration generally offers no services for this community of hosts, and safe defaults are a good design.

EMERGING THREATS TO THE INTERNET

Although one can connect safely to the Internet without a topological defense, like skinny dipping, this involves an element of danger. Host security has to be robust, and the exposed machines have to be monitored carefully. This is exactly the situation faced by web servers. A server like *www.budweiser.com* must have end-to-end connectivity so customers can reach it. Often commercial web services have to connect to sensitive corporate networks (think of FedEx and its shipping database or a server for online banking.) These web sites must be engineered with great care to balance connectivity and security.

The Internet is a collaboration of thousands of Internet service providers (ISPs) using the TCP/IP protocols. There is somewhat of a hierarchy to the interconnections; major backbone providers interconnect at important network access points and feed smaller, more local ISPs.

The graphical and dynamic properties of the Internet are popular research topics. Some recent studies have suggested that the Internet is a scale-free network, which would have important implications for sensitivity to partitioning. But this conclusion overlooks two issues. First, the raw mapping data used to build graphical descriptions are necessarily limited. Internet mapping is not perfect, although it is getting better (Spring et al., 2002). Second, the most critical interconnections were not formed by random processes or by accident, as perhaps are other "six degrees of separation" networks. Internet interconnections, especially backbone connections, have been carefully engineered by their owners over the years and have had to respond to attacks and a variety of misadventures, occasionally backhoes. These experiences have helped harden critical points in the network. (Network administrators do not like wake-up calls at 3 a.m., and unreliability is a bad component of a business model.)

But there are weaknesses, or at least obvious points of attack, in the core of the Internet. These include routing announcements, the domain name system (DNS), denial-of-service (DOS) attacks, and host weaknesses in the routers themselves.

Routing Announcements

A typical Internet packet goes through an average of 17 routers; the longest paths have more than 30 hops. A router's job is to direct an incoming packet toward the next hop on its way to its destination. For a router near an edge, this is simple—all packets are directed on a default path to the Internet. But routers

on the Internet need a table of all possible Internet destinations and the path to forward packets for each one.

A routing table is built from information exchanged among Internet routers using the border gateway protocol (BGP). When a network connects or disconnects from the Internet, an announcement is sent toward the core routers; dozens of these routing announcements are generated every minute. If a core router accepts an announcement without checking it, it can propagate trouble into the Internet. For example, a router in Georgia once announced a path to MIT's networks through a local dial-up user. As a result, many parts of the Internet could not reach MIT until the announcement was rescinded. Such problems often occur on a smaller scale, and most ISPs have learned to filter routing announcements. Only MIT should be able to announce MIT's network announcements.

One can easily imagine an intentional attempt to divert or disrupt the flow of Internet traffic with false announcements or by interfering with BGP in some way. Because ISPs frequently have to deal with inadvertent routing problems, they ought to have the tools and experience to deal with malicious routing attacks, although it may take them a while, perhaps days, to recover.

The Domain Name System

Although the domain name system (DNS) is technically not part of the TCP/IP protocol, it is nearly essential for Internet operation. For example, one could connect to *http://207.171.181.16/*, but nearly all Internet users prefer *http://www.amazon.com*. DNS translates from name to number using a database distributed throughout the Internet. The translation starts at the root DNS servers, which know where to find *.go*, *.eddo*, *.czar*, etc. Servers' databases contain long lists of subdomains—there are tens of millions in the *.com* database—that point to the name servers for individual domains. For example, DNS tells us that one of the servers that knows about the hosts in *amazon.com* is at IP address *207.171.167.7*.

DNS can be and has been attacked. Attackers can try to inject false responses to DNS queries (imagine your online banking session going to the wrong computer). Official databases can be changed with phony requests, which often are not checked very carefully. Some domains are designed to catch typographical errors, like *www.anazon.com*.

In October 2002, the root name servers came under massive denial-of-service (DOS) attacks (see below). Nine of the 13 servers went out of service—the more paranoid and robust servers kept working because they had additional redundancy (a server can be implemented on multiple hosts in multiple locations, which can make it very hard to take down). Because vital root DNS information is cached for a period of time, it takes a long attack before people notice problems.

Denial-of-Service Attacks

Any public service can be abused by the public. These days, attackers can marshal thousands of hacked computers, creating "botnets" or "zombienets." They can subvert large collections of weak edge hosts and run their own software on these machines. In this way, they can create a large community of machines to direct packets at a target, thus flooding target computers, or even their network pipes, to degrade normal traffic and make it useless. These packets can have "spoofed" return addresses that conceal their origins, making trace-back more difficult. DOS attacks occur quite often, frequently to politically unpopular targets; Amazon, Microsoft, and the White House are also popular targets. Some PC viruses spread and then launch DOS attacks.

The Internet infrastructure does not provide a means for trace-back or suppression of DOS attacks, although several techniques have been proposed. Targeted hosts can step out of the way of attacks by changing their network addresses. Added host and network capacity at the target is the ultimate protection.

Routers as Edge Hosts

Security problems for routers are similar to those for hosts at the edge of the network. Although routers' operating systems tend to be less well known and not as well understood as those of edge hosts, they have similar weaknesses. Cisco Systems produces most of the Internet routers, and common failures in their software can subject those routers to attack. (Of course, other brands also have weaknesses, but there are fewer of these routers.)

Attacks on routers are known, and underground groups have offered lists of hacked routers for sale. The holes in routers are harder to exploit, but we will probably see more problems with hacked routers.

Botnets

Attackers are looking for ways to amplify their packets. The Smurf attacks in early 1998 were the first well known attacks that used packet amplification (CERT® Coordination Center, 1998). The attackers located communities of edge computers that would all respond to a single "directed-broadcast" packet. These "tickling" packets had spoofed return addresses of the intended target. When community members all responded to the same tickling packet, the target was flooded with hundreds or even thousands of packets.

Most amplification networks now block directed-broadcast packets. At this writing, *netscan.org* has a list of some 10,000 broken networks with an average amplification of three, barely enough for a meaningful attack. So the hacking community began collecting long lists of compromised hosts and installed slave software on each one. These collections have many names—botnets and zom-

bienets, for example. A single anonymous host can instruct these collections to attack a particular site. Even though the target knows the location of these bot hosts (unless they send packets with spoofed return addresses), there are too many to shut down.

Worse, the master can download new software to the bots. They can "sniff" local networks, spread viruses or junk e-mail, or create havoc in other ways. The master may use public key encryption and address spoofing to retain control over the botnet software and hide the source of the master.

The latest development is the use of botnets to hide the identity of web servers. One can put up an illegal web server, and traffic to the server can be laundered through random corrupted hosts on the Internet, making the actual location of the server quite difficult to find.

CONCLUSION

Most hypothesized Internet security problems on the Internet have eventually appeared in practice. A recent one, as of this writing, was a virus that patches security problems as its only intended action—in other words, a "good" virus. (Of course, it had bugs, as most viruses do.) I first heard proposals for this approach some 15 years ago. Although well intentioned, it is a bad idea.

The good news is that nearly all Internet attacks, real or envisioned, involve flawed software. Individual flaws can be repaired fairly quickly, and it is probably not possible to take the Internet down for more than a week or so through software attacks. Attacks can be expensive and inconvenient, but they are not generally dangerous. In fact, the threats of cyberterrorism devalue the meaning of "terror."

There's more good news. The Internet continues to be a research project, and many experts continue to work on it. When a dangerous, new, and interesting problem comes along, many experts drop what they are doing and attempt to fix it. This happened with the Morris worm in 1988 and the SYN packet DOS attacks in 1996, for which there was no known remediation at the time. Major attacks like the Melissa virus were quelled within a week.

The success of the Internet has fostered huge economic growth. Businesses have learned to control the risks and make successful business models, even in the face of unreliable software and network connections. Insurance companies are beginning to write hacking insurance, although they are still pondering the possible impact of a widespread, Hurricane Andrew-like Internet failure. I think we have the tools and the experience to keep the Internet safe enough to get our work done, most of the time.

REFERENCES

Bellovin, S.M. 2002. A Technique for Counting NATted Hosts. Pp. 267–272 in Proceedings of the Second Internet Measurement Workshop, November 6–8, 2002, Marseilles, France. Available online at: *<http://www.research.att.com/~smb/papers/fnat.pdf>*.

CERT® Coordination Center. 1998. CERT® Advisory CA-1998-01 Smurf IP Denial-of-Service Attacks. Available online at: *<http://www.cert.org/advisories/CA-1998-01.html>*.

Saltzer, J.H., D.P. Reed, and D.D. Clark. 1984. End-to-end arguments in system design. ACS Transactions on Computer Systems 2(4): 277–288.

Spring, N., R. Mahajan, and D. Wetherall. Measuring ISP Technologies with Rocketfuel. 2002. Presented at ACM-SIGCOMM '02, August 19–22, 2002, Pittsburgh, Pennsylvania. Available online at: *<http://www.acm.org/sigcomm/sigcomm2002/papers/rocketfuel.pdf>*.

Wagner, D. 2003. Software Insecurity. Presentation at 2003 NAE Symposium on Frontiers of Engineering, September 19, Irvine, California.

BIOMOLECULAR COMPUTING

DNA Computing by Self-Assembly

ERIK WINFREE
Departments of Computer Science and Computation and Neural Systems
California Institute of Technology
Pasadena, California

INFORMATION AND ALGORITHMS IN BIOCHEMISTRY

Information and algorithms appear to be central to biological organization and processes, from the storage and reproduction of genetic information to the control of developmental processes to the sophisticated computations performed by the nervous system. Much as human technology uses electronic microprocessors to control electromechanical devices, biological organisms use biochemical circuits to control molecular and chemical events. The engineering and programming of biochemical circuits, *in vivo* and *in vitro*, would transform industries that use chemical and nanostructured materials. Although the construction of biochemical circuits has been explored theoretically since the birth of molecular biology, our practical experience with the capabilities and possible programming of biochemical algorithms is still very young.

In this paper, I will review a simple form of biochemical algorithm based on the molecular self-assembly of heterogeneous crystals that illustrates some aspects of programming *in vitro* biochemical systems and their potential applications. There are two complementary perspectives on molecular computation: (1) using the astounding parallelism of chemistry to solve mathematical problems, such as combinatorial search problems; and (2) using biochemical algorithms to direct and control molecular processes, such as complex fabrication tasks. The latter currently appears to be the more promising of the two.

Some major theoretical issues are common to both approaches—how algorithms can be encoded efficiently in molecules with programmable binding interactions and how these algorithms can be shown to be robust to asynchronous and unreliable molecular processes. Proof-of-principle has been experimentally

demonstrated using synthetic DNA molecules; how well these techniques scale remains to be seen.

ALGORITHMIC SELF-ASSEMBLY AS GENERALIZED CRYSTAL GROWTH

The idea of algorithmic self-assembly arose from the combination of DNA computing (Adleman, 1994), the theory of tilings (Grunbaum and Sheppard, 1986), and DNA nanotechnology (Seeman, 1982; Seeman, 2003). Conceptually, algorithmic self-assembly naturally spans the range between maximal simplicity (crystals) and arbitrarily complex information processing. Furthermore, it is amenable to experimental investigation, so we can rigorously probe our understanding of the physical phenomena involved. This understanding may eventually result in new nanostructured materials and devices.

DNA Computing

Leonard Adleman's original paper on DNA computing contained the seed of the idea we'll pursue here—that the programmability of DNA hybridization reactions can be used to direct self-assembly according to simple rules. In the first combinatorial-generation step of Adleman's procedure, DNA molecules representing all possible paths through the target graph were assembled by DNA hybridization in a single step. The basic idea (Figure 1) is for a set of molecules with unique sequences to represent the vertices and edges of the graph, thus governing which vertices can follow which other vertices. Each possible sequence of hybridization reactions, occurring spontaneously in any order, produces a double-stranded DNA molecule whose sequence encodes a valid path through the graph. By thus generalizing one-dimensional polymerization to include programmable binding, Adleman coaxed the DNA to generate patterns that follow certain mathematical rules. This is an elegant idea—and it works! The problem is that only simple computations can be performed with linear self-assembly. Paths through graphs correspond to regular languages, which have the complexity of finite-state machines—thus more sophisticated aspects of computation cannot be reached by this technique.

Tiling Theory

A tiling is an arrangement of a few basic shapes (called tiles) that fit together perfectly in the infinite plane. For each tiling, the set of shapes must be finite; for example, the tile set could consist of an octagon and a square, both with unit-length sides. One motivation for studying tiling is that the tiles correspond to the periodic arrangement of atoms in crystals. A remarkable result is that all possible periodic arrangements can be classified according to their

FIGURE 1 Linear self-assembly of DNA can be directed to follow valid paths through a graph. Sequences used in practice would have 15–30 nucleotides for each domain, rather than 3 nucleotides as shown here. Source: Adapted with permission from Winfree et al., 1998b.

fundamental symmetries; in three dimensions there are 230 symmetries, and in two dimensions there are 17 symmetries. This suggests that, given a finite set of polygonal tiles, one should be able to determine whether they can be arranged according to one of the known symmetries, or whether there is no way to arrange them on the plane.

This is what Hao Wang thought in the 1960s, but when he looked into the question, known as the tiling problem, he discovered that it is provably unsolvable (Wang, 1963)! That is to say, aperiodic tilings are also possible. In addition, it can be incredibly difficult to determine whether a given set of tiles can tile the plane aperiodically or whether every attempt will ultimately fail. To prove this result, Wang developed a way to create a set of tiles that fit together uniquely to reproduce the space-time history of any chosen Turing[1] machine, in such a way that, if the Turing machine halts (with an output), then the attempted tiling has to get stuck; if the Turing machine continues computing forever, then a consistent global tiling is possible.

Thus, the tiling problem reduces to the halting problem, the first problem proved to be formally undecidable. This result shows that tiling is theoretically as powerful as general-purpose computers. In fact, the tiles Wang used were all essentially square, distinguished only by labels on their sides that had to match up when the tiles were juxtaposed. Thus, the complexity arises from the logical constraints in how the tiles fit together, rather than from the tiles themselves.

Given the intimate relation between crystals and tiling theory, it is natural to ask if crystal growth has the potential to compute as powerfully. To answer this question, we need two things: (1) the ability to design molecular Wang tiles; and (2) precise rules for crystal growth that can be implemented reliably.

DNA Nanotechnology

We now turn to DNA nanotechnology, the brainchild of Nadrian Seeman's vision of using DNA as an architectural element (Seeman, 1982). Like RNA, DNA can make structures other than the usual double helix. These other structures include hairpins and three- and four-way branch points, which are important for biological function. Seeman, however, pictured these structures as hinges and joints, bolts and braces that could be programmed to fold and bind to each other by careful design of the DNA base sequence. Seeman and his students constructed a wide variety of amazing nanostructures: a wire-frame cube and truncated octahedron; single-stranded DNA and RNA knots, including the tre-

[1]Turing machines, invented by Alan Turing in 1936, are extremely simple computers that consist of a finite-state compute head that can move back and forth on an infinite one-dimensional memory tape. Turing showed that these machines are universal in the sense that they can perform any computation that can be performed by any other mechanical device—there is no fundamental need to use a more complicated kind of computer!

foil, the figure-eight, and Borromean rings; and rigid building-block structures, such as triangles and four-armed "bricks" known as double-crossover (DX) molecules; and more (Seeman, 2003).

The idea, then, is to use these "bricks" as molecular Wang tiles (Winfree et al., 1998a). The four arms of the DX molecules can be given sequences corresponding to the labels on the four sides of the Wang tiles. Thus, any chosen Wang tile can be implemented as a DNA molecule. Appropriate design of the molecule will encourage assembly into two-dimensional sheets.

The problem, then, is to ensure that the growth process results in tile arrangements in which all tiles match with their neighbors. It is easy, however, to envision ways of putting the tiles together so that the tiles match at each step but soon create a configuration for which there is no way to proceed without creating a mismatch or having to remove offending tiles. This situation is analogous to the distinction between uncontrolled precipitation, which occurs rapidly when there is a strong thermodynamic advantage to aggregation, and quality crystal growth, which occurs slowly when there is a slight thermodynamic advantage for molecules that bind in the preferred orientation, but other possible ways to bind are disadvantageous.

A formalization of this notion for Wang tiles, the Tile Assembly Model, supposes that each label on a Wang tile binds with a certain strength (typically, 0, 1, or 2) and that tiles will only stick to a growing assembly if they bind (possibly via multiple bonds) with a total strength greater than some threshold τ (typically 1 or 2); tiles that bind with a weaker strength immediately fall off (Winfree, 1998). Under these rules, growth from a "seed tile" can result in a unique, well defined pattern. Because Turing machines and cellular automata can be simulated by this process, the Turing-universality of tiling is retained.

As an example, consider the seven tiles shown in Figure 2 assembling at $\tau = 2$. These tiles perform a simple computation—they count in binary. Starting with the seed tile, labeled *S*, the tiles with strength-2 bonds polymerize to form a V-shaped boundary for the computation. There is a unique tile that can fit into the nook of the V; because it makes *two* strength-1 bonds, it can in fact be added. Two new nooks are created, and again a unique tile can be added in each location. The assembly thus grows forever, counting and counting with unabated madness.

Tiles can be added in any order, but the resulting pattern is the same. The same basic self-assembly mechanisms used here are sufficient to perform more sophisticated computations. No new ideas or mechanisms are necessary to obtain fully programmable Turing-universal behavior.

EXPERIMENTAL ADVANCES

The first demonstration of these ideas—two-dimensional, periodic arrays of DNA tiles—could hardly be called "algorithmic," but it did show that the se-

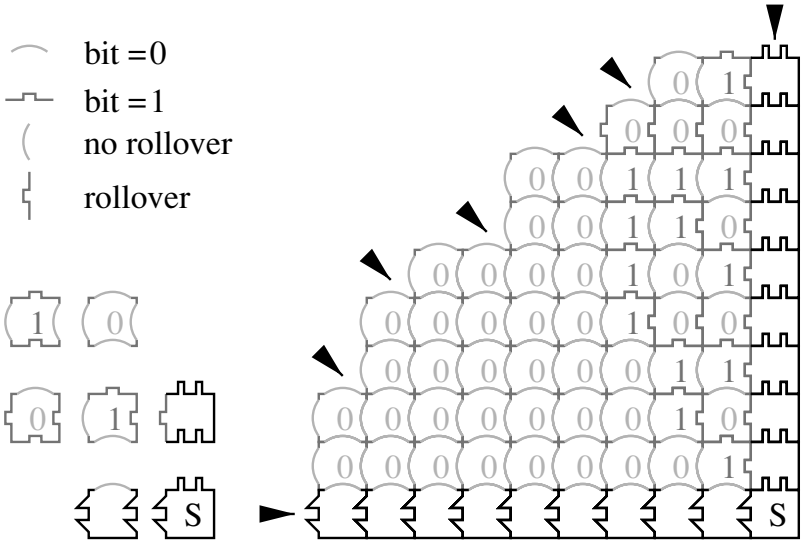

FIGURE 2 A set of seven tiles that implement a binary counter when started with the seed tile *S*. Strength-2 bonds are indicated by tile sides with *two* projections (or indentations); other bonds have strength 1. Arrows indicate sites where a tile may be added at $\tau = 2$. Source: Adapted with permission from Winfree, 2000.

quences given to the tiles' sticky ends could be used to program different periodic arrangements of tiles (Winfree et al., 1998a). The encoding of tiles as DNA DX molecules is illustrated in Figure 3; Figure 4 shows small crystals of DX molecules adsorbed on mica, as they appear in the atomic force microscope. Subsequent studies have shown that DNA tiles can be made from a variety of different molecular structures. Thus, the principle that the arrangement of two-dimensional tiles can be directed by programmable, sticky-end interactions appears to be quite robust.

The goal of creating three-dimensional, periodic arrays of DNA tiles, originally formulated by Seeman more than 20 years ago, remains an open problem in the field. Once solved, it will allow for more sophisticated information-processing techniques in algorithmic self-assembly, roughly analogous to the increase in power from one-dimensional to two-dimensional cellular automata or Turing machines.

For the time being, experimental demonstration of algorithmic self-assembly has been confined to one- and two-dimensional assemblies. The first use of one-dimensional algorithmic self-assembly appeared as the first step in Adleman's original DNA-based computing demonstration; this process formally corresponds to the generation of languages by finite-state machines. Further-

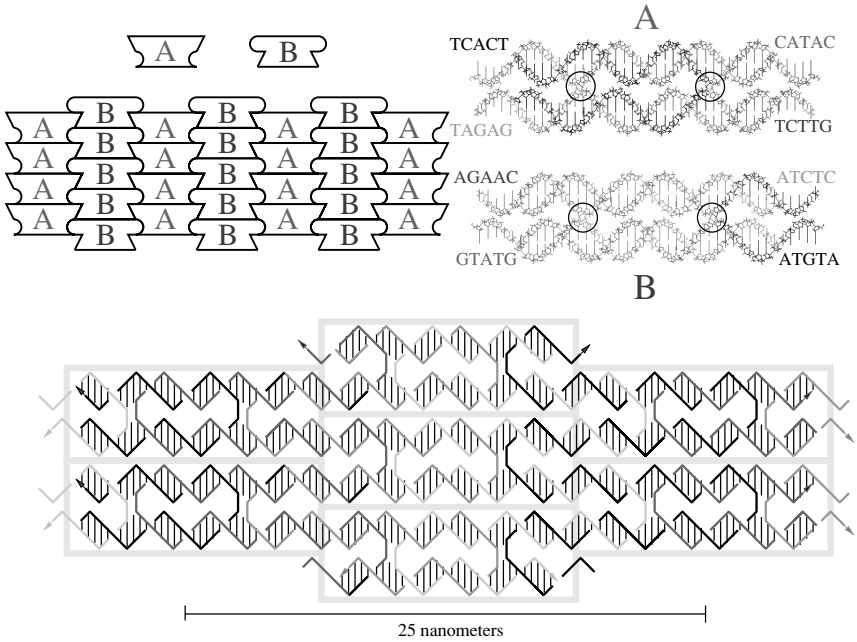

FIGURE 3 DNA double-crossover molecules can implement abstract Wang tiles, producing a two-dimensional lattice of DNA with binding interactions dictated by the DNA sticky ends. Source: Adapted with permission from Winfree, 2000.

more, using one-dimensional, tile-based assembly, it is possible to read an input string (encoded as a one-dimensional tile assembly) and generate an output string consisting of the cumulative[2] exclusive-OR (XOR) of the input string (Mao et al., 2000); this formally corresponds to a finite-state transducer.

The first two-dimensional, algorithmic self-assembly process to be experimentally demonstrated with DNA is a generalization of the one-dimensional XOR example (Rothemund and Winfree, in preparation). Beginning with an input row consisting of a single 1 in a sea of 0's, the next layer grows by placing a 0 where both neighbors in the layer below are the same and a 1 where they are different. This process, an example of a one-dimensional cellular automaton, generates a fractal pattern known as the Sierpinski gasket.

In addition to the DNA required to construct the input, only four DNA tiles are required (in principle) to grow arbitrarily large Sierpinski triangles. Experimentally, error-free Sierpinski triangles as large as 8 x 16 have been observed by atomic force microscopy. However, error rates (the frequency with which the

[2]The n^{th} bit of the cumulative XOR gives the parity of the first n bits of the input sequence.

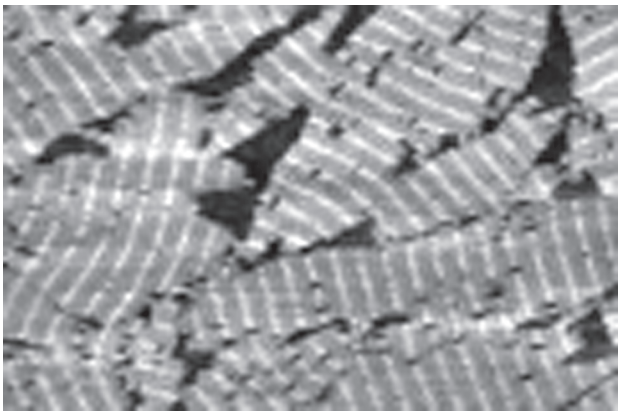

FIGURE 4 Atomic force microscope image of DNA double-crossover crystals. Stripes are spaced at 25 nm; individual $2 \times 4 \times 13$ nm tiles are visible. Source: Image taken by Nick Papadakis, Winfree Lab.

wrong tile was incorporated into the crystal) ranged from 1 to 10 percent, and many fragments appeared to have grown independently of the input structure. It is clear that controlling nucleation and finding mechanisms to reduce the error rates are critical challenges for making algorithmic self-assembly practical.

POTENTIAL TECHNOLOGICAL APPLICATIONS

Combinatorial Optimization Problems

Solving combinatorial optimization problems, in the spirit of Adleman's original paper, was the first application considered for algorithmic self-assembly. Adleman's essential insight is based on the fact that a class of hard computational problems, the NP-complete problems, share a common generate-and-test form—does a sequence exist that satisfies easy-to-check properties X, Y, . . . , and Z. All known algorithms for NP-complete problems require exponential[3] time or exponential parallelism. The basic idea is to use combinatorial chemistry techniques to simultaneously generate all potential solutions and then to filter them, based on chemical properties related to the information they encode, leaving at the end possibly only a single molecule that has all of the desired properties. If the final solution to the problem is defined by satisfying a small number of simple properties—as is the case for all NP-complete problems—then this approach can be used to find the solution in a short amount of time, if the

[3]Exponential in the length of the problem description, in bits.

parallelism is sufficient. That a single cc of DNA in solution at reasonable concentrations can contain 2^{60} bits of information—which can be acted on simultaneously by chemical operations—gives us hope that the parallelism could be sufficient.

By exploiting the situation in which multiple different tiles could be added at a given location—much like Adleman's assembly step that produced all possible paths through a graph—self-assembly can generate a combinatorial set of possible assemblies and then continue growing according to a process that tests the information to see if it has the desired properties. Theoretical schemes have been worked out that use a single self-assembly step to solve the Hamiltonian path problem (HPP) (Winfree et al., 1998b), solve the Boolean formula satisfiability problem (SAT) (Lagoudakis and LaBean, 2000), and perform other math calculations (Reif, 1997). How much computation could be done this way? If assembly were to proceed with few errors, solving a 40-variable SAT problem would require 30 milliliters of DNA at a tile concentration of 1 micromolar and might be completed in a few hours. This "best possible" estimate corresponds to 10^{12} bit operations per second—not bad for chemistry but still low compared to electronic computers.

The sheer speed and flexibility of silicon-based electronic computers make them preferable to DNA computing, even if self-assembly were to proceed without errors. We can conclude, then, that the low-hanging fruit are not to be found in the field of combinatorial search. But the ability of self-assembly to perform sophisticated computations suggests that we are making progress toward our goal of understanding (and potentially exploiting) autonomous biochemical algorithms. A more promising application is suggested by examining how self-assembly is used in biology.

Programmable Nanofabrication

Biology uses algorithmically controlled growth processes to produce nanoscale and hierarchically structured materials with properties far beyond the capability of today's human technology. Does DNA-based algorithmic self-assembly give us access to new and useful technological capabilities? The simplest applications would make use of self-assembled DNA as a template or scaffold for arranging other molecular components into a desired pattern. This could be used for biochemical assays, novel materials, or devices. Seeman has envisioned, for example, using periodic three-dimensional DNA lattices to assist with difficult protein crystallization or to direct construction of molecular electronic components into a memory (Robinson and Seeman, 1987).

The potential of self-assembly for fabricating molecular electronic circuits is intriguing, given the limitations of conventional silicon-circuit fabrication techniques. Photolithography is unable to create features significantly smaller than the wavelength of light, and even if it could, for several-nanometer line widths

the unspecified atomic positions within the silicon substrate would lead to large stochastic fluctuations in device function. For these reasons, many researchers are investigating electrical computing devices created from molecular structures, such as carbon nanotubes, in which the location of every atom is well defined. However, an outstanding problem is how to arrange these chemical components into a desired pattern.

DNA self-assembly could be used in a variety of ways to solve this problem: molecular components (e.g., AND, OR, and NOT gates, crossbars, routing elements) could be chemically attached to DNA tiles at specific chemical moieties, and subsequent self-assembly would proceed to place the tiles (and hence circuit elements) into the appropriate locations. Alternatively, DNA tiles with attachment moieties could self-assemble into the desired pattern, and subsequent chemical processing would create functional devices at the positions specified by the DNA tiles. None of these approaches has yet been convincingly demonstrated, but it is plausible that any of them could eventually succeed to produce two- or three-dimensional circuits with nanometer resolution and precise control of chemical structure.

Using self-assembly to direct the construction of circuits as large and complex as those found in modern microprocessors is daunting. The question arises, therefore, of whether there are useful circuit patterns that can be generated by a feasibly small number of tiles. Any circuit pattern that has a concise algorithmic description is a potential target for this approach. Small tile sets have been designed for demultiplexers, such as the ones necessary to access a RAM memory (shown in Figure 5), and for signal-processing primitives, such as the Hadamard matrix transform (Cook et al., 2004). Regular gate arrays, such as those used in cellular automata and field programmable gate arrays (FPGAs), are another natural target for algorithmic self-assembly of circuits.

Many technical hurdles will have to be overcome before algorithmic self-assembly can be developed into a practical commercial technology. It is not clear if real circuits will ever be built this way, but the sheer range of possibilities opened up by algorithmic growth processes suggests that algorithmic self-assembly will be used in the future for technologies that place molecular components in a precisely defined complex organization.

SUMMARY AND PROSPECTS

DNA-based self-assembly appears to be a robust, readily programmable phenomenon. Periodic two-dimensional crystals have been demonstrated for tens of distinct types of DNA tiles, illustrating that in these systems the sticky ends drive the interactions between tiles. Several factors limit immediate applications, however. Unlike high-quality crystals, current DNA tile lattices are often slightly distorted, with the relative position of adjacent tiles jittered by a nanometer and lattice defect rates of 1 percent or more. Some DNA tiles

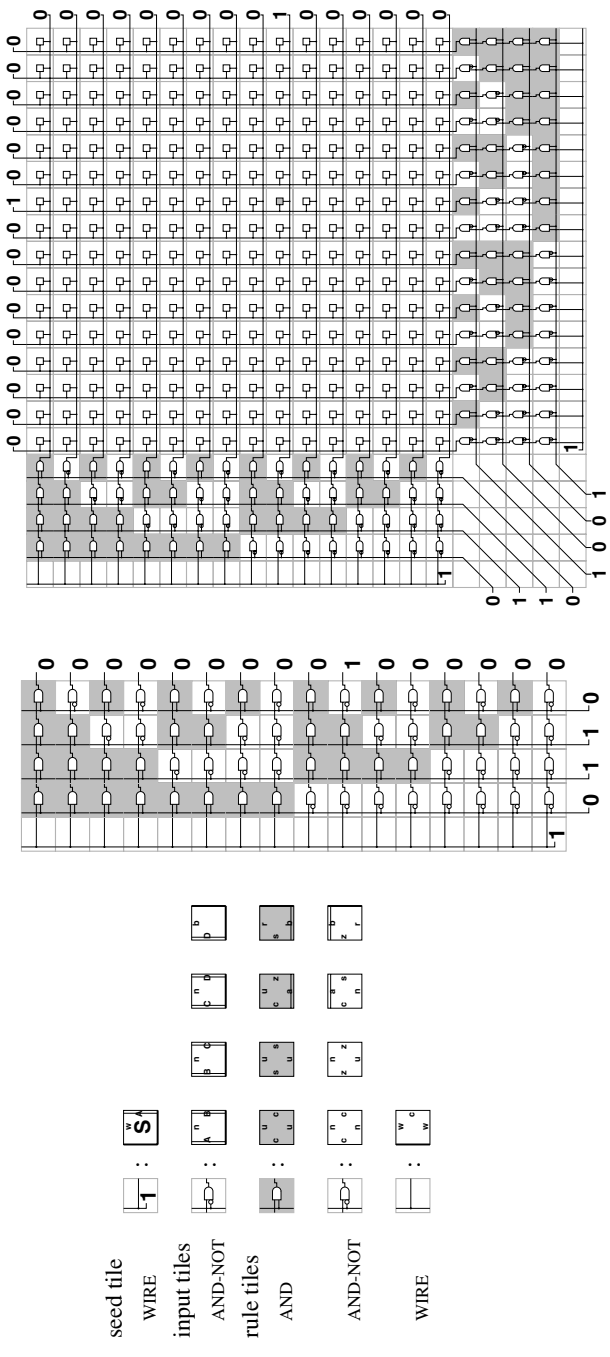

FIGURE 5 Using self-assembly of DNA tiles to create a molecular-scale pattern for a RAM memory with demultiplexed addressing. The tile set is closely related to the binary counter. Source: Adapted with permission from Cook et al., in press.

designed to form two-dimensional sheets appear to prefer tubes, for better or worse. Furthermore, procedures have yet to be worked out for reliably growing large (greater than 10 micron) crystals and depositing them nondestructively on the substrate of choice.

Although one- and two-dimensional algorithmic self-assembly has been demonstrated, per-step error rates between 1 and 10 percent preclude the execution of complex algorithms. Recent theoretical work has suggested the possibility of error-correcting tile sets for self-assembly, which, if demonstrated experimentally, would significantly increase the feasibility of interesting applications. A second prevalent source of algorithmic errors is undesired nucleation (analogous to programs starting by themselves with random input). Thus controlling nucleation, through careful exploitation of supersaturation and tile design, is another active topic of research. Learning how to obtain robustness to other natural sources of variation—lattice defects, ill-formed tiles, poorly matched sticky-end strengths, changes of tile concentrations, temperature, and buffers—will also be necessary.

Presuming that algorithmic self-assembly of DNA can be made more reliable, it then becomes important that we understand the logical structure of self-assembly programs and how that structure relates to and differs from existing models of computation. At the coarse scale of what can be computed—at all—by self-assembly of DNA tiles, there is a natural parallel to the Chomsky hierarchy of formal language theory. Recent theoretical work by Adleman, Goel, Reif, and others, has focused on two issues of efficiency: (1) the kinds of shapes and patterns that can be assembled using a small number of tiles; and/or (2) the kinds of shapes and patterns that can be assembled with rapid assembly kinetics.

To what extent has this investigation enlightened us about how information and algorithms can be encoded in biochemical systems? First, it is intrinsically interesting that self-assembly can support general-purpose computation, although it looks very different from conventional electronic computational circuits. At first glance, other biochemical systems, such as *in vivo* genetic regulatory circuits, appear to have a structure more similar to conventional electronic circuits. But we should be prepared for differences that dramatically alter how the system can be efficiently programmed. Ever-present randomness, pervasive feedback, and a tendency toward energy minimization are unfamiliar factors for computer scientists to consider. Nevertheless, functional computation can be hidden in many places!

Thus, DNA self-assembly can be seen as one step in the quest to harness biochemistry in the same way we have harnessed the electron. Electronic computers are good at (and pervasive at) embedded control of macroscopic and microscopic electromechanical systems. We don't yet have embedded control for chemical and nanoscale systems. Programmable, algorithmic biochemical systems may be our best bet.

REFERENCES

Adleman, L.M. 1994. Molecular computation of solutions to combinatorial problems. Science 266(5187): 1021–1024.

Cook, M., P.W.K. Rothemund, and E. Winfree. 2004. Self-assembled circuit patterns. Pp. 91–107 in DNA Computing: 9th International Workshop on DNA Based Computers. New York: Springer-Verlag.

Grunbaum, B., and G.C. Shephard. 1986. Tilings and Patterns. New York: Freeman.

Lagoudakis, M.G., and T.H. LaBean. 2000. 2D DNA Self-Assembly for Satisfiability. Pp. 141–154 in DNA Based Computers V, E. Winfree and D.K. Gifford, eds. Providence, R.I.: American Mathematical Society.

Mao, C., T.H. LaBean, J.H. Reif, and N.C. Seeman. 2000. Logical computation using algorithmic self-assembly of DNA triple-crossover molecules. Nature 407(6803): 493–496.

Reif, J. 1997. Local Parallel Biomolecular Computing. Pp. 217–254 in DNA Based Computers III, H. Rubin and D.H. Wood, eds. Providence, R.I.: American Mathematical Society.

Robinson, B.H., and N.C. Seeman. 1987. The design of a biochip: a self-assembling molecular-scale memory device. Protein Engineering 1(4): 295–300.

Seeman, N.C. 2003. Biochemistry and structural DNA nanotechnology: an evolving symbiotic relationship. Biochemistry 42(24): 7259–7269.

Seeman, N.C. 1982. Nucleic-acid junctions and lattices. Journal of Theoretical Biology 99: 237–247.

Wang, H. 1963. Dominoes and the AEA Case of the Decision Problem. Pp. 23–55 in Mathematical Theory of Automata, J. Fox, ed. Brooklyn, N.Y.: Polytechnic Press.

Winfree, E. 2000. Algorithmic/self-assembly of DNA: theoretical motivation and 2D assembly experiments. Journal of Biomolecular Structure and Dynamics 11: 263–270.

Winfree, E. 1998. Simulations of Computing by Self-Assembly. Caltech Computer Science Technical Report 1998.22. Pasadena: California Institute of Technology.

Winfree, E., F. Liu, L.A. Wenzler, and N.C. Seeman. 1998a. Design and self-assembly of two-dimensional DNA crystals. Nature 394(6693): 539–544.

Winfree, E., X. Yang, and N.C. Seeman. 1998b. Universal Computation via Self-Assembly of DNA: Some Theory and Experiments. Pp. 191–214 in DNA Based Computers II, L.F. Landweber and E.B. Baum, eds. Providence, R.I.: American Mathematical Society.

Natural Computation as a Principle of Biological Design

WILLEM P. C. STEMMER
Avidia Research Institute
Redwood City, California

The evolutionary process by which proteins and genetic sequences are designed in nature involves recursive cycles of diversification (by recombination and/or mutagenesis), followed by differential amplification based on the fitness of each sequence in a complex environment. In this process, each individual sequence "computes" its own fitness through repeated and diverse interactions with its local environment, similar to the way a steel ball computes its own path in a pinball or pachinko machine. The summation of all of the positive and negative effects of this large number of molecular interactions into a single fitness parameter is a form of molecular computation that occurs in nature, which we therefore call natural computation.

Natural computation is a form of molecular computation in which the outcome is a sequence or a structure rather than a numerical solution. Natural computation provides a real-time, real-environment calculation of the fitness parameter in biology and is an important theoretical and practical aspect of biological evolution. Because the output of one interaction is the input for the next and because of the complexity of the natural environment, natural molecular computations are virtually impossible to simulate by standard computation.

Natural computation occurs at a variety of scales, depending on whether the selectable unit is a (selfish) gene, a whole organism, or a community of organisms. The basic process is the same at all scales, involving the self-computation of each selectable unit's fitness, and thereby the composition of the entire population, based on diverse and repeated interactions with the local environment. Improved understanding of the natural design processes has greatly helped us develop better methods of designing proteins and whole genomes, as well as better methods of numerical molecular computation.

Existing directed-evolution processes are intended to improve biological sequences (genes, genomes) rapidly for complex but specific performance characteristics in environments that closely resemble the conditions of final application. All of the host genes other than the protein(s) or gene(s) being evolved are kept constant so that all of the computational power of the population (typically a molecular library created in vitro) is focused on optimizing the target protein for specific performance criteria and in a specific environment. This kind of laboratory-directed molecular computation, called screening, is used to measure the sum of the performance alterations obtained by multiple mechanisms, including DNA transcription; mRNA folding; translation and degradation; and protein expression, folding, interaction, spacial distribution, specific activity, and degradation.

Although molecular computation is widely regarded as futuristic, the natural computation methods used in directed evolution of biological molecules have proven to be practical and have generated strong commercial support. Laboratory-directed derivatives of the natural design process have been used to create a variety of valuable commercial products and product candidates in the areas of pharmaceuticals and vaccines (Maxygen, Inc.), chemicals, chemical processes and fermentation strains (Codexis, Inc.), and agriculture (Verdia, Inc.). These high-value examples are important for the widespread recognition of molecular computation as a powerful and commercially proven design process.

Natural computing and evolutionary optimization can, in principle, be extended to the design of a wide variety of nonbiological objects. Developing better methods of universal (i.e., numerical) molecular computation is the goal of most scientists in the field. However, nonnumerical applications of molecular computation may, in aggregate, prove to be just as useful and can be realized more immediately. Until there are practical formats for numerical molecular computation, these nonnumerical formats will serve as examples of the power promised by molecular computation.

Challenges and Opportunities in Programming Living Cells

RON WEISS
Department of Electrical Engineering
Princeton University
Princeton, New Jersey

With recent advances in our understanding of cellular processes and DNA synthesis methods, we can now regard cells as *programmable matter*. Cells naturally process internal and environmental information in complex fashions and interact with neighboring cells to achieve coordinated behavior. Through genetic engineering, we can now equip cells with new sophisticated capabilities for gene regulation, information processing, and communication. These new capabilities serve as catalysts for *synthetic biology*, an emerging engineering discipline to program cell behaviors as easily as we program computers.

Synthetic biology will benefit a wide variety of existing fields and enable us to harness cells for applications that are not feasible today. Applications include tissue engineering, molecular fabrication of biomaterials and nanostructures, synthesis of pharmaceutical products, biosensing, and will surely lead to quantitative insights into the operating principles that govern living organisms. Research so far has focused on building an enabling infrastructure for synthetic biology applications. A particular emphasis has been on constructing prototype synthetic gene networks that perform digital computation, analog computation, signal processing, and communications. In this paper, I will describe the building blocks for these genetic circuits, several intracellular and intercellular prototype genetic circuits that have been implemented recently, some of the challenges in designing such circuits, and the long-term significance of this work.

SYNTHETIC GENE NETWORKS

Genetic circuits are collections of basic elements that interact to produce a particular behavior. These elements include: DNA regions where RNA poly-

merase molecules bind and initiate *transcription* of DNA into messenger RNA (mRNA); mRNA sequences where ribosomal RNA molecules bind and initiate *translation* of mRNA into proteins; proteins that regulate the activity and production of other proteins; DNA regions that terminate transcription; and motifs that determine protein and mRNA stability. Each element serves a particular function that helps accomplish the overall behavior of the genetic circuit.

The basic elements can be grouped into units (or "devices") that perform logic operations. For example, Figure 1 shows a logic unit that implements the digital NOT operation using the biochemical reactions of transcription, translation, and protein decay. Based on this mechanism, we have constructed and tested synthetic gene networks with units that implement the NOT, AND, and IMPLIES logic functions (Weiss, 2001; Weiss and Knight, 2000). In addition, we also proposed and modeled a biochemical NAND gate that consists of two NOT gates whose outputs are connected with a wire-OR (Weiss et al., 1999). Theoretically, any arbitrary digital logic function can be implemented with genetic circuits using these gates.

By constructing biochemical logic circuits and embedding them in cells, one can extend or modify the behavior of cells. Consider how computer designers program the behavior of computers or robots by fabricating silicon-based logic circuits. To program cells in an analogous way, the genetic-circuit designer constructs a DNA sequence that encodes a particular circuit and then embeds this DNA molecule in cells (a process called transformation). Typically, the purpose of this network is to regulate precisely the production of proteins. Because proteins essentially perform all the "work" and information processing in the cell, cell behavior can be programmed by controlling when and under what conditions proteins are produced and degraded in the cell.

To date, several small synthetic gene networks have been built that accomplish specific genetic regulatory functions *in vivo*: the autorepressor (Figure 2a), in which a repressor regulates its own production to reduce noise in gene expression (Becskei and Serrano, 2000); the toggle-switch (Figure 2b), in which two repressors inhibit each other's production to achieve a bistable system (Gardner et al., 2000); the repressilator (Figure 2c), in which three repressors are connected in a ring topology to produce repeated oscillation (Elowitz and Leibler, 2000); the genetic clock and toggle switch constructed from transcriptional repressors and activators (Atkinson et al., 2003); and our synthetic gene networks used for engineering digital logic gates and circuits in cells (Weiss and Basu, 2002; Weiss et al., 1999).

Despite their logically simple functions, the difficulties in building these networks revealed that many challenges will be faced in building circuits of increasing complexity. For example, designers of genetic circuits will have to cope with the significant noise inherent in gene expression (Becskei and Serrano, 2000; Elowitz et al., 2002; McAdams and Arkin, 1997) and will have to match carefully the "impedances" of constituent devices to achieve compound circuits

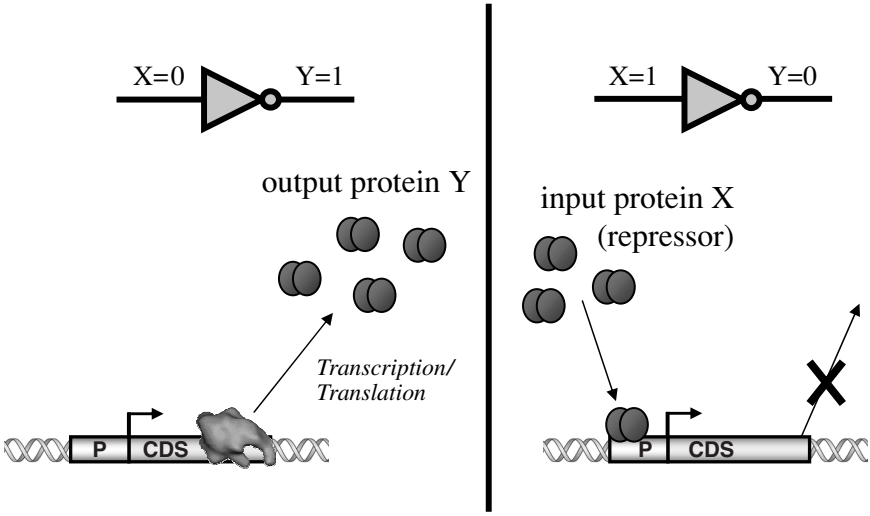

FIGURE 1 A simplified view of the two cases for a biochemical inverter, a logic device whose output value is always the inverse of its input value. The concentration of two particular proteins represent the input and output logic signals. In the first case (left), input protein X is absent, and the cell transcribes the coding sequence for output protein Y (labeled CDS). In the second case (right), input protein X is present and binds specifically to the gene at the promoter site (labeled P), preventing the cell from expressing output protein Y.

that operate properly and reliably (Weiss and Basu, 2002). Because these circuits operate in the context of a living organism that already has an existing "program," understanding the interactions of the exogenous circuit with the endogenous elements will be critical. When designing these circuits, such interactions will have to be minimized to avoid undesired cross talk and interference with normal cellular processes. The exception to this design rule applies when control over endogenous cell behavior is intentional and desired (e.g., controlling cellular metabolism or growth).

COMMUNICATION AND SIGNAL PROCESSING

Intercellular communication allows individual cells to coordinate their behavior and accomplish sophisticated tasks they simply cannot perform alone. In higher level organisms, such as mammals, eukaryotic cells send and receive signals to perform a wide range of activities, from differentiation and growth during embryogenesis to immune and stress responses during adult life. However, cell-to-cell communication is not exclusive to higher level organisms. Bacterial cells, for example, are known to regulate gene expression based on their

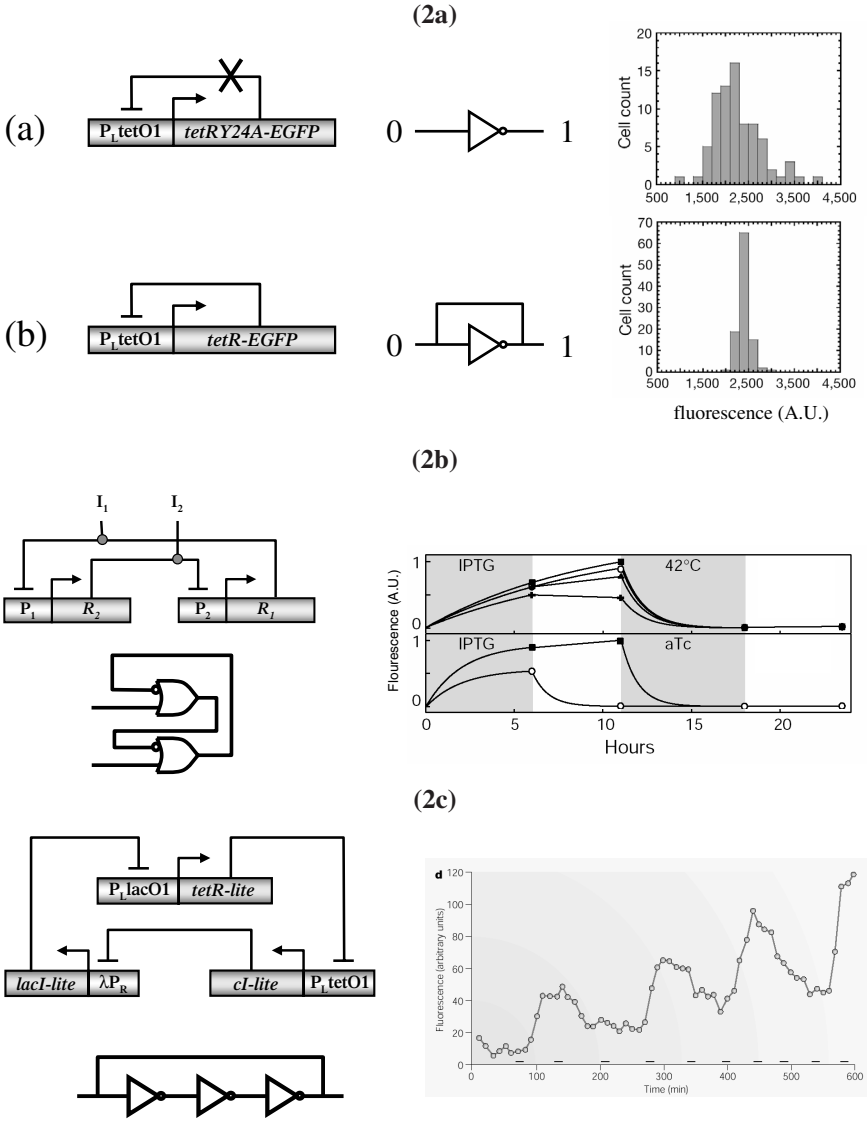

FIGURE 2 An overview of three synthetic gene networks, their corresponding logic circuits, and experimental behavior. These networks have been implemented to demonstrate specific genetic regulatory tasks. **2a**. The auto-repressor circuit reduces variations in gene expression. Source: Becskei and Serrano, 2000. Reprinted with permission. **2b**. The toggle switch circuit is a bistable system. Source: Gardner et al., 2000. Reprinted with permission. **2c**. The repressilator circuit exhibits periodic oscillation in gene expression. Source: Elowitz and Leibler, 2000. Reprinted with permission.

own cell density by secreting and then detecting concentrations of unique biochemical signals in a behavior known as *quorum sensing* (Bassler, 1999). This coordinated behavior gives bacterial cells a competitive advantage both as pathogens and in symbiotic relationships with their hosts.

To explore potential applications that would benefit from coordinated, multicellular behavior, we have begun to integrate communication capabilities with various synthetic genetic regulatory and information-processing networks. Toward this end, we first programmed *Escherichia coli* cells to communicate with each other by connecting several transcriptional regulatory elements with previously unrelated signaling elements from the marine bacterium *Vibrio fischeri* (Weiss and Knight, 2000). In the experimental system, we externally induced "sender" cells to synthesize the production of a small molecule ($30C_6HSL$) that then diffused outside the cell membrane and entered the cytoplasm of nearby "receiver" cells. The receiver cells responded to this chemical message by expressing a green fluorescent protein that was visible under a microscope (Figure 3). We are now extending this work to achieve programmed, two-way communications by constructing new circuits that use multiple signal synthesis and response elements from various bacterial sources. Two important challenges in engineering sophisticated and robust, multisignal, cell-to-cell communication networks will be matching response sensitivities and reducing cross talk between the signals.

Cells naturally analyze cell-to-cell communication and various environmental conditions with elaborate signal-processing circuitry that includes both digital and analog components. The ability of cells to detect and subsequently react to environmental and internal signals is a principal characteristic of many biological phenomena. Examples include the movement of bacteria towards higher concentrations of nutrients through chemotaxis, the release of fuel molecules in response to hormones that signal hunger, and cell differentiation during embryogenesis based on chemical gradients. To extend a cell's ability to respond to

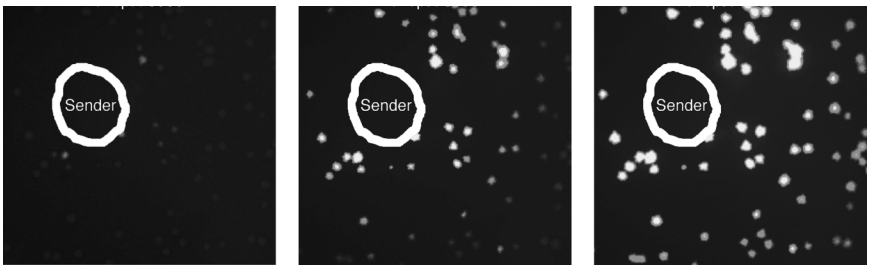

FIGURE 3 Time-series fluorescence images illustrating the response of colonies of engineered receiver cells to communication from nearby sender cells on a petri dish.

internal and environmental stimuli beyond the digital realm, we began to engineer analog signal-processing capabilities using synthetic gene networks. In our laboratory, we have built genetic circuits that enable cells to detect various chemical concentrations (below and above prespecified thresholds or only within certain ranges) and other circuits that allow cells to respond to multiple environmental signals (Basu et al., 2003; Weiss et al., 2003).

As a prototype for exploring issues in engineering signal-processing capabilities, we designed a new genetic *chemical-source pinpointing* circuit. The circuit is able to detect the presence of a particular extracellular molecule and then distinguish between various chemical concentrations of the molecule (Basu et al., 2003). Consider a toxin analyte whose location or even presence in the environment is unknown. The analyte is secreted from a particular pathogen and forms a chemical gradient centered on the source. A potential method of determining the location of the pathogen is to spread engineered sentinel cells in the suspected environment that can detect prespecified chemical concentrations. For a given concentration range, these cells will fluoresce in a ring pattern around the source. If the cells are engineered to detect multiple ranges distinguishable by different fluorescent colors, a bullseye pattern will emerge with several concentric rings around the analyte source (Figure 4).

The source-pinpointing circuit consists of several components that first detect the external analyte concentration and then determine whether the concentration falls above a prespecified low threshold and below a high threshold. Figure 5 shows two separate experiments of "sentinel" bacterial cells that have been programmed to respond to either low or high concentration thresholds of a biochemical secreted from nearby sender cells (Basu et al., 2003). By combining and inverting the low and high threshold outputs, one can realize a circuit that responds with a high fluorescent output only when the analyte concentration is within the prespecified range. With the aid of simulation tools, we have also been able to fine tune the threshold responses of these circuits by modifying the DNA sequences of chosen genetic elements. An important long-term goal for this type of research is to be able to tune the responses of genetic circuits with the same predictability and reliability as we can when we design electrical devices. By combining this type of analog information processing with digital computation and programmed cell-to-cell communication, we may be able to create a flexible and powerful engineering discipline for programmed cell behaviors.

KEY CHALLENGES

The initial modeling and experimental efforts in this field have generated a great deal of excitement but have also revealed many challenges that must be faced. Some of the significant challenges are described below.

Programming complex behavior will require the assembly of a large, well characterized library of intracellular and intercellular components. The library

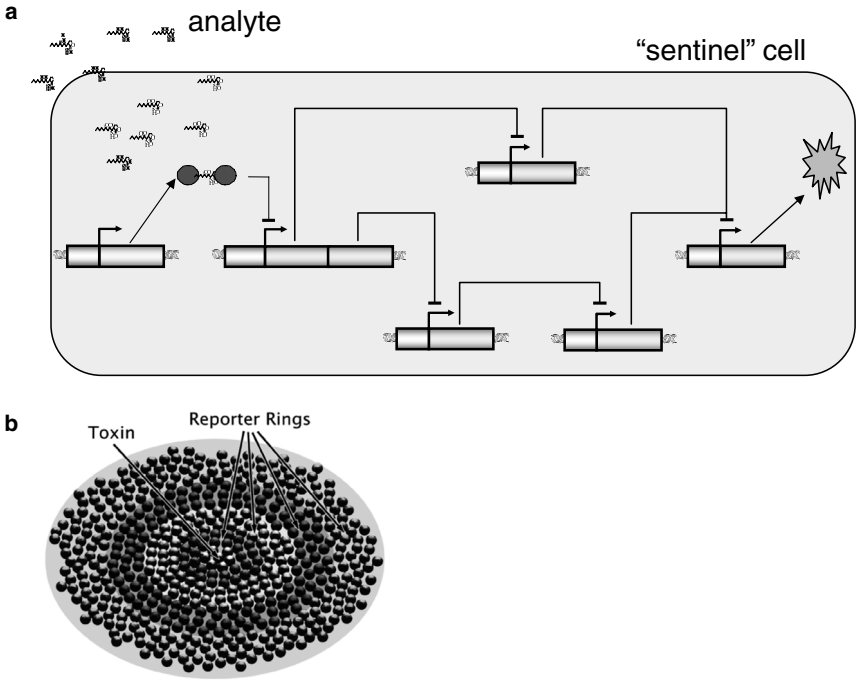

FIGURE 4 **4a.** Chemical-source pinpointing can be accomplished using cells with a genetic circuit that expresses a green fluorescent protein only when a given analyte concentration falls within a particular range. Because the chemical concentration is highest at the source and forms a gradient as the distance from the chemical source increases, cells with this circuit will form a fluorescent ring around the source. **4b.** Given a small library of circuits that can detect different chemical concentration ranges with unique fluorescent colors, cells will form separate rings centered on the source that result in a bullseye pattern.

should include elements for regulating transcription and translation, as well as elements for regulating protein-to-protein interactions, such as phosphorylation-based signal-transduction cascades. Most of the library elements must exhibit minimal cross talk and must not affect the host's behavior. However, an important subset of this component library should be devoted to interfacing with the host to control desired cellular functions, such as the production of specific enzymes during predefined conditions, control over cell replication, and programmed secretion of various chemicals.

Designing operational and efficient genetic circuits will require models and simulation tools that can provide accurate quantitative predictions of circuit behavior. Engineered genetic circuits operate within a highly complex environ-

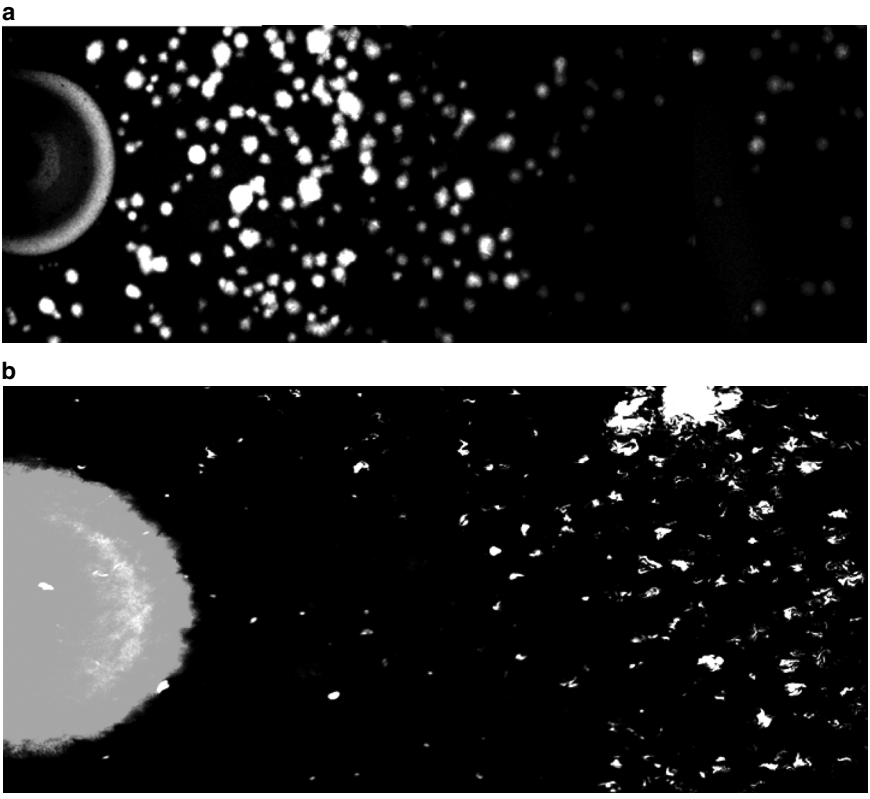

FIGURE 5 Fluorescence microscope images of "sentinel" bacterial cells programmed to detect whether the concentration of a chemical secreted by another type of bacteria is above or below particular thresholds. The bacteria that secrete the biochemical are fluorescing in red, while the sentinel bacteria scattered throughout the entire viewing area are fluorescing in yellow when the prespecified detection conditions are satisfied. **5a**. Concentration above a high threshold. **5b**. Concentration below a low threshold.

ment that is not well characterized and whose effects on the operation of the circuit have not been quantified. Models are still unable to predict the precise concentration averages and population statistics of the molecular components of even relatively simple systems. Solutions to this challenge will require more precise kinetic rates for describing the relevant biochemical reactions, models that incorporate additional cellular state (e.g., accurate RNA polymerase and ribosomal RNA levels), accurate predictions of noise in the biochemical reactions, a physical model of the cells and their surroundings, and potentially completely different modeling techniques. Furthermore, as the complexity of engineered circuits increases, they begin placing a significant metabolic burden on

the host. Molecular interactions between the exogenous components and the endogenous cellular circuitry can also affect host behavior. Effects on the host's environment must be understood and modeled to design complex synthetic circuits.

Genetic circuits must integrate specific components or network motifs that make them robust to fluctuations in the kinetics of biochemical reactions. Gene expression tends to be noisy because of the stochastic nature of the constituent biochemical reactions (Elowitz et al., 2002; McAdams and Arkin, 1997). In addition, fluctuations in environmental conditions, such as temperature and nutrient levels, affect cellular metabolism and consequently the operation of genetic circuits. Circuits that achieve reproducible, reliable behavior must do so despite components whose behavior fluctuates considerably. Mitigating the effects of gene expression noise will probably require a solution that incorporates positive and negative feedback loops.

A major bottleneck in genetic-circuit engineering is the difficulty of synthesizing DNA constructs. Anecdotal evidence suggests that the construction of new strands of DNA consumes a significant portion, if not most, of the time spent in circuit engineering. Currently, several ongoing efforts are trying to remove this bottleneck by using various *de novo* DNA-synthesis methods.

Genetic-circuit design will require novel approaches that may be fundamentally different from existing computer-circuit design methodologies. In constructing our circuits, we have used "rational" design in which detailed models are used to guide the circuit engineering, helping to select appropriate components and suggesting how to mutate them, when necessary, to achieve correct circuit behavior (Weiss and Basu, 2002). We have also used another technique called *directed evolution* to optimize circuit behavior (Yokobayashi et al., 2002). Building on nature's fundamental principle of evolution, this unique process directs cells to mutate their own DNA until they find gene network configurations that exhibit the desired system characteristics. Because our understanding of cellular processes is incomplete, efficient circuit design will most likely require a combination of rational design and directed evolution, or perhaps some other completely different approach.

LONG-TERM SIGNIFICANCE

The field of synthetic gene networks is still in its infancy. Researchers are currently trying to build small genetic-regulatory systems that exhibit a particular behavior reliably and predictably. Solving the existing challenges in building complex genetic circuits will take considerable effort. However, once the major challenges are solved, the construction of synthetic gene networks will enable us to direct cells to perform sophisticated *digital* and *analog* functions, both as individual entities and as part of larger cell communities. An engineering discipline to program cell behaviors and its associated tools will advance the capabili-

ties of genetic engineering and allow us to harness cells for a myriad of new applications. As a result, someday we will be able to modify and extend the behavior of cells in almost arbitrary ways. Because this technology will allow us to modify the instructions that govern the capabilities of organisms that live around us, as well as our own body's instructions, this technology has tremendous potential for improving control over the environment, our surroundings, and our quality of life.

REFERENCES

Atkinson, M.R., M.A. Savageau, J.T. Myers, and A.J. Ninfa. 2003. Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in *Escherichia coli.* Cell 113(5): 597–607.

Bassler, B.L. 1999. How bacteria talk to each other: regulation of gene expression by quorum sensing. Current Opinion in Microbiology 2(6): 582–587.

Basu, S., D. Karig, and R. Weiss. 2003. Engineered signal processing of cells: towards molecular concentration band detection. Natural Computing, an International Journal 2:4(463–478).

Becskei, A., and L. Serrano. 2000. Engineering stability in gene networks by autoregulation. Nature 405(6786): 590–593.

Elowitz, M.B., and S. Leibler. 2000. A synthetic oscillatory network of transcriptional regulators. Nature 403(6767): 335–338.

Elowitz, M.B., A.J. Levine, E.D. Siggia, and P. Swain. 2002. Stochastic gene expression in a single cell. Science 297(5584): 1183–1186.

Gardner, T.S., C.R. Cantor, and J.J. Collins. 2000. Construction of a genetic toggle switch in *Escherichia coli*. Nature 403(6767): 339–342.

McAdams, H.H., and A. Arkin. 1997. Stochastic mechanisms in gene expression. Proceedings of the National Academy of Sciences 94(3): 814–819.

Weiss, R. 2001. Cellular Computation and Communications Using Engineered Genetic Regulatory Networks. Ph.D. thesis, Massachusetts Institute of Technology, September 2001.

Weiss, R., and S. Basu. 2002. The device physics of cellular logic gates. Pp. 54–61 in NSC-1: The First Workshop of Non-Silicon Computing, Boston, Massachusetts, February 2002. Available online at: <*www.hpcaconf.org/hpca8*>.

Weiss, R., S. Basu, A. Kalmbach, S. Hooshangi, D. Karig, R. Mehreja, and I. Netravali. 2003. Genetic circuit building blocks for cellular computation, communications, and signal processing. Natural Computing, an International Journal 2(1): 47–84.

Weiss, R., G. E. Homsy, and T. F. Knight, Jr.. 1999. Pp. 275–295 in Dimacs Workshop on Evolution as Computation. Princeton, N.J.: Springer Verlag.

Weiss, R., and T.F. Knight, Jr. 2000. Engineered Communications for Microbial Robotics. Pp. 1–16 in DNA6: Sixth International Workshop on DNA-Based Computers, DNA2000, Leiden, The Netherlands, June 2000. New York: Springer-Verlag.

Yokobayashi, Y., R. Weiss, and F.H. Arnold. 2002. Directed evolution of a genetic circuit. Proceedings of the National Academy of Sciences 99(26): 16587–16591.

Dinner Speech

The Most Important Lessons You Didn't Learn in Engineering School

WILLIAM F. BALLHAUS, JR.
Aerospace Corporation
Los Angeles, California

Before I begin my comments, I want to show a brief video to set the stage. As leaders in your field, you are expected to make very tough decisions. I'd like you to watch this, and put yourself in the shoes of the characters.

Summary of "Challenger" Video

The clip opens several months prior to the Challenger disaster in the cafeteria of Morton Thiokol Inc., where Roger, an engineer with the company, discusses with his manager his concern about a potential problem with hardware aboard the rocket that carries the space shuttle into orbit. Previous launches have revealed that outside air temperatures below 53 degrees can damage the primary O rings, impacting the way they seal and resulting in massive hot gas blow-by. Roger has submitted a memo to his Thiokol management expressing his concern about the sealing problem, saying that failure of an O-ring seal could cause a "catastrophe of the highest order."

Months later, on the evening before the Challenger launch, the O-ring issue becomes the central topic of debate during a teleconference between the engineering team at Morton Thiokol and NASA teams at Marshall Space Flight Center and Kennedy Space Center. The forecast calls for temperatures well below 50 degrees during the launch window, and Morton Thiokol has recommended that the launch be postponed until temperatures rise above 53 degrees. Realizing that the launch could be postponed for months, a NASA manager questions the timing of the announcement and the data that support it. Roger makes a final plea, arguing that allowing Challenger to launch with what engineers have observed during previous flights would be "an act away from good-

ness." Although frustrated and angry, NASA management is willing to follow Thiokol's recommendations, but "We've got to make a decision here, and it's got to be based on quantifiable data," the NASA manager insists. Under pressure, the Morton Thiokol team goes off line to discuss a final recommendation.

Off line, Roger stands by his opinion that it is better to play it safe and not recommend that the launch proceed. Thiokol managers in the room are more concerned about embarrassment to the company because of the manner in which the issue was raised to NASA. Ultimately, the decision is left to Roger's manager, who reconnects to the NASA teleconference and recommends that NASA go through with the Challenger launch.

This video gives us some specific lessons on the Challenger accident. It illustrates the importance of *people*, *technical* and *management judgments*, *validated processes*, *personal* and *organizational accountability*, *effective communication*, and a *focus on quality* and *mission success*.

I personally learned a great deal from the Challenger accident. Although I was not in the shuttle program chain of command, at the time of the accident I was director of NASA's Ames Research Center; during the return-to-flight effort, I was at NASA headquarters as associate administrator for aeronautics and space technology. Most of you are 30–45 years old. The Challenger accident occurred on my forty-first birthday.

I would like to broaden the discussion and share with you the "Most Important Lessons You Didn't Learn in Engineering School," lessons I have learned during my 35 years in engineering and engineering management. Some of the lessons were easy to learn; others were brutally painful. Some of you may have your own lists. As current and future leaders, I hope you learn from the mistakes of others and apply the lessons learned in your own careers.

1. Learn from the mistakes of others, and never make the same mistake twice.

- Learn from mistakes, your own and those made by others. Don't repeat either. When there is a failure, anomaly, or out-of-family condition, it's important to follow a process. First, find the root cause. Second, develop and implement corrective action. Third, share the lessons learned. Fourth, revise processes to prevent repeating the mistake.
- An organization's processes and practices are usually the result of lessons learned from corrective actions developed in response to past anomalies and failures. Processes are designed to yield predictable, repeatable results and eliminate the reliance on heroic actions to avoid failure.
- General Electric Corporation, which recognizes the importance of learning from mistakes, instituted the practice of putting high-potential

people on audit teams early in their careers so they could learn from mistakes made by business unit managers. In that way, GE hoped these high-potential people would avoid making the same mistakes when they became general managers.

• I have found that organizations often are not as good at *inventing* new mistakes as they are at *repeating* the old ones.

2. Focus on mission success.

• What is mission success? Quite simply, it is making sure that your product—whatever it is—works as intended. A focus on mission success is especially critical in high-consequence businesses (e.g., space, nuclear, homeland security, etc.). These are "one strike and you're out" businesses in which "failure is not an option."

• In any project, you must manage cost, schedule, performance, and risk. If the first three are fixed, as they were for the NASA Mars missions that failed in the 1990s, then the only relief valve available to a program manager is risk. In such a case, the probability of failure increases.

• Do you understand what the risk tolerance is in your business? How many of you have risk-management plans that specifically address your area of responsibility? In program management, you need a risk-management plan that considers the probability, consequences, and difficulty of managing each of the risks you have identified. You should monitor progress in retiring risk as the program progresses.

• Experience in the space business shows that if quality and schedule are managed well, good cost performance results. If you manage only to cost, experience shows numerous examples of program execution failures.

• Trade-offs on schedule, cost, and technical risk must be made at a high level in the management chain. In my business, the failure of a two-cent part can cause the loss of a billion-dollar mission. You don't want lower management levels controlling cost and schedule by increasing risk.

• Remember that, in spite of our best efforts, humans will make mistakes. We design our systems to be robust and fault-tolerant and attempt to eliminate single-point failures. We should also make sure that processes are human-error tolerant.

• Develop a motto of "no surprises" by staying on top of details through an effective review process. Stress early problem identification and the timely, aggressive application of remedial measures. Create an environment that is conducive to the reporting of problems. Sam Goldwyn, the late move mogul of MGM, said, "I don't want 'yes' men around me. I want people to tell me the truth, even if it costs them their job." We want people with moral courage, but we don't want the type of environment that shoots the messenger.

3. Put the right people in the right positions and hold them accountable for accomplishing agreed-upon goals.

- People are the key ingredients in any successful organization. Make sure they have the right skills, motivation, and experience, and then give them the authority and accountability to get the job done.
- I cannot overemphasize the importance of individual and organizational accountability. The reason for mission failures are very often that the people we relied on failed to do properly what we were relying on them to do. Make sure accountability flows down to those who are doing the work.
- With human beings, you generally get the behavior that you incentivize, so pay attention to the incentives you put in place. People need to be tied directly to agreed-upon goals that will lead to success; otherwise you get the wrong behavior, and that can have unintended consequences.

4. Communicate effectively.

- An outstanding technical perspective is worthless without effective communication—whether you are communicating research results, managing a project, or trying to influence NASA management the night before a launch.
- I have heard papers presented at conferences where you know the person has spent years doing excellent work, but only 10 minutes preparing the presentation of the work. They are often surprised that others don't immediately see the value and cleverness of their accomplishment.
- Also, as we learned from the video, an organization must be sure that lines of communication are in place that allow important information to get to decision makers. Effective communication doesn't just happen in the lunch room by chance.
- You must communicate in a way that your audience can receive the message. We live in a sound-bite world, so you must be crisp and to the point.
- Engineers are developing systems solutions to meet the needs of modern society. Hence, they must be able to communicate with a wide variety of stakeholders with diverse interests (e.g., economic, sociological, environmental, medical, political, military, etc.). You must know your audience—and their interests and backgrounds.
- It's interesting to note that very few members of Congress have technical degrees. You may have heard about a congressman who asked a former NASA administrator, "Why do we need meteorological satellites when we have the Weather Channel?" Where did he think the weather pictures and data came from?

5. Optimism is admirable but "hope is not a strategy."

- When faced with seemingly unmanageable challenges, hoping those challenges will vanish is not a viable strategy. Always face reality, develop effective strategies and plans, and manage your way through the crisis.

• After the Berlin Wall came down in 1989, Department of Defense procurement budgets also came down, by 70 percent over several years. Hoping that defense budgets would go back up was not a viable strategy. Defense contractors struggled to develop strategies that would enable them to continue to grow, or at least survive. At that time, the CEO of Lockheed approached the CEO of Martin Marietta. Not surprisingly, both had already considered a potential merger as part of their internal strategic planning processes. They recognized an opportunity to expand their business bases, eliminate excess capacity, and increase shareholder value. Of the many major defense contractors at the beginning of the 1990s, only a handful managed to survive and prosper.

• Timing is everything. I joined the defense industry five months before the Berlin Wall came down and ended up going through four major mergers and acquisitions. I learned a great deal from leading transition teams and participating in the development of today's Lockheed Martin Corporation.

• The bottom line is that you must think creatively, be ready for changes, and work harder than the other guy. I'm reminded of something James Lofton, the great NFL receiver, said when he was inducted into the Pro Football Hall of Fame. First let me tell you that he has a dry sense of humor. When he was asked to reveal the "secrets" of his success, here's how he answered. "One trick is to work harder than the other guy. The second trick, always hustle. Third trick, study and know what you're doing. Fourth trick, always be prepared. Fifth, never give up. *Those are my tricks*."

• Incidentally, Lofton has a degree in industrial engineering from Stanford University.

6. Be accountable for delivering continuously increasing value to your customers.

• Customers are those whom you are in business to serve, however you earn your living. I once mentioned to a friend of mine, Ed Crawley, who was head of the Aeronautics and Astronautics Department at MIT, that I had eaten dinner earlier in the week with his "boss," the MIT president. Ed told me I didn't understand. He was a tenured MIT professor, he said, and so didn't have a boss. Ed may not have a boss, but I know that Ed knows who his customers are (his students and research-funding sources), and he works very hard to bring value to them.

• Identify your customers' needs, and assume accountability for meeting those needs. Provide your customer with single-point accountability whenever possible.

• If you want to test the value of a person or organization, ask what they are accountable for. If they cannot state the answer crisply, if there is confusion, the next obvious question is, "if you aren't clear about what you are accountable for, then what value do you add?" This was on my mind constantly during the massive downsizing in the aerospace business during the 1990s when we really

had to focus on the value-to-cost ratio in determining which organizations and people to retain.

• Personally, I've always tried to identify my customers. Earlier in my career—when I was doing "useful" (i.e., nonmanagement) work—I developed computational fluid dynamic algorithms and codes for use in airplane design. My primary customers were aircraft designers who were using wind tunnels and linear analyses but needed a more cost-effective way to get better designs. My other "customers" were the people at NASA headquarters who sponsored my research.

• Incidentally, years later—when I was briefed into the program—I found out that the tools we developed were used extensively in the design of the B-2 Stealth Bomber. You can also see what we did on a program called "HiMAT" in an interview on display at an exhibit at the National Air and Space Museum in Washington, D.C.

7. Make decisions based on "integrity to the business."

• High standards of personal integrity and ethical behavior are fundamental. If you ever get caught in a lie, you are finished because it breaks the bonds of trust. I won't do business with you again if you lie to me. But I am talking about something different here. Follow "principles-based decision-making" in business. Decisions should be made based on the value structure of the organization. As a leader, you must define your organization's value structure and get consensus around those values.

• Take, for example, the Tylenol tampering incident. In 1982, bottles of Tylenol that had been tampered with appeared on store shelves. Cyanide injected into the pills led to seven deaths. The CEO of Johnson & Johnson was faced with a dilemma.

• You would imagine that in such a situation your attorneys would advise you not to recall all of the product because the shareholders would sue because of the impact on shareholder value.

• Your marketing director would probably advise you to recall all of the product or risk permanent damage to the brand image.

• The resolution was based on J & J's core values—they are in the business of *healing*. The CEO decided to recall the product, taking full responsibility, and thereby regaining public trust in the company and product. Any other approach would have violated the integrity of the business.

8. Set a tone of speed and agility.

• Time equals money. For a commercial business, the price point and time to market are key factors. Everything you do takes time away from other things you could be doing. So place a high value on your time, parallel process when

you can to make the most effective use of your time, and don't let problems fester. Take decisive action to fix things promptly. Your example will set the tone for the rest of the organization.

9. Volunteer to take on tough and important assignments that address your organization's top problems.

- When you are a manager fully accountable for meeting aggressive objectives, whom are you going to select for key positions in your organization? Most likely, you will select people who have demonstrated they can deliver by taking on important and challenging problems and succeeding. That is often how people in senior management got there.
- Problem selection is one of the most important qualities of a successful researcher. Each project should serve the stakeholder's needs and should be visible. Ask yourself what the benefits will be if you are 100 percent successful in solving a problem. Can the problem be solved in a reasonable amount of time? Is it time to move to new areas where you can add greater value?
- Professional societies provide a low-risk environment for developing and demonstrating leadership skills and potential. They also help you develop a network of contacts with peers and senior people outside your organization. Learn from observing successful leaders in business, government, and academia.

10. Expand your knowledge and become learning-agile.

- The more experience you have and the more you are able to learn quickly, the more effective you can be in solving tough problems. And your knowledge base must extend beyond engineering, because developing appropriate system solutions to society's problems often requires cross-disciplinary technical knowledge, business expertise, and political and social awareness. The higher you go, the broader the knowledge base required. Knowledge you gain today may not be relevant to your current job, but it can prepare you for more senior positions in the future.
- Consider lifelong learning as a professional obligation.

Summary of Lessons Learned

- So there you have it, 35 years of experience delivered to you in slightly more than 35 minutes. I hope you will think about these lessons and incorporate them into your working environment. Also, I encourage you to develop *your own lessons* as you progress through your careers.
- I conclude with a video that demonstrates how your engineering talents and enthusiasm for your job can benefit both you personally and society as a whole.

Summary of "School 2" Video

Demonstrative little Bobby proudly tells his class that someday he wants to be an engineer. When asked why by his teacher, Bobby excitedly paints a picture of the futuristic and humanistic high-tech world he will help create with "power plants, robots that can think, and leaning trains" that usher people here and there quickly and efficiently. Grabbing the class's imagination with his growing enthusiasm, Bobby reveals that he will create "advanced technologies that will reduce pollution and save the world!" The class erupts in cheers.

- When I first saw this video it struck me as capturing why many of us became engineers. Personally, I'm excited about going to work every day just like this young man.
- The video is actually a commercial by the ABB Group of Companies, but I think you'll agree it makes a very positive statement to the general public about the excitement of engineering and the contributions we make to society.
- Engineering has brought outstanding benefits to our society in the last century (as illustrated in the National Academy of Engineering (NAE) publication, *The Greatest Engineering Achievements of the 20th Century*).
- In your discussions tomorrow, I'd like you to ask yourself two questions. What can we do to improve the public's perception of engineering? What can we do to enhance the technical literacy of the general population?

APPENDIXES

Breakout Sessions

On the second afternoon of the symposium, breakout sessions were held to give participants an opportunity to talk in seven smaller groups about public policy-oriented engineering issues. Each group was asked to focus on one aspect of public understanding of engineering or engineering outreach to the general public. The discussions in each group are summarized below. Recommendations represent the opinions of the participants and not necessarily of the National Academy of Engineering (NAE).

Q1: Why should engineering matter to the public? Why should public awareness matter to engineers? Does the public have a negative view of engineering? If so, how can we improve it?

Group 1

The group first posed preliminary questions and made some observations. First, why should engineers care about public awareness? The group concurred that public awareness is necessary to high-quality engineering as well as to diversity in the profession. What "public" should we target? The public can be divided into three groups: policy makers, people who are not engineers, and people who may become engineers (e.g., students). The way we promote awareness may vary depending on which group we are trying to reach. For example, policy makers for the most part are not technically well versed. The question then arises whether engineers should play a role in public policy making. The nonengineering public is generally not educated about technology.

Although approximately 30 percent of Americans have completed at least one college-level course, the majority attended schools that did not have engineering programs. This makes reaching the public difficult because most people show little interest unless something affects them directly. It seems clear that public perceptions also are dependent on current events. For example, placing a man on the moon was a widely applauded feat that contributed to a positive image. The explosion of the shuttle Challenger was a huge blow to the standing of engineers in the public mind. Public expectations of engineers continue to be high, even when funding and public support diminishes.

Many potential engineers (students) are not introduced to engineering before they select a critical career path. Often the information students are given about engineering does not portray engineering as an exciting or important career option. Engineers have to become involved in educating people and not rely on the educational system to do the job. However, engineers have typically worked through professional societies and have not presented a unified front. Perhaps as a society we need to redefine an "educated person," to include an understanding of technology. But first, engineers themselves must agree on a definition of "engineer." We need to work together to formulate a convincing, unified message that can be communicated to the public.

1. Does the public have a negative view of engineering?

Public perceptions may reflect "guilt by association," that is, if the public considers a specific product as negative, then they may also consider the engineers who developed it negatively. These feelings are usually limited in duration and limited to a specific engineering discipline. People may also be influenced by what they read in the media (e.g., public outcry about biomedical engineering followed negative feedback by patients with breast implants; public praise of biomedical engineering followed the revelation that a celebrity had received hip implants). The consensus of the group was that, on the whole, the public is indifferent or ambivalent about engineering.

There is a certain level of ignorance as to what engineers actually do, and engineers are at fault for this. People who are aware of engineering view engineers as nerdy, smart, sometimes arrogant problem solvers who have good salaries and stable jobs. A person's perspective on engineering often depends on a specific teacher or exposure to engineering in K–12. Unlike other careers, such as teaching, sales, or performing, students do not generally have day-to-day exposure to engineering. In addition, they have limited, if any, exposure to engineering in school. Although everyone is surrounded by products that are the results of engineering, the connection is rarely spelled out.

2. Why do public views of engineering matter?

The public should be concerned because the future and availability of qual-

ity jobs is at stake. Quality engineering is necessary to maintaining a high standard of living and is crucial to national security. Engineering enhances the quality of health and life. Engineers can also provide impartial input in debates that require scientific reasoning.

3. Why should the public view of engineering matter to engineers?

Many of the same reasons apply. Unless the public understands and takes an interest in engineering, we run the risk of losing industries, as well as high-tech jobs. Engineers need public approval to ensure the availability of funds that lead to solutions to problems (i.e., if public perceptions are negative, research money could disappear). Engineers protect the standard of living and provide crucial advice on national security issues. A higher level of public awareness of engineering can lead to an increase in the interest of students in engineering as a profession and an increase in the number, diversity, and quality of individuals who select engineering as a profession. Public awareness also contributes to the attraction of talented students from abroad, some of whom return to their home countries as positive ambassadors of the United States.

The consensus of the group was that, except in cases of negative events specific to engineering, the public perception of engineering is indifferent. Therefore, the last question was modified.

4. What can we do about public perceptions that are indifferent (rather than "negative" as in the original question)?

In general, we need a more astute Congress in terms of engineering. The question then becomes whether engineers should be trained to participate in public policy or whether politicians should be educated about engineering and technology issues. Politicians could be educated through a technology course sponsored by NAE specifically for policy makers. The general public could be educated through a national media effort, such as a PBS special, that provides an overview of engineering and its benefits to, and roles in, society. Many engineering fields already provide science-based television programming, but a documentary that provides a general introduction to engineering, clearly defines the field, and targets audiences with different educational levels would be beneficial. Grass-roots efforts/outreach programs geared toward integrating engineering in the schools should be continued.

The definition of an "educated person" should be expanded to include technological expertise; this should be done in conjunction with K–12 governing bodies. It is important that engineering be introduced to children early in their careers; high school may be too late because opinions about particular subject areas have already been set. Engineering can be included in the curriculum by using methods as part of the regular curriculum so that there is no time burden on teachers (teachers are already constrained by standardized testing schedules and related testing criteria and do not have time to incorporate seemingly "extrane-

ous" information into their class schedules). A history of engineering, for example, could be incorporated into the high school curriculum as a module in an existing course or as a stand-alone course. In either case, the emphasis would be on engineering contributions that have influenced the development of society and would define engineering as a career option. All levels of engineering professions (from blue collar workers with technical certificates to research scientists with advanced degrees) should be promoted. But first, we (i.e., engineers) must define who we are, deliver our message to the public convincingly, and continue to be proactively and interactively involved.

Q2: How can engineering make the world safer? How can we make sure that technological capabilities are used to their fullest advantage?

Group 2A

The group first discussed what "safer" means. Does it mean safety from terrorism, natural disasters, disease, hunger, etc.? We can define safety specifically or broadly. Defined broadly, safety concerns are related to inequalities in the way basic human needs across the world are met. These inequalities make the world an unsafe place. Engineers need to be aware of political and ethical issues of inequality in the United States and abroad and become more involved in policy making to address these issues at the local and national levels.

The group raised some specific safety concerns:

Enabling Technology. The development and strategic placement of sensors for the rapid detection of microbiological and chemical agents will be critical. Environmental sensors might be placed in cell phones, for example. Regulations and privacy issues will have to be addressed. Engineers can be involved in the development of sensors, the identification of deployment technologies, and making policy decisions.

Critical Infrastructures

— Electrical power plants. Energy generation and distribution networks were not initially designed to be interconnected; they were connected after the fact. Engineers can help in redesigning and retrofitting them so they can operate more reliably.

— Water systems. Not only are present water systems aging, clean water will be scarce in the future. We will need new technologies for detecting water quality, locating water sources, and restoring/cleaning water. The issue will be urgent for small, remote villages. Engineers can help in the scaling of technologies for cleaning large water sources/supplies to apply to small systems.

— Air transport industry. Engineers can help determine what it will take to

be 100 percent sure within a reasonable time that a particular passenger plane is completely safe and free of explosives and other weapons. How far are we from this certainty? Are there privacy issues to be addressed?

— Public health. Health care systems need engineering input in developing protocols and processes, medical devices, and real-time measurements of physiological and biological parameters (sensors and communication technology).

— The Internet and software. Engineers can help ensure that defective software with known bugs is not released and can help establish regulations to ensure that this does not happen. Engineers can also make the public less tolerant of software with bugs.

Engineers are problem solvers and driven by (1) economics (e.g., maximizing profits for a company), and (2) geopolitical issues (e.g., political factors that lead to investment in specific areas of technological development). Some in the group noted that engineers in academe might be less affected by these factors than engineers in industry. Engineers can provide both engineering solutions and policy input. To this end, we need mechanisms to help us identify how the world could be made safer. The National Academy of Engineering, the National Academy of Sciences, and the Institute of Medicine could identify problems that the engineering community could address to make the world a safer place instead of waiting to be solicited by the government to advise on issues identified by the government. In addition, moving from the development of a technology to its implementation can often be impacted by politics and policy making. Engineers can and should make a difference in both.

Group 2B

The discussion by this group can be summed up in five main points:

- One way engineers can make the world safer is through modeling and simulation. The group discussed whether the focus should be on terrorism or accidents. The consensus was that antiterrorism measures are too costly to make every building safe, so it is important that the focus be on resources.
- Engineers should focus on making systems that have dual use (e.g., masks for fire and chemical agents and HVAC for biological attacks and allergens).
- Engineers should communicate better with the public about risks and provide more information to public officials and decision makers.
- Engineers should educate the public about safety, which would help create market demand for safe products, which would then drive engineering design.
- Risk analysis should be a standard part of engineering curricula.

Q3: What is the proper balance between the free flow of information and national security in a free and technologically sophisticated society?

Group 3A

The group discussed how we can continue to hire qualified foreign nationals for our national laboratories and still satisfy security concerns. Our scientific and technical biases and reward systems tend toward openness. Many agreed that it is important not to keep the "good guys" ignorant, because the "bad guys" already know what the vulnerabilities are. It is not easy to find the proper balance between openness and secrecy. After 9/11, for example, visas were granted to some dead terrorists and denied to some legitimate applicants.

Bureaucrats tend to be cautious because bureaucracies exact a high price for those who not follow the letter of the law. Is there a middle ground between giving out too much information and not enough? The group agreed that guidelines would be helpful. In hindsight, it would have been better to publish articles on how a plane could be hijacked. But should information about how to inject cyanide be published?

In general, the group thought we should follow a layered approach to security issues and acknowledge that success is not dependent on protecting information. Tactical issues, such as not exporting high-end technology, and strategic issues, such as how to attract the best and brightest and maintain technological superiority, also affect security. Personal judgment is an important component in the layered approach.

One criterion for deciding whether or not to release information is "ease of use" for positive or negative ends. Should we have clear guidelines or should we rely on personal judgment or review boards? Engineers might elect a panel to review and issue guidelines to clarify what should and should not be published, but this may be possible only in select industries.

Group 3B

The group discussed many types of information controlled in the United States, ranging from information classified or otherwise restricted by the U.S. government, to information that is proprietary and copyrighted by private individuals or corporations. The bulk of the discussion, however, centered on information that is controlled by the U.S. government. Examples from World War II and current events, such as 9/11, were examined in the context of mechanisms, such as classification or "ITAR-restricted" labeling systems, used for information control. The general consensus was that the current mechanisms are not overly burdensome for engineers and scientists, and that they do provide some degree of protection to citizens from foreign attack. The group agreed that it is necessary to control information concerning methods and sources for

information-gathering, as well as information concerning military operations or planning. However, censorship of basic research was generally felt to be unnecessary and possibly even harmful to U.S. interests, because prohibiting scientists from communicating could have the unintended consequence of inhibiting technological progress.

Q4: Who becomes an engineer? How can we make our profession more representative of the country as a whole? What are the societal benefits of a diverse engineering workforce?

Group 4A

The group began by defining "engineer." According to the definition on page 6 of *Raising Public Awareness of Engineering,* "Engineers do applied research aimed at improving the quality of life." Many reasons were given for why the people in the group had become engineers: an early love for computers; a father who was an engineer or in a profession where he worked with engineers; a desire to learn problem-solving methods; a desire to be in a field that did not require knowledge of a foreign language; a desire to participate in the space race and a passion for space exploration; the default profession because the individual did not want to become a doctor (which parents wanted); prestige of the profession; appreciation for what engineers do; a desire for job security; a fascination with video games and a desire to learn how to design them; discovery of a talent for proving theorems.

The group also addressed what factors excited people's interest in math/science: K–12 science fairs; exposure to "cool" role models; early exposure via a parent, friend, other role model; the prestige factor. The group concluded that "cool" movies with engineers as the heroes (i.e., Hollywood glamorizing engineers) would help stimulate interest in math/science.

There are many reasons people are sometimes turned off by engineering: the perception that it is not creative; a fear of difficult math; unwillingness to "hang in there" and become comfortable with math; a lack of exposure to what engineers do (there is no formal system in the United States for exposing people to engineering); a lack of publicity about engineering; the perception that engineering is not fashionable or newsworthy; a strong focus today on biology; a short-term perspective on what is fun and little understanding that a long-term career in engineering can be a lot of fun; too much work required; the perception of engineers as "geeks;" the perceived inaccessibility of engineering accomplishments.

The group agreed that several things could be done to remedy the situation: universities and K–12 could be linked; a requirement for outreach could be built into National Science Foundation proposals; Hollywood could make a movie/show highlighting positive aspects of engineering.

Why should we care about a declining number of engineers? Services are being outsourced because of a shortage of trained domestic workers. The national interest won't be served. Lack of understanding of basic math/science will eventually come back to haunt us. The decline in engineers will eventually affect research funding and hence the advances that result from research.

What are the benefits of a diverse engineering workforce? If different perspectives are brought to bear on a problem, the day-to-day dynamic changes. People from different backgrounds can lead to changes in patterns of interacting/thinking. People from different cultures/backgrounds can make everyone more aware of how an engineering product will fit into different cultures. Even in pure science, the questions one asks are influenced by one's background, and there is a presumption that a diversity of questions is inherently beneficial. Education should play a role in advancing diversity (diversity in thinking is not always coupled to diversity in gender/race), because a diverse workforce will eventually lower the barriers to entry within the profession; once there is a critical mass, people who might have been deterred may decide to enter the field. People who become accustomed to working with people from a variety of cultures become more inclusive, which is a virtue.

The following can be done to make the engineering workforce more representative of the population as a whole: initiate more K–12 outreach programs; require students to take more math/science; educate teachers in presenting science/math/engineering in a way that is fun and accessible; quadruple the salaries of K–12 teachers; encourage interaction by having role models come into K–12 schools; develop a "carrot/stick" approach to encourage people to interact in a sustained way, perhaps by increasing vacation days for each K–12 visit; arrange for a K–12 teacher to do a sabbatical at a university or in a company then return to his/her school to share the lessons from the experience.

Group 4B

The following motivating factors for becoming an engineer were mentioned:

- Role models (including middle school teachers who can keep girls interested in math and science), college professors, and role models in the media.
- Families.
- Personality traits (e.g., independence, enjoyment of tinkering).

Many obstacles to becoming an engineer were mentioned:

- Math is considered competitive at an early age, which scares people off. Social pressure against women excelling in math is still strong.
- Math and science are not taught well.

- "Engineering" is not part of the curriculum like science is.
- Many believe that monetary rewards are not commensurate with the work engineers do.
- Opportunities in engineering are not generally well-known.
- The retention rate of women and minority students is low.
- The European Union and Japan graduate many more engineers than the United States; something about U.S. culture discourages people from pursuing careers in engineering.
- Some engineering fields (e.g., materials and industrial engineering) are not well recognized.

Some steps that might help to overcome these obstacles:

- Broadcast targeted advertisements in the media.
- Initiate outreach programs to elementary and middle school teachers.
- Undertake a national initiative (e.g., response to Sputnik) to stimulate interest.
- Exert pressure by industry not to hire from institutions that do not have enough minorities on the faculty or in the student body.
- Implement institutional reforms to retain women and minority faculty.
- Change the engineering curriculum to appeal to a broader base of students.
- Adopt more innovative approaches to education (e.g., self-taught electrical engineering courses at UC-Berkeley).
- Adopt a project-based educational format.
- Encourage families to promote engineering as a career.

Contributors

William F. Ballhaus, Jr. is president and chief executive officer of The Aerospace Corporation, an independent, nonprofit organization dedicated to the objective application of science and technology toward the solution of critical issues in the nation's space program. Dr. Ballhaus joined Aerospace as president in September 2000, after an 11-year career with Lockheed Martin Corporation, where he served as corporate officer and vice president, Engineering and Technology. Prior to his tenure with Lockheed Martin, Dr. Ballhaus served as president of two Martin Marietta businesses, Aero and Naval Systems, and Civil Space and Communications. He also was vice president and program director of Titan IV Centaur operations at Martin Marietta Space Launch Systems. Before joining Martin Marietta, Dr. Ballhaus served as director of the National Aeronautics and Space Administration's Ames Research Center. He also served as acting associate administrator for Aeronautics and Space Technology at NASA headquarters in Washington, D.C. Dr. Ballhaus is a graduate of the University of California at Berkeley, where he earned a doctorate in engineering and bachelor's and master's degrees in mechanical engineering. He is a member of the Defense Science Board and the National Academy of Engineering, where he is also a councillor. He is a member of the International Academy of Astronautics and a fellow of the American Institute of Aeronautics and Astronautics and the Royal Aeronautical Society. He serves on the Jet Propulsion Lab Advisory Council and on the board of directors of the Space Foundation. At Charles Stark Draper Lab, Dr. Ballhaus is one of the 47 members of the corporation who oversee the governance of the laboratory. In addition, he is a member of the engineering advisory boards at the University of California, Berkeley, the Massachusetts Institute of Technology, Johns Hopkins University, and the Univer-

sity of Southern California. Dr. Ballhaus is the recipient of numerous awards, including the presidential ranks of distinguished executive and meritorious executive, conferred by President Ronald Reagan, the NASA Distinguished Service Medal, the Air Force Exceptional Civilian Service Medal, the Distinguished Engineering Alumnus Award from the University of California, Berkeley, and the AIAA's Lawrence Sperry Award.

Gregory W. Characklis is an assistant professor in the Department of Environmental Sciences and Engineering at the University of North Carolina, Chapel Hill (UNC). His areas of research are integrated planning of water supply/treatment systems through use of engineering and economic criteria, impacts of water quality on resource value and allocation, and market-based reform of environmental regulation. From 1997 to 1999, Dr. Characklis was a fellow at the National Academy of Engineering. Prior to joining the faculty at UNC, he was director of resource development and management at Azurix Corp. where his responsibilities centered around assessing the technical and financial merits of water supply development projects throughout the United States, including most of the western states. Dr. Characklis received a B.S. in materials science and engineering from Johns Hopkins University, and M.S. and Ph.D. degrees in environmental science and engineering from Rice University.

William R. Cheswick is chief scientist at Lumeta, a Lucent Technologies' Bell Labs spin-off in Somerset, New Jersey. Prior to assuming his current position, Mr. Cheswick was a member of the technical staff at Bell Labs/Lucent and a systems programmer and consultant with Systems and Computer Technology Corporation. While at Bell Labs, Mr. Cheswick did early work on firewall design and implementation, including the first circuit-level gateway, for which he coined the term "proxy." He also worked on PC viruses, mailers, Internet munitions, and the Plan 9 operating system. In 1994, he co-authored the first book on firewalls, *Firewalls and Internet Security: Repelling the Wiley Hacker*, which was reissued in second edition in Spring 2003. In 1998, Mr. Cheswick started the Internet Mapping Project, which became the core technology for Lumeta. The Internet Mapping Project explores the extent of corporate and government intranets and checks for host leaks that violate perimeter policies. Mr. Cheswick holds five patents, has been interviewed broadly by the media on security issues, and has served on numerous program committees for security conferences. He has a B.S. from Lehigh University. Mr. Cheswick has a wide interest in science, and his avocations include home automation and security, time lapse photography, high power model rocketry, radio-controlled model aircraft, and interactive exhibits for science museums.

James R. Heath is the Elizabeth W. Gilloon Professor of Chemistry at the California Institute of Technology, where his research focuses on understanding

how to fabricate, assemble, and utilize nanometer scale structures, learning how to chemically synthesize a computer, and understanding the nanocircuitry of biological systems. Dr. Heath is the recipient of numerous awards, including the Packard Fellowship (1994), Sloan Fellowship (1997), Fellow of the American Physical Society (1999), Feynman Prize in Nanotechnology (2000), Raymond and Beverly Sackler Prize in the Physical Sciences (2001), and *Scientific American* Top 50 (2002). He received a B.S. from Baylor University and a Ph.D. from Rice University.

Jack Hergenrother is a research staff member in Silicon Technology at the IBM T.J. Watson Research Center in Yorktown Heights, New York, where he focuses on the design and fabrication of advanced silicon CMOS devices. Before joining IBM in August 2003, Dr. Hergenrother was a technical manager at Agere Systems and a distinguished member of the technical staff at Lucent Technologies' Bell Laboratories. From 1995 to 1997, he studied the physics and applications of single-electron transistors as a postdoctoral fellow at Harvard University. Dr. Hergenrother is the recipient of numerous awards, including a Marshall Scholarship, a National Science Foundation Fellowship, a Joint Services Electronics Program Fellowship, and the White Award for Excellence in Teaching at Harvard University. He is one of the few Americans to earn a Cambridge University "Blue" in soccer. Dr. Hergenrother's work on the Vertical Replacement-Gate (VRG) MOSFET, a novel silicon transistor in which all critical dimensions are controlled precisely without lithography, was recognized by *Discover Magazine* as a semifinalist for Technological Innovation of the Year in 2000. He is a member of the IEEE and the American Physical Society, and he serves on the CMOS Devices Subcommittee for the International Electron Devices Meeting. Dr. Hergenrother received a B.S.E. in chemical engineering and engineering physics from Princeton University, an M.A. in applied mathematics and theoretical physics from Cambridge University, and a Ph.D. in applied physics from Harvard University.

Christopher Jarzynski is a technical staff member in the Theoretical Division at Los Alamos National Laboratory. His current research interests include nonequilibrium statistical physics, computational biology, and computational chemistry. Dr. Jarzynski received an A.B. with high honors from Princeton University and a Ph.D. from the University of California, Berkeley, both in physics. He is the recipient of a Fulbright Scholarship and an Outstanding Teaching Assistant Award from the University of California, Berkeley.

Mohamed Athher Mughal is a biological weapons response subject matter expert with the U.S. Army Soldier and Biological Chemical Command, Homeland Defense Business Unit. He has over 17 years of research and technical experience with chemical and biological warfare and terrorism. Prior positions include co-

leader of the Biological Weapons Improved Response Team; strategic planner, Treaty Verification Team; team leader, Nondestructive Equipment Team; and branch chief, Chemical Detectors/Protection Branch. In 1998, Dr. Mughal helped organize and lead a team of over 60 response professionals from around the country that worked through a structured series of five technical workshops, focusing on city- and state-level response strategies for biological terrorism. He has published on biological terrorism preparedness and homeland defense in numerous journals and co-authored two Department of Defense technical reports on bioterrorism preparedness. Dr. Mughal received a B.S. in chemical engineering from the University of Maryland, College Park, and a Ph.D. in policy science from the University of Maryland, Baltimore County. He is a branch-qualified U.S. Army chemical officer and an honor graduate of the U.S. Army Chemical School.

Dianne K. Newman is the Clare Boothe Luce Assistant Professor of Geobiology and Environmental Science at the California Institute of Technology where her research focuses on the molecular basis of microbe/mineral interactions. Dr. Newman received a B.A. from Stanford University in German studies, a Ph.D. from the Massachusetts Institute of Technology in civil and environmental engineering, and was a postdoctoral fellow in the Department of Microbiology and Molecular Genetics at Harvard Medical School. Her awards include an ONR Young Investigator Award (2002) and a Packard Fellowship in Science and Engineering (2002).

Gregory A. Norris founded and directs Sylvatica, a life cycle assessment (LCA) research consulting firm in North Berwick, Maine. Dr. Norris is program manager for the United Nations' Environment Program's (UNEP) global Life Cycle Initiative, directing the Program on Life Cycle Inventory Analysis. He teaches graduate courses on LCA and Industrial Ecology at the Harvard School of Public Health (HSPH), where he also advises graduate students from HSPH and visiting research fellows from abroad. He consults on LCA and sustainable consumption to UNEP, to federal and state agencies in the United States, and to the private and nonprofit sectors. Dr. Norris has developed several software tools to assist analysis and decision making related to life cycle assessment and sustainable enterprise. Recent research integrates socio-economic pathways to human health within the LCA framework and develops the function-based approach to sustainable consumption analysis. He is adjunct research professor at the Complex Systems Research Center, University of New Hampshire; he directs the life cycle assessment activities of the Athena Sustainable Materials Institute in Ottawa, Canada; and he is a program associate in the Center for Hazardous Substance Research at Kansas State University and an editor of the *International Journal of Life Cycle Assessment*. Dr. Norris has an S.B. in mechanical engineering from the Massachusetts Institute of Technology, an M.S. in aeronautics and astronautics from Purdue University, and a Ph.D. in natural resources from the University of New Hampshire.

Alan J. Russell is director of the McGowan Institute for Regenerative Medicine and professor of surgery at the University of Pittsburgh School of Medicine, professor in the Departments of Chemical Engineering and Bioengineering at the University of Pittsburgh School of Engineering, and executive director of the Pittsburgh Tissue Engineering Initiative. After spending two years as a NATO Research Fellow at the Massachusetts Institute of Technology, Dr. Russell joined the University of Pittsburgh as an assistant professor of chemical engineering in 1989. He was named Nikolas DeCecco Professor and chairman of the department in 1995. In 1999, Dr. Russell co-founded the biotechnology company, Agentase LLC. Dr. Russell's research has focused on the symbiotic interface between enzymes and materials. Recently, his group has focused on the biocatalytic detoxification of chemical weapons, the study of proteins in extreme environments, biocatalytic polymer synthesis, biotechnology in supercritical fluids, and biomaterial design synthesis for tissue engineering. His awards include: R&D 100 Award, Carnegie Award for Excellence in Science and Technology, Fellow of the American Institute for Medical and Biological Engineering, and Presidential Young Investigator Award. Dr. Russell received a Ph.D. in chemistry from Imperial College, the University of London.

Thomas J. Silva is project leader of the Magnetodynamics Project, Magnetic Technology Division, National Institute of Standards and Technology (NIST) in Boulder, Colorado. Before assuming his current position in 2000, Dr. Silva was a staff scientist in the Electronics and Electrical Engineering Lab at NIST. He is recipient of the Presidential Early Career Award in Science and Engineering and the IEEE Distinguished Lectureship as well as co-recipient of a NIST Competence Award for research on atomic-scale metrology for ultra-high-density magnetic recording. Dr. Silva is a member of the IEEE Magnetics Society and the American Physical Society, and he is program chair/co-chair for the Metallic Multilayers Symposium in 2004 and the MMM/Intermag Joint Conference 2004, respectively. He received a B.S. in engineering science from the University of California, Berkeley, and M.S. and Ph.D. degrees in electrical and computer engineering from the University of California, San Diego.

Willem P. C. Stemmer is the vice president of research and founder of Avidia Research Institute and founder of Maxygen in Redwood City, California. He is the inventor of DNA shuffling, also called molecular breeding, which gave rise to Codexis, Inc. and Verdia, Inc. Dr. Stemmer received a Ph.D. from the University of Wisconsin-Madison, where his research was on bacterial virulence mechanisms. This was followed by a postdoctoral fellowship in medical genetics, where he worked on antibody engineering for radioimmunotherapy of cancer.

David Wagner is an assistant professor of computer science at the University of California, Berkeley. As a graduate student researcher, he co-founded UC

Berkeley's ISAAC security research group, which has made substantial contributions in computer, network, and wireless security and in online privacy. Dr. Wagner is a co-author of *The Twofish Encryption Algorithm: A 128-Bit Block Cipher* and *System Security: A Management Perspective.* Recent awards include being named one of *Popular Science's* Brilliant 10 (2002), honorary mention for the ACM Dissertation Award, Computer Science Division Information Technology Faculty Award, and Computing Research Association Digital Government Fellow. He serves on numerous program committees, steering committees, and study groups related to cryptography and software security. Dr. Wagner received an A.B. in mathematics from Princeton University, and M.S. and Ph.D. degrees in computer science from the University of California, Berkeley.

Ron Weiss is an assistant professor of electrical engineering at Princeton University, where his research focuses on programming biological organisms by embedding synthetic biochemical logic circuits into cells, as well as embedding sensors, actuators, and intercellular communication mechanisms. Application areas include drug and biomaterial manufacturing, programmed therapeutics, embedded intelligence in materials, environmental sensing and effecting, and nanoscale fabrication. As a graduate student at the Massachusetts Institute of Technology (MIT), Dr. Weiss worked on the MIT Project on Mathematics and Computation, associated with both the MIT Artificial Intelligence Lab and the MIT Laboratory for Computer Sciences. He completed a Ph.D. at MIT in the area of programming biological cells in 2001.

Erik Winfree is an assistant professor in computer science and computation and neural systems at the California Institute of Technology. His research concerns the theory and engineering of autonomous biochemical algorithms using in vitro systems of DNA and enzymes, including programmable DNA self-assembly, DNA and RNA conformational switches and devices, and RNA transcriptional circuits. Such systems are envisioned as embedded information processing and control for bottom-up nanofabrication, nanorobotics, biochemical diagnostics, and other biochemical processes. Dr. Winfree is the recipient of numerous awards, including the NSF PECASE/CAREER Award (2001), the ONR Young Investigators Award (2001), a MacArthur Fellowship (2000), and *Technology Review's* first TR100 list of "top young innovators" (1999). Prior to joining the faculty at Caltech in 1999, Dr. Winfree was a Lewis Thomas Postdoctoral Fellow in Molecular Biology at Princeton, and a visiting scientist at the Massachusetts Institute of Technology's Artificial Intelligence Lab. Dr. Winfree received a B.S. in mathematics and computer science from the University of Chicago in 1991, and a Ph.D. in computation and neural systems from Caltech in 1998.

Program

NATIONAL ACADEMY OF ENGINEERING

Ninth Annual Symposium on
Frontiers of Engineering
September 18–20, 2003

ENVIRONMENTAL ENGINEERING

Organizers: Jane Bare and Joseph Hughes

Microbial Mineral Respiration
Dianne K. Newman, California Institute of Technology

Water-Resource Engineering, Economics, and Public Policy
Gregory W. Characklis, University of North Carolina

Life Cycle Development: Expanding the Life Cycle Framework to Address Issues of Sustainable Development
Gregory A. Norris, Sylvatica

* * *

FUNDAMENTAL LIMITS OF NANOTECHNOLOGY: HOW FAR DOWN IS THE BOTTOM?

Organizers: Gang Chen and Robert Schoelkopf

Status, Challenges, and Frontiers of Silicon CMOS Technology
Jack Hergenrother, IBM T.J. Watson Research Center

Molecular Electronics
James R. Heath, California Institute of Technology

Limits of Storage in Magnetic Materials
Thomas J. Silva, National Institute of Standards and Technology

Thermodynamics of Nanosystems
Christopher Jarzynski, Los Alamos National Laboratory

* * *

DINNER SPEAKER

The Most Important Lessons You Didn't Learn in Engineering School
William F. Ballhaus, Jr., The Aerospace Corporation

* * *

COUNTERTERRORISM TECHNOLOGIES AND INFRASTRUCTURE PROTECTION
Organizers: Matt Blaze and Stephen Lee

Biological Counterterrorism Technologies

Using Biotechnology to Detect and Counteract Chemical Weapons
Alan J. Russell, University of Pittsburgh

An Engineering Problem-Solving Approach to Biological Terrorism
Mohamed Athher Mughal,
U.S. Army Soldier and Biological Chemical Command
(talk given by Laurel O'Connor)

Infrastructure Protection

Software Insecurity
David Wagner, University of California, Berkeley

Internet Security
William R. Cheswick, Lumeta Corporation

* * *

BIOMOLECULAR COMPUTING

Organizers: Lila Kari and Mitsunori Ogihara

DNA Computing by Self-Assembly
Erik Winfree, California Institute of Technology

Natural Computation as a Principle of Biological Design
Willem P. C. Stemmer, Avidia Research Institute

Challenges and Opportunities in Programming Living Cells
Ron Weiss, Princeton University

Participants

NATIONAL ACADEMY OF ENGINEERING

Ninth Annual Symposium on
Frontiers of Engineering
September 18–20, 2003

Ali Adibi
Assistant Professor
Department of Electrical and Computer Engineering
Georgia Institute of Technology

Terry L. Alford
Associate Professor
Department of Chemical and Materials Engineering
Arizona State University

Luis A. Nunes Amaral
Associate Professor
Department of Chemical Engineering
Northwestern University

Kristi S. Anseth
Professor and HHMI Assistant Investigator
Department of Chemical and Biological Engineering
University of Colorado

Mauro J. Atalla
Principal Research Engineer/Project Leader
Otis Program Office
United Technologies Research Center

Ricardo S. Avila
CAD Project Manager
Imaging Technologies Laboratory
GE Global Research

Lisa Axe
Associate Professor
Department of Civil and Environmental Engineering
New Jersey Institute of Technology

Ronald Azuma
Senior Research Staff Computer Scientist
HRL Laboratories

Jeffrey Baclaski
Senior Engineer
Amgen, Inc.

David Baraff
Senior Animation Scientist
Pixar Animation Studios

Jane Bare *(unable to attend)*
Chemical Engineer
Office of Research and Development
U.S. Environmental Protection Agency

Lee A. Barford
Senior Scientist and Project Manager
Agilent Laboratories

Roger S. Barga
Researcher
Microsoft Research
Microsoft Corporation

Rashid Bashir
Associate Professor
School of Electrical Engineering
Purdue University

William D. Batchelor
Associate Professor
Department of Agricultural and Biosystems Engineering
Iowa State University

Matt Blaze
Research Scientist
AT&T Laboratories

Craig A. Blue
Research Staff Member
Oak Ridge National Laboratory

Diann Brei
Associate Professor
Department of Mechanical Engineering
University of Michigan

Cindy Bruckner-Lea
Staff Scientist
Pacific Northwest National Laboratory

Karen J. L. Burg
Assistant Professor of Bioengineering
Department of Bioengineering
Clemson University

Barrett S. Caldwell
Associate Professor
School of Industrial Engineering
Purdue University

Gregory W. Characklis
Assistant Professor
Department of Environmental Sciences and Engineering
University of North Carolina

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

William R. Cheswick
Chief Scientist
Lumeta Corporation

Paul Chodak
Director, Engineering and Environmental Services
American Electric Power

Steven A. Cummer
Assistant Professor
Department of Electrical and Computer Engineering
Duke University

Jennifer Sinclair Curtis
Professor
School of Chemical Engineering
Purdue University

Marie-Claire Cyrille
Research Staff Member
San Jose Research Lab
Hitachi Global Storage Technologies

Drew Dean
Computer Scientist
Computer Science Laboratory
SRI International

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

Serge Dubovitsky
Principal Engineer
Photonics Research and System Group
Jet Propulsion Laboratory

Brian D. Edgecombe
Research Scientist
ATOFINA Chemicals, Inc.

Kenneth B. Elliott
Associate Director
Trusted Computer Systems Department
Aerospace Corporation

James A. Ferwerda
Research Associate
Program of Computer Graphics
Cornell University

Claire F. Gmachl
Associate Professor
Department of Electrical Engineering
Princeton University

J. Scott Goldstein
Assistant Vice President
Science Applications International Corporation

Daniel A. Green
Research Associate
DuPont Central Research and Development

Aparna Gupta
Assistant Professor
Department of Decision Sciences and Engineering Systems
Rensselaer Polytechnic Institute

Eric S. Hamby
Member of the Research Staff
Xerox Innovation Group
Xerox Corporation

Jonah Harley
Research Scientist
Agilent Laboratories

James R. Heath
Elizabeth W. Gilloon Professor of Chemistry
Division of Chemistry and Chemical Engineering
California Institute of Technology

Grant S. Heffelfinger
Deputy Director
Materials and Process Sciences Center
Sandia National Laboratories

Rich Helling
Research Leader
Engineering Sciences
Dow Chemical Company

Jack Hergenrother
Research Staff Manager
Silicon Technology Research
IBM T.J. Watson Research Center

Julia W. P. Hsu
Member of Technical Staff
Sandia National Laboratories

Joseph B. Hughes
Professor and Chair
School of Civil and Environmental Engineering
Georgia Institute of Technology

Chris Jarzynski
Staff Member
Los Alamos National Laboratory

Robert L. Johnson
Staff Engineer
Environmental Assessment Division
Argonne National Laboratory

Lila Kari *(unable to attend)*
Associate Professor and Canada Research Chair in Biocomputing
Department of Computer Science
University of Western Ontario

Chester N. Kennedy
Director, Electronics Technology
Lockheed Martin Corporation

Kelly J. Knight
Engineering Specialist
Bechtel National, Inc.

Martin Krucinski
Member Research Staff II
Wilson Center for Research and Technology
Xerox Corporation

Jeffrey W. Kysar
Assistant Professor
Department of Mechanical Engineering
Columbia University

Richard Lai
Manager for Advanced Microelectronics
Northrop Grumman Space Technology

Kelvin H. Lee
Associate Professor
School of Chemical and Biomolecular Engineering
Cornell University

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

Darren Lytle
Environmental Engineer
U.S. Environmental Protection Agency

Hari C. Manoharan
Assistant Professor
Geballe Laboratory for Advanced Materials
Stanford University

Costas D. Maranas
Associate Professor
Department of Chemical Engineering
Pennsylvania State University

Dimitrios Maroudas
Professor
Department of Chemical Engineering
University of Massachusetts, Amherst

Hiroshi Matsui
Assistant Professor
Department of Chemistry
City University of New York, Hunter College

Steven H. McKnight
Chief, Polymers Research Branch
U.S. Army Research Laboratory

Charles Meneveau
Professor and Vice-Chair
Department of Mechanical Engineering
Johns Hopkins University

Veena Misra
Associate Professor
Department of Electrical and Computer Engineering
North Carolina State University

Urbashi Mitra
Associate Professor
Department of Electrical Engineering
University of Southern California

Arianna T. Morales
Staff Research Scientist
Materials and Processes Laboratory
General Motors Research and Development Center

Mohamed Mughal (*unable to attend*)
Homeland Defense Business Unit
U.S. Army Soldier and Biological Chemical Command (SBCCOM)

Dianne K. Newman
Clare Booth Luce Assistant Professor of Geobiology and Environmental Science
Department of Geology and Planetary Sciences
California Institute of Technology

Gregory Norris
President
Sylvatica

Laurel O'Connor
Principle Engineer
Battelle Eastern Regional Technology Center

Mitsunori Ogihara
Professor
Department of Computer Science
University of Rochester

Miodrag Oljaca
Director, Nanotechnology
MicroCoating Technologies

Teresa L. Olson
Senior Staff Engineer
Missiles and Fire Control
Lockheed Martin Corporation

Gautham Parthasarathy
Senior Research Engineer
Solutia, Inc.

Assimina A. Pelegri
Associate Professor
Department of Mechanical and Aerospace Engineering
Rutgers University

Santosh Ranganath
Senior Research Engineer
Delphi Corporation

Yuan Qiao Rao
Senior Research Scientist
Eastman Kodak Company

Jonathan H. Raub
Senior Technical Advisor
Cummins, Inc.

Vilupanur A. Ravi
Associate Professor
Department of Chemical and Materials Engineering
California State Polytechnic University

Jennifer Rumsey
Technical Specialist
Cummins, Inc.

Alan J. Russell
Director
McGowan Institute for Regenerative Medicine
University of Pittsburgh

Neeraj Saxena
Section Manager
PGS Technology
BOC Group

David V. Schaffer
Assistant Professor
Department of Chemical Engineering
University of California, Berkeley

Robert J. Schoelkopf
Assistant Professor
Applied Physics
Yale University

Steven Shaw
Assistant Professor
Department of Electrical and Computer Engineering
Montana State University

Po-Jen Shih
Research Scientist
Eastman Kodak Company

Thomas Silva
Project Leader, Magnetodynamics
Magnetic Technology Division
National Institute of Standards and Technology

Steven Slayzak
Senior Mechanical Engineer
National Renewable Energy Laboratory

Willem Stemmer
Vice President of Research and Founder
Avidia Research Institute

Chih-Jen Sung
Associate Professor
Department of Mechanical and Aerospace Engineering
Case Western Reserve University

Chi Tang
Research Associate
PPG Industries, Inc.

Joseph W. Tringe
Lead, Electronics Foundation Research Group
Space Vehicles Directorate
Air Force Research Laboratory

Leendert van Doorn
Manager, Secure Embedded Systems Group
IBM T.J. Watson Research Center

Peter J. Vikesland
Assistant Professor
Department of Civil and Environmental Engineering
Virginia Polytechnic Institute and State University

David Wagner
Assistant Professor
Computer Sciences Division
University of California, Berkeley

Christoph Wasshuber
Member, Group Technical Staff
Texas Instruments Inc.

John W. Weatherly
U.S. Army Cold Regions Research and Engineering Laboratory

Ron Weiss
Assistant Professor
Electrical Engineering Department
Princeton University

Andrew B. Williams
Assistant Professor
Department of Electrical and Computer Engineering
University of Iowa

Erik Winfree
Assistant Professor
Computer Science and Computation and Neural Systems
California Institute of Technology

Shumin Zhai
Research Staff Member
IBM Almaden Research Center

Yin Zhang
Senior Technical Staff Member
AT&T Labs - Research

Jorge G. Zornberg
Assistant Professor
Civil Engineering Department
University of Texas, Austin

Dinner Speaker

William F. Ballhaus, Jr.
President and CEO
The Aerospace Corporation

Guests

Christian Bailey
Staff Reporter
Wall Street Journal-Dow Jones & Co., Inc.

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

Bruce Hensel
KNBC

Thomas Kim
Air Force Office of Scientific Research

Thomas S. Maddock
Consulting Engineer
Boyle Engineering Corporation

Marty Perreault
Program Director
National Academies Keck Futures Initiative

Irving Reed
Emeritus Professor
Department of Electrical Engineering
University of Southern California

Sarah Tegen
Editorial Associate
The National Academies

National Academy of Engineering

Wm. A. Wulf
President

Lance A. Davis
Executive Officer

Janet Hunziker
Program Officer

Jenny Hardesty
Senior Project Assistant

Penny Gibbs
Program Associate

Tess Samuel
Development Officer

Randy Atkins
Senior Public Relations Officer