Free Executive Summary

Getting Up to Speed: The Future of Supercomputing

Susan L. Graham, Marc Snir, and Cynthia A. Patterson, Editors, Committee on the Future of Supercomputing, National Research Council

ISBN: 0-309-09502-6, 222 pages, 6 x 9, paperback (2004)

This free executive summary is provided by the National Academies as part of our mission to educate the world on issues of science, engineering, and health. If you are interested in reading the full book, please visit us online at http://www.nap.edu/catalog/11148.html . You may browse and search the full, authoritative version for free; you may also purchase a print or electronic version of the book. If you have questions or just want more information about the books published by the National Academies Press, please contact our customer service department toll-free at 888-624-8373.

This executive summary plus thousands more available at www.nap.edu.

Copyright 2004 © National Academy of Sciences. Permission is granted for this material to be shared for noncommercial, educational purposes, provided that this notice appears on the reproduced materials, the Web address of the online, full authoritative version is retained, and copies are not altered. To disseminate otherwise or to republish requires written permission from the National Academies Press.

Executive Summary

Supercomputing is very important to the United States for conducting basic scientific research and for ensuring the physical and economic well-being of the country. The United States has a proud history of leadership in supercomputing, which has contributed not only to its international standing in science and engineering and to national health and security but also to the commercial strength of many industries, including the computing industry. Supercomputing has become a major contributor to the economic competitiveness of our automotive, aerospace, medical, and pharmaceutical industries. The discovery of new substances and new techniques, as well as cost reduction through simulation rather than physical prototyping, will underpin progress in a number of economically important areas. The use of supercomputing in all of these areas is growing, and it is increasingly essential to continued progress.

However, in recent years our progress in supercomputing has been slowing, as attention turns to other areas of science and engineering. The advances in mainstream computing caused by improved processor performance have enabled some former supercomputing needs to be addressed by clusters of commodity processors. Yet important applications, some vital to our nation's security, require technology that is only available in the most advanced custom-built systems. We have been remiss in our attention to conducting the long-term research and development, and sustaining the industrial capabilities that we will need in the future. The Japanese Earth Simulator has served as a wake-up call, reminding us that complacency can cause us to lose not only our competitive advantage but also, and more importantly, the national competence that we need to achieve our own goals.

To maintain our level of achievement in supercomputing and its applications, as well as to keep us from falling behind relative to other nations and to our own needs, a renewed national effort is needed. That effort must have the following components:

- Government leadership in maintaining a national planning activity that is sustained, ongoing, and coordinated and that drives investment decisions.
- Continuing progress in creating hardware, software, and algorithmic technologies that enable the application of supercomputing to important domain-specific problems. Such progress will require continuing government investment.
- International collaborations in all aspects of supercomputing except those that would demonstrably compromise national security.

Supercomputing has always been a specialized form of computing at the cutting edge of technology. As the computing field has grown and matured, computing has become broader and more diverse. From an economic perspective, there are large new markets that are distinct from supercomputing—for example, personal computing devices of various kinds, computers invisibly embedded in many kinds of artifacts, and applications that use large amounts of computing in relatively undemanding ways. As a consequence, potential providers of supercomputing systems and software and potential creators of future supercomputing technology are fewer in number than they once were. In the face of continuing need and the competing demands that weaken supply, the committee recommends that the following actions and policies be initiated.

Overall Recommendation: To meet the current and future needs of the United States, the government agencies that depend on supercomputing, together with the U.S. Congress, need to take primary responsibility for accelerating advances in supercomputing and ensuring that there are multiple strong domestic suppliers of both hardware and software.

Copyright © 2004 National Academy of Sciences. All rights reserved.
This executive summary plus thousands more available at http://www.nap.edu

The government is the primary user of supercomputing. Government-funded research that relies on supercomputing is pushing the frontiers of knowledge and bringing important societal benefits. Because supercomputing is essential to maintain U.S. military superiority, to achieve the goals of stockpile stewardship, and to maintain national security, the government must ensure that the U.S. supercomputing infrastructure advances sufficiently to support our needs in the coming years. These needs are distinct from those of the broad information technology industry. They involve platforms and technologies that are unlikely on their own to have a broad enough market in the short term to satisfy government needs.

To guide the government agencies and Congress in assuming that responsibility, the committee makes eight recommendations.

Recommendation 1. To get the maximum leverage from the national effort, the government agencies that are the major users of supercomputing should be jointly responsible for the strength and continued evolution of the supercomputing infrastructure in the United States, from basic research to suppliers and deployed platforms. The Congress should provide adequate and sustained funding.

A small number of government agencies are the primary users of supercomputing, either directly, through acquisitions, or indirectly, by awarding contracts and grants to other organizations that purchase supercomputers. At present, those agencies include: the National Nuclear Security Administration and the Office of Science at the Department of Energy (DOE), the Department of Defense (DoD), the National Security Agency (NSA), the National Aeronautics and Space Administration (NASA), the National Oceanic and Atmospheric Administration (NOAA), and the National Science Foundation (NSF). (The increasing use of supercomputing in biomedical applications suggests that the National Institutes of Health (NIH) should be added to the list.) Although the agencies have different missions and different needs, they could benefit from the synergies of coordinated planning and acquisition strategies and coordinated support for R&D. For instance, many of the technologies, in particular software technology, need to be broadly available across all platforms. Therefore, those agencies must be jointly responsible and jointly accountable. Moreover, for the agencies to meet their own mission responsibilities and also take full advantage of the investments made by other agencies, collaboration and coordination must become much more long range. The agencies that are the biggest users of supercomputing must develop and execute an integrated plan.

The committee emphasizes the need for developing an integrated plan rather than coordinating distinct supercomputing plans through a diffuse interagency structure. An integrated plan is not an integrated budget. The use of such a plan would not preclude agencies from individual activities, nor would it prevent them from setting their own priorities. Also, it must not be used to the exclusion of unanticipated needs and opportunities. Rather, the intent is to identify common needs at an early stage, and to leverage shared efforts for meeting those needs, while minimizing duplicative efforts. Different agencies should pick the activities that best match their missions; for example, long-term basic research best matches NSF's mission, while industrial supercomputing R&D is more akin to DARPA's mission.

Recommendation 2. The government agencies that are the primary users of supercomputing should ensure domestic leadership in those technologies that are essential to meet national needs.

Current U.S. investments in supercomputing and current plans are not sufficient to provide the supercomputing capabilities that our country will need. It needs supercomputers that satisfy critical requirements in areas such as cryptography and stockpile stewardship, as well as systems that will provide breakthroughs for the broad scientific and technological progress

Copyright © 2004 National Academy of Sciences. All rights reserved.
This executive summary plus thousands more available at http://www.nap.edu

underlying a strong and robust U.S. economy. The committee is less concerned that the top-ranked computer in the TOP500 list (as of June 2004) was located in Japan. U.S. security is not necessarily endangered if a computer in a foreign country is capable of doing some computations faster than U.S.-based computers. The committee believes that had the United States at that time made an investment similar to the Japanese investment in the Earth Simulator, it could have created a powerful and equally capable system. The committee's concern is that the United States has not been making the investments that would guarantee its ability to create such a system in the future.

Leadership is measured by the ability to acquire and exploit effectively machines that can best reduce the time to solution of important computational problems, not by the ability to produce a machine that is at the top of some list. From this perspective, it is not the Earth Simulator system per se that is worrisome but rather the fact that the construction of this system might turn out to be a singular event. It appears that custom high-bandwidth processors such as those used by the Earth Simulator are not viable products without significant government support. Two of the three Japanese companies that were manufacturing such processors do not do so anymore, and the third (NEC) may also bow to market realities in the not-too-distant future, since the Japanese government seems less willing now to subsidize the development of cutting-edge supercomputing technologies. Only by maintaining national leadership in these technologies can the U.S. government ensure that key supercomputing technologies, such as custom high-bandwidth processors, will be available to satisfy its needs. The U.S. industrial base must include suppliers on whom the government can rely to build custom systems to solve problems arising from the government's unique requirements. Since only a few units of such systems are ever needed, there is no broad market for the systems and hence no commercial off-the-shelf suppliers. Domestic supercomputing vendors can become a source of both the components and the engineering talent needed for building these custom systems.

Recommendation 3. To satisfy its need for unique supercomputing technologies such as high-bandwidth systems, the government needs to ensure the viability of multiple domestic suppliers.

Supercomputers built out of commodity components satisfy a large fraction of supercomputing applications. These applications benefit from the fast evolution and low cost of commodity technology. But commodity components are designed for the needs of large markets in data processing or personal computing and are inadequate for many supercomputing applications. The use of commodity clusters results in lower sustained performance and higher programming costs for some demanding applications. This is especially true of some security-related computations where shorter time to solution is of critical importance, justifying the use of custom-built, high-bandwidth supercomputers even at a higher cost per solution.

It is important to have multiple suppliers for any key technology in order to maintain competition, to prevent technical stagnation, to provide diverse supercomputing ecosystems that will address diverse needs, and to reduce risk. However, it is unrealistic to expect that such narrow markets will attract a large number of vendors. As is true for many military technologies, there may be only a few suppliers.

To ensure their continuing existence, domestic suppliers must follow a viable business model. For a public company, that means having a predictable and steady revenue stream recognizable by the financial market. A company cannot continue to provide leadership products without R&D. At least two models have been used successfully in the past: (1) an implicit guarantee for the steady purchase of supercomputing systems, giving the companies a steady income stream with which to fund ongoing R&D and (2) explicit funding of a company's R&D. Stability is a key issue. Suppliers of such systems or components are often small companies that can cease to be viable; additionally, uncertainty can mean the loss of skilled personnel to other

Copyright © 2004 National Academy of Sciences. All rights reserved.
This executive summary plus thousands more available at http://www.nap.edu

sectors of the computing industry or the loss of investors. Historically, government priorities and technical directions have changed more frequently than would be justified by technology lifetimes, creating market instabilities. The chosen funding model must ensure stable funding.

Recommendation 4. The creation and long-term maintenance of the software that is key to supercomputing requires the support of those agencies that are responsible for supercomputing R&D. That software includes operating systems, libraries, compilers, software development and data analysis tools, application codes, and databases.

Supercomputer software is developed and maintained by the national laboratories, by universities, by vertically integrated hardware vendors, and by small independent companies. An increasing amount of the software used in supercomputing is developed in an open source model. Many of the supercomputing software vendors are small and can disappear from the marketplace. The open source model may suffer from having too few developers of supercomputing software with too many other demands on their time.

The successful evolution and maintenance of complex software systems are critically dependent on institutional memory—that is, on the continuous involvement of the few key developers that understand the software design. Stability and continuity are essential to preserve institutional memory. Whatever model of support is used, it should be implemented so that stable organizations with lifetimes of decades can maintain and evolve the software. At the same time, the government should not duplicate successful commercial software packages but should instead invest in new technology. When new commercial providers emerge, the government should purchase their products and redirect its own efforts toward technology that it cannot otherwise obtain.

Barriers to the replacement of application programming interfaces are very high, owing to the large sunk investments in application software. Any change that significantly enhances our nation's ability to program very large systems will not bring about merely a local improvement but will entail the radical coordinated change of many technologies to create a new ecosystem. To facilitate this change, the government needs long-term, coordinated investments in a large number of interlocking technologies.

Recommendation 5. The government agencies responsible for supercomputing should underwrite a community effort to develop and maintain a roadmap that identifies key obstacles and synergies in all of supercomputing.

The challenges in supercomputing are very significant, and the amount of ongoing research is limited. To make progress, it is important to identify and address the key roadblocks. Furthermore, technologies in different domains are interdependent: Progress on a new architecture may also require specific advances in packaging, interconnects, operating system structures, programming languages and compilers, and the like. Thus, investments need to be coordinated. To drive decisions, one needs a roadmap of all the technologies that affect supercomputing. The roadmap needs to have quantitative and measurable milestones. Its creation and maintenance should be an open process that involves a broad community. It is important that a supercomputing roadmap be driven both top-down from application needs and bottom-up from technology barriers and that mission needs as well as science needs be incorporated. It should focus on the evolution of each specific technology and on the interplay between technologies. It should be updated annually and undergo major revision at suitable intervals.

The roadmap should be used by agencies and by Congress to guide their long-term research and development investments. Those roadblocks that will not be addressed by industry without government intervention need to be identified, and the needed research and development

Copyright © 2004 National Academy of Sciences. All rights reserved.
This executive summary plus thousands more available at http://www.nap.edu

must be initiated. Metrics must be developed to support the quantitative aspects of the roadmap. It is important also to invest in some high-risk, high-return research ideas that are not indicated by the roadmap, to avoid being blindsided.

Recommendation 6. Government agencies responsible for supercomputing should increase their levels of stable, robust, sustained multiagency investment in basic research. More research is needed in all the key technologies required for the design and use of supercomputers (architecture, software, algorithms, and applications).

The peak performance of supercomputers has increased rapidly in the last decades, but their sustained performance has lagged, and the productivity of supercomputing users has lagged. Over the last decade the advance in peak supercomputing performance was largely due to the advance in microprocessor performance driven by increased miniaturization, with some contribution from increased parallelism. Perhaps because a large fraction of supercomputing improvements resulted from these advances, few novel technologies were introduced in supercomputer systems, and supercomputing research investments decreased. However, many important applications have not benefited from these advances in mainstream computing, and it will be harder for supercomputing to benefit from increased miniaturization in the future. Fundamental breakthroughs will be needed that will require an increase in research funding.

The research investments should be informed by the supercomputing roadmap but not constrained by it. It is important to focus on technologies that have been identified as roadblocks and that are beyond the scope of industry investments in computing. It is equally important to support long-term speculative research in potentially disruptive technical advances. The research investment should also be informed by the "ecosystem" view of supercomputing—namely, that progress is often needed on a broad front of interrelated technologies rather than as individual breakthroughs.

Research should include a mix of small, medium, and large projects. Many small individual projects are necessary for the development of new ideas. A smaller number of large projects that develop technology demonstrations are needed to bring these ideas to maturity and to study the interaction between various technologies in a realistic environment. Such demonstration projects (which are different from product prototyping activities) should not be expected to be stable platforms for exploitation by users, because the need to maintain a stable platform conflicts with the ability to use the platform for experiments. It is important that the development of such demonstration systems have the substantial involvement not only of academic researchers but also of students, to support the education of the new generation of researchers and to increase the supercomputing workforce. It is also important that the fruits of such projects not be proprietary. The committee estimated the necessary investments in such projects at about $140 million per year.

In its early days, supercomputing research generated many ideas that eventually became broadly used in the computing industry. Such influences will continue in the future. Many of the technical roadblocks faced today by supercomputing are roadblocks that will affect all computing over time. There can be little doubt that solutions developed to solve this problem for supercomputers will eventually influence the broader computing industry, so that investment in basic research in supercomputing is likely to be of widespread benefit to all of information technology.

Recommendation 7. Supercomputing research is an international activity; barriers to international collaboration should be minimized.

Research has always benefited from the open exchange of ideas and the opportunity to build on the achievements of others. The national leadership advocated in these

Copyright © 2004 National Academy of Sciences. All rights reserved.
This executive summary plus thousands more available at http://www.nap.edu

recommendations is enhanced, not compromised, by early-stage sharing of ideas and results. In light of the relatively small community of supercomputing researchers, international collaborations are particularly beneficial. The climate modeling community, for one, has long embraced that view.

Collaboration with international researchers must include giving them access to supercomputing systems; they often spend time in the United States to work closely with resident scientists. Many of the best U.S. graduate students come from other countries, although they often remain as permanent residents or new citizens. Access restrictions based on citizenship hinder collaboration and are contrary to the openness that is essential to good research.

Restrictions on the import of supercomputers to the United States have not benefited the U.S. supercomputing industry and are unlikely to do so in the future. Some kinds of export controls—on commodity systems, especially—lack any clear rationale, given that such systems are built from widely available commercial components. It makes little sense to restrict sales of commodity systems built from components that are not export controlled. Because restrictions on the export of supercomputing technology may damage international collaboration, the benefit of using export controls to prevent potential adversaries or proliferators from accessing key supercomputing technology has to be carefully weighed against that damage.

Since supercomputer systems are multipurpose (nuclear simulations, climate modeling, and so on), their availability need not compromise the domestic leadership needed for national defense, so long as safeguards are in place to protect critical applications.

Recommendation 8. The U.S. government should ensure that researchers with the most demanding computational requirements have access to the most powerful supercomputing systems.

Access to the most powerful supercomputers is important for the advancement of science in many disciplines. A model in which top supercomputing capabilities are provided by different agencies with different missions is healthy. Each agency is the primary supporter of certain research or mission-driven communities; as such, each agency should have a long-term plan and budget for the acquisition of the supercomputing systems that are needed to support its users. The planning and funding process followed by each agency must ensure stability from the viewpoint of its users.

The users should be involved in the planning process and should be consulted in setting budget priorities for supercomputing. The mechanisms for allocating supercomputing resources must ensure that almost all of the computer time on capability systems is allocated to jobs for which that capability is essential. Budget priorities should be reflected in the HEC plan proposed in Recommendation 5. In Chapter 9, the committee estimated the cost of a healthy procurement process that would satisfy the capability supercomputing needs (but not their capacity needs) of the major agencies using supercomputing, and would include the platforms primarily used for research, at about $800 million per year. This estimate includes both platforms used for mission-specific tasks and platforms used to support science.

The NSF supercomputing centers have traditionally provided open access to a broad range of academic users. However, some of these centers have increased the scope of their activities in order to support high-speed networking and grid computing and to expand their education mission. The increases in scope have not been accompanied by corresponding increases in funding, so less attention is paid to supercomputing, and support for computational scientists with capability needs has been diluted.

It is important to repair the current situation at NSF, in which the computational science users of supercomputing centers appear to have too little involvement in programmatic and budgetary planning. All the research communities in need of supercomputing capability have a shared responsibility to provide direction for the supercomputing infrastructure they use and to

Copyright © 2004 National Academy of Sciences. All rights reserved.
This executive summary plus thousands more available at http://www.nap.edu

ensure that resources are available for sustaining the supercomputing ecosystems. Funding for the acquisition and operation of the research supercomputing infrastructure should be clearly separated from funding for computer and computational science and engineering research. It should compete on an equal basis with other infrastructure needs of the science and engineering disciplines. That is not now the case.

Copyright © 2004 National Academy of Sciences. All rights reserved.
This executive summary plus thousands more available at http://www.nap.edu

Getting Up to Speed: The Future of Supercomputing
http://books.nap.edu/catalog/11148.html

Prepublication Copy—Subject to Further Editorial Corrections

Getting Up to Speed
The Future of Supercomputing

Susan L. Graham, Marc Snir, and Cynthia A. Patterson, Editors

Committee on the Future of Supercomputing

Computer Science and Telecommunications Board
Division on Engineering and Physical Sciences

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

The National Academies Press
Washington, D.C.
www.nap.edu

Copyright © 2004 National Academy of Sciences. All rights reserved.
This executive summary plus thousands more available at http://www.nap.edu

Prepublication Copy—Subject to Further Editorial Corrections

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

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

Support for this project was provided by the Department of Energy. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the organizations that provided support for the project.

International Standard Book Number XXXXX
Library of Congress Control Number XXXXX

Cover designed by Jennifer Bishop.

Copies of this report are available from the National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, D.C. 20055, (800) 624-6242 or (202) 334-3313 in the Washington metropolitan area. Internet, http://www.nap.edu

Copyright XXXX by the National Academy of Sciences. All rights reserved.

Printed in the United States of America

Copyright © 2004 National Academy of Sciences. All rights reserved.
This executive summary plus thousands more available at http://www.nap.edu

Prepublication Copy—Subject to Further Editorial Corrections

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

Copyright © 2004 National Academy of Sciences. All rights reserved.
This executive summary plus thousands more available at http://www.nap.edu

COMMITTEE ON THE FUTURE OF SUPERCOMPUTING

SUSAN L. GRAHAM, Co-chair, University of California, Berkeley
MARC SNIR, Co-chair, University of Illinois at Urbana-Champaign
WILLIAM J. DALLY, Stanford University
JAMES W. DEMMEL, University of California, Berkeley
JACK J. DONGARRA, University of Tennessee, Knoxville, and Oak Ridge National Laboratory
KENNETH S. FLAMM, University of Texas, Austin
MARY JANE IRWIN, Pennsylvania State University
CHARLES KOELBEL, Rice University
BUTLER W. LAMPSON, Microsoft Corporation
ROBERT F. LUCAS, University of Southern California
PAUL C. MESSINA, Distinguished Senior Computer Scientist, Consultant
JEFFREY M. PERLOFF, University of California, Berkeley
WILLIAM H. PRESS, Los Alamos National Laboratory
ALBERT J. SEMTNER, Naval Postgraduate School
SCOTT STERN, Northwestern University
SHANKAR SUBRAMANIAM, University of California, San Diego
LAWRENCE C. TARBELL, JR., Technology Futures Office, Eagle Alliance
STEVEN J. WALLACH, Chiaro Networks

Staff
CYNTHIA A. PATTERSON, Study Director
PHIL HILLIARD, Research Associate (through May 2004)
MARGARET MARSH HUYNH, Senior Program Assistant
HERBERT S. LIN, Senior Scientist

Copyright © 2004 National Academy of Sciences. All rights reserved.
This executive summary plus thousands more available at http://www.nap.edu

COMPUTER SCIENCE AND TELECOMMUNICATIONS BOARD

DAVID LIDDLE, U.S. Venture Partners, Co-chair
JEANNETTE M. WING, Carnegie Mellon University, Co-chair
ERIC BENHAMOU, Benhamou Global Ventures, LLC
DAVID D. CLARK, Massachusetts Institute of Technology, CSTB Chair Emeritus
WILLIAM DALLY, Stanford University
MARK E. DEAN, IBM Almaden Research Center
DEBORAH ESTRIN, University of California, Los Angeles
JOAN FEIGENBAUM, Yale University
HECTOR GARCIA-MOLINA, Stanford University
KEVIN KAHN, Intel Corporation
JAMES KAJIYA, Microsoft Corporation
MICHAEL KATZ, University of California, Berkeley
RANDY H. KATZ, University of California, Berkeley
WENDY A. KELLOGG, IBM T.J. Watson Research Center
SARA KIESLER, Carnegie Mellon University
BUTLER W. LAMPSON, Microsoft Corporation, CSTB Member Emeritus
TERESA H. MENG, Stanford University
TOM M. MITCHELL, Carnegie Mellon University
DANIEL PIKE, GCI Cable and Entertainment
ERIC SCHMIDT, Google Inc.
FRED B. SCHNEIDER, Cornell University
WILLIAM STEAD, Vanderbilt University
ANDREW J. VITERBI, Viterbi Group, LLC

CHARLES N. BROWNSTEIN, Director
KRISTEN BATCH, Research Associate
JENNIFER M. BISHOP, Program Associate
JANET BRISCOE, Manager, Program Operations
JON EISENBERG, Senior Program Officer
RENEE HAWKINS, Financial Associate
MARGARET MARSH HUYNH, Senior Program Assistant
HERBERT S. LIN, Senior Scientist
LYNETTE I. MILLETT, Senior Program Officer
JANICE SABUDA, Senior Program Assistant
GLORIA WESTBROOK, Senior Program Assistant
BRANDYE WILLIAMS, Staff Assistant

For more information on CSTB, see its Web site at <http://www.cstb.org>, write to CSTB, National Research Council, 500 Fifth Street, N.W., Washington, DC 20001, call (202) 334-2605, or e-mail the CSTB at cstb@nas.edu.

Copyright © 2004 National Academy of Sciences. All rights reserved.
This executive summary plus thousands more available at http://www.nap.edu

Copyright © 2004 National Academy of Sciences. All rights reserved.
This executive summary plus thousands more available at http://www.nap.edu

Preface

High-performance computing is important in solving complex problems in areas from climate and biology to national security. Several factors have led to the recent reexamination of the rationale for federal investment in research and development in support of high-performance computing, including (1) continuing changes in the various component technologies and their markets, (2) the evolution of the computing market, particularly the high-end supercomputing segment, (3) experience with several systems using the clustered processor architecture, and (4) the evolution of the problems, many of them mission-driven, for which supercomputers are used.

The Department of Energy's (DOE's) Office of Science expressed an interest in sponsoring a study by the Computer Science and Telecommunications Board (CSTB) of the National Research Council (NRC) that would assess the state of U.S. supercomputing capabilities and relevant research and development. Spurred by the development of the Japanese vector-based Earth Simulator supercomputer, the Senate's Energy and Water Development Appropriations Committee directed Advanced Simulation and Computing (ASC) of the National Nuclear Security Administration (NNSA) at DOE to commission, in collaboration with DOE's Office of Science, a study by the NRC. Congress also commissioned a study by the JASONs[1] to identify the distinct requirements of the stockpile stewardship program and its relation to the ASC acquisition strategy.

CSTB convened the Committee on the Future of Supercomputing to assess prospects for supercomputing technology research and development in support of U.S. needs, to examine key elements of context—the history of supercomputing, the erosion of research investment, the changing nature of the problems demanding supercomputing, and the needs of government agencies for supercomputing capabilities—and to assess opportunities for progress. The 18 distinguished members of the study committee (see Appendix A for their biographies) were drawn from academia, industry, and government research organizations in the United States. Several committee members have had previous government and/or industry service. Their collective expertise includes software, computer architecture, performance assessment, applications using supercomputing, economics, and policy matters.

The committee did its work through its own expert deliberations and by soliciting input from key officials in its sponsoring agency (DOE) and numerous experts in both the United States and Japan, including government officials, academic researchers, supercomputer manufacturers, software vendors, supercomputer center managers, and application users of supercomputing systems (see Appendix B). In addition to meeting six times, the committee hosted a workshop attended by more than 20 scientists from a broad range of disciplines to explore the supercomputing needs and opportunities of key scientific domains in the coming decade and to discuss the supercomputing technologies that will facilitate supercomputer use in these domains. Many of the workshop participants provided white papers (see Appendix C for a list) expressing their views on computational challenges in supercomputing, which informed both the workshop and this report.

The committee also visited five DOE supercomputer centers and the National Security Agency's (NSA's) Supercomputer Center (see Appendix B). A subset of the committee received classified briefings from the Department of Energy on stockpile stewardship and from the NSA on signals intelligence that helped illuminate how these mission requirements drive supercomputing needs now and in the future. Given that a significant fraction of government funding of supercomputing is for classified national security programs, the committee believed such briefings were needed to ensure that its report would be useful for the entire supercomputing community. Having received the briefings, the committee believes that the needs of the classified supercomputing applications reinforce, but do not change, the committee's findings and recommendations for the future of supercomputing. This unclassified report does not have a classified annex, nor is there a classified version.

---

[1]Formed in 1959, the JASONs are a select group of scientific advisors who consult with the federal government, chiefly on classified research issues.

Copyright © 2004 National Academy of Sciences. All rights reserved.
This executive summary plus thousands more available at http://www.nap.edu

To facilitate communication within the broader community, the committee hosted a town hall meeting at the annual 2003 Supercomputing Conference in Phoenix, Arizona. In addition, a subset of the committee spent one week in Japan meeting with senior colleagues from the Japanese government, industry, and academia to discuss scientific, technical, and policy issues of mutual interest and to better understand both the similarities and the differences in how the two countries approach supercomputing. They visited several sites in Japan, including the Earth Simulator; the government ministry responsible for funding the Earth Simulator; a university supercomputer center; Japan's Aerospace Exploration Agency; and an auto manufacturer. On the committee's behalf, the National Academy of Engineering co-sponsored with the Engineering Academy of Japan a 1-day forum in Tokyo on the future of supercomputing. Twenty-five Japanese supercomputing experts participated in the forum. The sharing of ideas in those meetings provided important perspectives that contributed to the completeness and accuracy of this report. It is the hope of the committee that activities such as the Tokyo forum will lead to future collaboration between Japan and the United States in areas that will advance supercomputing in both countries.

In July 2003, the committee released an interim report[2] that provided a high-level description of the state of U.S. supercomputing, the needs of the future, and the factors that contribute to meeting those needs. That report generated a number of comments that helped to guide the committee in its work for this final report. Additional inputs helpful to committee members and staff came from professional conferences, the technical literature, and government reports.

The committee is grateful to the many people who contributed to this complex study and its comprehensive report. First and foremost, the committee thanks the sponsors, DOE's Office of Science (Fred Johnson and Dan Hitchcock) and DOE's NNSA (Dimitri Kusnezov, Edgar Lewis, and José Muñoz), not only for their financial support but also for their help in facilitating meetings with people with whom its members wished to speak.

The committee appreciates the thoughtful testimony received from many individuals at its plenary sessions (see Appendix B for a complete list of briefers). The NSA and DOE site visits provided critical input to the committee deliberations. These site visits would not have been possible without the assistance of people at each locale. The committee and staff thank the following people for their help: Gary D. Hughes (NSA), Lynn Kissel (Lawrence Livermore National Laboratory), Horst D. Simon (Lawrence Berkeley National Laboratory), Robert Thomas (Sandia National Laboratories), James S. Peery (Los Alamos National Laboratory), Rick Stevens (Argonne National Laboratory), and Thomas Zacharia (Oak Ridge National Laboratory).

The committee thanks the workshop participants for the insights they contributed through their white papers (see Appendix C for a list of papers), discussions, breakout sessions, and subsequent interactions. The committee is particularly grateful to Warren Washington (National Center for Atmospheric Research), Charles McMillan (Lawrence Livermore National Laboratory), Jeffrey Saltzman (Merck Research Laboratory), and Phillip Colella (Lawrence Berkeley National Laboratory) for their thoughtful plenary presentations.

Many people were instrumental in making the trip to Japan a success. The committee is extremely grateful to Kenichi Miura (Fujitsu fellow) and Tadashi Watanabe (NEC) for their assistance before and during the trip. The 1-day Japan–U.S. Forum on the Future of Supercomputing would not have been possible without the support of the Engineering Academy of Japan and the National Academy of Engineering. The committee learned a lot from insightful presentations and discussions from all the Japanese forum participants. The committee and staff also thank the individuals at each site who took time to meet with the committee. In particular, they thank Tetsuya Sato at the Earth Simulator Center and Harumasa Miura at the Ministry of Education, Culture, Sports, Science and Technology. Maki Haraga provided excellent translation services and logistical help for the committee's entire trip.

[2]National Research Council (NRC). 2003. *The Future of Supercomputing: An Interim Report*. Washington, D.C.: National Academies Press.

Copyright © 2004 National Academy of Sciences. All rights reserved.
This executive summary plus thousands more available at http://www.nap.edu

The committee was fortunate to receive many thoughtful and perceptive comments from the reviewers as well as from the Monitor and the Coordinator of this report. These comments were instrumental in helping the committee to sharpen and improve its report.

Finally, the committee thanks the various members of the NRC staff who helped to move this report from vision to reality. Cynthia Patterson provided continuing wisdom, guidance, encouragement, and friendship, in concert with her hard work on the report. Margaret Huynh's skills in organizing the committee's meetings and supporting its efforts and Phil Hilliard's research support were key contributions to the work of the committee. Liz Fikre edited the final manuscript for publication. Kevin Hale and Machelle Reynolds successfully facilitated the security clearances and security review necessary to complete this study in a timely manner. Janice Mehler and Liz Panos were very helpful in facilitating and expediting the review process.

Susan L. Graham and Marc Snir, Co-chairs
Committee on the Future of Supercomputing

Copyright © 2004 National Academy of Sciences. All rights reserved.
This executive summary plus thousands more available at http://www.nap.edu

Prepublication Copy—Subject to Further Editorial Corrections

Copyright © 2004 National Academy of Sciences. All rights reserved.
This executive summary plus thousands more available at http://www.nap.edu

Getting Up to Speed: The Future of Supercomputing
http://books.nap.edu/catalog/11148.html

Acknowledgment of Reviewers

This report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the National Research Council's (NRC's) Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making the published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following individuals for their review of this report:

Mark E. Dean, IBM,
Steven Gottlieb, Indiana University
Shane Mitchell Greenstein, Northwestern University,
Sidney Karin, University of California, San Diego,
Ken Kennedy, Rice University,
Richard Loft, National Center for Atmospheric Research,
J. Andrew McCammon, University of California, San Diego,
Kenichi Miura, Fujitsu Laboratories Limited
Michael Norman, University of Illinois, Urbana-Champaign
Richard Proto, National Security Agency (retired),
Daniel A. Reed, University of North Carolina, Chapel Hill,
Ahmed Sameh, Purdue University,
Gary Smaby, Smaby Group, Inc.,
Burton J. Smith, Cray Inc.,
Allan Edward Snavely, San Diego Supercomputing Center,
William W. Stead, Vanderbilt University, and
Paul Tackley, University of California, Los Angeles.

Although the reviewers listed above have provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations, nor did they see the final draft of the report before its release. The review of this report was overseen by Elsa Garmire of Dartmouth University and Samuel H. Fuller of Analog Devices, Inc. Appointed by the NRC, they were responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the authoring committee and the institution.

Copyright © 2004 National Academy of Sciences. All rights reserved.
This executive summary plus thousands more available at http://www.nap.edu

Copyright © 2004 National Academy of Sciences. All rights reserved.
This executive summary plus thousands more available at http://www.nap.edu

Contents

EXECUTIVE SUMMARY **1**

1 INTRODUCTION AND CONTEXT **8**
Study Context 9
Computenik 11
About the Interim Report 12
Organization of the Report 13

2 EXPLANATION OF SUPERCOMPUTING **14**

3 BRIEF HISTORY OF SUPERCOMPUTING **20**
The Prehistory of U.S. Supercomputing 20
Supercomputers Emerge as Market 22
Control Data and Cray 23
Enter Japan 25
Innovation in Supercomputing 26
Recent Developments in SupercomputING 29
The U.S. High-Performance Computing Industry Today 30
An Industrial Revolution 33
Impacts 35
Direct Contributions 36
Spillover Effects 37

4 THE DEMAND FOR SUPERCOMPUTING **49**
Compelling Applications for Supercomputing 51
Common Themes and Synergies Across Applications Areas 52
Selected Application Areas 53
Stockpile Stewardship 53
Signals Intelligence 55
Defense 56
Climate Modeling 57
Plasma Physics 60
Transportation 60
Bioinformatics and Computational Biology 62
Societal Health and Safety 64
Earthquakes 65
Geophysical Exploration and Geoscience 66
Astrophysics 67
Materials Science and Computational Nanotechnology 67
Human/Organizational Systems Studies 69
Projected Computing Needs for Applications 70

5 WHAT IS TODAY'S SUPERCOMPUTING TECHNOLOGY? **74**
Supercomputer Architecture 74
Scaling of Technology 74
Types of Supercomputers 76
Performance Issues 78

Copyright © 2004 National Academy of Sciences. All rights reserved.
This executive summary plus thousands more available at http://www.nap.edu

Trade-offs 81
Trends in Supercomputer Architecture 83
Supercomputing Algorithms 85
Solving Partial and Ordinary Differential Equations 85
Mesh Generation 86
Dense Linear Algebra 87
Sparse Linear Algebra 87
Discrete Algorithms 88
Fast Transforms 88
New Algorithmic Demands Arising from Supercomputing 89
Disciplinary Needs 89
Interdisciplinary Needs 89
Synthesis, Sensitivity Analysis, and Optimization Replacing Analysis 90
Huge Data Sets 90
Changing Machine Models 90
Supercomputing Software 91
Operating Systems and Management Software 92
Programming Models, Programming Languages, and Tools 93
Libraries 96
Applications Software 96
Reliability and Fault Tolerance 98
Performance Estimation 98
Performance Benchmarks 99
Performance Monitoring 100
Performance Modeling and Simulation 100
Performance Estimation and the Procurement Process 102
The Imperative to Innovate and Barriers to Innovation 102
Systems Issues 102
Issues for Algorithms 103
An Example from Computational Fluid Dynamics 104
Software Issues 105

6 SUPERCOMPUTING INFRASTRUCTURES AND INSTITUTIONS 115
Supercomputing Ecosystem Creation and Maintenance 118
How Ecosystems Get Established 118
Potential Barriers for New Ecosystems 121
Ecosystem Workforce 124
Consumer Institutions 126
Supercomputing Centers 126
Industrial Supercomputing 127

7 SUPERCOMPUTING ABROAD 131
Japan 132
Similarities 132
Differences 132
The Earth Simulator 133
Other Japanese Centers 134
China 135
Europe 135
The United Kingdom 136
Germany 136
France 137
Spain 137
Application Software 137

Copyright © 2004 National Academy of Sciences. All rights reserved.
This executive summary plus thousands more available at http://www.nap.edu

8 A POLICY FRAMEWORK 140
The Government as the Leading User and Purchaser of Supercomputer Technology 140
Supercomputer Technology Investments as Public Goods 141
Potential Costs of Government Intervention 142
Alternative Modes for Government Intervention 143
Government Incentives 143
Government Research 144
Competing Government Objectives 146
Coordination Versus Diversification 146
Commitment Versus Flexibility 146
Secrecy Versus Spillovers 147

9 STEWARDSHIP AND FUNDING OF SUPERCOMPUTING 150
Satisfying Current Supercomputing Needs 151
Ensuring Future Supercomputing Leadership 152
The Need for Hardware and Software Producers 152
The Need for Stability 152
The Need for a Continuum from Research to Production 153
The Need for Money 155
The Need for People 157
The Need for Planning and Coordination 157
A Supercomputing Roadmap 158
Responsibility and Oversight 160

10 FUTURE OF SUPERCOMPUTING—CONCLUSIONS AND RECOMMENDATIONS 164
Conclusions 164
Recommendations 167

A COMMITTEE MEMBER AND STAFF BIOGRAPHIES 178

B SPEAKERS AND PARTICIPANTS AT MEETINGS AND SITE VISITS 187

C WHITE PAPERS PREPARED FOR THE APPLICATIONS WORKSHOP 197

D GLOSSARY AND ACRONYM LIST 199

Copyright © 2004 National Academy of Sciences. All rights reserved.

This executive summary plus thousands more available at http://www.nap.edu

Copyright © 2004 National Academy of Sciences. All rights reserved.
This executive summary plus thousands more available at http://www.nap.edu