Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017-2020

DETAILS

156 pages | 6 x 9 | PAPERBACK
ISBN 978-0-309-38961-7 | DOI 10.17226/21886

CONTRIBUTORS

Committee on Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science in 2017-2020; Computer Science and Telecommunications Board; Division on Engineering and Physical Sciences; National Academies of Sciences, Engineering, and Medicine

Future Directions for NSF ADVANCED COMPUTING INFRASTRUCTURE to Support U.S. Science and Engineering in 2017–2020

Committee on Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science in 2017-2020

Computer Science and Telecommunications Board

Division on Engineering and Physical Sciences

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

THE NATIONAL ACADEMIES PRESS
Washington, DC
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

This activity was supported by Award No. OCI-1344417 from the National Science Foundation. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

International Standard Book Number-13: 978-0-309-38961-7
International Standard Book Number-10: 0-309-38961-5
Digital Object Identifier: 10.17226/21886

Additional copies of this report are available for sale from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu.

Printed in the United States of America

Suggested citation: National Academies of Sciences, Engineering, and Medicine. 2016. *Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017-2020*. Washington, DC: The National Academies Press. doi:10.17226/21886.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Ralph J. Cicerone is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. C. D. Mote, Jr., is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the **National Academies of Sciences, Engineering, and Medicine** at **www.national-academies.org**.

OTHER RECENT REPORTS OF THE COMPUTER SCIENCE AND TELECOMMUNICATIONS BOARD

Privacy Research and Best Practices: Summary of a Workshop for the Intelligence Community (2016)

Bulk Collection of Signals Intelligence: Technical Options (2015)
Cybersecurity Dilemmas: Technology, Policy, and Incentives: Summary of Discussions at the 2014 Raymond and Beverly Sackler U.S.-U.K. Scientific Forum (2015)
Interim Report on 21st Century Cyber-Physical Systems Education (2015)
A Review of the Next Generation Air Transportation System: Implications and Importance of System Architecture (2015)
Telecommunications Research and Engineering at the Communications Technology Laboratory of the Department of Commerce: Meeting the Nation's Telecommunications Needs (2015)
Telecommunications Research and Engineering at the Institute for Telecommunication Sciences of the Department of Commerce: Meeting the Nation's Telecommunications Needs (2015)

At the Nexus of Cybersecurity and Public Policy: Some Basic Concepts and Issues (2014)
Emerging and Readily Available Technologies and National Security: A Framework for Addressing Ethical, Legal, and Societal Issues (2014)
Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017-2020: An Interim Report (2014)
Interim Report of a Review of the Next Generation Air Transportation System Enterprise Architecture, Software, Safety, and Human Factors (2014)

Geotargeted Alerts and Warnings: Report of a Workshop on Current Knowledge and Research Gaps (2013)
Professionalizing the Nation's Cybersecurity Workforce? Criteria for Future Decision-Making (2013)
Public Response to Alerts and Warnings Using Social Media: Summary of a Workshop on Current Knowledge and Research Gaps (2013)

Computing Research for Sustainability (2012)
Continuing Innovation in Information Technology (2012)
The Safety Challenge and Promise of Automotive Electronics: Insights from Unintended Acceleration (2012, with the Board on Energy and Environmental Systems and the Transportation Research Board)

The Future of Computing Performance: Game Over or Next Level? (2011)
Public Response to Alerts and Warnings on Mobile Devices: Summary of a Workshop on Current Knowledge and Research Gaps (2011)
Strategies and Priorities for Information Technology at the Centers for Medicare and Medicaid Services (2011)
Wireless Technology Prospects and Policy Options (2011)

Limited copies of CSTB reports are available free of charge from

Computer Science and Telecommunications Board
National Academies of Sciences, Engineering, and Medicine
Keck Center of the National Academies
500 Fifth Street, NW, Washington, DC 20001
(202) 334-2605/cstb@nas.edu
www.cstb.org

COMMITTEE ON FUTURE DIRECTIONS FOR NSF ADVANCED COMPUTING INFRASTRUCTURE TO SUPPORT U.S. SCIENCE IN 2017-2020

WILLIAM D. GROPP, University of Illinois, Urbana-Champaign, *Co-Chair*
ROBERT J. HARRISON, Stony Brook University, *Co-Chair*
MARK R. ABBOTT, Woods Hole Oceanographic Institution
ROBERT L. GROSSMAN, University of Chicago
PETER M. KOGGE, University of Notre Dame
PADMA RAGHAVAN, Pennsylvania State University
DANIEL A. REED, University of Iowa
VALERIE TAYLOR, Texas A&M University
KATHERINE A. YELICK, University of California, Berkeley

Staff

JON EISENBERG, Director, Computer Science and Telecommunications Board, and Study Director
SHENAE BRADLEY, Administrative Assistant

Preface

Advanced computing, a term used in this report to include both compute- and data-intensive capabilities, is used to tackle a rapidly growing range of challenging science and engineering problems. The National Science Foundation (NSF) requested that the National Academies of Sciences, Engineering, and Medicine carry out a study examining anticipated priorities and associated trade-offs for advanced computing in support of NSF-sponsored science and engineering research. The study encompasses advanced computing activities and programs throughout NSF, including, but not limited to, those of its Division of Advanced Cyberinfrastructure. The statement of task for the full study is given in Box P.1. In response to this request, the Academies established the Committee on Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science in 2017-2020 (see Appendix C).

The first phase of the study culminated in an interim report issued in 2014, *Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017-2020: An Interim Report*, that identified key issues and discussed potential options. The interim report set forth nine major areas where the committee sought input from the scientific computing community (Box P.2). The committee received over 60 comments from individuals, research groups, and organizations (listed in Appendix A) in response to its call for comments. It gathered further input through additional data-gathering sessions convened by the committee and listed in Appendix B. This is the committee's final report. As this study was being completed, an executive order was issued estab-

lishing a National Strategic Computer Initiative (NSCI), a measure that underscores the importance of advanced computing for the nation in general—and for science in particular. This report briefly discusses NSF's role in the NSCI; see Section 2.7 and Box 2.5.

William D. Gropp and Robert J. Harrison, *Co-Chairs*
Committee on Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science in 2017-2020

BOX P.1
Statement of Task

A study committee will examine anticipated priorities and associated trade-offs for advanced computing in support of National Science Foundation (NSF)-sponsored science and engineering research. Advanced computing capabilities are used to tackle a rapidly growing range of challenging science and engineering problems, many of which are compute-, communications-, and data-intensive as well. The committee will consider:

1. The contribution of high-end computing to U.S. leadership and competiveness in basic science and engineering and the role that NSF should play in sustaining this leadership;
2. Expected future national-scale computing needs: high-end requirements, those arising from the full range of basic science and engineering research supported by NSF, as well as the computing infrastructure needed to support advances in modeling and simulation as well as data analysis;
3. Complementarities and trade-offs that arise among investments in supporting advanced computing ecosystems; software, data, communications;
4. The range of operational models for delivering computational infrastructure, for basic science and engineering research, and the role of NSF support in these various models; and
5. Expected technical challenges to affordably delivering the capabilities needed for world-leading scientific and engineering research.

An interim report will identify key issues and discuss potential options. It might contain preliminary findings and early recommendations. A final report will include a framework for future decision making about NSF's advanced computing strategy and programs. The framework will address such issues as how to prioritize needs and investments and how to balance competing demands for cyberinfrastructure investments. The report will emphasize identifying issues, explicating options, and articulating trade-offs and general recommendations.

The study will not make recommendations concerning the level of federal funding for computing infrastructure.

BOX P.2
Questions Posed to the Scientific Community in the Committee's Interim Report

The committee explored and sought comments on the following:

- How to create advanced computing infrastructure that enables integrated discovery involving experiments, observations, analysis, theory, and simulation;
- Technical challenges to building future, more capable advanced computing systems and how NSF might best respond to them;
- The computing needs of individual research areas;
- How to balance resources and demand for the full spectrum of systems, for both compute- and data-intensive applications, and the impacts on the research community if NSF can no longer provide state-of-the-art computing for its research community;
- The role of private industry and other federal agencies in providing advanced computing infrastructure;
- The challenges facing researchers in obtaining allocations of advanced computing resources and suggestions for improving the allocation and review processes;
- Whether wider and more frequent collection of requirements for advanced computing could be used to inform strategic planning and resource allocation, how these requirements might be used, and how they might best be collected and analyzed;
- The tension between the benefits of competition and the need for continuity as well as alternative models that might more clearly delineate the distinction between performance review and accountability and organizational continuity and service capabilities; and
- How NSF might best set overall strategy for advanced computing-related activities and investments as well as the relative merits of both formal, top-down coordination and enhanced, bottom-up process.

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 Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its 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:

Daniel E. Atkins III, University of Michigan,
David A. Bader, Georgia Institute of Technology,
Robert Brammer, Brammer Technology, LLC,
Andrew A. Chien, University of Chicago,
Jeff Dozier, University of California, Santa Barbara,
Dennis Gannon, Microsoft Research (retired),
Gary S. Grest, Sandia National Laboratories,
Laura M. Haas, IBM,
Anthony ("Tony") John Grenville Hey, University of Washington eScience Institute,
David Keyes, King Abdullah University of Science and Technology,
Michael L. Klein, Temple University,
David A. Lifka, Cornell University,

Jeremiah P. Ostriker, Columbia University,
Terrence J. Sejnowski, Salk Institute for Biological Studies,
Marc Snir, Argonne National Laboratory,
Warren M. Washington, National Center for Atmospheric Research, and
John West, Texas Advanced Computing Center.

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 Marcia J. Rieke, University of Arizona, and Butler W. Lampson, Microsoft Research, who 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.

Contents

Summary

The National Science Foundation (NSF) asked the National Academies of Sciences, Engineering, and Medicine to provide a framework for future decision making about NSF's advanced computing strategy and programs. Advanced computing refers here to the advanced technical capabilities, including computer systems, software, and expert staff, that support a wide range of science and engineering research and that are of a large enough scale and cost that they are typically shared among multiple researchers, institutions, and applications. Advanced computing encompasses support for data-driven research as well as modeling and simulation.

The recommendations of the Committee on Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science in 2017-2020 are aimed at achieving four broad goals: (1) positioning the United States for continued leadership in science and engineering, (2) ensuring that resources meet community needs, (3) aiding the scientific community in keeping up with the revolution in computing, and (4) sustaining the infrastructure for advanced computing.

POSITION THE UNITED STATES FOR CONTINUED LEADERSHIP IN SCIENCE AND ENGINEERING

Large-scale simulation and the accumulation and analysis of massive amounts of data are revolutionizing many areas of science and engineering research. Increased advanced computing capability has historically

enabled new science, and many fields today rely on high-throughput computing for discovery. Modeling and simulation, the historical focus of high-performance computing, is a well-established peer of theory and experiment. Data-driven research, a complementary "fourth paradigm" for scientific discovery, needs data-intensive computing capabilities and resources. To support this research, NSF is a major provider of the advanced computing used for U.S. basic science, not only for its own grantees but also in support of research sponsored by other agencies, such as the National Institutes of Health and the Department of Energy.

Meeting future needs will require systems that support a wide range of advanced computing capabilities, including large-scale parallel systems and data-intensive systems. Approaches that combine large-scale computing and data resources in "converged" systems can play a role; more specialized systems may also be needed to meet some requirements. Commercial cloud computing offers certain advantages and can play a role in NSF's advanced computing strategy. However, NSF computing centers already exploit economies of scale and load sharing, and commercial cloud providers do not currently support very large, tightly coupled parallel applications, especially for high-end simulation workloads. For other applications, especially data-centric workloads and communities that share data sets, cloud computing is positioned today to play a growing role.

Recommendation 1. The National Science Foundation (NSF) should sustain and seek to grow its investments in advanced computing—to include hardware and services, software and algorithms, and expertise—to ensure that the nation's researchers can continue to work at frontiers of science and engineering.

Recommendation 1.1. NSF should ensure that adequate advanced computing resources are focused on systems and services that support scientific research. In the future, these requirements will be captured in its roadmaps.

Recommendation 1.2. Within today's limited budget envelope, this will mean, first and foremost, ensuring that a predominant share of advanced computing investments be focused on production capabilities and that this focus not be diluted by undertaking too many experimental or research activities as part of NSF's advanced computing program.

Recommendation 1.3. NSF should explore partnerships, both strategic and financial, with federal agencies that also provide advanced

computing capabilities, as well as federal agencies that rely on NSF facilities to provide computing support for their grantees.

Recommendation 2. As it supports the full range of science requirements for advanced computing in the 2017-2020 time frame, the National Science Foundation (NSF) should pay particular attention to providing support for the revolution in data-driven science along with simulation. It should ensure that it can provide unique capabilities to support large-scale simulations and/or data analytics that would otherwise be unavailable to researchers and continue to monitor the cost-effectiveness of commercial cloud services.

Recommendation 2.1. NSF should integrate support for the revolution in data-driven science into NSF's strategy for advanced computing by (a) requiring most future systems and services and all those that are intended to be general purpose to be more data-capable in both hardware and software, (b) expanding the portfolio of facilities and services optimized for data-intensive as well as numerically intensive computing, and (c) carefully evaluating inclusion of facilities and services optimized for data-intensive computing in its portfolio of advanced computing services.

Recommendation 2.2. NSF should (a) provide one or more systems for applications that require a single, large, tightly coupled parallel computer and (b) broaden the accessibility and utility of these large-scale platforms by allocating high-throughput as well as high-performance workflows to them.

Recommendation 2.3. NSF should (a) eliminate barriers to cost-effective academic use of the commercial cloud and (b) carefully evaluate the full cost and other attributes (e.g., productivity and match to science workflows) of all services and infrastructure models to determine whether such services can supply resources that meet the science needs of segments of the community in the most effective ways.

Maintaining leadership in advanced computing will be challenging. The resources available for advanced computing are inherently limited by research budgets, even as the demand for computing is growing and changing rapidly across the scientific enterprise and as the gap between supply and demand grows. If NSF is unable to increase or better leverage its resources for advanced computing, it seems inevitable that it will be unable to meet future demand for computational resources and will have

to reduce the size of the very largest research projects that are supported by its advanced computing facilities.

ENSURE THAT RESOURCES MEET COMMUNITY NEEDS

Despite various ongoing efforts to collect and understand requirements from some science communities and occasional efforts to chart strategic directions, the overall planning process for advanced computing resources and programs is not systematic or uniform and is not visibly reflected in NSF's strategic planning, despite its foundation-wide importance. The creation of an ongoing and more regular and structured process would make it possible to collect requirements, roll them up, and prioritize advanced computing investments based on science and engineering priorities.

Recommendation 3. To inform decisions about capabilities planned for 2020 and beyond, the National Science Foundation (NSF) should collect community requirements and construct and publish roadmaps to allow it to better set priorities and make more strategic decisions about advanced computing.

Recommendation 3.1. NSF should inform its strategy and decisions about investment trade-offs using a requirements analysis that draws on community input, information on requirements contained in research proposals, allocation requests, and foundation-wide information gathering.

Recommendation 3.2. NSF should construct and periodically update roadmaps for advanced computing that reflect these requirements and anticipated technology trends to help it set priorities and make more strategic decisions about science and engineering and to enable the researchers that use advanced computing to make plans and set priorities.

Recommendation 3.3. NSF should document and publish on a regular basis the amount and types of advanced computing capabilities that are needed to respond to science and engineering research opportunities.

Recommendation 3.4. NSF should employ this requirements analysis and resulting roadmaps to explore whether there are more opportunities to use shared advanced computing facilities to sup-

port individual science programs such as Major Research Equipment and Facilities Construction projects.

The roadmaps would reflect the visions of the science communities supported by NSF, including both large users and those (in the "long-tail") with more modest needs. The goal is to develop brief documents that set forth the overall strategy and approach rather than high-resolution details. They would look roughly 5 years ahead and provide a vision that extends about 10 years ahead. The roadmaps would help inform users about future facilities, guide investment, align future procurements and services with requirements, and enable more effective partnerships within NSF and with other federal agencies.

The roadmapping and requirements process could be strengthened by developing a better understanding of the relationships among requirements, the costs of different approaches (roadmap choices), and science benefits. Such information would inform program managers about the total cost of proposed research, help focus researcher attention on effective use of these valuable shared resources, and encourage more efficient software and research techniques.

Recommendation 4. The National Science Foundation (NSF) should adopt approaches that allow investments in advanced computing hardware acquisition, computing services, data services, expertise, algorithms, and software to be considered in an integrated manner.

Recommendation 4.1. NSF should consider requiring that all proposals contain an estimate of the advanced computing resources required to carry out the proposed work and creating a standardized template for collection of the information as one step of potentially many toward more efficient individual and collective use of these finite, expensive, shared resources. (This information would also inform the requirements process.)

Recommendation 4.2. NSF should inform users and program managers of the cost of advanced computing allocation requests in dollars to illuminate the total cost and value of proposed research activities.

AID THE SCIENTIFIC COMMUNITY IN KEEPING UP WITH THE REVOLUTION IN COMPUTING

Computer architectures are changing rapidly along with programming models to use the hardware, creating challenges for the science

community, which depends on and has invested significantly in science codes written for yesterday's systems. The rise of data-intensive science brings with it new software and systems. Better software tools, technical expertise, and more flexible service models (ways of delivering software and computing resources) can improve the productivity of researchers both today and in the future.

Recommendation 5. The National Science Foundation (NSF) should support the development and maintenance of expertise, scientific software, and software tools that are needed to make efficient use of its advanced computing resources.

Recommendation 5.1. NSF should continue to develop, sustain, and leverage expertise in all programs that supply or use advanced computing to help researchers use today's advanced computing more effectively and prepare for future machine architectures.

Recommendation 5.2. NSF should explore ways to provision expertise in more effective and scalable ways to enable researchers to make their software more efficient; for instance, by making more pervasive the XSEDE (Extreme Science and Engineering Discovery Environment) practice that permits researchers to request an allocation of staff time along with computer time.

Recommendation 5.3. NSF should continue to invest in and support scientific software and update the software to support new systems and incorporate new algorithms, recognizing that this work is not primarily a research activity but rather is support of software infrastructure.

If NSF was to invest solely in production, it would miss some key technology shifts and its facilities would quickly become obsolete. By taking a leadership role in defining future advanced computing capabilities and helping researchers use them more effectively, NSF can help ensure that its software and systems remain relevant to its science portfolio, that researchers are prepared to use the systems, and that investments across the foundation are aligned with this future.

Recommendation 6. The National Science Foundation (NSF) should also invest modestly to explore next-generation hardware and software technologies to explore new ideas for delivering capabilities that can be used effectively for scientific research, tested, and transitioned

into production where successful. Not all communities will be ready to adopt radically new technologies quickly, and NSF should provision advanced computing resources accordingly.

SUSTAIN THE INFRASTRUCTURE FOR ADVANCED COMPUTING

Expertise and other long-lived assets, such as the physical infrastructure for computing centers, are an essential part of a robust and sustainable advanced cyberinfrastructure. In recent years, NSF has adopted a strategy for acquiring computing facilities and creating centers and programs to operate and support them that relies on irregularly scheduled competition among host institutions roughly every 2 to 5 years and on equipment, facility, and operating cost sharing with those institutions. Mounting costs and budget pressures suggest that a strategy that relies on state, institutional, or vendor cost sharing may no longer be viable. Repeated competition can lead to proposals designed to win a competition rather than maximize scientific returns. Moreover, it is important to ensure the development and retention of the talent that is needed to effectively manage systems, support users, and evolve software to make effective use of today's and tomorrow's architectures.

Recommendation 7. The National Science Foundation (NSF) should manage advanced computing investments in a more predictable and sustainable way.

Recommendation 7.1. NSF should consider funding models for advanced computing facilities that emphasize continuity of support.

Recommendation 7.2. NSF should explore and possibly pilot the use of a special account (such as that used for Major Research Equipment and Facilities Construction) to support large-scale advanced computing facilities.

Recommendation 7.3. NSF should consider longer-term commitments to center-like entities that can provide advanced computing resources and the expertise to use them effectively in the scientific community.

Recommendation 7.4. NSF should establish regular processes for rigorous review of these center-like entities and not just their individual procurements.

Managing its advanced computing investments in a more predictable and sustainable way, as it does for other long-term (10 years or more) infrastructure, not only would benefit the researchers currently supported by NSF's advanced computing programs, but also would provide opportunities to apply the same expertise more broadly within NSF, such as the large-scale science projects that have long-term needs for advanced computing. It would also create new opportunities for NSF's advanced computing programs to address long-term storage, preservation, and curation challenges for data.

1

Overview and Recommendations

The National Science Foundation (NSF) requested that the National Academies of Sciences, Engineering, and Medicine carry out a study examining anticipated priorities and associated trade-offs for advanced computing in support of NSF-sponsored science and engineering research. In this study, advanced computing is defined as the advanced technical capabilities, including both computer systems and expert staff, that support research across the entire science and engineering spectrum and that are of a scale and cost so great that they are typically shared among multiple researchers, institutions, and applications.[1] As used here, the term encompasses support for data-driven research as well as modeling and simulation.[2] Data have always been an important element of advanced computing, but the emergence of "big data" has created new opportunities for research and stimulated new demand for data-intensive capabilities. The scope of the study encompasses advanced computing activities and programs throughout NSF, including, but not limited to,

[1] Also critical to NSF-supported advanced computing activities are wide-area and campus networks, which provide access and the infrastructure necessary to bring together data sources and computing resources where they cannot practically be colocated. Both types of networks have been supported by NSF programs. Understanding future networking needs would involve examination of a much wider range of activities across NSF—not just advanced computing, including many aspects of cyberinfrastructure, but also planned major experimental facilities—and is therefore not addressed in this report.

[2] Throughout this report, "computing" should be read broadly as encompassing data analytics and other data-intensive applications as well as modeling and simulation and other numerically intensive or symbolic computing applications.

those of its Division of Advanced Cyberinfrastructure. The statement of task for the Committee on Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science in 2017-2020 is given in Box P.1. This final report from the study follows the committee's interim report issued in 2014.[3]

The committee's recommendations are aimed at achieving four broad goals: (1) position the United States for continued leadership in science and engineering, (2) ensure that resources meet community needs, (3) aid the scientific community in keeping up with the revolution in computing, and (4) sustain the infrastructure for advanced computing.

1.1 POSITION THE UNITED STATES FOR CONTINUED LEADERSHIP IN SCIENCE AND ENGINEERING

NSF's investments in advanced computing are critical enablers of the nation's science leadership. Advanced computing at NSF has been used to understand the formation of the first galaxies in the early universe and to analyze the impacts of cloud-aerosol-radiation on regional climate change. Advanced computing has been a key to award-winning science, including the 2011 Nobel Prize in physics and the 1998 and 2013 Nobel Prizes in chemistry (see Box 3.2). Its use has moved outside of traditional areas of science to understanding social phenomenon captured in real-time video streams and the connection properties of social networks.

Large-scale simulation, the accumulation and analysis of massive amounts of data, and other forms of advanced computing are all revolutionizing many areas of science and engineering research. Modeling and simulation, the historical focus of high-performance computing systems and programs, is a well-established peer of theory and experimentation. Increased capability has historically enabled new science, and many fields increasingly rely on high-throughput computing.

Data-driven research has emerged as a complementary "fourth paradigm" for scientific discovery[4] that needs data-intensive computing capabilities and resources configured for the transfer, search, analysis, and management of scientific data, often under real-time constraints. Even in modeling and simulation applications, data-intensive aspects are increasingly important as large data sets are produced by or incorporated into the simulations. Both data-driven and computationally driven sci-

[3] National Research Council, *Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017-2020: An Interim Report*, The National Academies Press, Washington, D.C., 2014.

[4] J. Gray, T. Hey, S. Tansley, and K. Tolle, "Jim Gray on eScience: A Transformed Scientific Method," in *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, Redmond, Wash., 2009.

entific processes involve a range of algorithms and workflows that may be compute-intensive or bandwidth-intensive, making simple machine characterizations difficult, especially given that science and engineering discovery frequently integrates all of these. As a result, leadership in frontier science also requires that the United States maintain leadership in both simulation science and data-driven science.

NSF has been very successful in making advanced computing resources, especially in support of modeling and simulation, available to an expanding set of disciplines supported by NSF, and has an opportunity to assert similar leadership in data-driven science. NSF is a major provider of computing support for the nation's science enterprise, not just for the research programs it directly supports. For example, about half of the computer resources allocated under the Extreme Science and Engineering Discovery Environment (XSEDE) program are to non-NSF-supported researchers, including 14 percent for work supported by the National Institutes of Health. Moreover, the science and engineering community and other federal agencies that support scientific research look to NSF to provide leadership and to play crucial roles in developing and applying advanced computing, including advancing the intellectual foundations of computation, creating practical tools, and developing the workforce.

An exponential rate of growth in demand is now observed that is outpacing the rate of growth in advanced computing resources. At the same time, the cost of provisioning facilities has risen because demand is rising faster than technology improvements are now able to deliver at fixed price. The rise in data-driven science and increasing need for both numerically intensive and data-intensive capabilities (Recommendation 2) create further demand for resources.

Production support is needed for software (including pre-installed popular applications and libraries) as well as hardware, to include community software as well as frameworks, shared elements, and other supporting infrastructure. NSF's Software Infrastructure for Sustained Innovation (SISI) program is a good foundation for such investments. However, SISI needs to be grown in partnership with NSF's science directorates to a scale that matches need, and then be sustained essentially indefinitely; the United Kingdom's Collaborative Computational Projects (CCPs) provide examples of the impact and successful operation of community-led activities that now span nearly four decades. Production support is further needed for data management. Curation, preservation, archiving, and support for sharing all need ongoing investment.

Recommendation 1. The National Science Foundation (NSF) should sustain and seek to grow its investments in advanced computing—to include hardware and services, software and algorithms, and exper-

tise—to ensure that the nation's researchers can continue to work at frontiers of science and engineering.

An important element of fulfilling its role of maintaining the nation's science leadership and achieving the vision in NSF's Cyberinfrastructure Framework for 21st Century Science is providing the research community with access to the needed advanced computing capabilities. This will include

- Providing access to sufficient computing facilities and services to support NSF's portfolio of science and engineering research, including both aggregate capacity and large-scale parallel computers and software systems;
- Assuming leadership in providing access to general-use hardware and software that integrate support for data-driven science as well as large hardware and software systems focused on data-driven science; and
- Assuming leadership for data-driven science, first by integrating support for data-driven science into most or all of the systems it provides support for on behalf of the research community and next by deploying advanced computing systems focused on data-driven science.

Recommendation 1.1. NSF should ensure that adequate advanced computing resources are focused on systems and services that support scientific research. In the future, these requirements will be captured in its roadmaps.

Recommendation 1.2. Within today's limited budget envelope, this will mean, first and foremost, ensuring that a predominant share of advanced computing investments be focused on production capabilities and that this focus not be diluted by undertaking too many experimental or research activities as part of NSF's advanced computing program.

Recommendation 1.3. NSF should explore partnerships, both strategic and financial, with federal agencies that also provide advanced computing capabilities, as well as federal agencies that rely on NSF facilities to provide computing support for their grantees.

Today's landscape for advanced computing is far richer in terms of an expanding range of needs and in terms of technical opportunities for meeting those needs. Key elements of this landscape include the following:

- Scientists supported by NSF advanced computing increasingly include a "long tail" of users with more modest requirements for advanced computing than those with research applications that require parallel computers with a large number of tightly coupled processors. The latter applications cannot be run (or run with acceptable efficiency) on smaller systems or on current commercial cloud systems.
- Increased capability has historically enabled new science (see examples in Box 3.1). Without at least some growth in capability, researchers pursuing science that requires capability computing will have difficulty making advances.
- Many fields increasingly rely on high-throughput computing that requires a greater aggregate amount of computing than a typical university can be expected to provide. Such applications can be run efficiently on both large and medium-size machines. Although a large-scale system can run many smaller jobs with good efficiency, systems capable of running only smaller jobs cannot run large-scale jobs with acceptable efficiency. It is not necessary or more efficient to restrict large, tightly coupled systems to run only large, highly scalable applications. Modestly sized jobs may still require tight connections, even though at smaller scale, and the utilization of large systems is improved with a mixture of job sizes.
- The rise in the volume and diversity of scientific data represents a significant disruption and opportunity for science and engineering and for advanced computing. Data-intensive advanced computing represents a significant opportunity for U.S. science and engineering leadership. Some data-intensive applications can be accommodated in more data-capable general-purpose platforms; other applications will require specifically configured systems. Supporting data-driven science also places additional demands on wide-area networking to share scientific data and raises challenges around long-term storage, preservation, and curation. It also requires diverse and hard-to-find expertise.
- Large systems are more accessible to a larger group of users; both cloud technologies and science gateways lower the barriers to access applications at scale.
- Cloud computing has shown that access can be "democratized": many users can access a large system for small amounts of total time in a fashion not supported by current approaches to allocating supercomputer time. Moreover, cloud computing users can leverage extensive libraries of software tools developed by both commercial providers and individual scientists. In many ways, this ability for a far larger community to access the power of large-scale systems, whether it is a conventional supercomputer or a commercial cloud configured to support some aspects of scientific discovery, represents a qualitative change in the computing landscape. However, NSF computing centers already exploit economies of

scale and load sharing, and commercial cloud providers do not currently support very large, tightly coupled parallel applications, especially for high-end simulation workloads. However, this area is under rapid development, and the price (i.e., cost to NSF) and types of services are likely to change. The cost of commercial cloud services could be greatly reduced by reducing or eliminating the overhead charged on these services, bulk purchase by NSF of cloud resources, and/or partnering with commercial cloud providers.

The greater complexity of the landscape means that it will be especially important, as recommended in Section 1.2, to derive future requirements for advanced computing platforms from an analysis of science needs, workload characteristics, and priorities.

To maximize performance, NSF could deploy systems that were optimal for each class of problem. But as a practical matter and for cost-effectiveness, NSF must secure access to capabilities that will represent compromises with respect to individual applications but reasonably support the overall research portfolio. Put another way, it will require careful resource management driven by an understanding of the science and engineering returns on investments in advanced computing. Understanding which compromises to make requires a comprehensive understanding of science requirements and priorities; see the discussion of requirements and roadmapping below.

Recommendation 2. As it supports the full range of science requirements for advanced computing in the 2017-2020 time frame, the National Science Foundation (NSF) should pay particular attention to providing support for the revolution in data-driven science along with simulation. It should ensure that it can provide unique capabilities to support large-scale simulations and/or data analytics that would otherwise be unavailable to researchers and continue to monitor the cost-effectiveness of commercial cloud services.

Recommendation 2.1. NSF should integrate support for the revolution in data-driven science into NSF's strategy for advanced computing by (a) requiring most future systems and services and all those that are intended to be general purpose to be more data-capable in both hardware and software, (b) expanding the portfolio of facilities and services optimized for data-intensive as well as numerically intensive computing, and (c) carefully evaluating inclusion of facilities and services optimized for data-intensive computing in its portfolio of advanced computing services.

To support data-driven science, advanced computing hardware and software systems will need adequate data capabilities, in most cases more than is currently provided. Some research will need large-scale data-centric systems with data-handling capabilities that are quite different from traditional high-performance computing systems. For example, data analytics often requires that data reside on disk for extended periods. Several factors suggest that meeting these needs will require one or more large investments, rather than just multiple small projects, including the following: (1) the scale of the largest problems, (2) the opportunities for new science when disparate data sets are colocated, and (3) the cost efficiencies that come from consolidating facilities. Indeed, the growth in data-driven science suggests that investments will ultimately be needed on a scale comparable to those that support modeling and simulation At the very least, the systems should be better balanced for data (input/output and perhaps memory size), thereby allowing the same systems to be used for different problems without needing to double the size of the resources. As data play a growing role in scientific discovery, long-term data management will become an important aspect of all planning for advanced computing. A partnership with a commercial cloud provider could provide access to larger systems than NSF could afford to deploy on its own. Of course, even as it moves to provide better support for data-driven research, NSF cannot neglect simulation and modeling research.

> **Recommendation 2.2.** NSF should (a) provide one or more systems for applications that require a single, large, tightly coupled parallel computer and (b) broaden the accessibility and utility of these large-scale platforms by allocating high-throughput as well as high-performance workflows to them.

Simply meeting current levels of demand will require continuing to provide at least the capacity currently provided by the XSEDE program and the capability currently provided by Blue Waters. Even as NSF develops its future requirements (Recommendation 3) that can be used to develop long-term plans, the observed growth in demand suggests that some growth be included in NSF's short-term plans.

> **Recommendation 2.3.** NSF should (a) eliminate barriers to cost-effective academic use of the commercial cloud and (b) carefully evaluate the full cost and other attributes (e.g., productivity and match to science workflows) of all services and infrastructure models to determine whether such services can supply resources that meet the science needs of segments of the community in the most effective ways.

For 2020 and beyond, many of these recommendations may well still hold true, but NSF should rely on the requirements process outlined in the next section.

1.2 ENSURE RESOURCES MEET COMMUNITY NEEDS

At a time when resources are tight and demand for advanced computing resources continues to grow, it is especially important for NSF to maximize the return on investment in terms of science and engineering outcomes by improving the efficiency of advanced computing facility use. One part of this is ensuring that the resources provided match the requirements of the science applications, and this aspect is discussed separately below; another is to ensure that the resources are effectively used. How NSF can help the community use the computing infrastructure effectively is discussed in Sections 1.3 and 1.4.

The resources available for advanced computing are inherently limited by research budgets as compared to the potentially ever-expanding demand for advanced computing. Despite various ongoing efforts to collect and understand requirements from some science communities and occasional efforts to chart strategic directions, the overall planning process for advanced computing resources and programs is not systematic or uniform and is not visibly reflected in NSF's strategic planning, despite its foundation-wide importance. Further, much of what quantification there is makes use of measurements related to floating-point performance; this is misleading both because the performance of many applications is not well modeled using just floating-point performance and because the sustained as opposed to peak performance of some processors (especially most highly parallel processors) is low on many of those applications.

The creation of an ongoing and more regular and structured process would make it possible to collect requirements, roll them up, and prioritize advanced computing investments based on science and engineering priorities. It would reflect the visions of science communities and support evaluation of potential scientific advances, probability of success, advanced computing requirements and their costs, and their affordability. Such a process needs to be nimble enough to respond to new science opportunities and computing technologies but have a long-enough time horizon to provide continuity and predictability to both users and resource providers. The process also needs to involve the growing body of researchers from a growing number of disciplines who use NSF infrastructure.

Requirements established for future systems and services must also address trade-offs—for example, within a given budget envelope for hardware, more memory implies less compute or input/output capacity.

The criteria established for future procurements should reflect scientific requirements rather than simplistic or unrepresentative benchmarks.

One way to capture requirements and enable the science community to participate in the process is to establish roadmaps. Roadmaps do not suggest a single path to a destination but rather multiple routes to a variety of goals. Such roadmaps would help make science requirements concrete and relate them to future computing capabilities and facilitate planning by researchers, program directors, and facility and service operators at centers and on campuses over a longer time horizon. By capturing anticipated technology trends, the roadmaps can also provide guidance to those responsible for scientific software projects. The roadmaps can also address dependencies between investments by federal agencies through consultation with agencies that use NSF advanced computing facilities or provide computing to the NSF-supported research community.

The goal is to develop fairly brief documents that set forth the overall strategy and approach rather than high-resolution details, looking roughly 5 years ahead with a vision that extends perhaps for 10 years ahead. Roadmaps would help inform users about future facilities, guide investment, align future procurements with requirements and services, and enable more effective partnerships within NSF and with other federal agencies. If researchers are given information about the capabilities they can expect, they can make better plans for their future research and the software to support it. By describing what types of resources NSF will and will not provide, roadmaps would permit other agencies, research institutions, and individual principal investigators to make complementary plans for investments. They would also encourage reflection within individual science communities about their future needs and the challenges and opportunities that arise from future computing technologies. By establishing predictability over longer timescales, roadmaps would help those proposing or managing major facilities to rely on shared advanced computing resources, helping reduce the overall costs of advanced computing. The provision in 2015 of such a roadmap for the Department of Energy (DOE) by its Office of Advanced Scientific Computing Research has already enabled the community and science programs to direct their investments and software development efforts toward systems that, in some detail, they know will appear in 2018-2019 and, in less detail, toward a path that extends into the exascale era of around 2023 and beyond. The NSF academic community presently lacks this ability to plan.

The roadmapping process would also be an opportunity to address data curation and storage requirements and link them to individual programs developing data capabilities such as the Big Data Regional Innovation Hubs. In essence, it could provide ingredients of an NSF-wide data

plan that supports the needs of NSF's grantees and the science communities NSF supports.

Requirements-setting and roadmapping efforts could be built or modeled on activities undertaken to define requirements for large scientific facilities such as the Academies' astronomy and astrophysics decadal surveys or DOE's Particle Physics Project Prioritization Panel. However, the requirements will need to be aggregated at a higher level given that advanced computing facilities generally serve many scientific disciplines. In addition, because of the wide use of computing and data at all scales of resources, it is critical that any such requirements gathering include input from the whole community, including those with more modest (midrange) computing and data needs. Sometimes called "the long tail of science," these users have more modest requirements (but still beyond that available in a group, departmental, or campus system) and make up the majority of researchers.

Recommendation 3. To inform decisions about capabilities planned for 2020 and beyond, the National Science Foundation (NSF) should collect community requirements and construct and publish roadmaps to allow it to better set priorities and make more strategic decisions about advanced computing.

Recommendation 3.1. NSF should inform its strategy and decisions about investment trade-offs using a requirements analysis that draws on community input, information on requirements contained in research proposals, allocation requests, and foundation-wide information gathering.

Recommendation 3.2. NSF should construct and periodically update roadmaps for advanced computing that reflect these requirements and anticipated technology trends to help it set priorities and make more strategic decisions about science and engineering and to enable the researchers that use advanced computing to make plans and set priorities.

Recommendation 3.3. NSF should document and publish on a regular basis the amount and types of advanced computing capabilities that are needed to respond to science and engineering research opportunities.

Recommendation 3.4. NSF should employ this requirements analysis and resulting roadmaps to explore whether there are more opportunities to use shared advanced computing facilities to sup-

port individual science programs such as Major Research Equipment and Facilities Construction projects.

The roadmapping and requirements process could be strengthened by developing a better understanding of the relationship among the cost of different approaches (roadmap choices), requirements, and science benefits. For example, the information would inform program managers about the total cost of proposed research and help focus researchers' attention on effective use of these valuable shared resources, encouraging more efficient software and research techniques. NSF's XSEDE program has adopted this practice, which could be expanded to cover all aspects of NSF-supported advanced computing including campus-level resources.

Recommendation 4. The National Science Foundation (NSF) should adopt approaches that allow investments in advanced computing hardware acquisition, computing services, data services, expertise, algorithms, and software to be considered in an integrated manner.

Recommendation 4.1. NSF should consider requiring that all proposals contain an estimate of the advanced computing resources required to carry out the proposed work and creating a standardized template for collection of the information as one step of potentially many toward more efficient individual and collective use of these finite, expensive, shared resources. (This information would also inform the requirements process.)

Recommendation 4.2. NSF should inform users and program managers of the cost of advanced computing allocation requests in dollars to illuminate the total cost and value of proposed research activities.

1.3 AID THE SCIENTIFIC COMMUNITY IN KEEPING UP WITH THE REVOLUTION IN COMPUTING

However, even with a good match to the science requirements, getting the most out of modern computing systems is difficult. Better software tools and more flexible service models (ways of delivering software and computing resources) can improve the productivity of researchers.

Improvements to software and new algorithms can often significantly reduce computational and data-processing demands. One class of improvements increases performance on current computer architectures; another takes better advantage of new architectures. There is considerable

uncertainty about future architectural directions for computing in general and for advanced computing for science and engineering specifically. Architectures are already changing in response to power density issues, which have had limited clock speed growth since 2004, even as transistor density continued to grow. As a result, the creation and evolution of software for scientific applications have become more difficult, especially for those problems that do not readily lend themselves to massive parallelism.

The service model, application programming interfaces, and software stacks offered by cloud computing complement the existing supercomputing batch models and software stacks. Both the economics and applicability across the full range of science applications will need careful examination.

Production support is needed for software as well as hardware, to include community software as well as frameworks, shared elements, and other supporting infrastructure. NSF's SISI program is a good foundation for such investments. Production support is further needed for data management. Curation, preservation, archiving, and support for sharing all need ongoing investment.

Recommendation 5. The National Science Foundation (NSF) should support the development and maintenance of expertise, scientific software, and software tools that are needed to make efficient use of its advanced computing resources.

Recommendation 5.1. NSF should continue to develop, sustain, and leverage expertise in all programs that supply or use advanced computing to help researchers use today's advanced computing more effectively and prepare for future machine architectures.

Recommendation 5.2. NSF should explore ways to provision expertise in more effective and scalable ways to enable researchers to make their software more efficient; for instance, by making more pervasive the XSEDE (Extreme Science and Engineering Discovery Environment) practice that permits researchers to request an allocation of staff time along with computer time.

Recommendation 5.3. NSF should continue to invest in supporting science codes and in continuing to update them to support new systems and incorporate new algorithms, recognizing that this work is not primarily a research activity but rather is support of software infrastructure.

If NSF was to invest solely in production, it would miss some key technology shifts and its facilities would quickly become obsolete. Some innovation takes the form of fine-tuning of production systems, yet nontrivial but small investments in exploratory or experimental facilities and services are also needed to create, anticipate, and prepare for technology disruptions. NSF needs to play a leadership role in both defining future advanced computing capabilities and enabling researchers to effectively use those systems. This is especially true in the current hardware environment, where architectures are diverging in order to continue growing computing performance. Leadership by NSF will help ensure that its software and systems remain relevant to its science portfolio, that researchers are prepared to use the systems, and that investments across the foundation are aligned with this future.

It will be especially important for NSF to be not only engaged in but helping to lead the national and international activities that define and advance future software ecosystems that support simulation and data-driven science, including converging the presently distinct tools and programming paradigms, and the software required for exascale hardware technologies. NSF may be especially well positioned to collaborate internationally compared to the mission science agencies given its long track record of open science collaboration. DOE is currently investing heavily in new exascale programming tools that, through the scale of investment and buy-in from system manufacturers, could plausibly define the future of advanced programming even though the design may not reflect the needs of all NSF science because the centers and researcher communities it supports are not formally engaged in the specification process. It is also important for NSF to be engaged with the private sector and academia for insights into data analytics.

Recommendation 6. The National Science Foundation (NSF) should also invest modestly to explore next-generation hardware and software technologies to explore new ideas for delivering capabilities that can be used effectively for scientific research, tested, and transitioned into production where successful. Not all communities will be ready to adopt radically new technologies quickly, and NSF should provision advanced computing resources accordingly.

Investments by other federal agencies in new computing technologies and NSF's own computing research programs will both be sources of advanced hardware and software architectures to consider adopting in NSF's advanced computing programs. Achieving continued growth in NSF's aggregate computing performance on a fixed budget will likely require new architectural models that are more energy efficient. The

requirements-gathering and roadmapping process can be used to obtain long-term predictions of available capabilities and their energy requirements. That process will also provide insights into which technology advances are suitable for future production services.

1.4 SUSTAIN THE INFRASTRUCTURE FOR ADVANCED COMPUTING

A hard but essential part of managing advanced computing in a fixed budget envelope will be discontinuing activities in order to start or grow other activities. The requirements analysis will provide a rational and open basis for these decisions, and the roadmaps will enable communities to plan and adapt in advance of future investments. Even in a favorable budget environment for science and engineering generally and for advanced computing specifically, NSF will need to manage exponentially growing demand and rising costs (see Section 1.1).

One response to these challenges is to take advantage of the opportunities described in Sections 1.2 and 1.3 to increase efficiency and productivity in the use of advanced computing facilities, to use the requirements process to inform trade-offs, and to exploit new technologies. In addition, there are several possibilities for finding more or better leveraging resources. These include the following:

- Making a case for additional resources based on the requirements analysis. For example, the 2003 report *A Science-Based Case for Large-Scale Simulation*[5] is widely credited with developing the rationale and science case for a major expansion of DOE's Advanced Scientific Computing Research program. It may also be useful to look retrospectively at what computing capabilities were needed to achieve past science breakthroughs.
- Seeking funding mechanisms that ensure consistent and stable investments in advanced computing.
- Adopting approaches that make it easier to accommodate the costs of large facilities within annual budgets, such as leasing to smooth costs across budget years.
- Exploring partnerships, both strategic and financial, with federal agencies that also provide advanced computing capabilities as well as federal agencies that rely on NSF facilities to provide computing support for their grantees. For example, NSF might enter into a financial agreement with other federal (or possibly private) providers of advanced computing

[5] Office of Science, U.S. Department of Energy, *A Science-Based Case for Large-Scale Simulation*, Washington, D.C., 2003.

services for access to a fraction of a large system, thus maintaining some ability to support research that involves a large system without incurring its full acquisition cost.

Chapter 7 of this report provides more details on these options and discusses several others for NSF to consider.

In recent years, NSF has adopted a strategy for acquiring computing facilities and creating centers and programs to operate and support them that relies on irregularly scheduled competition among host institutions roughly every 2 to 5 years and on equipment, facility, and operating cost sharing with those institutions. Mounting costs and budget pressures suggest that a strategy that relies on state, institutional, or vendor cost sharing may no longer be viable. Moreover, there are reasons to consider models that provide a longer funding horizon for service providers that operate facilities and the expertise needed for their effective utilization.

In particular, one key reason is to ensure the development and retention of the advanced computing expertise that is needed to effectively manage systems, support their users, address the increasing complexity of hardware and software, and manage the needed transition of software to make effective use of today's and tomorrow's architectures. Doing so requires sustained attention to the workforce and more viable career pathways for its members. A longer funding horizon would also better match the depreciation period for buildings, power, and cooling infrastructure.

Another reason to consider other models is that repeated competition can lead to proposals designed to win a competition rather than maximize scientific returns. For example, it can unduly favor unproven technology over more proven, production-quality technology. By contrast, a model with longer time horizons may be better positioned to deliver systems that meet the scientific requirements established by the requirements definition and roadmapping activities. Supporting at least two entities will provide healthy competition as well as stability. The acquisition of individual systems from commercial vendors would remain competitive. Such longer-term entities can take the form of distributed organizations; XSEDE, for example, has evolved in this direction in providing the scientific research community with expertise and services.

A longer funding horizon would also better match the duration of major scientific facilities and the useful lifetime of scientific data, creating new opportunities to address long-term challenges of storage, preservation, and curation. Greater continuity would also foster greater leveraging of advanced computing expertise and facilities across NSF. For instance, long-lived experimental or observational facilities could better manage the risk of standing up their own cyberinfrastructure by partnering with centers.

Recommendation 7. The National Science Foundation (NSF) should manage advanced computing investments in a more predictable and sustainable way.

Recommendation 7.1. NSF should consider funding models for advanced computing facilities that emphasize continuity of support.

Recommendation 7.2. NSF should explore and possibly pilot the use of a special account (such as that used for Major Research Equipment and Facilities Construction) to support large-scale advanced computing facilities.

Recommendation 7.3. NSF should consider longer-term commitments to center-like entities that can provide advanced computing resources and the expertise to use them effectively in the scientific community.

Recommendation 7.4. NSF should establish regular processes for rigorous review of these center-like entities and not just their individual procurements.

2

Background

2.1 STUDY TASK AND SCOPE

The National Science Foundation (NSF) requested that the National Academies of Sciences, Engineering, and Medicine carry out a study examining anticipated priorities and associated trade-offs for advanced computing in support of NSF-sponsored science and engineering research. The scope of the study encompasses advanced computing activities and programs throughout NSF, including, but not limited to, those of its Division of Advanced Cyberinfrastructure. The statement of task for the study is given in Box P.1. This final report from the study follows the committee's interim report issued in 2014.[1]

In this study, advanced computing is defined as the advanced technical capabilities, including both computer systems and expert staff, that support research across the entire science and engineering spectrum and that are so large in scale and so expensive that they are typically shared among multiple researchers, institutions, and applications. The term also encompasses higher-end computing for which there are economies of scale in establishing shared facilities rather than having each institution acquire, maintain, and support its own systems. At the midscale, the demarcation between institutional and NSF responsibility is not well established (Box 2.1). For compute-intensive research, it includes not

[1] National Research Council, *Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017-2020: An Interim Report*, The National Academies Press, Washington, D.C., 2014.

BOX 2.1
Who Is Responsible for Midscale Computing Infrastructure?

One of the consequences of the exponential growth in computing power is that today's smart phones are more powerful than supercomputers of decades ago. For many researchers, a laptop or desktop system provides all of the computing power that they might need. Other researchers may need slightly more, while others depend on the capabilities only available in current supercomputer systems. Who should be responsible for providing this computing infrastructure?

At the very high end (national scale in terms of cost, operation, and use), the computing infrastructure is like other national-scale research facilities and supports research that is not possible without it. At the very low end (individual desktops or laptops), it can be argued that this should now be the responsibility of individual institutions, just like the other basic research support that they provide. What about the midrange? How capable of a system should individual institutions or regional consortia be expected to provide for their researchers? What about researchers who need large amounts of computing in the aggregate, but where each individual run could be done on a small machine?

Some institutions are already providing significant computing resources for their researchers; this is often viewed as a competitive advantage both in attracting and retaining faculty and staff and in winning grants. But many institutions, notably public universities, are finding their budgets squeezed. Others are creating ways for their researchers to pool funds into a shared computing infrastructure (creating what in many ways is a private cloud), which may also be partly supported by institutional funds.

As the National Science Foundation (NSF) considers how it supports advanced computing, it will need to consider how much computing is the responsibility of the institution, how much may be supported at individual institutions and regional consortia (in part through grants from NSF or other agencies), and how much is provided as a national resource. This is a complex issue, and one that will require more study and engagement with stakeholders. Among the issues to consider are the following:

- How best to take advantage of economies of scale;
- How to ensure that all researchers, not just those at the best-funded research institutions, have access to the computing resources needed for their research;
- How to avoid wasted or unused cycles and ensure systems are well-managed and secure;
- How to ensure that the systems match the needs of researcher—that is, their configuration provides data and compute capabilities needed by the software used by the researchers, and the network connectivity provides sufficient access to the system for all collaborators; and
- How to encourage and help institutions to provide a basic level of computing support, taking advantage of ways to share infrastructure and expertise.

The requirements analysis recommended by the committee (see Chapter 4) will provide valuable data in addressing these issues.

only today's supercomputers, which are able to perform more than 10^{15} floating-point operations per second (known as "petascale"), but also high-performance computing (HPC) platforms that share the same components as supercomputers but may have lower levels of performance. As used here, the term encompasses support for data-intensive research that involves analyzing terabytes (and increasingly petabytes) of data as well as modeling and simulation.

Historically, and even now, NSF advanced computing centers have focused on high-performance computing primarily for simulation. Although these applications are essential and growing, the new and very rapidly growing demand for more data-capable services still needs to be addressed. This chapter looks chiefly at traditional HPC, while the new opportunities and challenges of the "data revolution" are emphasized in Chapters 4 and 5.

2.2 PAST STUDIES OF ADVANCED COMPUTING FOR SCIENCE

In the early 1980s, the science community developed several reports regarding the lack of access to advanced computing resources. The 1982 report *Large-Scale Computing in Science and Engineering*, known as the "Lax report,"[2] was jointly sponsored by the Department of Defense (DOD) and NSF, with cooperation from the Department of Energy (DOE) and the National Aeronautics and Space Administration. It focused on the growing importance of supercomputing in the advancement of science and the looming gap in access to and capability of these resources. The Lax report noted that the United States was at risk of losing its lead in supercomputing and that the development of new systems (especially those relying on new architectures such as massively parallel machines) would require continued investment by the federal government and that the commercial sector could not be expected to provide the necessary research and development (R&D) support. The report proposed four thrusts for a national program:

1. Increased access to supercomputer resources through a nationwide network,
2. Research in software and algorithms for the expected changes in hardware architectures,
3. Training of staff and graduate students, and
4. R&D for future generations of supercomputers.

[2] Panel on Large Scale Computing in Science and Engineering, *Report of the Panel on Large-Scale Computing in Science and Engineering*, National Science Foundation, Washington, D.C., 1982, http://www.pnl.gov/scales/docs/ lax_report1982.pdf.

The Lax report led to the first round of NSF supercomputer centers established in 1985-1986. While a subset of these centers continued through 1997, the director of NSF commissioned the Task Force on the Future of the NSF Supercomputer Centers Program in 1994, chaired by Edward Hayes.[3] The report of the task force, issued in 1995, put forth many of the points from the Lax report, noting that supercomputing would enable progress across many areas of science and this progress would depend on continuing development of highly trained personnel as well new algorithms and software. The Hayes report made many recommendations that focused on both "leading-edge sites" and broader partnerships that would include experimental and regional facilities. The net result was that the report recognized that there would be fewer leading-edge sites to accommodate more systems below the apex of the computational pyramid. This was manifested as the Partnership for Advanced Computational Infrastructure (PACI) from 1997 to 2004. PACI was supplemented by the terascale initiatives in 2000, which led to the creation of the TeraGrid in 2004, which transitioned to the present-day Extreme Science and Engineering Discovery Environment program.

The 2003 Atkins report[4] articulated a more ecological, holistic view of cyberinfrastructure-enabled research, including computing, data stewardship, sensing, activation, and collaboration, to create a comprehensive platform for discovery. It was followed by a series of workshops and reports exploring the role of cyberinfrastructure to particular research communities.[5]

In 2005, NSF's Office of Cyberinfrastructure released the solicitation "High Performance Computing System Acquisition: Towards a Petascale Computing Environment for Science and Engineering" (NSF 05-625). This was the first in a series of solicitations along different tracks, culminating in the Blue Waters petascale facility at the National Center for Supercomputing Applications (NCSA) that began operating in 2013.

The past reports present common themes, many of which persist today, as this report will show. Today, advanced computing capabilities are involved in an even wider range of scientific fields and challenges, and the rise of data-driven science requires new approaches. The gap between

[3] Task Force on the Future of the NSF Supercomputer Centers Program, *Report of the Task Force on the Future of the NSF Supercomputer Centers Program*, National Science Foundation, Washington, D.C., September 15, 1995, http://www.nsf.gov/pubs/1996/nsf9646/nsf9646.pdf.

[4] National Science Foundation, *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*, 2003, http://www.nsf.gov/cise/sci/reports/atkins.pdf.

[5] National Science Foundation, "Reports and Workshops Relating to Cyberinfrastructure and Its Impacts," http://www.nsf.gov/cise/aci/reports.jsp, accessed January 27, 2016.

supply and demand, noted in the Lax report, remains an important issue. The need to maintain and grow the workforce, especially in regard to the needed skills, remains a persistent issue. The evolution in hardware and the subsequent impacts on algorithms and software has been a recurring concern. However, changes in architectures have been far more disruptive over the past decade, and broad commercial trends influence the HPC market more than ever. Finally, increasing use of large-scale computing by the commercial sector offers some new opportunities and challenges.

2.3 HIGH-PERFORMANCE COMPUTING TERMINOLOGY

This report refers to a number of concepts from HPC. These terms do not have precise definitions but are valuable in referring to qualitative properties of different kinds of computing and computing systems.

- *Capability computing* refers to computing that requires the most capable systems, typically the most powerful supercomputers.
- *Capacity computing* refers to computing with large numbers of applications, none of which require a "capability" platform but in their aggregate require large amounts of computing power.
- *High-throughput computing* refers to the use of many computing resources over a period of time to attack a particular set of computational tasks.
- *Leadership class* is a term for the most powerful computing systems. This has typically been based on the floating-point performance of the computing system, though a more comprehensive metric can be used. See Figure 4.1 (Branscomb pyramid) for one (though dated) ranking of computer systems from desktop through leadership class.
- *High-end computing* covers computing from systems larger than a system that a single research group might operate through leadership class systems. There is no accepted definition for how powerful a system must be to be considered a high-end computing system. The terms "supercomputer" and "high-performance computer" have similar, imprecise meanings.
- *Ensemble computing* often refers to the use of many runs with different input data or parameters to explore the sensitivity of the problem to small changes.
- *Tightly coupled computing* refers to computations where each computing element must exchange data with some other computing elements very frequently, such as once per simulation time step. Such computations require a high-performance internode interconnect.
- *Memory capacity limited* refers to applications that have more demanding requirements than others. For example, simulations in three

dimensions of large domains can require very large amounts of memory; a 10,000 × 10,000 × 10,000 cube requires 10^{12} points or roughly 1 TB of storage per variable stored.

- *Peak and sustained performance*. Peak performance refers to the performance of a computing system that is theoretically possible. It usually refers to floating-point performance and assumes that the maximum number of floating-point operations is performed in every clock cycle. No applications run at the peak rate. Sustained performance is the performance that an application or a collection of applications can sustain over the course of the entire application.

This report avoids the terms "capability computing" and "capacity computing" because they are too imprecise and have also historically been too focused on floating-point performance.

2.4 STATE OF THE ART

The past several decades have seen remarkable progress in computer hardware, algorithms, and software. This section reviews the state of the art in hardware, software, and algorithms, with a particular emphasis on the challenges created by the disruptive changes in computer architecture driven by the need to increase computing power.

2.4.1 Hardware

The past decade has seen an enormous disruption in computer hardware throughout the computing industry, as processor clock speed increases have stalled and parallel processing has moved on-chip with multicore processors.[6] The primary drivers have been power density and total energy consumption—concerns that are important in portable devices and increasingly in large data and compute centers due to fundamental cooling limits of packaging and overall facility infrastructure and operations costs. The continued growth in transistor density had been used primarily to add more processor cores, starting with dual-core chips in the mid-2000s to 20-core chips a decade later. But these processors were historically designed to maximize performance without a strong constraint on energy use; a second trend has been the growth of many-core architectures that involve a larger number of smaller and simpler cores, each more energy efficient than a traditional processor. In aggregate, a

[6] For more on this challenge and its implications, see National Research Council, *The Future of Computing Performance: Game Over or Next Level?* The National Academies Press, Washington, D.C., 2011.

processing chip with hundreds of simpler cores can often provide much higher computational performance than a smaller number of more powerful cores. The many-core designs include graphical processing units (GPUs) and were initially designed as accelerators to a traditional CPU, whereby software primarily ran on the CPU but could offload computing-intensive kernels to the accelerator. More recent many-core designs provide for stronger integration between the accelerator and CPU, allowing for shared memory between the two or stand-alone processors made entirely of many-core chips. Box 2.2 contains further discussion of these architectural challenges.

One consequence of the growth in the use of computing by all aspects of society and not just for science research is that much of the investment by both computer hardware and software vendors is directed at the larger commercial market for computing. An example of this is the use of GPUs in computational science. These processors have been adapted to support computational science, but the initial innovations were made to serve the gaming market. As the commercial markets continue to grow and new applications are developed, advanced cyberinfrastructure will need to continue to figure out how best to exploit innovations and advancements in the greater commercial market.

Looming ahead is the end of transistor scaling, which will mean an end to the current strategy of improving computing performance by adding more cores per chip. The result is unlikely to be a discrete stopping

BOX 2.2
Computer Architecture and Hardware in Transition

Moore's law has driven the technology behind high-performance computing (HPC) systems for decades, by doubling the number of transistors on a die at regular intervals, with the speed of these smaller transistors getting faster at essentially an equal rate. Although transistor density will continue to increase for some time to come, the year 2004 represented a watershed where HPC architectures were forced to change direction dramatically. Getting heat out of chips hit a limit, so that increasing the inherent transistor speed no longer translates into faster core clocks. The only alternative was to use extra transistors in more, but slower, cores and require applications to use that resulting parallelism explicitly. This development, combined with a rapid growth in the number of racks for a system, permitted benchmark performance for the LINPACK kernel (solution of a dense system of linear equations by Gaussian elimination) to continue on its near doubling of growth per year. The emergence of "lightweight" processors that were even slower than the power-limited, high-end servers paradoxically

continued

BOX 2.2 Continued

added to this increase by allowing many more nodes to be physically packaged in the same volume. Such architectures took over the bulk of the Top 10 (of the TOP500) ranking[1] until 2008, when a second architectural transition occurred with the introduction of numeric-intensive chips with very large numbers of even simpler cores derived from high-end graphics processors. "Hybrid systems" that join such chips with conventional cores have yet again changed the complexion of the Top 10 systems (Figure 2.2.1). It appears, however, that even these changes have hit at least a temporary roadblock, with no growth in the top system for dense linear algebra since 2013.

The same phenomena can be seen in benchmarking of HPC systems for applications that are decidedly non-numeric and have many of the properties one might expect for big data. Figure 2.2.2 is similar to Figure 2.2.1, except that the Graph500 benchmark involves a breadth-first search through very large graphs. A rapid rise in year-over-year performance hit a wall in 2013, with very little growth since then. In addition, unlike LINPACK, this benchmark has proven somewhat difficult for hybrid systems.

In between the dense linear algebra of LINPACK (and the "classical" scientific computing it represents) and the non-numeric Graph500 is a third benchmark where reported data are becoming available and which represents problems that lie between these two. The High-Performance Conjugate Gradients (HPCGs),

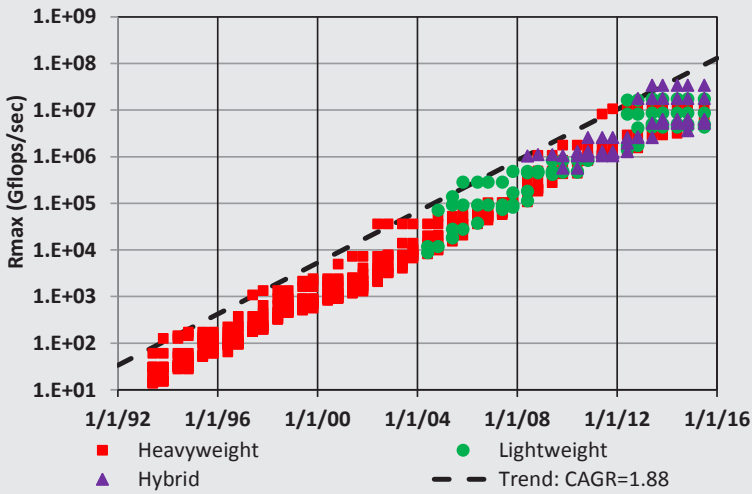

FIGURE 2.2.1 Speed of Top 10 systems from TOP500 ranking. SOURCE: Updated from Peter Kogge, "Updating the Energy Model for Future Exascale Systems," in *High Performance Computing: 30th International Conference, ISC High Performance 2015, Frankfurt, Germany, July 12-16, 2015, Proceedings*, using data from http://top500.org.

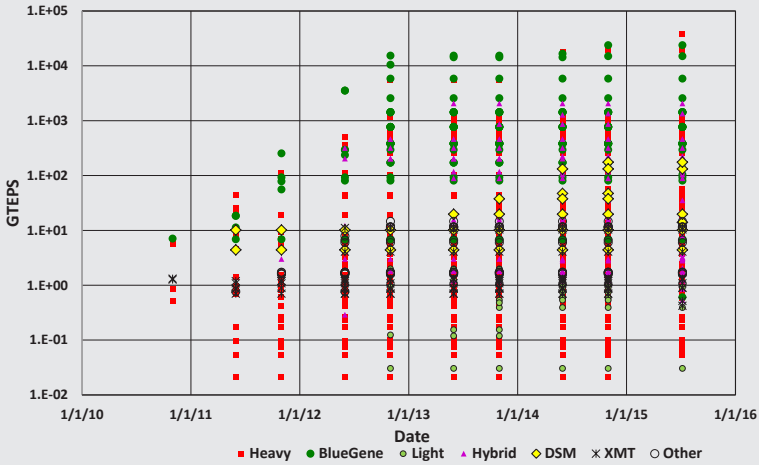

FIGURE 2.2.2 Performance on Graph500 breadth-first search benchmark by date, in giga-traversed edges per second. NOTE: GTEPS, giga-traversed edges per second. SOURCE: Updated from Peter Kogge, "Updating the Energy Model for Future Exascale Systems," *in High Performance Computing: 30th International Conference, ISC High Performance 2015, Frankfurt, Germany, July 12-16, 2015, Proceedings*, using data from http://graph500.org.

benchmark represents the solution of a large matrix equation (as does LINPACK), but one that is extremely sparse and is solved using a different and iterative algorithm. The data to be involved in the computations are now embedded in a sparse graph-like data structure through which the program must spend significant time traversing before a computation can be performed. While there are insufficient reports to look at trends, the data that are available can be compared to LINPACK numbers on the same machines. Figure 2.2.3 diagrams the ratio of HPCG computation rates to peak computational rates over a variety of systems, with a clear indication that solving such problems is far more challenging to today's architectures, especially for the hybrid systems that dominate LINPACK. At best, a few percent of the floating-point computational capability of systems is usable for HPCG, where efficiencies of as much as 90 percent are common for LINPACK. This has been well known in the HPC community, where memory performance is often more important for performance on such problems than peak floating-point performance.

Efficient use of computational hardware is not the only problem facing today's architectures. Memory capacity is also becoming a constraint. Figure 2.2.4 displays the ratio of memory to floating-point performance for LINPACK over the past 20 years. Again, until about 2004, ratios of 1 byte per floating-point operation (FLOP) were common but went into a precipitous decline after that, especially for hybrid systems. The average supercomputer today has between 1/100th and 1/10th the memory per FLOP of a decade ago.

continued

BOX 2.2 Continued

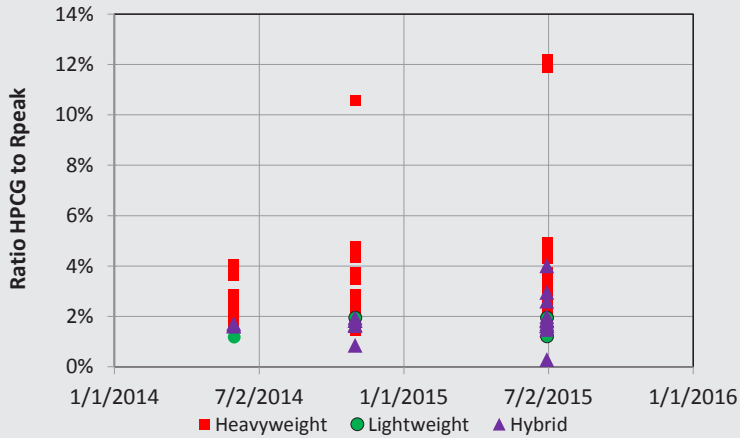

FIGURE 2.2.3 Ratio of High Performance Conjugate Gradients (HPCG) computation rates to peak computational rates over a variety of systems. SOURCE: Updated from Peter Kogge, "Updating the Energy Model for Future Exascale Systems," in *High Performance Computing: 30th International Conference, ISC High Performance 2015, Frankfurt, Germany, July 12-16, 2015, Proceedings.*

The reason for this constraint goes back to architecture and the way commercial memory chips are attached to modern processors. The basic memory cell has in fact continued to get smaller, in accordance with Moore's law, and has not suffered the power issue that changed processor chip architectures. Instead, the need to keep such chips cheap has meant that vendors have downsized the size of memory chips to provide better yield, giving up memory size per chip as a result. Also, the way memory is connected to modern processors has hit a wall of its own. There are only so many pins available on modern processor chips to connect to memory, regardless of the number or speed of cores on the processor. This means that the maximum number of memory chips that may be attached to a processor chip is relatively limited, and with the slower growth rate of memory chip capacity relative to processor performance, the result is exactly what has been observed.

This issue of the path between processor and memory is also most probably at the root of the poor performance observed for both Graph500 and HPCG, as the rate at which commands can be sent from the processor chip to the memory chips has also largely flattened. For problems where the data to be processed next must be located by looking up some indices first, all the complex caching designed into modern processors becomes largely wasted.

These observations do not doom our capability to advance toward exascale; instead they warn us that a major upheaval in architecture is likely, one that will end up having as much effect on programming and algorithms as the advent of

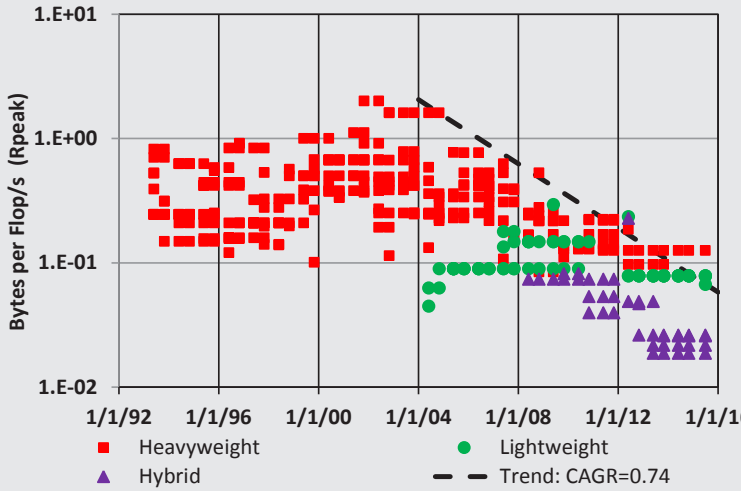

FIGURE 2.2.4 Ratio of memory to floating-point performance for LINPACK benchmark over the past 20 years. SOURCE: Updated from Peter Kogge, "Updating the Energy Model for Future Exascale Systems" in *High Performance Computing: 30th International Conference, ISC High Performance 2015, Frankfurt, Germany, July 12-16, 2015, Proceedings.*

the single-chip microprocessor in the early 1990s and the rise of multicore in the mid-2000s. This change is visible today with the introduction of "3D stacked memory" components, where multiple memory die are placed on top of a logic die. The path between the two offers both significant increases in memory bandwidth and decreases in the energy of such accesses. Today such "stacks" are still tied to conventional processor chips, enabling just a "faster" memory path. In the near term, however, combinations of lightweight and hybrid architectures will move cores onto the logic die along with the network interface controller, resulting in a stand-alone compute node. Hundreds of these may be placed in the space of a modern compute node, breaking the barriers presented today.

The upshot of this is that the advanced computing facilities of the near future are liable to look significantly different from today. Consideration must be given to ensuring that the programs and algorithms being written today that need to scale into these new regimes are designed with these differences in mind and that early facilities should be available as such machines come online to allow validation of the portability of such codes.

[1] See the TOP500 website at http://top500.org, accessed January 27, 2016.

point in chip density, but rather a continued slowing of improvements based on technical and cost challenges, as well as diminishing returns on investments if the density improvements do not immediately equate to improvements in cost performance of computing devices.

The problem of declining performance improvements is not limited to science and engineering applications, but high-end computing with its emphasis on benchmarks and scaling may be the place where slowing the rate of performance improvements will be most obvious. One place where this slowing of performance improvement can be seen is in the bottom of the TOP500 list, which is based on the performance of a simple dense linear algebra algorithm. Since the late 2000s, the rate of performance improvement for the systems at the bottom of the list (still very fast) has fallen considerably.

Memory system design is also undergoing rapid changes, as new forms of on-package dynamic random-access memory (DRAM) memory provide enormous bandwidth improvements but currently less capacity than off-chip DRAM. At the same time, new forms of non-volatile memory have been developed with much higher bandwidths than disks but somewhat different performance characteristics than DRAM. These features may be of particular interest to data analysis applications, although many simulations are also limited by data sizes and could benefit. These new types of memory may be added to the hierarchy in a current system design, but they may be under software rather than hardware or operating system control. In general, data movement between processors or to memory is expensive in both time and energy, so hardware mechanisms that automatically schedule and move data may be replaced by simpler mechanisms that leave data movement under software control.

Although each of these innovations is designed to increase performance while minimizing energy use, they pose significant challenges to software. The scientific modeling and simulation community has billions of dollars invested in software based on message passing between serial programs, with only isolated examples of applications that can take advantage of accelerators. Shrinking memory size per core is a problem for some applications, and explicit data movement may require significant code rewriting because it requires careful consideration of which data structures should be allocated in each type of memory, keeping track of memory size limits, and scheduling data movement between memory spaces as needed.

Further disruptive innovation is on the horizon. For example, processor-in-memory technology has been advanced as a way to reduce memory latency and increase bandwidth, and memristors could potentially be used for non-volatile memory with a very high density and fast access times.

The scientific computing community therefore must balance (1) leaving software and programming models unchanged and giving up on opportunities for more computing performance that come from these hardware changes with (2) developing new codes based on new programming models, such as those being researched within the DOE exascale initiative, that can exploit the new hardware. Some type of energy-efficient processing and memory system will be necessary for building an aggregate exascale capability that NSF can afford to operate, whether that is in a single system, in many systems, or partially based on commercial cloud resources. The breadth of NSF's workload and the number of architectural options complicate this decision. On the surface, the many-core processors may be best suited to compute-intensive simulation problems, yet some data analysis workloads, such as image analysis and neural net algorithms, run effectively on GPUs, while highly irregular simulation problems so far do not. Non-accelerator many-core options such as the Intel Phi may provide more familiar programming support and more workload flexibility, but may not achieve the same performance benefits. Further, they are relatively untested and had yet to demonstrate high performance across a wide range of applications at the time this report was prepared.

Data storage has also undergone its own exponential improvement, with both data densities (bits per unit area) and bit per unit cost doubling every 1 to 2 years. New technologies are providing revolutionary advances and blurring the line between "storage" and "memory." However, while the technology continues to improve, the rate of improvement has fallen off in recent years. Historically, external storage has primarily meant magnetic hard disk drives (HDDs) in which data are encoded on spinning platters of magnetic media. The vast majority of the world's online data (some 1-2 zettabytes) are stored on HDD, and this is projected to be the case for at least the next 5 years. Over the course of six decades and driven in part by advances in fundamental material science, HDDs have gone from devices the size of washing machines, storing 3.75 MB, to modern 2.5-in. disks holding 8 TB and up. This expansion in capacity is projected to continue. But capacity is just one of several figures of merit—others include bandwidth, latency, and input/output operations per second (IOPS), which have all advanced at a much slower pace than capacity—and none are anticipated to advance significantly over current HDD technologies that have effective bandwidths of circa 1-200 MB/s, latencies of a few milliseconds, and IOPS of 1-200. This is in part due to the physical constraints of spinning media, but also because investments are focusing on new technologies that are already delivering 1,000-fold advances over HDD in some performance metrics. Parallelism to many disks is required to provide very high data rates. Latencies have not improved as much;

for spinning disks, the latencies are dominated by the disk revolutions per minute and head seek time, which have advanced much more slowly than the densities and transfer rates (bandwidth). In contrast, solid-state disks (SSDs) provide much lower latency and greater data transfer rates. SSDs are presently based on various non-volatile (meaning data persists even without power) silicon memory technologies that will continue to benefit from advances in silicon manufacturing technologies. In the past, SSDs were regarded as both small and expensive, but in the past few years, the capacity of SSDs has approached that of HDDs, and while presently about 3 to 10 times more expensive per byte than HDDs, price parity is expected within several years. With new standards for connecting SSDs to computer systems (e.g., non-volatile memory express), SSDs are now capable of delivering bandwidths of several gigabytes per second, latencies of a few microseconds, and 100,000 IOPS. In addition to use in storage, the price, performance, persistence, and power characteristics of non-volatile memory technologies enable innovations in computer architectures to complement regular DRAM, such as in the proposed DOE pre-exascale systems. In summary, over the next few years, HDD storage capacity will continue to decrease slowly in cost, but various performance metrics will see revolutionary change as non-volatile memory technologies become even more price competitive, and eventually storage capacity itself will fall in cost once silicon technologies dominate.

Advances in storage capacity were critical enablers of the data-intensive Nobel Prize-winning work of Perlmutter (see Box 3.2), as well as the discovery of the Higgs boson at the Large Hadron Collider by an international collaboration storing and analyzing more than 100 PB of data. Diverse other fields of science have been transformed by the ability to manipulate massive data sets from genomics, social networks, video and images, satellite data, and the results of simulations. Looking forward, continued advances in capacity and revolutionary advances in other aspects of data technologies promise new revolutions in science across many fields presently constrained by their ability to store, explore, or analyze their data at sufficient scale or speed.

Because of these relatively high latencies, as well as the limits in bandwidth compared with semiconductor memory, a wide range of memory architectures are being developed with intermediate performance. Some of these will be used closer to the compute elements and have been mentioned above. Others may be used to boost the apparent performance of disks, for example, by providing a higher-bandwidth, lower-latency, temporary buffer that can absorb bursts of data to write to disk. All of these new input/output (I/O) and memory products will need new software and, in many cases, new algorithms that fit their performance characteristics.

The last major component of high-end computers is the internode interconnect; that is, the network that is used to move data between compute elements or between centralized data storage and the compute system. Although the performance of these interconnects has also increased significantly, with bandwidths for proprietary networks used in HPC systems of 40-80 GB/s per link being typical, the latencies have not improved much in recent years, with high-performance interconnects having latencies on the order of 1 microsecond. Commodity interconnects are one to two orders of magnitude slower, with link speeds of 1 GB/s being common, and with 10 GB/s available at the high end of commodity interconnects.

The manner in which the links are connected is also important. There are three separate but related decisions. One is the topology of the connections. High-end supercomputers link nodes directly together in an n-dimensional torus. For example, the IBM BlueGene/Q uses a five-dimensional (5D) torus; the Cray Gemini network uses a three-dimensional (3D) torus, with two compute nodes connected to each torus node. A second is the switch radix—how many ports each switch has. A third is whether the network uses switch notes that are distinct from processor nodes. Recently, interconnect design principles from HPC, such as more highly connected networks with better bisection bandwidth and latencies, have been adopted for commercial applications.[7]

Also of importance is wide-area networking, which is critical to the success of NSF's advanced computing, especially in terms of providing access and the infrastructure necessary to bring together data sources and computing resources. The size of some data sets is forcing some data off-line or onto remote storage, so storage hierarchies, storage architectures, and WAN (wide area network) architectures are increasingly important to overall infrastructure design. NSF has made significant investments in wide-area networking. The Internet2 network plays an important role in connecting researchers. It carries multiple petabytes of research data and also connects researchers globally with peering to more than 100 international research and education networks. Wide-area networks have a distinct set of technical, managerial, and social complexities that are beyond the scope of this report.

[7] See, for example, A. Singh, J. Ong, A. Agarwal, G. Anderson, A. Armistead, R. Bannon, et al., "Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network," presented at the Association for Computing Machinery Special Interest Group on Data Communication (SIGCOMM), 2015, http://conferences.sigcomm.org/sigcomm/2015/pdf/papers/p183.pdf.

2.4.2 Software

Although computer hardware will to continue to improve, the rate of improvement has been slowing down and producing increasingly disruptive programming features. Software for scientific simulations for parallel systems with more than a handful of processing cores has largely been written in a message passing model (e.g., with MPI) using domain decomposition, where the physical domain or other major data structures are divided in pieces assigned to each processor. This works especially well for problems that can be decomposed statically and where communication between processes is predictable, involving a limited number of neighbors along with global operations. The assumptions underlying this model are that (1) locality is critical to scaling, so the application programmer needs to do the data decomposition, (2) the network and processors are reliable, and (3) the performance is roughly uniform across the machine. At the same time, many of the data analysis workloads processed on cloud computing platforms have used a map-reduce style in which independent tasks are spread across nodes and results are aggregated using global communication operations at intermediate points. This model allows for hardware heterogeneity or variable-speed processors, but does not permit point-to-communication between tasks. Both models have proven powerful in their own setting.

The relative stability until recently of the hardware platforms has allowed a rich set of libraries and frameworks for simulation to emerge, many supported by NSF (Box 2.3). This includes libraries for sparse and dense linear algebra, spectral transforms, and application frameworks for

BOX 2.3
Volume and Complexity of Scientific Software

The total volume and complexity of scientific software that runs on today's high-performance computing (HPC) systems have grown enormously in the past two decades. And while some scientific fields are just beginning to build analysis pipelines for their experiments, in fields like high-energy physics and biology these have existed for many years. Large community codes for modeling problems in materials and climate, for example, have many different models to simulate different conditions, options for algorithm choices, and multiple implementations for specific hardware. These applications are written in a variety of languages and libraries and, in many cases, involve multiple languages mixed together. They may use FORTRAN for numerical kernels, C++ for complex data structures, and Python to manage the steps in a software pipeline, and they may call multiple scientific libraries that are themselves written in other languages. Parallelism is typically expressed using message passing, typically MPI, possibly with threading used for on-

BOX 2.3 Continued

node parallelism. But applications at a large center may take advantage of many different languages, libraries, and programming abstractions as well as tools to help with debugging, performance analysis, data management, and visualization.

The diversity of libraries used in scientific computing gives some indication of the software investment needed to sustain a broad program of scientific discovery using HPC. Tables 2.3.1 and 2.3.2 show the usage of some of the most popular scientific libraries and programming models in one center based on a survey of users and weighted by the number of hours each project uses. These are based on data reported in categories chosen by those responding to the survey. As a result, there is some overlap in categories, and some used a general category (e.g., PGAS, or partitioned global address space) where others used a specific category (e.g., UPC, a PGAS language). These should be used only (1) to see the breadth of libraries, languages, and systems and (2) as a very rough guide to the amount of use of each item.

TABLE 2.3.1 Scientific Libraries Used at One Center

Tier	Library
1st	LAPACK, FFTW, ScaLAPACK, PETSc, NCAR, hypre, SuperLU, MUMPS, Chombo, Trilinos, Root
2nd	METIS, BOOST, CERNLIB, BLAS, SLEPc, BoxLib, PSPLINE, GSL, CHROMA, QDP++, MKL, pARPACK, SCOREC, gotoBlas, FFTPACK

NOTE: Libraries are grouped by usage in terms of number of compute hours used by the projects that listed the library. The tiers are subjective but represent, roughly, clusters of usage. Libraries in each tier are used by roughly 10 times the number of compute hours as those in the next tier (measured by the total time used by the project, not necessarily the library). Acronyms are defined in Appendix D.
SOURCE: Survey of National Energy Research Scientific Computing Center (NERSC) users, Sudip Dosanjh, NERSC director, personal communication.

TABLE 2.3.2 Programming Systems Used at One Center

Tier	Programming System
1st	MPI, Fortran, C++, OpenMP, C
2nd	Shellscript, Python
3rd	Posix Threads, Tcl/TK, Java, Perl, Assembler, Charm++, OpenCL, IDL, PGAS, SHMEM
4th	GASnet, MATLAB, UPC, Global Arrays, CoArray Fortran, Lua, Ruby, UPC++, CUDA, OpenCL

NOTE: Systems are grouped by usage in terms of the number of hours used by the projects that listed the programming system. The tiers are subjective but represent rough clusters of usage. The first tier are systems used by the majority of applications. The second tier are systems that use far fewer compute hours but still have significant use. Programming systems in the first two tiers are used by jobs that consume roughly 10 times the number of compute hours as those in the third tier (the table does not reflect the fraction of time each job spends using each programming system). Acronyms are defined in Appendix D.

climate modeling, astronomy, fluid dynamics, mechanical modeling, and many more. To manage the overall power consumption of larger future systems, it will not be viable to carry out larger computations simply by scheduling threads on more cores. The processors themselves will need to become more energy efficient. As a result, scientific software will need to be revised to take advantage of power-conserving processor features like software-managed memory, wider serial instructions, and multiple data architectures. Scientific libraries face these challenges but are also a point of leverage, allowing multiple applications to benefit from optimizations to new architectures. Looking ahead, substantial investments in software will also be required to take advantage of future hardware, as will research to address new models of concurrency and correctness concerns.

The virtual machine abstractions in the commercial cloud have enabled a different class of applications, with complex workflows for data analysis built and distributed as an integrated software stack. These are particularly popular in biology and particle physics.

2.4.3 Algorithms

The situation is even more complicated for algorithms, where improvements in algorithmic complexity are harder to predict. Not all of the improvements fall into a general category, but some of the common approaches include hierarchical algorithms, exploiting sparseness or symmetry, and reducing data movement. In simulation problems, both the mathematical models of a given physical system and the algorithms to solve them may be specialized to a problem domain, allowing for more efficient computations. The same is true for data analysis, where some pre-existing knowledge of the data may permit faster analysis techniques. Machine characteristics may also affect the choice of algorithms, as the relative costs of computation, data movement, and data storage continue to change across generations, along with the types and degrees of hardware parallelism. Minimizing the total work performed is generally a desirable metric, but on machines with very fast processing and limited bandwidth, recomputation or other seemingly expensive computations may pay off if data movement is reduced, and memory size limits can make some algorithms impractical. Future algorithmic innovations will still be essential for addressing more complex simulation problems—for example, modeling problems with enormous ranges of time- or space scale, or problems that combine multiple physical models into a single computation. They will also be needed for new problems in data-driven science, such as enabling multimodal analysis across disparate types of data, interpreting data with a low signal-to-noise ratio, and handling enormous data sets where only samples of the data may be analyzed. New algorithms will

also be needed to take advantage of future hardware with its new forms of parallelism and different cost metrics, including algorithms that can detect or tolerate various types of errors. Finally, scientific discovery at the boundary of simulation and observation will require new algorithms to measure uncertainties, adjust models dynamically to fit observed data, and interpret data that are incomplete or biased.

Although research into algorithms will continue to have large payoffs in some domains, it does not replace the need for increasingly capable machines. Algorithmic improvements have historically gone hand-in-hand with hardware improvements, provided that the algorithmic advances can be effectively implemented on the advanced hardware. Machine learning algorithms based on neural networks, for example, are only effective because of the performance of modern hardware, and the massive high-throughput computations of the Materials Genome Initiative would not be possible on the hardware available two decades ago. So while hardware performance gains will be increasingly difficult in the future, substantial algorithmic improvements for some problems are probably impossible. For these problems, decades of work on algorithms have led to optimal solutions, and further improvements must come from hardware and operating system software (Box 2.4).

BOX 2.4
Algorithms and Moore's Law Challenges

The rate of improvement in hardware performance, whether measured by clock rate or even by concurrency, has been slowing down. Although the situation is much more complicated for algorithms, there are cases where year-to-year improvement in algorithms is also becoming more difficult.

One example that is often used to demonstrate the essential contribution of algorithms is the solution of the large, linear systems of equations that arise when approximating the solution to a three-dimensional partial differential equation on a grid of size n-by-n-by-n. Figure 2.4.1 is a typical example. It shows that the improvement in performance for this problem is comparable to the improvement indicated by Moore's law.[1] In other words, for this particular problem with a size of n = 64, using the most modern algorithm on a 35-year-old computer system would be as effective (by this simple measure) as running the 35-year-old algorithm on a state-of-the-art system. This is true, and it emphasizes the tremendous advancements in numerical algorithms. However, note that the most modern algorithm, Full Multigrid, requires only O(1) work per solution value. As this problem is defined, there is no longer much room for improvement. Full Multigrid is an optimal algorithm for this problem at any size; a size of n = 1,000 (i.e., a matrix with a billion

continued

BOX 2.4 Continued

Method	Storage	Flops
GE (banded)	n^5	n^7
Gauss-Seidel	n^3	$n^5 \log n$
Optimal SOR	n^3	$n^4 \log n$
CG	n^3	$n^{3.5} \log n$
Full MG	n^3	n^3

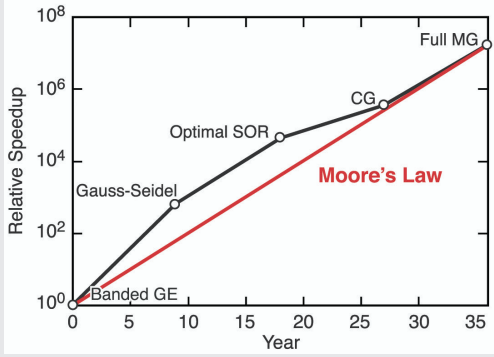

FIGURE 2.4.1 *Top:* A table of the scaling of the memory and processing requirements for the solution of the electrostatic potential equation on a uniform cubic grid of n × n × n cells for n = 64. *Bottom:* The relative gains of some solutions algorithms for this problem and Moore's law for the improvement of processing rates over the same period. SOURCE: After Office of Science, U.S. Department of Energy, *A Science-Based Case for Large-Scale Simulation*, Volume 1, Washington, D.C., 2003, p. 32.

rows) is easily handled today. Any further improvements in performance can come only from faster hardware or by the relatively small reductions in the constant term in the time complexity of the algorithm.

This example is not meant to say that all linear systems can now be solved in optimal time; it applies only to one well-studied and relatively simple problem. Optimal algorithms for solving other types of systems of linear equations have yet to be found, and seeking such algorithms remains an active and important area of research. And for particular problems, alternative formulations may provide a route to a solution without needing to solve this particular linear system of equations. But this example does point out that there is a limit to the use of better algorithms—in some cases, there is no alternative but to run the current optimal algorithm on faster hardware.

---

[1] For a more thorough discussion, see Office of Science, U.S. Department of Energy, *A Science-Based Case for Large-Scale Simulation*, Washington, D.C., 2003, p. 32.

2.5 NSF INVESTMENTS IN ADVANCED COMPUTING

Since the beginning of NSF's supercomputing centers program in the 1980s, its Division of Advanced Cyberinfrastructure (ACI) and its predecessor organizations have supported computational research across NSF with both supercomputers and other high-performance computers and provided services to a user base that spans work sponsored by all federal research agencies. Although a large fraction of the leadership-class investments have been driven by the mission-critical requirements of DOE and DOD, NSF has played a pivotal role in moving forward the state of the art in HPC software and systems.

ACI supports and coordinates a range of activities to develop, acquire, and provision advanced computing and other cyberinfrastructure for science and engineering research together with research and education programs. A significant fraction of ACI's investments have been for two tiers of advanced computing hardware; a petascale computing system, Blue Waters, deployed in 2013 at the University of Illinois, and a distributed set of systems deployed under the eXtreme Digital program and integrated by the Extreme Science and Engineering Discovery Environment (XSEDE). XSEDE makes eight compute systems located at six sites available to researchers along with a distributed Open Science Grid and visualization, storage, and management services. Resource allocations for both tiers are made through competitive processes managed by the Petascale Computing Resource Allocations Committee (PRAC) and the XSEDE Resource Allocation Committee (XRAC), respectively. As things stand currently, roughly half of all available computing capacity will shut down in 2018 with the anticipated end-of-life decommissioning of Blue Waters.

One of the major contributions of NSF to computational science has been the development of software: application codes, libraries, and tools. NSF's implementation of the Cyberinfrastructure Framework for 21st Century Science and Engineering vision[8] identifies three classes of software investments: software elements (targeting small groups seeking to advance one or more areas of science), software frameworks (targeting larger, interdisciplinary groups seeking to develop software infrastructure to address common research problems), and software institutes (to establish long-term hubs serving larger or broader research areas). Investments at the larger/broader end are supported under the cross-foundation Software Infrastructure for Sustained Innovation program, while those at

[8] National Science Foundation, "Implementation of NSF CIF21 Software Vision," http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504817, accessed January 27, 2016.

the smaller/narrower end are supported by the relevant science and engineering divisions.[9]

Not included in the ACI portfolio are investments in computer science research infrastructure, such as the GENI (Global Environment for Network Innovations) testbed. Such resources are important research resources but belong more properly to the specific research program within NSF. Also not included is basic research into algorithms and software, which while also vital, is supported by other research programs in NSF (both within Computer and Information Science and Engineering [CISE] and the other science divisions) and at other federal agencies such as DOE.

Trends in the overall investment in advanced computing can be seen by looking at the spending amounts reported by federal agencies to the Networking and Information Technology Research and Development program's National Coordination Office. Figure 2.1 shows the total federal investment in all categories tracked by Networking and Information Technology Research and Development (NITRD) including high-end computing infrastructure and applications (HECIA), a category that shows both long-term growth over the period 2000-2015 as well as a significant fall-off from a mid-2000s investment spike. Note that advanced computing systems have a relatively short useful lifetime. However, NSF's investments in HECIA have fallen off from nearly 40 percent to less than 20 percent of the total (Figure 2.2a-b), even as demand has grown.

2.6 DEMAND FOR AND USE OF NSF ADVANCED COMPUTING RESOURCES

The use of advanced computing resources cuts across research funded by all the divisions of NSF, as shown in Figure 2.3. Data obtained from XSEDE indicate that the number of active users has quintupled over the past 8 years, and the use[10] grew exponentially through about 2009. Use increases less rapidly after that, matching the slower growth in available resources (cf. Figure 2.5). The usage patterns over the years indicate significant usage by all of the NSF directorates, including Mathematical and Physical Sciences, Biological Sciences, Geosciences, Engineering, CISE, and Social, Behavioral and Economic Sciences. Notably, use by the Direc-

[9] National Science Foundation, "Implementation of NSF CIF21 Software Vision," http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504817, accessed January 27, 2016.

[10] XSEDE use is measured in service units (SUs), which are defined locally for each XSEDE machine and normalized across machines based on High-Performance Linpack benchmark results. SUs do not account for other relevant system parameters such as memory or storage use. Also, a large fraction of available SUs in the current XSEDE resources comes from coprocessors that can be used only after significant changes to software and, sometimes, to algorithms as well.

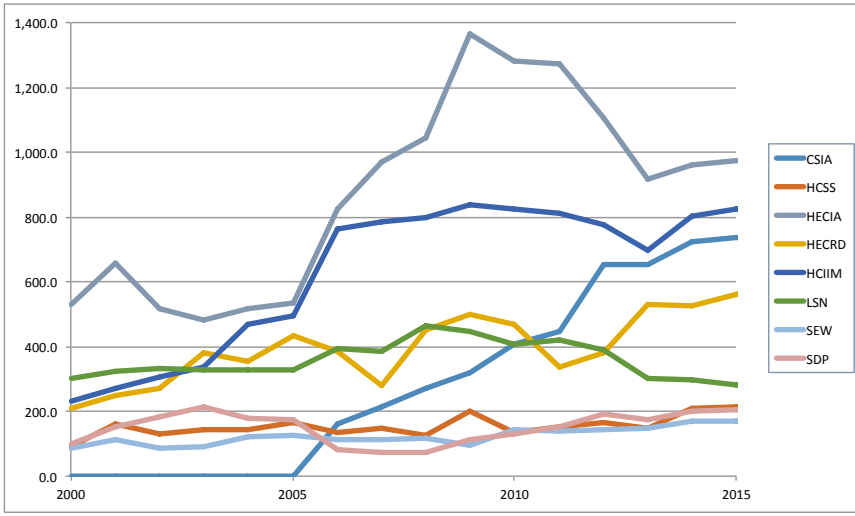

FIGURE 2.1 Total federal investment ($ millions) in the Networking and Information Technology Research and Development program categories. NOTE: CSIA, Cyber Security and Information Assurance; HCIIM, Human Computer Interaction and Information Management; HCSS, High Confidence Software and Systems; HECIA, High-End Computing Infrastructure and Applications; HECRD, High-End Computing Research and Development; LSN, Large Scale Networking; SDP, Software Design and Productivity; SEW, Social, Economic, and Workforce Implications of IT and IT Workforce Development. SOURCE: Compiled from data provided in annual supplements to the president's budget request, prepared by the National Coordination Office for the Networking and Information Technology Research and Development program, https://www.nitrd.gov/Publications/SupplementsAll.aspx.

torate for Social, Behavioral and Economic Sciences is continuing to grow exponentially and by 2014 exceeded the use by Mathematics and Physical Sciences in 2005, showing the broad growth in the use of computing across the foundation.

Further, for such infrastructure as XSEDE, NSF supports a significant fraction of non-NSF funded users. With XSEDE, the usage patterns indicate that for large allocations (e.g., over 10 million service units) approximately 47 percent of the allocations are for non-NSF funded users (Figure 2.4). That share includes 14 percent in support of research funded by the National Institutes of Health.

Although it is difficult to know exactly how much advanced computing is required by the nation's researchers, one available metric is the amount of computer time requested on the XSEDE resources. There is a growing gap between the amount requested, which continues to grow exponentially, and the amount available (Figure 2.5). The implication is

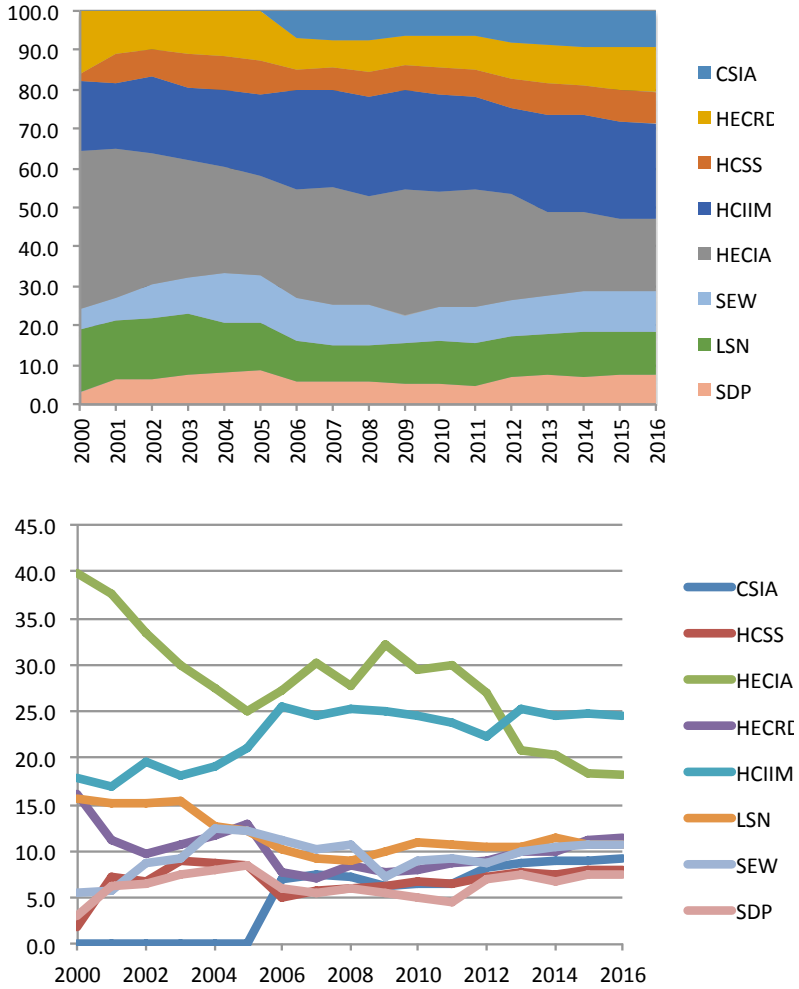

FIGURE 2.2 National Science Foundation investment by Networking and Information Technology Research and Development category from 2000 to 2016 as (a) a percent of total and (b) in millions of dollars. NOTE: CSIA, Cyber Security and Information Assurance; HCIIM, Human Computer Interaction and Information Management; HCSS, High Confidence Software and Systems; HECIA, High-End Computing Infrastructure and Applications; HECRD, High End Computing Research and Development; LSN, Large Scale Networking; SDP, Software Design and Productivity; SEW, Social, Economic, and Workforce Implications of IT and IT Workforce Development. SOURCE: Compiled from data provided in annual supplements to the president's budget request, prepared by the National Coordination Office for the Networking and Information Technology Research and Development program, https://www.nitrd.gov/Publications/SupplementsAll.aspx.

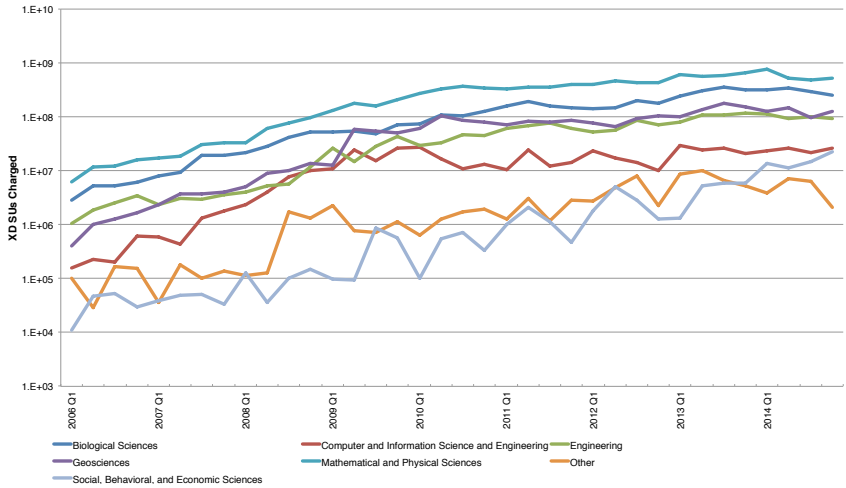

FIGURE 2.3 Use of XSEDE resources by NSF directorate funding research grantee, 2006-2014. NOTE: NSF, National Science Foundation; SU, service unit; XSEDE, Extreme Science and Engineering Discovery Environment. SOURCE: Derived from data obtained by querying Open XDMoD database, University at Buffalo.

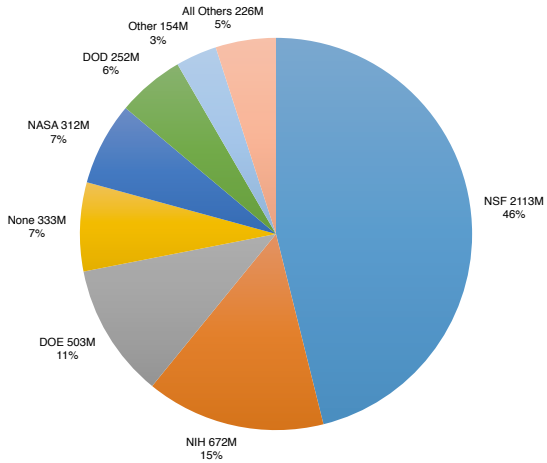

FIGURE 2.4 Estimated use in XSEDE service units of NSF advanced computing by grantees of other federal agencies, based on allocations of XSEDE resources over calendar year 2014. NOTE: NSF, National Science Foundation; XSEDE, Extreme Science and Engineering Discovery Environment. SOURCE: Derived from data obtained by querying Open XDMoD database, University at Buffalo (J.T. Palmer, S.M. Gallo, T.R. Furlani, M.D. Jones, R.L. DeLeon, J.P. White, N. Simakov, et al., Open XDMoD: A tool for the comprehensive management of high-performance computing resources, *Computing in Science and Engineering* 17.4(2015):52-62, 2015).

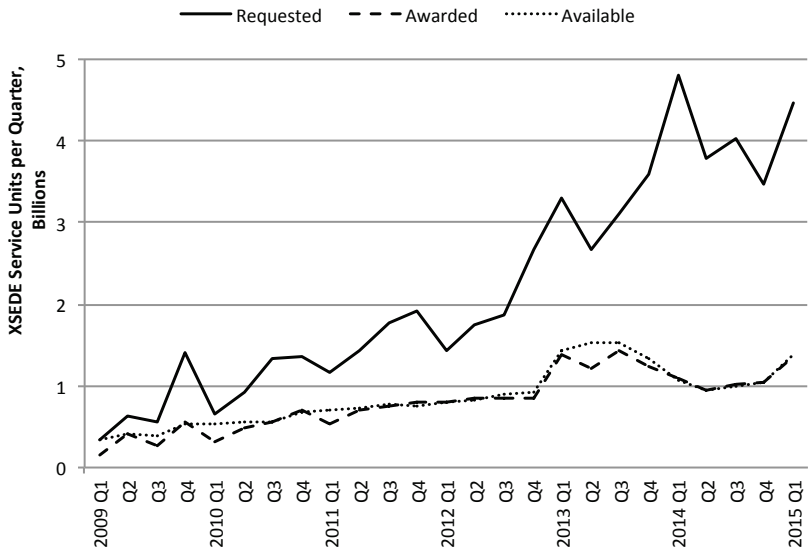

FIGURE 2.5 Requested XSEDE resources compared to awarded and available resources, illustrating the gap as well as growing divergence between available and requested resources. NOTE: XSEDE, Extreme Science and Engineering Discovery Environment. SOURCE: Data from Open XDMoD, University at Buffalo (J.T. Palmer, S.M. Gallo, T.R. Furlani, M.D. Jones, R.L. DeLeon, J.P. White, N. Simakov, et al., Open XDMoD: A tool for the comprehensive management of high-performance computing resources, *Computing in Science and Engineering* 17.4(2015):52-62, 2015). Custom query by Robert L. DeLeon.

that insufficient computing resources inhibits the effective execution and constrains the scale of accomplishment of already funded NSF science.

2.7 NATIONAL STRATEGIC COMPUTING INITIATIVE

As this study was being completed, an executive order[11] was issued establishing a National Strategic Computing Initiative. Section 3a of the order designates NSF as one of the three lead agencies for the initiative and calls for NSF to "play a central role in scientific discovery advances, the broader HPC ecosystem for scientific discovery, and workforce development." Box 2.5 compares items in the executive order with the major themes of this report.

[11] Executive Office of the President, "Executive Order—Creating a National Strategic Computing Initiative," July 29, 2015, https://www.whitehouse.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative.

BOX 2.5
Provisions of the Executive Order Establishing a National Strategic Computing Initiative[1] and Their Relationship to Major Themes of This Report

The following are themes in this report as well as the executive order establishing a National Strategic Computing Initiative (NSCI):

1. High-performance computing (HPC) remains critical for science and industry; if anything, the need and value continue to grow. (NSCI Section 1)
2. "Increasing coherence between the technology base used for modeling and simulation and that used for data analytic computing." (NSCI Section 2.2)
3. Building on its successes in cyberinfrastructure, the National Science Foundation (NSF) has an important role to play both in providing HPC (including data and compute) for basic science and in development of the science needed to advance HPC, including the algorithms, software, and hardware for extreme scale computing. (NSCI Section 3a)
4. NSF must also contribute to the development of an HPC workforce. (NSCI Section 3a)
5. Public-private partnerships should be explored. (NSCI Section 1.2)
6. HPC research must be transitioned into practice. (NSCI Section 1.4) This report's recommendations to NSF echo this need; in particular, NSF needs both to perform research in support of HPC and to support bringing that research into practice as needed by the NSF user community.
7. Embrace an integrated approach to providing effective HPC, combining hardware, software, and algorithms, as well as address the development of an HPC-capable workforce and the whole of HPC, including the midrange as well as the high end. (NSCI Section 2.4)

Several themes in this report are not specifically discussed in the executive order:

1. Although convergence of data-intensive and compute-intensive systems is important and will address many needs, some applications require more specialized approaches that may emphasize compute or data. (NSCI Section 2.2 focuses on convergence)
2. The demand for computing continues to outstrip supply; more needs to be done to (a) provide greater resources (especially systems and expertise in using them) and (b) make the best use of these resources (NSCI makes no statements on budgets; efficient use of the ecosystem is mentioned but without specific coordination).
3. A diversity of platforms and software will be needed to capture the long tail of science. NSCI calls for acceleration of the deployment of an exascale class system but says nothing about the acceleration needed for future science needs at all scales.

[1] Executive Office of the President, "Executive Order—Creating a National Strategic Computing Initiative," July 29, 2015, https://www.whitehouse.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative.

3

Maintaining Science Leadership

Advanced computing underpins virtually every discipline of science and engineering and is critical to the National Science Foundation's (NSF's) mission "to promote the progress of science; to advance the national health, prosperity, and welfare; to secure the national defense; and for other purposes."[1] The use of advanced computing enables discoveries in fundamental areas of physical sciences; provides new insights in biological sciences that have implications for national health; leads to improved engineering of devices; enables the development of new materials and systems with both commercial and defense implications; and aids in our understanding of the environment and society. Advanced computing has traditionally been used for modeling and simulation to interpret and project the implications of mathematical models of physical phenomena and, increasingly, to analyze the large and complex data sets from observations and experiments.

The impacts on science have been both broad and deep. Advanced computing supports the education and research of thousands of students and scientists across the country, and it has been essential in some of the most significant award-winning scientific discoveries. Advanced computing has been used for scientific discoveries across many disciplines, from cosmology and astrophysics to biology and medicine. For example, advanced computing at NSF has been used to understand the forma-

[1] National Science Foundation, "At a Glance," http://www.nsf.gov/about/glance.jsp, accessed March 31, 2016.

tion of the first galaxies in the early universe, to analyze the impacts of cloud-aerosol-radiation on regional climate change, and to understand the design and behavior of computing device technology as the end of Moore's law scaling approaches. The use of advanced computing systems, including those designed for data-intensive workloads, has expanded beyond traditional domains to understanding social phenomena captured in real-time video streams, connection properties of social networks, and voter redistricting schemes. Other examples of science impacts can be found in Box 3.1.

Advanced computing has been a key to multiple Nobel Prizes (Box 3.2), including the 2013 Nobel Prize in chemistry awarded jointly to Martin Karplus, Michael Levitt, and Arieh Warshel for "the development of multiscale models for complex chemical systems." The team used NSF TeraGrid resources for particle simulations to predict the structure of proteins and combine molecular dynamics with quantum mechanical calculations.

3.1 CRITICAL ROLE OF NSF

NSF plays a critical role in providing the advanced computational infrastructure, including advanced computing, necessary to keep the United States at the forefront in the areas of science and engineering. According to the Networking and Information Technology Research and Development (NITRD) reports[2] on investments in high-end computing infrastructure and applications (HECIA), NSF ranked second, behind the Department of Energy (DOE), for 2015 in investments in high-end computing facilities. DOE invested more than $350 million, while NSF was just over $200 million. With respect to investments in the NITRD high-end computing research and development program area, NSF is currently very close to DOE, with the Department of Defense (DOD)[3] leading with more than $200 million and NSF and DOE in the range of $125 million.[4] However, NSF investment in HECIA has declined significantly

[2] See fiscal year 2000 through fiscal year 2016 editions of Networking and Information Technology Research and Development National Coordination Office. The Networking and Information Technology Research and Development Program: Supplement to the President's Budget.

[3] DoD includes Office of the Secretary of Defense, National Security Agency, and the DoD Service research organizations.

[4] Note that the interpretation of the NITRD budget numbers is difficult and not consistent across agencies, as has been observed in President's Information Technology Advisory Committee reports (2010) and that DOE has been investing in extreme scale research both through its application communities with Scientific Discovery through Advanced Computing (SciDAC) and co-design projects and through research evaluation prototypes, some of which may not have been included in the HECRD (High End Computing Research and

BOX 3.1
Examples of the Science Impacts of Advanced Computing

Gaining New Insights About Earthquakes

The Southern California Earthquake Center (SCEC) and its lead scientist, Thomas Jordan, use NSF advanced computing resources to improve our understanding of earthquakes and provide more accurate hazard assessments. SCEC's PressOn project is creating more physically realistic, wave-based earthquake simulations using an earthquake model that calculates how earthquake waves ripple through a three-dimensional (3D) model of the ground. Given detailed information about the geological material in specific areas, physics-based 3D wave propagation simulations are able to calculate how earthquake waves will move through the Earth and how strong the ground motions will be when the waves reach the surface. In 2014, the SCEC team investigated the earthquake potential of the Los Angeles Basin, where the Pacific and North American plates run into each other at the San Andreas Fault. In this study, the simulation showed earthquake waves trapped, and reverberating, within the Los Angeles Basin, leading to high-shaking ground motions much greater than expected. In 2015, SCEC used the NSF-funded Blue Waters supercomputer at the National Center for Supercomputing Applications and the Department of Energy-funded Titan supercomputers at the Oak Ridge Leadership Computing Facility to carry out a simulation that doubled the maximum simulated frequency of the previous year's model, therefore also doubling the accuracy. Even though the number of calculations required increased as the maximum simulated frequency of the earthquake went up, the computing power of Blue Waters and Titan reduced the time needed for these calculations from months to weeks. Researchers believe seismic hazard analyses need to simulate earthquake frequencies above 10 hertz to realistically capture the full dynamics of a potential event. Physics-based 3D earthquake simulations at 10 hertz, once a distant dream, are now on the horizon.

SOURCE: Adapted from NSF, "Los Angeles Basin Jiggles Like Big Bowl of Jelly in Cutting-Edge Simulations," August 20, 2015, http://nsf.gov/discoveries/disc_summ.jsp?cntn_id=136013.

Simulating an Atomic-Resolution Model of the Protein Shell of the Human Immunodeficiency Virus

The HIV capsid project, headed by Klaus Schulten of the University of Illinois at Urbana-Champaign, constructed the first atomic-resolution model of a mature HIV capsid and simulated it on Blue Waters, representing steps toward a better understanding of the interactions of potential drugs and host cell factors with the capsid. The project expanded the frontier of molecular dynamics simulation capabilities from simulating just a few proteins to simulating full organelles. It involved simulations of about 65 million atoms using thousands of nodes on Blue Waters, and required several years to redesign computer codes to make them more scalable to petascale systems. Going from simulating organelles to full cells with 100 billion atoms will require much faster computers.

SOURCE: Adapted from National Center for Supercomputing Applications, "The Computational Microscope," *Blue Waters Highlights,* https://bluewaters.ncsa.illinois.edu/documents/10157/5a7f03ba-4a2c-45a2-b7b3-3fa1fa1293c7, accessed March 31, 2016.

Understanding Avian Lineages by Comparing the Genomes of 48 Bird Species

The Avian Phylogenomics Consortium project published eight papers in *Science* and 20 papers in other journals on its work reconstructing the evolutionary history of birds. The research involved processing hundreds of times more genetic data per species than previous studies. The size of the data sets and the complexity of the analysis required multiple XSEDE resources: Ranger, Lonestar, and Stampede at the Texas Advanced Computing Center (TACC); Nautilus at the National Institute of Computational Sciences; and Gordon at the San Diego Supercomputer Center. TACC resources as well as a cluster at the University of Texas were used to test and validate new computational techniques; Nautilus was used to generate phylogenetic trees at the chromosome level for all the bird genomes; and the Gordon cluster was used to infer phylogenetic trees at the genome level. The analysis allowed the researchers to realize the existence of new inter-avian relationships, redrawing the family "tree" for nearly all of the 10,000 species of birds alive today.

SOURCE: Adapted from Extreme Science and Engineering Discovery Environment (XSEDE), "2014-2015 XSEDE Highlights," 2015, https://www.xsede.org/documents/10157/169907/XSEDE_Highlights_2015.pdf, p. 14.

Discovering the Dark Side of the Universe and Testing General Relativity with Advanced LIGO

NSF's Advanced Laser Interferometer Gravitational Wave Observatories (aLIGO) have begun taking data. Their first direct detection of waves of astrophysical origin is imminent. Frequent observations of coalescing binaries of two neutron stars, a neutron star and a black hole, and two stellar-mass black holes are expected once the detectors have reached their design sensitivity in 2019.

The coalescence and merger of two black holes (binary black holes; BBHs) is the most extreme of aLIGO's gravitational wave sources. Because BBHs produce gravitational waves and no other kind of radiation, observation of merging BBHs makes it possible to probe the predictions of general relativity.

Numerical relativity simulations that implement general relativity without approximation (other than numerical truncation error) are essential for enabling discovery and testing of general relativity with aLIGO. Codes with exponential convergence (highly efficient via spectral methods) are being deployed on the nation's (and international) supercomputers to generate waveform predictions ("templates") for aLIGO. The tremendous challenge is that the BBH parameter space is nine-dimensional: mass ratio of the two black holes, six components of the holes' spin vectors, orbital eccentricity, and argument of periapsis. Thousands of numerical relativity simulations, each tracking the holes over many orbits before merger, are needed. These thousands of waveforms will populate carefully chosen discrete nodes that will allow accurate interpolation throughout the continuous parameter space.

Although a single numerical relativity BBH simulation runs on about 48 cores within 3 months, running thousands is a massive challenge at the interface of high-throughput and capability computing. NSF's Blue Waters is presently enabling the first coarse sweep (hundreds of simulations; only about 40 pre-merger orbits) through the BBH parameter space with hundreds of simulations. A next-generation machine and improved algorithms will be needed to carry out the thousands of very long coalescence simulations that aLIGO needs for discovery and for testing general relativity.

continued

BOX 3.1 Continued

SOURCE: Adapted from Christian D. Ott, email message to Bill Gropp, September 11, 2015.

Simulating Glacial Climate in Coastal South Africa: Developing the Climate Parameters to Model a Paleoscape During Modern Human Origins

There is widespread consensus that the modern human lineage evolved in Africa, and it has been hypothesized that the Cape region of South Africa may have been the refuge region for the progenitor lineage of all modern humans during harsh global glacial phases. During this phase of human origins, the economy was based on hunting and gathering, and hunter-gatherer adaptations are tied to the way that climate and environment shape the food and technological resource base. Curtis Marean, Arizona State University, leads a multinational, multidisciplinary team of researchers in a pioneering application of high-performance computing to the study of these interactions. "Our project began as a straight archaeological dig," Marean says. "Then I realized that we needed much better climate and environmental contextual data to understand the archaeological record we were excavating." Marean began using XSEDE resources—both the Pittsburgh Supercomputer Center's (PSC's) Blacklight and San Diego Computer Center's Trestles compute systems and assistance from XSEDE's Extended Collaborative Support Service (ECSS)—through the Novel and Innovative Projects program. The ECSS team of David O'Neal (PSC), expert in optimizing atmospheric physics codes for HPC systems, and Campus Champion Fellow Eric Shook (Kent State University) helped port the variable-resolution global climate model CCAM [Commonwealth Center for Advanced Manufacturing] to Blacklight and adapt the code to allow very-high-resolution paleoclimate simulations over the Cape South Coast region. Referring to the workshop where first results were presented, Marean observed that "people were totally blown away, and it was so exciting to see something that has never been accomplished before—the production of glacial climate from a regional climate model," providing a foundation for further study of the climate experienced by early humans.

SOURCE: Adapted from Ralph Roskies, Pittsburgh Supercomputing Center, email message to Robert Harrison, October 21, 2015.

as a percentage of its total investments in NITRD research and in total dollar amount (Figure 2.2), even as the gap between request and available computing resources has grown (Figure 2.5).

NSF has a critical role with advanced computing because of its mission to initiate and support "basic scientific research and research funda-

Development) category. However, the point here is that the levels of investment are roughly comparable. DOE also supports industry research in processor and memory design, interconnects, and programming environments through the Research and Evaluation Prototypes program.

mental to the engineering process."[5] NSF's Division of Advanced Cyberinfrastructure (ACI) and its predecessor organizations have supported computational research across NSF and provided services to a user base that spans work sponsored by all federal research agencies. While a large fraction of the leadership-class investments have been driven by the mission-critical requirements of DOE and DOD, NSF has played a pivotal role in moving forward both the state of the art in HPC software and systems and the scale and scope of impacts that are enabled through their use to address key scientific challenges. This is in complement to other agencies such as DOE, the National Institutes of Health, DOD, and the Defense Advanced Research Projects Agency, which are mission driven.

Currently, NSF-supported advanced computing investment includes a diversity of resources. NSF supports several large-scale hardware facilities, together with associated staff expertise (Blue Waters at the University of Illinois, Urbana-Champaign, Stampede at the University of Texas, Austin, and Yellowstone at the National Center for Atmospheric Research-Wyoming), long-tail and high-throughput resources (Comet at the University of California, San Diego), data-intensive resources (Wrangler at University of Texas, Austin, and Bridges at the Pittsburgh Supercomputing Center), and cloud resources (Jetstream at Indiana University).

In February 2012, NSF published *Cyberinfrastructure for 21st Century Science and Engineering: Advanced Computing Infrastructure Vision and Strategic Plan*.[6] The document addressed broadly the cyberinfrastructure needed by science, engineering, and education communities to address complex problems and issues. The Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21) strategic plan seeks to position and support the entire spectrum of NSF-funded communities at the cutting edge of advanced computing technologies, hardware, and software. The CIF21 vision is to position NSF to "be a leader in creating and deploying a comprehensive portfolio of advanced computing infrastructure, programs, and other resources to facilitate cutting-edge foundational research in computational and data-enabled science and engineering (CDS&E) and their application to all disciplines."[7] In addition, the vision calls for NSF to "build on its leadership role to promote human capital development and education in CDS&E to benefit all fields of science and engineering."[8] After the publication of the strategic plan,

[5] National Science Foundation Act of 1950, as amended, and related legislation, 42 U.S.C. 1861 et seq.

[6] National Science Foundation (NSF), *Cyberinfrastructure for 21st Century Science and Engineering: Advanced Computing Infrastructure Vision and Strategic Plan,* NSF 12-051, http://www.nsf.gov/pubs/2012/nsf12051/nsf12051.pdf, February 2012.

[7] Ibid., p. 4.

[8] Ibid., p. 4.

BOX 3.2
Recent Nobel Prizes Recognize Simulation and Data as the Third and Fourth Pillars of Scientific Discovery

The 1998 Nobel Prize in chemistry was shared by Walter Kohn (University of California, Santa Barbara) for his "development of the density functional theory," and John Pople (Northwestern University) for his "development of computational methods in quantum chemistry." Density functional theory is the workhorse of computational chemistry and materials science and is now, perhaps, the central predictive computational tool of the multiagency Materials Genome Initiative. Through his development and distribution of the popular and efficient Gaussian software, Pople put powerful simulation tools into the hands of both theoreticians and experimentalists and ushered in the modern era of computational chemistry.

In 2013, the Nobel Prize for chemistry was awarded to Martin Karplus (University of Strasbourg and Harvard University), Michael Levitt (Stanford University), and Arieh Warshel (University of Southern California) for "the development of multiscale models for complex chemical systems." Their theoretical innovations synthesized classical and quantum models, and by realizing these advances in powerful computer programs they enabled the predictive modeling of chemical reactions in complex systems relevant to combustion, drug design, and biological systems. Levitt remarked,[1] "Computational structural biology, the field that I pioneered with Martin Karplus and Arieh Warshel, has certainly grown and matured through access to NSF-funded programs like XSEDE. . . . Our 2013 Nobel Prize in chemistry represents a huge step forward in the perception that high-performance computing is now of clear importance in a field of study previously considered as being purely experimental. The importance of XSEDE lies in its ability to work across many disciplines with a broad spectrum of users extending from novices to the most experienced users and all this at no cost of the researcher."

NSF formed a foundation-wide committee with representatives from all directorates to move NSF in the direction of achieving the advanced computing infrastructure vision.

3.2 GLOBAL ISSUES

Advanced computing is arguably the most critical ingredient of international science leadership, affecting leadership in other disciplines, providing an essential and often unique role in international experiments, and literally tying international collaborations together through networking, data, and computing infrastructure. Nearly every developed country has some type of national program for computing because of its importance to economic growth, science, defense, and society.

In 2011, Saul Perlmutter (Lawrence Berkeley National Laboratory), Brian Schmidt (Australian National University), and Adam Riess (Johns Hopkins University) shared the Nobel Prize in physics for "the discovery of the accelerating expansion of the Universe through observations of distant supernovae,"[2] the unknown cause of which is termed "dark energy." Their search for type Ia supernovae by an extended survey of thousands of galaxies was enabled by advanced image processing techniques and fast computers.[3] Reflecting on the role of computation in this work, Perlmutter remarked that "we need more supercomputers to narrow down the history of expansion because there are subtle differences in the histories that you get when using the theory of dark energy or Einstein's theory of general relativity. . . . This is all data intensive. From the very moment that you're trying to find the supernova, you're hunting for a little, tiny spark of light embedded in collections of thousands of images where each image is many megapixels, or even gigapixels collected by the big mosaic cameras. And then, to analyze the data and compare your results to different cosmological models also requires large computers, as do the Monte Carlo and all the statistical models you need. Finally, to compare these many models derived from first principles requires simulations of exploding stars—so that's another large computer job that is part of the story."[4]

[1] National Science Foundation, "Computational Science Takes the Nobel Stage," February 11, 2014, http://www.nsf.gov/discoveries/disc_summ.jsp?cntn_id=130427.

[2] Royal Swedish Academy of Sciences, "Nobel Prize in Physics 2011," press release, October 4, 2011.

[3] C. Day, "Nobel prizes for computational science," *Computing in Science and Engineering* 14(6):88, 2012.

[4] R. Brueckner, "Interview: Universe has Some Surprises in Store, Says Nobel Laureate Saul Perlmutter," November 26, 2013, http://insidehpc.com/2013/11/universe-surprises-store-nobel-laureate-saul-perlmutter/.

Leadership is difficult to quantify by a single, good benchmark or by analysis of a single system. The TOP500 list[9] ranks computers globally by their performance on the High-Performance Linpack benchmark. It is sometimes criticized for using this measure, which is much more compute-intensive than most modeling and simulation applications and does not reflect data-intensive workloads. Moreover, it does not contain all advanced computing systems, either because the system is business confidential, classified, or, as in the case of the Blue Waters system, because owners did not wish to take time away from normal operations to run the benchmark. Nevertheless, the list is an excellent source of historical data, and taken in the aggregate gives insights into investments in advanced computing internationally. The United States continues to dominate the

[9] See the TOP500 website at http://top500.org, accessed January 27, 2016.

list, with 45 percent of the aggregate performance across all machines on the July 2015 list, but it has dropped substantially from a peak of over 65 percent in 2008. NSF has had systems either high on the list (e.g., Kraken, Stampede) or comparable to the top systems (i.e., Blue Waters), reflecting the importance of computing at this level to NSF-supported science. Although there are fluctuations across other countries, the loss in performance share across this period is mostly explained by the growth in Asia, with China's share growing from 1 percent to nearly 14 percent today and Japan growing from 3 to 9 percent.

The Association for Computing Machinery's Gordon Bell Prize may be a better metric of scientific talent and usable performance; it is awarded annually to teams who demonstrate the best performance on a real application. Of the 26 awards to date, 20 were awarded to U.S. teams using U.S. systems, some involving participants from other countries. The other 6 were from Japanese teams on the Earth Simulator system in the early 2000s and the K computer in 2013, both custom-designed systems. Although NSF systems and staff were involved in some awards, DOE laboratory staff and systems have largely dominated the U.S. awards, and the most recent award was to a commercial entity and custom system (D.E. Shaw Research's Anton 2).

China's Tianhe-2 supercomputer stands at the top of the TOP500 list. Barely visible in high-performance computing 15 years ago,[10] China's presence on the list has continued to grow. China had, however, not announced at the time this report was being prepared its new 5-year plan for high-performance computing, so it is difficult to be precise about its future plans.

Japan has a long history of strong support for advanced computing in support of both science and industrial competitiveness. By several measures, Japan has often deployed and operated the world's fastest machine for science, most recently with the Earth Simulator (#1 on the TOP500 list from 2002-2004) and the K Computer (#1 on the TOP500 list in 2011). Japan has plans for both a powerful, leadership-class system, called the Flagship 2020 project, and a roadmap for nine powerful systems at university centers. The Flagship 2020 project is expected to provide roughly 1 ExaFLOP/s (10^{18} floating-point operations per second), although the focus is on a leadership-class system for science rather than a particular peak performance target. This is likely to be one of the most powerful systems in the world when it becomes operational and a pow-

[10] J. Alspector, A. Brenner, R.F. Leheny, and J.N. Richmann, "China—A New Power in Supercomputing Hardware," Institute for Defense Analysis, March 27, 2013, https://www.ida.org/~/media/Corporate/Files/Publications/IDA_Documents/ITSD/ida-documents-ns-d-4857.ashx.

erful advantage for Japanese scientists. Perhaps more importantly, Japan has a roadmap for what might be considered its second-tier systems, which will deploy, by 2020, nine highly capable systems at its university HPC centers, including eight with performance exceeding 10 PetaFLOP/s (Figure 3.1). Although what Japan actually acquires is of course subject to change, the roadmap illustrates the depth of the Japanese government's commitment to advanced computing.

The situation in Europe is complex. First, although there is a cross-European Union (EU) consortium, the Partnership for Advanced Computing in Europe (PRACE), most investment decisions are made by individual countries. Germany, for example, has made significant investments to support both basic research and industrial competitiveness. The Juqueen system at the Juelich Research Center, an IBM Blue Gene/Q system, is one of the most powerful systems in the world. It is over half the size of the Mira system at the Argonne National Laboratory in Illinois, which is one of the two leadership-class systems operated by DOE's Office of Science. Germany has three other systems ranked in the top 25. Outside the EU, other European countries have their own powerful systems. For example, the Swiss National Supercomputing Center operates a system ranked #6 by the TOP500 list in June 2015.

In terms of data-intensive computing, the United Kingdom's eScience program identified the emergence of data-intensive computing—as a complement to the tradition simulation and modeling research activities—as long ago as the early 2000s. It would later prove influential in launching NSF's cyberinfrastructure program and as an inspiration for other national programs, such as the Australian eResearch and the Dutch eScience programs.[11]

Another country that has made significant recent investments is Saudi Arabia. Both its Shaheen system, an IBM Blue Gene, and its more recent Shaheen II, a Cray XC40, both installed at the King Abdullah University of Science and Technology and used for scientific research, have been among the world's fastest systems for science research. The Indian government has approved a 7-year supercomputing program worth $730 million (Rs. 4,500-crore) intended to revitalize its program and raise the nation's status as a world-class computing power.

Leadership means drawing the best talent nationally and internally and supporting training of the next generation of scientists in computer

[11] International Panel for the 2009 RCUK Review of the e-Science Programme, *Review of e-Science 2009: Building a UK Foundation for the Transformative Enhancement of Research and Innovation*, Research Councils UK and the Royal Society, 2009, https://www.epsrc.ac.uk/newsevents/pubs/rcuk-review-of-e-science-2009-building-a-uk-foundation-for-the-transformative-enhancement-of-research-and-innovation/.

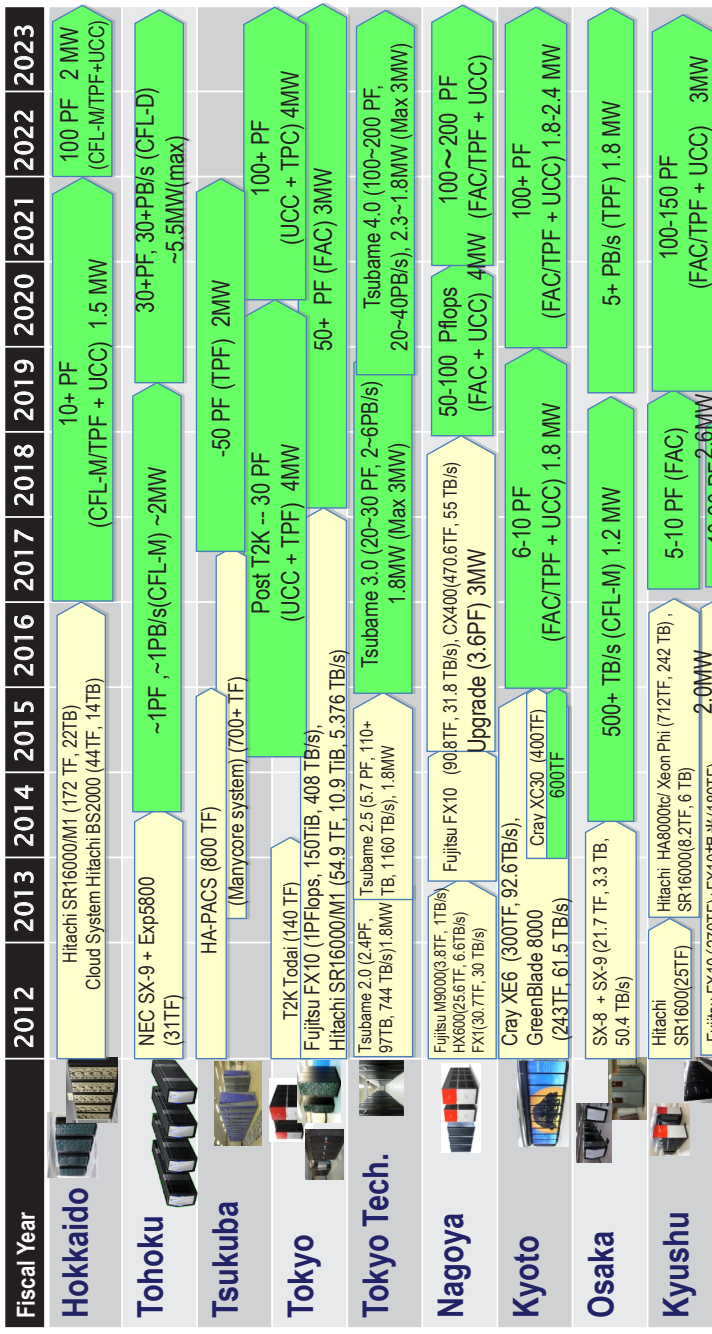

FIGURE 3.1 Snapshot of Japan's roadmap for its nine high-performance computing infrastructure university centers. SOURCE: High-Performance Computing Infrastructure Consortium in Japan, available at http://www.hpci-c.jp/news/20150527_center_summary.pdf.

science, data sciences, and scientific computing. This next generation includes the designers of future computer architectures, systems software, algorithms, and computational tools, as well as the applications. It is difficult to quantify these future impacts, except to point to the historical benefits that the United States has seen from computing and the continual demand by industry, government, and academia for experts in the design and use of advanced computing, networking, and data systems.

4

Future National-Scale Needs

Forecasting the future national needs for advanced computing is difficult. The recent revolution in data-centric computing emphasizes both the broad generality of computation and the difficulty in forecasting what future needs will emerge as new methods and opportunities arise. In addition, the end of Dennard (frequency) scaling and the move to massive parallelism and new computing architectures mean that many existing applications will need to be updated or replaced in order to make effective use of forthcoming systems. This chapter discusses previous approaches for discussing needs that were based primarily on floating-point performance and makes recommendations for how to think about the more complex, multidimensional requirements for computing and data systems in the future.

4.1 THE STRUCTURE OF NSF INVESTMENTS AND THE BRANSCOMB PYRAMID

For the past 30 years, National Science Foundation (NSF) investments in advanced computing have focused on the top two levels of the Branscomb pyramid (Box 4.1): leadership-class and center-class machines (Figure 4.1). Although the Branscomb pyramid has been invoked (and revised) for decades, it focuses only on the computational aspects of the portfolio. Variations have appeared over the years that include other axes, such as storage and bandwidth (see Chapter 5), but the pyramid remains as a useful, albeit incomplete, organizing principle.

The pyramid conveys more than just a portfolio of computational capability that spans five (or more) orders of magnitude in performance. It implies that there is substantial congruence in the programming models up and down the pyramid. Moreover, there is an expectation that the number of users roughly scales with the horizontal extent of each level. If the pyramid is to represent resource consumption, then these last two issues must be considered.

In the area of programming models, Kogge and Resnick[1] showed that there was a significant discontinuity in 2004 with a sudden growth in the diversity in architectures in the TOP500 systems. Kogge and Resnick noted that this was the result of barriers to the previous several decades of increases in single-core performance, more memory per chip, memory latency, and interconnect performance. In 2004, there was growth in multicore systems relying on simpler cores and slower clock speeds, slow growth in memory density, and complex interconnects. Thus, the advanced computing portfolio began to be a combination of heavyweight architectures (e.g., Cray XE6 nodes for National Center for Supercomputing Applications' Blue Waters),[2] lightweight architectures (e.g., IBM BlueGene/Q for Argonne National Laboratory's Mira), hybrid architectures (e.g., Cray XK7 for Oak Ridge National Laboratory's Titan or the Intel Xeon Phi nodes on Texas Advanced Computing Center's [TACC's] Stampede), and heterogeneous multicore systems on a chip (e.g., ARM Cortex).

The current expectations are that scientists can use the same codes across a broad range of systems and that U.S. vendors do not develop chips and packages uniquely for the highest-end systems. However, future performance improvements to the largest-scale systems may require even more exotic technologies to cope with such issues as resilience, power management, and energy efficiency. As a result, the high-performance community is currently exploring whether those seeking the very highest performance will have to adopt new programming models, tools, and practices sooner. On the other hand, users of both scientific and commercial systems see considerable value in maintaining a common programming model across the spectrum—from the single chip in handheld devices to the largest multi-processor systems.

Regarding the number of users at each level of the pyramid, there are economic and cultural pressures that sometimes work to increase the

[1] P.M. Kogge and D.R. Resnick, *Yearly Update: Exascale Projections for 2013,* Sandia Report SAND2013-9229, Sandia National Laboratories, October 2013, http://prod.sandia.gov/techlib/access-control.cgi/2013/139229.pdf.

[2] Strictly speaking, Blue Waters is a hybrid machine, but is predominantly lightweight because only about 16 percent of the nodes include GPUs.

BOX 4.1
The Branscomb Pyramid

What Is the Branscomb Pyramid?

The Branscomb pyramid was developed as part of a panel report to the National Science Board in 1993[1] and was named after the panel chair, Lewis Branscomb. It relates three things: computational power (y-axis, more is up), number of systems (also y-axis, fewer is up), and number of users (width of figure in x direction). It should really be a right triangle (so x-axis represents number of user in some possibly log scale).

Why Was It Defined and How Did It Keep Its Value?

The Branscomb pyramid was defined to graphically show the relationship between the three components described above, and in particular the fact that the more capable and powerful the system (up), the fewer you can afford (also up) and the fewer "people" (or research projects) that can be supported. It was this last relationship that was the important insight graphically presented by the Branscomb pyramid. Many parts of this picture are still true today, but there are also many changes that, if not rendering the Branscomb pyramid obsolete, require a much more careful interpretation of the current situation in computing.

In What Ways Does the Branscomb Pyramid Misrepresent the State of Computing Today?

There are several, and each is important and addressed in this report:

- Compute power is not simply measured. Even in high-performance computing (HPC), it has long been known that aspects such as sustained memory bandwidth (itself very different from peak memory bandwidth) are often far more indicative of application performance than peak floating-point operations per second (FLOP/s). The revolution in data science adds another set of dimensions to this, by adding data volume, bandwidth, latency, etc., to the list.
- The ability to solve a problem depends on far more than just hardware. As has been noted (see Box 2.3) for some applications, using modern algorithms on 35-year-old hardware will give a faster solution than using a 35-year-old algorithm on modern hardware. The Branscomb pyramid does not represent this aspect of computing—the combination of software, algorithms, and human expertise in solving the problem. Since the y-axis captures, in some sense, both the number of systems and the cost of those systems, the cost of this non-hardware expense also needs to be captured.
- New access modes make it possible for a much greater pool of users to have access to extremely large resources. In the old model, captured in the Branscomb pyramid, users of (especially) the peak systems typically used a large fraction of the system for a significant length of time. This is still the primary mode of operation for leadership-class systems around the world—research projects

are allocated tens of millions of node-hours (hundreds of millions to billions of CPU core hours). Due to the limited nature of the resource, this implies that there can be only a relatively small number of such projects. However, some research problems may require only a small amount of total time but be infeasible without access to the special capabilities of a leadership-class system. An example is an application that requires 1 PB of memory to run and has linear complexity in that amount of data. On today's leadership-class systems, this would take only a few seconds to minutes to run, assuming the data are already located on the system's high-performance disk. Although today's HPC systems are not set up to run this sort of workload, converged systems would be well suited, as would cloud service models, provided that the necessary data are already colocated with the cloud computing resources.

Is This a New Observation?

No, the growing gap between the Branscomb pyramid and present-day computing has been recognized in the community. For example, in the 2011 report *National Science Foundation Advisory Committee for Cyberinfrastructure: Task Force on Campus Bridging,*[2] there is this finding:

> The cyberinfrastructure environment in the US is now much more complex and varied than the long-useful Branscomb Pyramid. As regards computational facilities, this is largely due to continued improvements in processing power per unit of money and changes in CPU architecture, continued development of volunteer computing systems, and evolution of commercial Infrastructure/Platform/Software as Service (cloud) facilities. Data management and access facilities and user communities are also increasingly complex, and not necessarily well described by a pyramid.

If anything, this understates the situation.

So Why Do We Keep Talking About the Pyramid?

The Branscomb pyramid is a convenient way to represent those three original relationships. Even with greater access through new service models or application gateways, there will be only enough "leadership-class" computing for a relatively small number of users and user groups. While this may still span thousands or tens of thousands of users, the computer revolution has put significant computing power into the hands of everyone. As long as the many simplifications that go into this picture are understood, it remains a valuable way to express this relationship.

---

[1] National Science Foundation Blue Ribbon Panel on High Performance Computing, *From Desktop to Teraflop: Exploiting the U.S. Lead in High Performance Computing*, National Science Board, NSB-93-205, 1993.

[2] NSF Advisory Committee for Cyberinfrastructure Task Force on Campus Bridging, *Final Report*, March 2011, http://www.nsf.gov/od/oci/taskforces/TaskForceReport_CampusBridging.pdf.

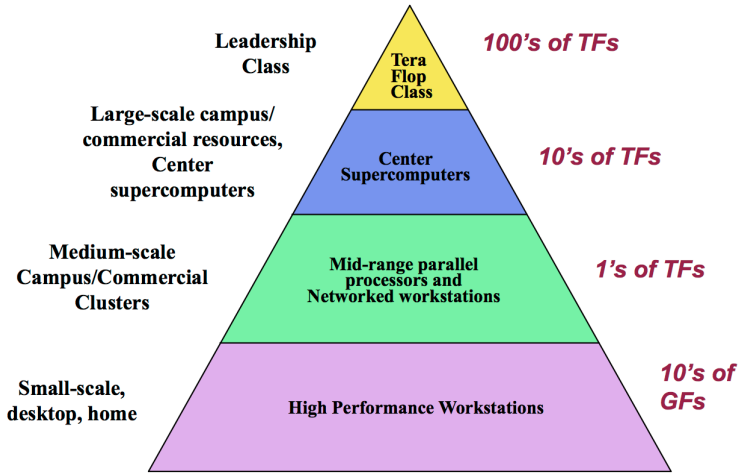

FIGURE 4.1 The Branscomb pyramid that appeared in the National Science Foundation Advisory Committee for Cyberinfrastructure Task Force on Campus Bridging *Final Report,* March 2011, http://www.nsf.gov/cise/aci/taskforces/TaskForceReport_CampusBridging.pdf. NOTE: GF, gigaflop; TR, teraflop. SOURCE: Image by Francine Berman, San Diego Supercomputer Center, "Beyond Branscomb," presentation at Clusters and Computational Grids for Scientific Computing, September 10-13, 2006, licensed under the Creative Commons 3.0 unported attribution license (http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode).

number of users at the top two levels of the pyramid (leadership class and center class). This can be done either through limiting the scope of resources allocated to a single job (number of processors, run time, etc.) or through explicit programs designed to increase the user base (e.g., the MATLAB portal at TACC or specific funding solicitations). Thus, a modern interpretation of the pyramid needs to clearly distinguish between the usage of services and the provisioning of services.

The modern interpretation also needs to recognize that there are significant pressures to build out the resources that are affordable rather than those that are needed. While budget realism is essential, it can lead to an acquisition process dominated by cost considerations rather than one driven by the science requirements. Previous studies of the advanced computing portfolio consistently cite similar science needs (computational fluid dynamics, astrophysics, cosmology, materials science, etc.) and the gap between these needs and the available computational resources. This

suggests that there remains a persistent gap between the science requirements and the advanced computing components that are eventually provided.

Looking forward, the constellation of resources to support science needs is much broader and more diverse than in the past, ranging from cloud-based systems to high-speed wireless networks to the more traditional centers and campus-based facilities. The range of usage models has also widened (Box 4.2), and the science community is generating a flood of new data. Thus, there are more demands from new user communities. Moreover, the adoption timescale is much shorter for new technologies and capabilities. When these forces are coupled, new workflows emerge and evolve much more rapidly.

In addition, although discussions of computing needs often focus on what systems need to be acquired and operated, the *effective use* of computing systems, particularly large-scale systems that address national needs, requires much more than just computer hardware. Expert staff are needed to operate the systems, diagnose performance problems, and help the user community. Software needs to be tuned and updated for each generation of system, and community codes, which encourage sharing of effort and efficient use of resources, need to be nurtured. New algorithms are needed to address new problems and to make better use of the hardware. Data need to be preserved and curated. These needs must not be forgotten when provisioning computing resources.

4.2 DATA-INTENSIVE SCIENCE AND THE NEEDS FOR ADVANCED COMPUTING

The current generation of advanced computing infrastructure focuses largely on meeting the requirements of workflows for *simulation* science that has fueled advances across many disciplines over the past two decades. However, the landscape of scientific workflows—the series of computational or data manipulation steps required to carry out a scientific analysis—is evolving rapidly to respond to the remarkable potential that data-driven science (or more colloquially "big data") holds for answering open scientific questions such as "How do we reliably detect a potential pandemic early enough to intervene?" or (in combination with simulation) "Can we predict new materials with advanced properties before these materials have ever been synthesized?"[3]

Advances in sensing and measurement from empirical approaches have resulted in a wealth of scientific data that can be utilized to develop

[3] National Institute of Standards and Technology, Big Data Program, see http://bigdatawg.nist.gov/home.php.

BOX 4.2
Supporting Different Usage Models of Computing Resources

Computational science needs span a wide spectrum. Some applications require a single, large, tightly coupled distributed memory parallel computer (such as the earthquake, human immunodeficiency virus, and general relativity examples in Box 3.1). Others may require large numbers of runs, each of which may require only a few compute nodes but run for hours, days, or even weeks (such as the avian lineage example in Box 3.1); the total compute requirement of these applications can be very large. Other applications may require real-time or continuous access to computing resources—for example, if computing is required to process data from an active experiment such as a telescope or sensor network.

Each of these types of computing (and there are others) requires a different service model. For example, applications requiring a large, tightly coupled system need the resources to be made available in a coordinated fashion and may need the resources to be allocated to ensure efficient communication between the processes in the application. Applications using large numbers of runs may need to run without interruption for days or even weeks. It can be difficult to accommodate different types of applications on the same system—for example, long-running applications using a few nodes can fragment the available nodes, making it impossible to schedule the resources for a tightly coupled application to run efficiently.

In the short term, providers of computing resources can develop different service models that match modern science workflows, possibly by adaptively partitioning the computing system into groups of nodes that support development and tuning of applications, real-time applications, long-running applications requiring only a few nodes, and highly parallel applications (adaptive to support requirements that change with time, such as the need to use an entire system to run a single tightly coupled parallel application).

In the longer term, enhancements to the applications could help relax some of the usage constraints. For example, unless one is analyzing a real-time stream from an experiment, there may be no actual requirement that a single code be able to run uninterrupted for weeks. This is an artifact of how the program has been designed and written and the availability of tools that can provide ways to stop and restart an application (e.g., through system-level checkpoint-restart or through a user-level software library for checkpoint-restart). In the case of highly parallel applications, it is often possible to make the code more flexible both with respect to requirements about when and on which compute nodes the programs begin running. However, doing so can require significant changes to the programs and may require expertise that the computational scientists who have developed the applications may not have. More fundamentally, both the algorithms used and the implementations of those algorithms need to be examined. In some cases, the methods were appropriate for smaller problems but not for the size of problems to which they are now being applied.

new models or refine existing models in order to gain new insights. As the costs of sensors continues to decline, experimental and observational data are being generated not only by large instruments (assembled from many small sensors), but also from large arrays of geographically distributed sensors. Social media feeds are important new data sources for social science research.

More generally, as data accrue from experiments and simulations, and as data from multiple experiments and simulations are integrated, scientific discoveries are increasingly being made from the accumulated and integrated data using advanced computing. This is sometimes known as the "fourth paradigm" of scientific discovery, because it supplements discovery paradigms based on theory, experiment, and simulation.[4] Further, there are additional opportunities for scientific insights at the interfaces of each of these paradigms of discoveries.

A good example is provided by the aspirations of the genomics community. Microarray data sets—in which several hundred to several thousand genes were measured under different experimental conditions, resulting in data sets that were megabytes in size—have given way to data sets in which the expression level of the entire genome is measured, resulting in data sets that are gigabytes in size. Similarly, gene chips produce data sets that are kilobytes in size, while whole genome sequencing is producing data sets that are hundreds of gigabytes in size. As a rough rule of thumb, genomic and related clinical data for a cancer patient (assuming normal tissue is sequenced, a tumor is sequenced, and a tumor after relapse is sequenced) are approximately 1 TB in size. A cohort of 10,000 patients requires 10 petabytes of storage, and a cohort of 1 million patients (a goal for the community over the next several years) would require 1 EB of storage.

Another example is the Large Synoptic Survey Telescope (LSST), which will produce a wide-field astronomical survey of the universe using a 8.4-meter telescope and 3-gigapixel camera. LSST will collect 15 TB of raw image data every night that will be processed in near-real time to provide scientists with alerts about new and unexpected astronomical events and reprocessed annually. It will yield a 200 PB data archive by the end of the decade-long survey.

Today, both scientific researchers and businesses use a wide array of data analytics tools. In some areas, large-scale analytics companies such as Google and Amazon—rather than the scientific community—are in a leadership position. In some other areas, the needs of science may not

[4] T. Hey, S. Tansley, and K. Tolle, eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, 2009, http://research.microsoft.com/en-us/collaboration/fourthparadigm/.

overlap the needs of industry. For example, the statistical analysis of large, in memory data sets (such as the anticipated output from the LSST) is more similar to a scientific computation that to the type of analyses that have generally interested Google or Amazon. Looking ahead, there may be opportunities for researchers to make better use of the tools and concepts developed by industry or for the industrial and scientific communities to partner more effectively where their needs overlap.

One challenge with respect to the private sector is that salaries for those trained in data analytics can be far higher in the private sector than in the academic research community. This makes it difficult for academic researchers to stay abreast of emerging technical tools that enable data-intensive science. For NSF, this creates two challenges. The first challenge is to act strategically to develop the needed workforce to support both science and business applications. The second is to find ways to keep people with these skills in the science community despite lower salaries—for example, by offering reasonably secure, stable career paths as well as exciting work.

From a technical requirements perspective, infrastructure for data-intensive science needs to consider data acquisition, storage and archiving, search and retrieval, analytics, and collaboration (including publish/subscribe services). Recent NSF requirements to submit data management plans as part of proposals signal recognition that access to data is increasingly important for interdisciplinary science and for research reproducibility. Although the focus is sometimes on the hardware infrastructure (amount of storage, bandwidth, etc.), the human and software infrastructure is also important. Understanding the software frameworks that are enabled within the various cloud services and then mapping scientific workflows onto them requires a high level of both technical and scientific insight. Moreover, these new services enable a deeper level of collaboration and software reuse that are critical for data-intensive science.

When considering cyberinfrastructure requirements, the needs of data-intensive science are often considered as separate from those of the more traditional computationally intensive science problems, such as climate modeling. However, as new massive data sets become available (e.g., the LSST project), the line separating these two types of research becomes blurred. As a specific example, today's climate models have dramatically increased temporal and spatial resolution compared to models from 10 years ago. Although this has greatly improved model performance, many processes that once could be parameterized simply at coarse resolution now must be included explicitly in high-resolution models. One example is the representation of cloud processes, where the physics are poorly understood and critical parameters cannot be measured.

The June 2014 issue of *Philosophical Transactions of the Royal Society*[5] was devoted to a generation of coupled deterministic/probabilistic models that illustrate a possible convergence between compute-intensive models and data-intensive models.

4.3 FORECASTING FUTURE REQUIREMENTS

Developing science requirements for any large project takes enormous experience and insight (and creativity). Establishing requirements that can be achieved within a cost and schedule framework is even harder. By its very nature, science is unbounded. There are always pressures to improve our understanding or to make better predictions.

Past NSF efforts have, in general, implicitly constrained requirements, either through budget caps or by technical feasibility. Obviously, there is some iteration between these elements, although the NSF petascale program[6] was driven largely by cost and desired sustained computational performance. There was an implicit assumption that the acquired systems would enable a certain class of scientific models and analyses to be addressed. With this approach, there was a risk that the science requirements could have been only loosely coupled to the systems that were acquired, and key areas of science could have been left unserved.

Today, with growing demand for computing and constrained budgets, it has become especially important to understand the relative benefits and risks of different technical approaches for the science portfolio. This section describes some of the challenges NSF will face in developing science requirements for advanced computing.

- *Quantifying science benefits.* It remains an unsolved (and probably unsolvable) problem to accurately quantify the return on investment in scientific research, and certainly it is not possible to predict the return. But it may be possible to consider the likely costs and risks of different approaches, as well as the possible opportunities, and use these to guide the setting of objectives and priorities.
- *Suitable measures of advanced computing performance.* It is also important to avoid reducing the requirements to a too simplistic measure, such as peak floating-point operations per second (FLOP/s). The system with the best FLOP/s per dollar may not provide the best value for science

[5] For example, T. Palmer, P. Düben, and H. McNamara, Stochastic modelling and energy-efficient computing for weather and climate prediction, *Philosophical Transactions of the Royal Society A* 372:20140118, 2014, doi:10.1098/rsta.2014.0118.

[6] National Science Foundation, "High Performance Computing System Acquisition: Towards a Petascale Computing Environment for Science and Engineering," Program Solicitation NSF 05-0625.

applications, where sustained, rather than peak, performance is far more important. Key areas of science may have different requirements, such as sustained I/O performance for data-centric applications.

- *Rapidly evolving science needs.* Any science-driven requirements process must also confront the issue that the science itself is changing rapidly on the timescale associated with large-scale advanced computing acquisition and deployment. Past experience has shown that although a procurement can be completed in several years, large systems sometimes take as long as 10 years from initial concept to full availability to users. A rolling decadal roadmapping process could help inform users about plans for the upgrade and replacement of existing systems and, more generally, the performance characteristics of expected future systems.
- *Responding to the rapid evolution of data-driven science.* For example, new classes of weather forecasting models combine the tools of computational fluid dynamics along with data-driven parameterizations to improve forecasts for small-scale (but intense) events. Moreover, these data-driven approaches often rely on ensembles of many model runs. As the network of real-time sensors connected through high-speed links to the Internet grows, these data-driven models will require new capabilities in regard to computation, storage, and bandwidth. Much as with business analytics, these data-intensive methods will be based on near real-time streaming data. Moreover, new technologies could have profound benefits for data-intensive science. One can envision networks of sensors where each sensor node has local compute capabilities that rival the supercomputer performance of only a few years ago. The impact on adaptive computing and sensing could be significant, realizing one of Jim Gray's admonitions to move computation to the data.[7]

Such workflows will require autonomous tools to assess data quality and model performance; human intervention and control will not scale up to these new models. Planning for these changing (and often poorly formulated) requirements will require considerable insight. These changing scientific workflows extend to the human side of scientific computing as well. Especially in regards to data-intensive science, reproducibility will be challenging. These requirements will often be as important as the traditional technical requirements of CPU performance, latency, storage, and bandwidth.

- *Complex and rapidly changing technology landscape.* Just a few years ago, mainstream high-performance computing was limited to commodity x86 processors from Intel and Advanced Micro Devices and IBM Power

[7] A. Szalay and J.A. Blakeley, "Gray's Laws: Database-centric Computing in Science," in *The Fourth Paradigm: Data-Intensive Scientific Discovery* (T. Hey, S. Tansley, and K. Tolle, eds.), Microsoft Research, Redmond, Wash., 2009.

processors. Today, high-performance computing is making use of general-purpose graphical processing units and accelerators, and some designs such as Intel's Xeon Phi are focused on scientific computing. Looking ahead, more diversity is likely to come in the form of things like integrated non-volatile random-access memory and processors integrated with memory. At the same time, much of the broader commercial industry is focused on the needs of mobile devices.

The technical landscape now has a range of new service providers beyond the hardware/software companies. Much has been made of cloud services, although most of the discussion has focused on its elastic computation and storage model along with an aggressive pricing strategy. However, a key capability of cloud services is the rich software framework that is available for users. Not only can these services and frameworks be leveraged to support changing science workflows, they can be extended to include new components that can then be made available to other users. The science community rapidly adopts these new "providers," such as Dropbox, until a new and improved service appears on the market.

Along with the challenges of a changing scientific and technical landscape, any requirements process must recognize that there will always be gaps. For example, one cannot predict with any certainty the technical or business directions of the major hardware and software vendors beyond several years. To give another example, the International Technology Roadmap for Semiconductors makes evident the major technical challenges faced by industry in maintaining the pace of performance improvements several years out. Widely used proprietary software such as CUDA is also subject to rapid change.[8] Adoption rates of (or resistance to) new technologies is another challenge. The requirements process must at least consider the economic forces that are driving the technology market as well as the political and cultural forces that either speed or resist adoption. Moreover, it must also recognize that the science community must be capable of using the advanced computing portfolio, which means one cannot follow a "build it and they will come" approach.

4.4 THINKING ABOUT A NEW APPROACH TO DEVELOP REQUIREMENTS FOR ADVANCED COMPUTING

At its heart, there needs to be a rigorous process for development and assessment of the science requirements for advanced computing. The process needs to ensure that these science requirements have substantive

[8] Language standards such as C++, Fortran, and OpenMP are less subject to unexpected changes over the short term.

feedbacks between the science, the technical approach, and cost. It also needs to make explicit what research can and cannot be done within a given budget envelope and with a particular set of acquisition decisions. Moreover, a clear and bounded vision for the types of science that advanced computing will support is needed. For NSF, this will likely mean developing an understanding of how much of the portfolio can be supported by a more data-capable general-purpose platform (and what specific data capabilities are needed), and what is left over that either needs specialized advanced computing supported as cyberinfrastructure, or perhaps topical computing supported in part by the science programs. A more productive view than just focusing on the hardware that can be afforded would be to describe and quantify a set of services that are needed to meet a class of science challenges. Such an approach would allow a more flexible investment strategy (build a specific center, work with cloud service providers, etc.) rather than trying to fit everything into a small set of infrastructure assets.

A process that relies on documented science objectives and assessment of the progress made toward achieving these objectives, rather than simply statements that greater computational capacity will improve understanding of a specific scientific process or phenomenon, can help improve future decisions. For example, such an assessment might show that the ability to run an ensemble of 1,000 short-term weather forecasting models will improve the quality of the forecasts by a specific percentage. These science objectives capture the value of the requirement as a function of benefit and affordability, where benefit is in turn a function of importance, quality, utility, and probability of success. This approach to cost-benefit analysis would allow science communities to understand the linkages between science, technical requirements, and cost, thus allowing more rational analysis of the trade space of science capability, technical requirements, and cost.

The more granularity that can be provided in terms of costs and benefits, the better the decisions that can be made when the inevitable trades need to be made between science, technology, and cost. For example, one must consider the full costs of the advanced computing components, including both the acquisition as well as operations and maintenance costs (including hardware, software, and staff costs). In doing so, one must also consider fixed costs (staff, maintenance contracts, etc.) as well as marginal costs (elastic costs of cloud services, etc.). This is especially important as NSF and the science community are moving toward a model of buying services as needed rather than recognizing the true fixed costs. For example, a scientific programmer may spend only 1 month on a particular project, but the employer needs to provide a full year of salary. The existing supercomputer centers repeatedly note the difficulty in maintain-

ing experienced and highly trained staff as the funding agencies move into a mode of buying talent by the month rather than providing stable support for the expertise the scientific community depends on.

Another component of the requirements analysis is to identify the linkages and dependencies across NSF's advanced computing portfolio. Such systems engineering across a diverse portfolio will not be simple, but it is essential to developing a resilient portfolio that can support a wide range of science areas. Advanced computing requirements should also take into account the science needs (because NSF provides advanced computing to research communities funded by other federal agencies) and contributions of other federal agencies (because some NSF-funded researchers make use of advanced computing provided by other federal agencies).

Today, most users of NSF's advanced computing infrastructure have no understanding of the value of the resource that they have been granted. While rationed (through the allocation process), advanced computing resources are for the most part "free" (there is no charge for them). This leads to a mindset that puts little value on making efficient use of these resources, particularly because there is no way in the current system to trade, for example, computer time for expert help in tuning applications. As a first step, building an awareness of the value and cost of computing resources may lead to a more holistic and comprehensive approach to using the advanced computing resources. One possible way to do this would be to provide a dollar value of the computational resource granted. There are some dangers in this approach; the goal is to build awareness of the costs and value, not create a chargeback mechanism. Section 6.3.8 describes a possible pilot project to explore the benefits, risks, and problems with such an approach.

This process will yield a much more thorough understanding of the complete costs and technical feasibility of the portfolio in the context of documented science objectives. This will inform an analysis of trade-offs that will modify the approach to fit within economic and political realities. Balancing its primary science mission with the need to operate infrastructure will require constant assessment by NSF, as noted in the recent decadal survey of the ocean sciences.[9]

4.5 ROADMAPPING

The Department of Energy (DOE) has created a roadmap for future advanced scientific computing research systems that provides research-

[9] National Research Council, *Sea Change: 2015-2025 Decadal Survey of Ocean Sciences*, The National Academies Press, Washington, D.C., 2015.

ers with a view of what capabilities to expect (Figure 4.2). By describing the next 3 to 5 years of leadership computing systems, DOE has given the community useful information about the general characteristics and organization of the next DOE leadership-class systems. The longer-term roadmaps are less concrete but still provide information about the general intentions of DOE: continue increasing single machine performance, which contrasts with keeping the single machine performance about the same but increasing the total number of machines.

NASA and other mission agencies have regularly employed a roadmap process that outlines a small set of science themes that will engage the scientific enterprise over the next few decades.[10] These themes then serve as a framework for a series of notional missions or activities that address specific questions in the theme. Some of these questions may need to be addressed sequentially (e.g., the approach to one question may depend on the knowledge gained from answering another question), while others may proceed roughly in parallel. Taken together, the notional missions lay out a roadmap that is based on scientific progress at each stage. However, unlike a decadal survey, the roadmap also lays out options and multiple pathways and identifies the scientific and technical challenges.

A fundamental aspect of the roadmapping process is that it is driven by the science themes, rather than simply a quest for a certain level of technical capability. Also, the process lays out options and challenges. Lastly, it links scientific progress to technological capabilities, rather than a "build it and they will come" approach. Maintaining a linkage between science need and technological capability is an important aspect of effective roadmapping.

Implementing a roadmapping process that reflects all of the research supported by NSF advanced computing will not be easy. For one thing, as Dennard scaling has fallen off, there is growing pressure to use domain-specific hardware to achieve greater computing performance. For another, the requirements have become more diverse as the range of science using advanced computing has grown. Specialized accelerators, storage facilities, or other capabilities may be needed to enable some research objectives efficiently, and it will in any event be difficult to roll up requirements into a sufficiently small set. It may be necessary to develop separate road-

[10] For example, the end-to-end challenges in managing massive research data are considered in NASA Earth Science Technology Office/Advancd Information Systems Technology (ESTO/AIST) Big Data Study Roadmap Team, "NASA Earth Science Research in Data and Computational Science Technologies," September 2015, http://ieee-bigdata-earthscience.jpl.nasa.gov/references/aist-big-data-study-draft-summer-2015.

System Attributes	NERSC Now	OLCF Now	ALCF Now	NERSC Upgrade	OLCF Upgrade	ALCF Upgrades	
Name Planned Installation	**Edison**	**TITAN**	**MIRA**	**Cori 2016**	**Summit 2017-2018**	**Theta 2016**	**Aurora 2018-2019**
System peak (PF)	2.6	27	10	>30	**200**	>8.5	180
Peak power (MW)	2	9	4.8	<3.7	**13.3**	1.7	13
Total system memory	357 TB	710 TB	768 TB	~1 PB DDR4 + High Bandwidth Memory (HBM) + 1.5 PB persistent memory	**> 2.4 PB DDR4 + HBM + 3.7 PB persistent memory**	>480 TB DDR4 + HBM	>7 PB High Bandwidth On-Package Memory, local memory, and persistent memory
Node performance (TF)	0.460	1.452	0.204	>3	**>40**	>3	>17 times Mira
Node processors	Intel Ivy Bridge	AMD Opteron Nvidia Kepler	64-bit PowerPC A2	Intel Knights Landing many-core CPUs Intel Haswell CPU in data partition	**Multiple IBM Power9 CPUs and multiple Nvidia Voltas GPUS**	Intel Knights Landing Xeon Phi many-core CPUs	Knights Hill Xeon Phi many-core CPUs
System size (nodes)	5,600 nodes	18,688 nodes	49,152	9,300 nodes 1,900 nodes in data partition	**~4,600 nodes**	>2,500 nodes	>50,000 nodes
System interconnect	Aries	Gemini	5D Torus	Aries	**Dual Rail EDR-IB**	Aries	2nd Generation Intel Omni-Path Architecture
File system	7.6 PB 168 GB/s, Lustre®	32 PB 1 TB/s, Lustre®	26 PB 300 GB/s GPFS™	28 PB 744 GB/s Lustre®	**120 PB 1 TB/s GPFS™**	10 PB, 210 GB/s Lustre initial	150 PB 1 TB/s Lustre®

FIGURE 4.2 Advanced Scientific Computing Research (ASCR) computing upgrades at a glance. NOTE: Acronyms defined in Appendix D. SOURCE: U.S. Department of Energy, Office of Science, "ASCR Computing Upgrades at a Glance," http://science.energy.gov/~/media/ascr/pdf/facilities/ASCR_Computiing_Facility_Upgrades.pdf, accessed April 25, 2016.

maps by science area, and then aggregate similar needs across areas (e.g., the use of unstructured grids and iterative linear methods in simulations).

Another challenge is determining a good configuration for a computing system that requires more than just a measure of the number of operations per second (e.g., clock speed) or size of data (e.g., disk space). Research has shown that simple benchmarks are, individually, rarely predictive of the performance of an application, and even collections of benchmarks give only a rough estimate.[11] Highly accurate performance estimates, while possible, remain a difficult and time-consuming process. As a result, the community has relied on a very simple measure of computing performance, based on floating-point performance only. For example, XSEDE allocates resources in service units (SUs), which are related to the performance of High-Performance Linpack. This reflects the peak floating-point performance of a system but little else. Allocations under the PRAC program for Blue Waters are in node-hours, which is more easily related to the specific system but is not easily convertible to SUs or node-hours for a different system. A first step would be to gather more information about the needs of applications. Relevant measures include memory size and bandwidth, data size and bandwidth, interconnect bandwidth and application sensitivity to interconnect latency, integer and floating-point performance, and long-term data storage requirements. Some of this information could be gathered by tools designed for this purpose, applied to an application running on a current system, reducing the burden on the computational scientists. An example of what the first step in this process might be is presented in Box 4.3.

Despite the challenges and the likely imperfections in the roadmaps that are developed, it should be possible to develop roadmaps that provide enough guidance to the community to be worthwhile. By focusing on the overall picture rather than the high-resolution details, roadmaps can indicate to the community what the major investments will look like. This will allow researchers to make better decisions about future research. It will also allow researchers to start preparing their software to be ready for future systems—for example, by providing advance notice about significant changes in architecture or configuration.

[11] See, for example, L. Carrington, M. Laurenzano, A. Snavely, R. Campbell, and L. Davis, How Well Can Simple Metrics Represent the Performance of HPC Applications?, in *Proceedings of the ACM/IEEE SC 2005 Conference*, 2005, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=1560000.

BOX 4.3
Gathering Data About Computational System Needs from Proposals for Research Requiring Advanced Computing

As a first step in gathering more information about the computational system needs of applications, the National Science Foundation (NSF) could ask that all proposals for research that would require advanced computing include relevant measures such as memory size and bandwidth, data size and bandwidth, interconnect bandwidth and application sensitivity to interconnect latency, integer and floating-point performance, and long-term data storage requirements. Some of this information could be gathered by tools designed for this purpose, applied to an application running on a current system, reducing the burden on the computational scientists.

Understanding the computational requirements of applications is a complex task. Many studies[1] have shown that even for the subset of applications that are numerical simulation codes, there is no simple way to predict performance. What is well known is that just using floating-point performance, whether the peak performance from the processor manufacturer or the rate achieved by the High-Performance Linpack benchmark, is often a very poor way to predict performance. The situation has been made far worse in the past few years with the advent of systems using accelerators, which offer much higher floating-point performance but may not even be able to run some applications or may require significant code rewrites to make use of the accelerators. Advanced language, compilation, and execution systems could have transformative impact on both productivity and performance; however, these still largely represent frontiers of research rather than ready-to-deploy technologies. The situation is further complicated in the case of more data-intensive applications.

Given this complexity, what data ought to be collected from advanced computing proposals to understand the computational requirements of the applications? This is really an unsolved problem, and one for which the National Science Foundation could support research. In addition, for the purposes of requirements gathering, the data must be relatively easy for the code developers to provide and not require a significant analysis effort. Below are some examples of data that would provide more information than is currently collected (the number of service units requested), without requiring a detailed analysis by experts of each application. These examples have been selected because they are either relatively easy to determine, based on the code or algorithm, or they can be measured from a typical run of the application using widely available, open source tools. The data should apply to a typical execution of the application; if there are either multiple applications or widely different execution behaviors, data should be provided for each instance. The list below is targeted at parallel high-performance computing applications, but the approach can be applied to other areas; some items include examples relevant to some data science applications.

1. Basic performance characteristics, including floating-point operations, memory motion, and internode communication.

continued

BOX 4.3 Continued

2. Application performance characteristics, including code scaling and per-core efficiency and use of accelerators and vectorization.
3. Input/output (I/O) data sizes and number of I/O operations.
4. Algorithms used by the application; the Berkeley 13 motifs[2] may be a starting point for a list.
5. Application implementation, including programming languages and major libraries used.
6. Total application needs, such as the number of runs for an ensemble study.

In addition to these examples, it may be useful to collect application workflow requirements. Many computational science studies involve a workflow that includes multiple applications. These may be computationally intensive or they may be used to control other codes. For requests to use a known community code, for example, the data requested should instead be the name of the code and enough information about the running environment, such as the problem size, so that the compute needs in point 1 can be computed.

The list above illustrates useful data that could be obtained with relatively little effort. Only item 1 requires some code analysis, which can be obtained by running the application with tools such as PAPI [precision approach path indicator] and Darshan (instructions should be provided on how to use these tools; XSEDE [Extreme Science and Engineering Discovery Environment] and Blue Waters could provide tools and support to make this easy for applications that are already running on their systems). The other data are either descriptive or, in the case of scalability, ought to be readily available to the application's users.

This approach provides only the highest-level description of the application needs. While this would provide valuable data not currently being tracked, especially because it includes memory and I/O needs, as well as suitability for accelerators, it is not sufficient to predict performance on any system. Such a list could certainly be improved over time. However, it must be easy for the researcher to provide the information and not require a lengthy analysis of the code. If a shorter list was desired, the data in item 2 (application performance characteristics), combined with the number of SUs required, would provide valuable guidance in setting requirements for production computing systems.

---

1 See, for example, L. Carrington, M. Laurenzano, A. Snavely, R. Campbell, and L. Davis, How Well Can Simple Metrics Represent the Performance of HPC Applications?, in *Proceedings of the ACM/IEEE SC 2005 Conference*, 2005, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=1560000.

[2] K. Asanovic, R. Bodik, B.C. Catanzaro, J.J. Gebis, P. Husbands, K. Keutzer, D.A. Patterson, W.L. Plishker, J. Shalf, S.W. Williams, and K.A. Yelick, *The Landscape of Parallel Computing Research: A View from Berkeley*, Technical Report No. UCB/EECS-2006-183, December 18, 2006, http://www.eecs.berkeley.edu/Pubs/TechRpts/ 2006/EECS-2006-183.pdf.

5

Investment Trade-offs in Advanced Computing

Owing to the success of computing in advancing all areas of science and engineering, advanced computing is now an essential component in the conduct of basic and applied research and development. In a perfect world, investments in advanced computing hardware and software, together with investments in human expertise to support their effective use, would reflect the full range and diversity of science and engineering research needs. But, as discussed in Chapter 2, the gap between supply and demand is significant and continuing to grow. In addition, developments in data-intensive science are adding to the demand for advanced computing. From the smallest-scale system to the largest leadership-scale system, one of the challenges of advanced computing today is the capacity requirement along two well-differentiated trajectories—namely, high-throughput computing for "data volume"-driven workflows and high parallel processing for "compute volume"-driven workflows. Although converged architectures may readily support requirements at the small and medium scales, at the upper end, leadership-scale systems may have to emphasize some attributes at the expense of others.

Moreover, within a given budget envelope for hardware, the criteria for future procurements should reflect scientific requirements rather than simplistic or unrepresentative benchmarks. There is no single metric by which advanced computing can be measured. Although peak floating-point operations per second (FLOP/s) is the most common benchmark, even within the simulation community it has long been known that other

aspects of computer performance, including memory bandwidth and latency, are often more important.

Finally, advanced computing is more than hardware. Investments in software, algorithms, and tools can help scientists make more effective use of resources, effectively increasing the computing power available to the community.

The trade-offs to be considered are many, with different impacts on advanced computing cost and capability. This chapter starts by considering trade-offs associated with the volume of compute and data operations, applies them to investments in systems designed for simulation and data-intensive workload, and considers converged solutions (Section 5.1). This example was chosen because in the near term, it is perhaps the most critical trade-off that the National Science Foundation (NSF) must consider, as it balances the needs of existing computational users against a rapidly emerging data science community. The chapter then turns to another critical trade-off, between investments in production and investments to prepare for future needs (Section 5.2). Several investment trade-offs faced by NSF in simulation science, along with their impact, are discussed in Section 5.3. An example portfolio illustrating how NSF might address these trade-offs is sketched out in Section 5.4.

5.1 TRADE-OFFS AMONG COMPUTE, DATA, AND COMMUNICATIONS

Supporting both simulation and data-driven science requires making trade-offs among compute, data, and communications capabilities. At a conceptual level, workflows for the simulations of physics-based models are typically compute-volume driven in that they require a higher number of arithmetic or logical operations per unit of data moved. An illustrative example is many-body simulation of the electronic structure of molecules or materials, which is dominated by the contraction of dense, multidimensional tensors. On the other hand, workflows for developing or refining models by utilizing data from experiments or simulations are typically data-volume driven in that they require a larger number of units of data moved per arithmetic or logical operation. Examples include the analysis of genomic data from large studies or the analyses of streaming data. Further, in areas where scientific advances may be imminent at the confluence of both of these approaches—for example, in Earth systems science, where climate and weather models can be coupled to data from observatories—workflows will likely exhibit a complex mix of both of these aspects.

The communication-volume dimension refers to the speeds at which data chunks from very small to very large sizes can be moved efficiently

within the system. Such communication is accomplished through networks that may connect processors directly or via/across memory and storage subsystems; technology trends typically point to one or more orders of magnitude differences in the latencies per operation at a processor, memory, or storage element. Consequently, communication networks can be configured to serve efficiently the critical set of latencies at appropriate bandwidths. Now, high performance for a particular workflow will depend not only on how its data and compute-volume dimensions tap into the corresponding dimensions, but perhaps even more crucially on how the software implementation and algorithms underlying the workflow match the communication dimension.

A key question to consider is how the different types of workflows can inform advanced computing designs and specifications so that they can be provisioned appropriately to advance national priorities for discovery and innovation. Here, the major dimensions of advanced computing, as shown in Figure 5.1, play a critical role. The compute-volume and data-volume dimensions of advanced computing architectures are closely related to the corresponding compute and data dimensions of scientific workflows. However, the correspondence to the communication-volume dimension[1] is more complex, and it drives the space of trade-offs in regard to how desirable levels of performance may be obtained for specific types of workflows.

5.2 TRADE-OFFS FOR DATA-INTENSIVE SCIENCE

When making design and investment trade-offs for systems that support data-intensive science, one needs to consider the entire workflow, from instrument to scientific publication, and optimize the entire infrastructure, not just individual systems. One key trade-off is that investments in capabilities for data processing and long-term storage need to be balanced against each other accordingly. For example, in a project that collects data over several years, data are analyzed as they are collected and typically continue to be analyzed for several more years. In this case, the project is required to store the data, analyze them, and almost always to reanalyze them as algorithms improve and as new data arrive. As another example, a design that allocates more time to computing capabilities may complete its analysis faster but may not be able to store

[1] *Communication volume* is used here as shorthand for the more accurate and complex representation of internode communication, including latency, point-to-point bandwidth, bisection bandwidth, network topology and routing, and similar characteristics. Latency in particular is critical for many applications; some algorithms require high bisection bandwidth.

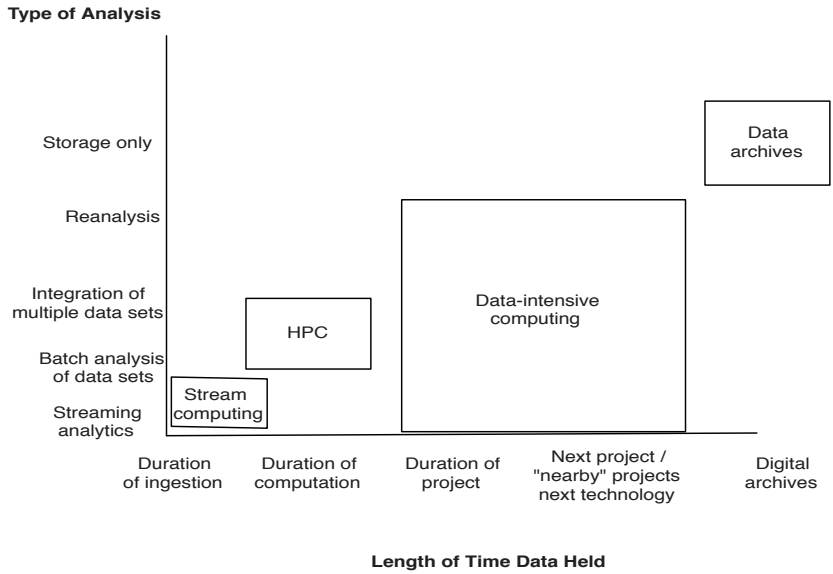

FIGURE 5.1 Length of time data are held for different forms of computing and types of analysis.

sufficient data to carry out the analysis of interest. Consider two designs with different allocations between the compute and storage allocated to a project. As Figure 5.2 illustrates, some projects may reanalyze all of the raw data throughout the project and so must keep the data online. Another important aspect of data-intensive science to keep in mind is that as different large data sets are integrated and the results analyzed, there are usually new types of scientific discoveries that are possible. Data-intensive projects often provide data to other projects that may use their data as part of a broader "integrative analysis." The trade-offs concern balancing how much data can be stored and for how long with how many processors can be used to analyze the data and how the communication network can be optimized for analysis and for efficient redistribution of the data to other interested parties. When instruments, computers, and archival sites are geographically distributed, the data they produce may be processed and consumed at multiple sites, requiring special attention to the wide-area networks needed to transfer the data, how data should be staged and consolidated, and so forth. For experiments, deciding how much data to save is a trade-off between the cost of saving and the cost

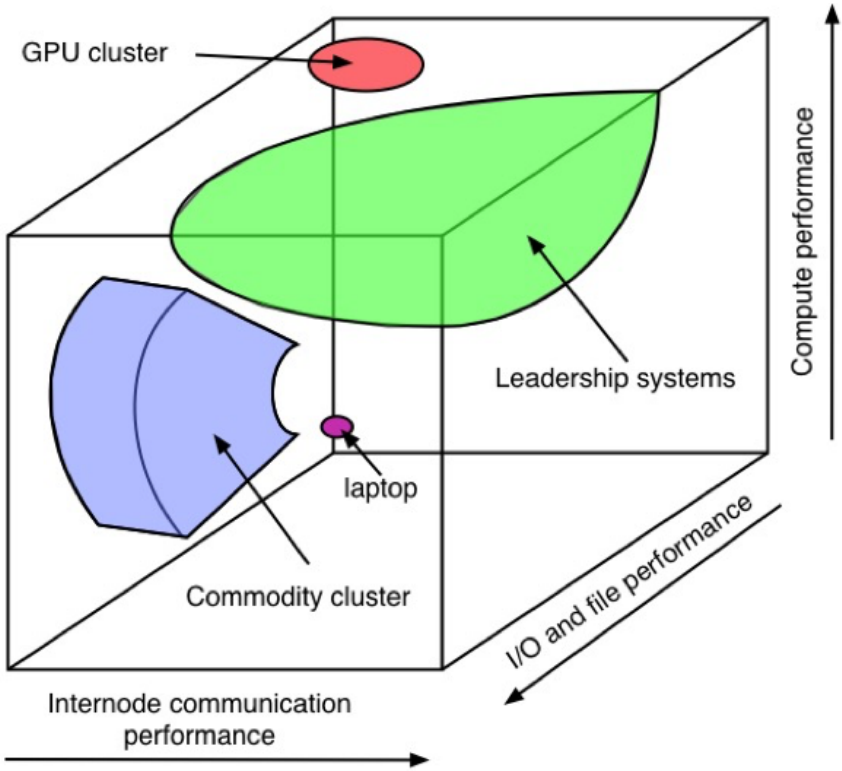

FIGURE 5.2 A simplified view of computing, including the three axes of compute performance, input/output (I/O) and file system performance, and internode communication (network) performance. Commodity clusters typically use commodity interconnects with performance less than typical high-performance computing (HPC) systems. However, they often include large amounts of fast file systems and large numbers of nodes, giving high aggregate compute performance. The best leadership-class systems (and the most expensive) will have the highest performance along all three axes. In reality, leadership-class systems are also a compromise, often trading I/O performance for greater floating-point performance. Missing from this view are important characteristics such as the type of compute or data architecture. The shapes are meant to capture common trade-offs in different types of systems and are not meant to be rigorous. For example, the shape for the "commodity cluster" describes systems where the I/O capacity is roughly proportional to the number of nodes (compute performance), and the system uses interconnects that typically have lower performance than is found in high-end HPC systems.

of reproducing, and this is potentially more significant than the trade-off between disks and processors.

In summary, for data-intensive science projects, one must balance the amount of data that can be stored with the capability and capacity of the workflows for analyzing the stored data. Second, trade-offs in the workflows themselves must be optimized. For data-volume-driven workflows, scientific outcomes are best achieved when the advanced computing is configured for efficient, high-throughput processing at scale with communication attributes directed toward efficiencies at the processing and storage layers for continuous updating and reanalysis of petabyte-sized data sets. Consequently, achieving U.S. leadership in this space requires achieving such capabilities at scale through an appropriate balance of advanced computing attributes in the networked storage elements in regard to the data-volume and communication-volume dimensions and in the processing elements in regard to the compute-volume dimension to achieve high throughput of data analyses workflows.

5.3 TRADE-OFFS FOR SIMULATION SCIENCE

Looking to the future, a variety of trade-offs will need to be examined with regard to the investments that NSF can make and their potential for enabling high-impact outcomes in simulation science. These trade-offs concern the scale of high-performance computing (HPC) systems and the fact that scale itself can become a tipping point for enabling new and unprecedented discoveries. The pivotal role of NSF in advancing simulation science and engineering through its HPC investments at different scales is readily demonstrated by using the NCSA Blue Waters project and the XSEDE program as illustrative examples. The Blue Waters project has enabled breakthrough scientific results in a range of areas, including an enhanced understanding of early galaxy formation, accelerating nanoscale device design, and characterizing Alzheimer's complex genetic networks.[2] NSF also supports the development and integration of mid-scale HPC resources through its XSEDE program, which provides HPC capacities to the broader scientific community along with resources for training, outreach, and visualization and supporting research in such areas as earthquake modeling and the simulation of black hole mergers.[3] Further trade-offs concern the maturation of simulation science from one-off simulations of a select few critical points of a high-dimensional modeling space to ensemble calculations that can manage uncertainties

[2] See Blue Waters, "Impact," https://bluewaters.ncsa.illinois.edu/impact-overview, accessed January 29, 2016.

[3] XSEDE, "Impact," https://www.xsede.org/impact, accessed January 29, 2016.

for increases in prediction accuracies[4] or enable high-fidelity modeling, simulation, and analyses for cost-efficient and innovative digital engineering and manufacturing.[5]

The scientific workloads supported by NSF through the Blue Waters and XSEDE programs are largely *compute-volume* and *communication-volume* driven, although aggregate memory capacity can be a key enabler for some frontier science applications. A notable point is that Blue Waters provides leadership capabilities in regard to all these dimensions, as shown in Figure 5.2, while some other NSF XSEDE investments provide capacities for such workflows at the midrange. For example, Stampede enables high throughput of low to midscale computations.

Historically, scientific workloads that are *compute-volume driven* have largely driven the balance of trade-offs in regard to the compute, communication, and storage components of such HPC advanced computing. Further, as described earlier, the trend toward multicore nodes with increasing core counts and very high degrees of thread and core-level parallelism require the use of high-bandwidth and low-latency networks to enable data exchange at the right scale. Additionally, the underlying algorithms and software implementations of the associated scientific workloads have been refined to define a more intricate relationship between how the elements of the advanced computing have to be tuned to provide the right balance along the compute-volume and communication-volume dimensions. Even within a single workflow, different algorithms with different trade-offs between compute and communication may be preferable on a given platform. This makes it difficult to evaluate which platforms are best suited for which applications.

As a consequence of trends in both hardware and software, including multicore nodes with high degrees of parallelism and sophisticated algorithms that require higher levels of data sharing while reducing the number of operations per unit data, the communication-volume dimension is a key differentiator in how trade-offs need to be managed. The advanced systems for simulation science often require that significant fractions of the cost budget are invested in the form of low-latency and high-bandwidth communication networks to couple multicore processor nodes. For example, much of the budget for a system to support simulation science workloads would be allocated to multicore processor nodes

[4] National Weather Service, "NMME: North American Multi-Model Ensemble," http://www.cpc.ncep.noaa.gov/products/NMME/, accessed January 29, 2016.

[5] National Digital Engineering and Manufacturing Consortium, *Modeling, Simulation and Analysis, and High Performance Computing: Force Multiplier for American Innovation*, Final Report to the U.S. Department of Commerce Economic Development Administration on the National Digital Engineering Manufacturing Consortium (NDEMC), 2015, http://www.compete.org/storage/documents/NDEMC_Final_Report_030915.pdf.

and the network that connects them to enable fast data exchange as the simulations proceed. In contrast, if the same budget was to be directed to serve streaming data science workloads, the bulk of it would go toward the storage network to enable high levels of data ingestion from storage by the multicore processor nodes. Both the Wrangler and Gordon systems funded by NSF are targeted in this manner toward data-intensive computation. This illustrates the contrasts between how trade-offs need to be balanced for serving different workflows.

In summary, for compute-volume-driven workflows, scientific outcomes are best achieved when the advanced computing is configured for efficient parallel processing at scale for a single analysis or simulation. The elements along the communication-volume dimension of the advanced computing (i.e., the interconnects) should be configured toward efficiencies at the processing and storage layers for continuous data exchange and the high-throughput output of data that are the results or outcomes of the processing. It is natural therefore to interpret performance in these HPC systems as they are traditionally known, to represent high levels of coupled parallel processing for compute-volume-driven applications.

5.4 DATA-FOCUSED, SIMULATION-FOCUSED, AND CONVERGED ARCHITECTURES

One of the features of the current era in computing is that there are several distinct architectures for the largest high-performance computers. At the same time, large systems for handling data, especially commercial data systems, are as large or larger in size—and even raw aggregate computing power—as the HPC leadership-class systems. Today, this suggests that there are two types of systems: HPC systems focused on simulation and systems focused on data. The true situation is almost certainly more complex. An issue complicating the discussion is that leadership-class systems for simulation science are operated mainly by research organizations and the government, while leadership-class systems for data science today are operated mainly by industry. Advances in HPC system architectures have generally been shared. In the case of data-intensive systems, some tools have been made open source while others have remained proprietary. Stronger ties between the computer science and computational science communities focused on new data analysis tools, techniques, and algorithms would help bridge this gap.

Some demands can be met by what is sometimes called convergence computing, in which high-performance systems are designed to meet the needs of both high-end simulation science and data science workflows. Indeed, high-performance systems for data-intensive computing in industry are almost always coupled with large-scale storage systems

(Figure 5.1). This coupling is still relatively infrequent in NSF-sponsored projects, and no NSF-supported project has data storage at the scale of a large Internet-scale company. As a simple example, the total online data storage for Blue Waters and XSEDE systems is in aggregate on the order of 100 PB, while online data storage systems at Google can be estimated at over tens of exabytes,[6] two orders of magnitude larger. In addition, the architectures at Internet-scale commercial companies are designed for the *continuous* updating and reanalysis of data sets that can be tens to hundreds of petabytes in size, something that is again rare in the research environment.[7] Presently costing several hundred million dollars, an exabyte of storage will become affordable for science applications within a few years because both disk and tape storage are still following an exponential increase in density and reduction in cost. However, bandwidth to the data will likely remain expensive, and it must be borne in mind that any significant analysis of an exabyte data set implies exascale computation. Over time, it seems reasonable to expect researchers to adopt industry use patterns as the necessary software is written. For example, researchers might analyze aggregated video streams to understand social behavior, with much of the large volumes of data not being retained for long.

5.5 TRADE-OFFS BETWEEN SUPPORT FOR PRODUCTION ADVANCED COMPUTING AND PREPARING FOR FUTURE NEEDS

Given the high demand for advanced computing, it will be essential for NSF to focus on and devote the majority of investments to provide production capabilities in support of its advanced computing roadmap. Production support is needed for software as well as hardware, to include community software as well as frameworks, shared elements, and other supporting infrastructure. NSF's Software Infrastructure for Sustained Innovation (SISI) program is a good foundation for such investments. However, SISI needs to be grown in partnership with NSF's science directorates to a scale that matches need, where it can then be sustained essentially indefinitely. The United Kingdom's Collaborative Computational Projects (CCPs) provide examples of the impact and successful operation of community-led activities that now span nearly four decades. Produc-

[6] Precise figures are not available, but a plausible estimate can be found in What If?, "Google's Datacenters on Punch Cards," https://what-if.xkcd.com/63/, accessed January 29, 2016.

[7] In high-energy physics analyses, enormous data sets are frequently reanalyzed in their entirety, but typically written only once.

tion support is further needed for data management; curation, preservation, archiving, and support for sharing all need ongoing investment. This balance is reflected in the example in Section 5.7.

However, if NSF invested solely in production, it would miss some key technology shifts, and its facilities would become obsolete quickly. Some innovation takes the form of fine-tuning of production systems, but modest, directed investments in exploratory or experimental facilities and services are also needed to create, anticipate, and prepare for technology disruptions.

NSF needs to play a leadership role in both defining future advanced computing capabilities and enabling researchers to effectively use those systems. This is especially true in the current hardware environment, where architectures are diverging in order to continue growing computing performance. Such investments would include (1) research into how to use and program novel architectures and (2) research into how applications might effectively use future production systems. In the first category, longer-term, curiosity-driven research likely belongs as part of the Computer and Information Science and Engineering directorate's research portfolio rather than NSF's advanced computing program, which would be focused on roadmap-driven experimentation. Leadership by NSF will help ensure that its software and systems remain relevant to its science portfolio, that researchers are prepared to use the systems, and that investments across the foundation are aligned with this future.

The range of possible options for advanced computing is growing as new architectures for analyzing data, increasing computing performance, or managing parallelism are introduced. One associated risk is that investments end up spread across too many emerging technologies, fragmenting the investment portfolio and reducing the ability to make investments at the scale needed for production capabilities. Another risk is that the criteria used to select among the technologies do not adequately reflect realistic science requirements, as can happen when overly simplistic benchmarks are used, leading to acquisition of systems that fall short in serving the research community.

As new technologies offering greater performance or other new capabilities begin to mature, decisions must be made about when to shift investments in the new direction. Many applications will benefit from higher performance, some applications may not need more performance, and one also expects new applications to emerge when higher performance thresholds are reached. It may be worthwhile to push aggressively into higher performance to enable some new applications, even if other applications take a long time to exploit the new architectures, or never do so.

Today, accelerators, including general-purpose GPUs and other tech-

nologies that are FLOP/s-rich but memory-poor and, possibly, hard-to-program can provide very high performance at reduced cost for a subset of applications. This can create tension between, on the one hand, moving forward aggressively with these technologies to obtain higher performance, thereby putting pressure on researchers to transition their software and algorithms to use the technologies more quickly, and, on the other hand, allowing sufficient time (and resources) for researchers to undertake such transitions. One possible indicator would be the level of active research on how to use a new architecture effectively. A high level might indicate that it is premature to consider the architecture ready for production systems. This indicator is not perfect; for example, there is still active research on how to use cache effectively.

More generally, the requirements expressed in the advanced computing roadmaps can serve as a guide to when technologies are ripe for transition from exploratory to production status. A requirements analysis is necessary to reveal the trade-offs implicit behind any such investment in NSF-wide infrastructure.

A 10-year roadmap would extend well into the exascale era. By focusing on its advanced computing roadmap rather than the first exascale system, NSF will ensure its investments have long-term benefit and will also assist the wider community in understanding and navigating the associated technology transitions. Although exascale systems may seem remote or even irrelevant to the majority of (but certainly not all) NSF users, technology advances in areas like energy efficiency needed for exascale capability will change the hardware and software landscape and have bearing on the purchase and operational costs of the aggregate capability NSF will need in the future. It will thus be important for NSF and the research users it supports to be involved in the national discussion around exascale and other future-generation computing, including through the recently announced National Strategic Computing Initiative, for which NSF has been designated as a lead agency.

At the same time, it will be especially important that NSF not only is engaged, but is actually helping to lead the national and international activities that define and advance future software ecosystems that support simulation and data-driven science. This includes active participation in and coordination of the development of tools and programming paradigms and the software required for exascale hardware technologies. The Department of Energy (DOE) is currently investing heavily in new exascale programming tools that, through the scale of investment and buy-in from system manufacturers, could plausibly define the future of advanced programming even though the design may not reflect the needs of all NSF science because the centers and researcher communities it supports are not formally engaged in the specification process.

5.6 CONFIGURATION CHOICES AND TRADE-OFFS

There are many choices and trade-offs to consider in allocating resources to computing infrastructure. This chapter has discussed several key trade-offs in detail, but there are many others. Whenever considering trade-offs, it is important to keep in mind that designing for a broader overall workflow almost certainly means configuring a system that is not perfect for all individual workflows; rather, it is able to run the entire workflow more effectively than other configurations. Thus, simply maximizing the performance or capability of one aspect, such as floating-point performance or data handling capacity, will not provide useful guidance.

5.6.1 Capability Can Be Used for Capacity but Not Vice Versa

Perhaps one of the most important items to consider is that not all computing resources are interchangeable. This may seem obvious, but it is often forgotten when computation is described in term of peak FLOP/s, cores, or memory size. In addition, some computations (again, both compute-centric and data-centric computations) are infeasible on systems smaller than a certain size. For example, many simulations require large amounts of memory (in the hundreds of terabytes to 1 petabyte [1 TB = 10^{12} bytes; 1 PB = 10^{15} bytes]), frequent exchanges of data, and terabytes to petabytes of data storage for both input and output data. Today, such simulations can only be run on leadership-class systems, such as NSF's Blue Waters or DOE's Mira and Titan systems. A simulation attempting to run on a system with a slower network will spend most of its time waiting on data to arrive (while still occupying most of the system memory); on a smaller system, there will not be enough memory to start the application.[8] Thus, without a system with these characteristics, such simulations cannot be performed.

On the other hand, large, capable systems can be used effectively for applications with smaller requirements. One argument that is sometimes made is that leadership-class systems ought to be used only for applications that require or can make good use of their unique capabilities. This is overly simplistic and is not looking at the overall objective, which, in essence, is to accomplish the greatest amount and most valuable science within the available budget (or with a minimum of cost and risk). Note that while it might be possible to run smaller jobs at slightly lower cost on a less capable system, the cost advantage is likely to be small given the

[8] In principle, out-of-core techniques, or even virtual memory approaches, could be used to address the lack of sufficient fast memory. But in practice this has the same problem as a too-slow network—the application might run, but it would run so inefficiently as to be impractical (and costly, since it would tie up the system for a very long time).

significant economies of scale that can be realized in large systems. The goal is to be cost-effective over the entire portfolio of applications, not to optimize for each individual application. Moreover, if a large system is running only large jobs, then there are likely to be many unutilized nodes, because a few large numbers of nodes are unlikely to sum up to the total system size. Small jobs can improve utilization and, thus, have a small marginal cost.

Although this report avoids the terms capability (instead referring to leadership-class systems) and capacity (systems that can run large numbers of jobs, none of which require a leadership-class system), this point can be most concisely expressed as "capability can be used for capacity but not vice versa." This critical point, reflected in Recommendation 2.2, calls for NSF to operate at least one leadership-class system so that the science that requires such systems can continue to be conducted.

As discussed above, a system that is optimized for data-driven science that requires processing (and reprocessing) large numbers of mostly independent data records needs capabilities not required on systems optimized for simulation. In particular, there is a greater need for a large amount of persistent storage; it may also be important to prefer higher bandwidth to independent storage devices—for example, having large numbers of compute nodes, each with several disks, over a unified system that provides access of all data to all nodes, as is common on "classic" supercomputers. In a perfect world, NSF could deploy several such systems, each optimized for a different workload. Unfortunately, in a budget-constrained environment, NSF will need to make some trade-offs and, in particular, consider alternative approaches to provisioning the necessary resources.

Although there are clearly applications that are dominated either by floating-point-intensive work or by data-intensive work, there are many problems that require a combination of capabilities. For example, some forms of graph analytics require the same sort of low-latency, high-bandwidth interconnect used in leadership-class HPC systems. In fact, the systems that dominate the Graph500 benchmark[9] are all large HPC systems, even though this benchmark involves no floating-point computation. Similarly, there are other features, such as large memory size, high memory bandwidth, and low memory latency, that are desirable in leadership-class systems for a wide range of problems, be they data-centric or simulation/compute-centric. Thus, it is best, as illustrated in Figure 5.2, to consider leadership-class systems as a spectrum of systems with different emphases. With enough funds, several large-scale systems could be deployed, each making different trade-offs in this space of con-

[9] See the GRAPH500 website at http://www.graph500.org, accessed January 29, 2016.

figuration parameters. One can have several large-scale systems making different trade-offs, or different subsystems of a coupled system that make different trade-offs. If simulations increasingly ingest experimental data and increasingly require in situ analysis, then the latter solution may work better.

Arguments like these suggest that an economy of scale implies that all resources should be centralized to gain maximum efficiency from the system. However, this is not correct, because it takes more than just hardware to provide effective advanced computing. Instead, a balance needs to be struck that provides resources large enough to tackle the critical science problems that the nation's researchers face while also providing systems tuned for different workloads and the expertise to ensure that these scarce and valuable resources are effectively used. It is also important to have several centers of expertise to ensure that the community has access to several different perspectives. An example of a possible set of trade-offs is given at the end of this chapter.

Deploying a flexible hardware platform capable of addressing a wide range of data-centric, high-performance, and high-throughput workflows is just the first step. Also essential are deploying and supporting the associated software stacks and addressing the challenges and barriers faced by researchers and their communities who will use the systems for their research and education.

5.6.2 Trading FLOP/s for Data Handling and Memory Size per Requirements Analysis

In the short run, even as it develops a more systematic requirements process, NSF needs to ensure continued access to advanced computing resources (which include both data and compute and the expertise to support the users), informed by feedback from the research communities it supports. In the longer run, it is essential that NSF use a robust requirements-gathering process to guide the selection of system configurations needed to ensure continued access. This will necessarily involve trade-offs of different capabilities, as each choice will have a significant cost. While this must be driven by the requirements analysis, one likely trade-off will be to improve data handling and memory size at the expense of peak FLOP/s. Some NSF systems have already done this; Blue Waters, with large amounts of memory, high input/output (I/O) performance, and a large number of conventional CPUs to support existing applications, is a good example. Wrangler is another good example of this trade-off in practice, although at a much smaller scale than Blue Waters. Note that the configuration of Blue Waters was guided by a process that required demonstrated performance on full applications, including reading input data and writing results to files, rather than just benchmarks and that this

significantly influenced the configuration of the system. The roadmapping process recommended in this report would ensure that future systems would be similarly aligned with the needs of the community.

5.6.3 Trade-offs Associated with Rewriting Code for New Architectures

When considering these trade-offs, it is important to consider the tension between maintaining compatibility with legacy applications and providing the highest performance for new applications. Note that there is a *huge* investment in scientific software that is not only written but also tested and (perhaps) understood. This code base cannot be rewritten without a significant investment in time and money. The financial cost is real and must be considered when evaluating the cost advantage of a new architecture. At the same time, if a new architecture is likely to persist, then that cost will only need to be paid once. An example of a new architecture that required many applications to be rewritten is the successful adoption of distributed memory parallel computers, along with message-passing programming, more than 20 years ago, which enabled an entire class of science applications.

A related issue is the one of scientist productivity versus achieved application performance balanced with efficient use of expensive, shared computational resources. As this report stresses, the goal is to maximize the *science* that is enabled and supported by advanced computing. An individual scientist may rightfully be focused on the fastest path to discovery and not be concerned about computational performance unless it is essential to completing the computations with available resources or time, such as is the case for the massively parallel applications running on Blue Waters. However, efficient utilization and maximum scientific productivity of a fully allocated, shared facility requires that the majority of cycles are consumed by well-optimized software. As systems become increasingly complicated and hard to use effectively, a burden has been put on the science teams as well as the computing facilities to create and maintain application codes that run efficiently on a range of systems. Many users are concerned about the difficulty in moving their codes to new architectures. In the short run, this means that production systems cannot be predicated on users needing to rewrite their applications to use new architectures. They also cannot depend on unproven software technologies to make existing or new applications run efficiently on new architectures. This concern with productivity also applies to new applications. Not all architectures are easy to use efficiently, and some algorithms remain very challenging to parallelize—for example, parallelization in time.

These observations relate to the relatively short run. However, NSF

also needs to be planning well into the future for a post-CMOS [complementary metal-oxide semiconductor] era. Here, the divisions of the Computer and Information Science and Engineering directorate other than the Division of Advanced Cyberinfrastructure (ACI) can play a role by ensuring that the requirements of the science community are included in computer science and engineering research on future device technologies and architectures.

5.6.4 Trade-offs Between Investments in Hardware and in Software and Expertise

Following on the theme of maximizing the science, today's hardware is challenging to use efficiently and, despite many attempts and interesting ideas, this is unlikely to change. NSF has already established several services that support application developers in making better use of the systems, both for XSEDE and for the PRAC teams on Blue Waters. The initial investments in the SISI program are a good start. They have the potential for broad impact if the investments reach sufficient scale, are sufficiently focused on the nitty-gritty of improving the engineering of codes, and are sustained over a sufficient period, if not indefinitely, with both external review and community-based requirements analysis being essential ingredients. Recent NSF-sponsored work[10] points to plausible mechanisms that could be adopted to assess the science impact of software as well as establish directions and locations for future investments. Investments in future hardware must continue to be considered together with support for using those systems, and that support must be organized for effective delivery.

5.6.5 Optimizing the Entire Science Workflow, Not the Individual Parts

Furthering the topic of getting the most science from the system, it is important to optimize for the entire scientific workflow, not just each part separately. This is for two reasons: first, as is well known, the global optimum is often not made up of a number of local optima. Second, it may not be possible to afford an optimal solution for each part of the problem. An example of a common yet incorrect trade-off is to design a system to meet the floating-point performance needs of a benchmark that is thought to represent an application. Yet in practice, the full application may require file I/O, memory bandwidth, or other characteristics. In

[10] J. Howison, E. Deelman, M.J. McLennan, R. Ferreira da Silva, and J.D. Herbsleb, Understanding the Scientific Software Ecosystem and Its Impact: Current and Future Measures, *Research Evaluation,* 2015, doi:10.1093/reseval/rvv014.

addition, the *science* may require running pre- and post-processing tools, visualization systems, or data analysis tools. It is critical that the entire workflow be considered. Note that the Blue Waters procurement was one of the few for leadership-class systems that required overall application performance, including I/O, as part of the evaluation criteria; as a result, this system has more I/O capability than most systems with the same level of floating-point performance and is, in fact, as powerful for I/O operations as the leadership-class systems planned by DOE for 2016-2017.

5.6.6 General-Purpose Versus Special-Purpose Systems

There are some applications that on their own use a significant faction of NSF's advanced computing resources. It may make sense, based on an assessment of the science impacts, to dedicate a system optimized for those applications (either together or singly) and provide a general-purpose system that can handle (most/many) of the remaining application areas that require a leadership-class system. For example, such systems may have smaller per-node memory requirements or per-node I/O performance; they may require simpler communication topologies but place a premium on the lowest possible internode communication latency. Similarly, as discussed above, an architecture focused on data volume will devote a much higher part of its cost to I/O than a system focused on compute or communication. While it may still be more cost effective to have a single machine that is good at all aspects of advanced computing (the convergence approach), it is essential that options that consider a small portfolio containing either specialized machines or access to time on specialized systems be considered.

5.6.7 Midscale Versus High-End Systems

Important scientific discoveries are made not just at the high end of the compute- and data-intensive scales, but also in the midscale and low end. Because of the improvement in software applications and tools and accessible training, there is a growing demand for the use of midscale advanced computing infrastructure. Work at the midrange produces a large number of scientific publications and supports a large scientific community. The requirements for midscale computing will only grow as improvements in software make it easier and easier to take advantage of these resources.

The Hadoop software ecosystem provides an interesting example of what is needed for midscale systems to become widely usable. Traditional HPC clusters, around since the mid-1990s, were challenging to use until the message passing interface (MPI, a standardized message-passing system that runs on a wide variety of parallel computers) matured, soft-

ware was developed that could leverage it, and students were trained to use it. Similarly, it was not until Hadoop emerged and began to mature that the same clusters could be easily used for data-intensive computing (especially of unstructured data). A Hadoop software ecosystem had to be developed and students trained to use it. Demand is just beginning to rise as this data-intensive ecosystem matures and researchers are trained to use it.

As advanced computing technology advances, midscale users will benefit from work to develop easily used software and standardized configurations that can be scaled to different sizes and thus readily reproduced to serve larger communities through foundation, university, and industry partnerships.

5.7 EXAMPLE PORTFOLIO

NSF needs to act now in acquiring the next generation of computing systems in order to continue supporting science. The following is just an example of the sort of portfolio for hardware, together with supporting expertise, that NSF could consider, along with some explanations for the choices. This is not a recommendation; rather, it is an illustration of some of the options with the rationale behind them.

1. *One or two leadership-class systems, configured to support data science, traditional simulation, and emerging uses of large-scale computation.* Such systems are needed to support current NSF science; by ensuring that there is adequate I/O support, as well as interconnect performance and memory, such a system can also address many data science applications. These systems must include support for experts to ensure that the science teams can make efficient use of these systems. Continuity of support for advanced computing expertise is essential because people with these skills are hard to find, train, and retain. Note also that these systems may not be optimal for any *one* workload, but can be configured to run the required applications more efficiently than other choices. Also, these systems should not be limited to running only applications that can run nowhere else; to ensure the most effective use of these resources, they should be used for a mixture of what might be called capability and capacity jobs, with priority given to the jobs that cannot be run on any other resources. In the case where funding is extremely tight and only one system is possible, that system must complement other systems that are available to the nation's scientists, such as those operated by DOE, and memoranda of understanding among agencies may help ensure that the aggregate needs of the research community are met.

2. *A cooperative arrangement with one or more operators of large-scale*

clouds. These are likely to be commercial clouds that can provide some access to a system at a different point in the configuration space for a leadership-class data system. This addresses the need for access to extremely large systems optimized for this class of data-intensive research. Conversely, because the commercial sector is rapidly evolving and scaling out these large-scale clouds, it makes sense to lease the service rather than attempt to build one at this time.[11] Note also that the commercial sector is investing heavily in applied research for these platforms, which suggest that NSF emphasize support for basic research.

3. *A number of smaller systems, optimized for different workloads, including support for the expertise to use them effectively.* It is important to have enough providers to provide distributed expertise as well as two types of workforce development: training for staff and training for students. Currently, XSEDE effectively provides this access to smaller systems optimized for different workloads. This capability is essential in supporting the breadth of use of advanced computing in NSF.

4. *A program to evaluate experimental computer architectures*. This effort would acquire small systems (or acquire access to the systems without necessarily taking possession of them) and work with the research community to evaluate the systems in the context of the applications that NSF supports.[12] This program will help inform the future acquisitions of the systems in the above three points, as well as inform basic research problems in computer and computational science, such as programming models, developer productivity, and algorithms. This approach differs from research testbeds for basic computer science; while important, those testbeds should be defined by the particular research divisions that need them.

5. A sustained SISI program. Continue to learn from the SISI program and apply lessons learned to long-term investments in software.

[11] Once NSF is using a large amount of time on a cloud, the cost of contracting with a service provider will need to be compared to the cost of operating its own cloud system. Many of the economies of scale that work for the cloud providers are applicable to NSF; the decision should be made based on data about the total costs.

[12] Some centers are already evaluating systems with NSF and external funding. TACC supports Hikari (funded by Hewlett Packard and NTT) for exploration of the effectiveness of direct high-voltage DC in data centers supplied by solar power, and Catapult (Microsoft-funded) evaluates the effectiveness of a specific field-programmable gate array-based infrastructure for science. Other centers are conducting similar activities. The Beacon system at the National Institute for Computational Sciences, partly funded by NSF, provided access to Intel Xeon Phi processes before they were deployed in production systems by TACC. The team that proposed Beacon included researchers from several scientific disciplines, including chemistry and high-energy physics.

6

Range of Operational Models

The National Science Foundation's (NSF's) current model of cyberinfrastructure, including advanced computing, is based on a mix of centralized and distributed funding, anchored by the Division of Advanced Cyberinfrastructure (ACI) within the Directorate of Computer and Information Science and Engineering (CISE). Previously, ACI was the Office of Cyberinfrastructure (OCI), reporting to the director. This central structure currently supports the Blue Waters facility (a leading-edge facility) and a set of smaller computing and storage resources via the Extreme Science and Engineering Discovery Environment (XSEDE). In addition to these centrally funded resources, the Geosciences Directorate operates advanced computing facilities at the National Center for Atmospheric Research (NCAR), and it and other NSF directorates fund cyberinfrastructure via a variety of programs.

Advanced computing shares many elements of other NSF infrastructure investments, but it also differs in some profound ways. First, unlike advanced telescopes or particle accelerators, where there is no competing commercial market, a vibrant computing industry develops new technologies and products and responds to market needs and opportunities that dwarf computing expenditures in academia and by federal research sponsors. Second, computing market shifts and the well-documented, rapid evolution of computing technology mean that researcher expectations and economically viable computing technologies change every few years. Consequently, advanced computing capital assets have a very short operational lifetime, in marked contrast to many other scientific instru-

ments. These shifts, however, do not mean that long-term planning is unnecessary or impossible. Businesses and academia regularly develop strategic information technology (IT) plans that accommodate technology shifts.

Third, advanced computing is distinguished by its universality; it is applicable to all scientific and engineering domains, spanning data capture and analysis, simulation and modeling, and communication and collaboration. Fourth, and consequently, demand for advanced computing continues to grow rapidly, placing increasing stress on the financial models and social processes used to support research cyberinfrastructure. Although states, universities, and companies have long subsidized the capital and operating costs of NSF's leading-edge advanced computing, those costs have now reached tens to hundreds of millions of dollars. Consequently, the willingness of these parties to engage in "pay to play" (i.e., accept losses in exchange for publicity or collateral institutional advantage) has declined accordingly.

6.1 GOALS AND OPPORTUNITIES

The unique attributes of advanced computing create both opportunities and challenges for any NSF strategy, requiring both nimbleness in the face of changing technologies and economics and stability to ensure sustained capabilities and research continuity. The following basic principles will help ensure the sustainability of NSF's advanced computing strategy:

- Realistic business assessment that exposes the true costs and subsidies of cyberinfrastructure deployment and operation at all scales;
- Identification and tracking of technology trends and economics, along with the research opportunities they create;
- Long-term planning and articulated strategy (a roadmap) that allows the broad research community and service providers to plan accordingly;
- Balanced support for computing hardware, storage systems, and networks, along with professional staff, software and tools, and operating budgets; and
- NSF-wide commitment to cyberinfrastructure investment, strategic directions, and operational processes.

Three crosscutting aspects of sustainability are particularly crucial: continuity, coverage, and skills.

6.1.1 Service Continuity and Adaptability

Service continuity encompasses long-term strategic planning and sustainability on a decadal or longer timescale. NSF's Major Research Equipment and Facilities Construction (MREFC) projects for scientific infrastructure typically involve years of planning. Today, NSF's cyberinfrastructure facilities are rarely used to support computational modeling and data analysis for MREFC projects. The former have lifetimes of just a few years, making it impractical for MREFC project leaders to reduce overall costs of advanced computing by including NSF's own cyberinfrastructure facilities on the MREFC operational plan. This must change if common cyberinfrastructure is to support MREFC projects and other long-term community research.

Historically, most research data has been produced by carefully planned experiments, and it has been both expensive to capture and highly guarded by the researchers who produced it. Ubiquitous, inexpensive sensors and a new generation of large-scale scientific instruments, including MREFC infrastructure, have changed the economics of data capture and are shifting scientific expectations about data retention and community sharing.

Although NSF's recent requirement that all NSF-funded research projects have a data management and accessibility plan is an explicit policy recognition of data's importance, there is no NSF-wide cyberinfrastructure strategy or program to support disciplinary or cross-disciplinary data sharing and preservation. Hence, much of the data preservation responsibility and financial burden rests on individual investigators and their home institutions. Today, when the cognizant investigators no longer perceive value in retaining the data, those data are often lost. This is increasingly problematic as the longer-term research value of data often accrues to those in other disciplines.

6.1.2 Service Coverage: Breadth and Depth

In its earliest form, cyberinfrastructure was synonymous with high-performance computing and computational science. Today it encompasses not only high-performance computing but also large-scale data archiving and analytics, software codes and tools, and human expertise and computing-mediated research and discovery. Orthogonally, cyberinfrastructure spans the capabilities and needs of individual investigator laboratories, campus sites, regional and national research facilities, and commercial cloud service providers.

Any comprehensive cyberinfrastructure strategy must include the entire spectrum of services and span the entire range of organizational

scales. It cannot be simply about leading-edge supercomputing platforms or just about big data analytics; it must integrate both at multiple scales. Nor can it focus on hardware infrastructure while neglecting both software development and maintenance and training and support of technical expertise. It must balance sustainability against adaptation, recognizing that community needs evolve and technology shifts drive new solutions.

The rise of "big data" as a cyberinfrastructure challenge that rivals the scale and complexity of advanced scientific computing is indicative of this need for community adaptation. To respond appropriately to this technology shift and opportunity, NSF must adapt its investments and infrastructure. Big data will require big infrastructure, just as leading-edge computational science does, and will likely involve a mix of both centralized facilities and decentralized repositories at universities. The Australian eResearch initiative and its Australian National Data Service is a relevant example.

In this context, the NSF community would benefit from a coherent, big data retention and preservation strategy and capability, one that balances investigator and disciplinary differences against communal benefit and research collaborations. Unfunded mandates for retention and preservation will not be workable. A balanced model is likely to require greater total funding, a better balance of capital and operating budgets, more focus on business practices and return on research investment, and greater coordination across NSF directorates.

6.1.3 Skills and Workforce

Sustainable and effective cyberinfrastructure depends critically on the skills and expertise of domain scientists and of committed and well-trained advanced computing professionals. Even if they are not directly responsible for code development and workflow management, scientists using advanced computing need to be generally knowledgeable about these matters. For their part, technical staff members not only deploy and operate facilities, but also support community toolkits and codes, serve as keepers of institutional knowledge and expertise, and manage and ensure data security and provenance. Unlike hardware, with a lifetime of a few years, the human infrastructure of people's experiences in operating such systems has a lifetime of decades. Despite their importance, these staff often lack clear academic career paths and are dependent on an uncertain stream of funding for support.

Given the global competition of computing and computational science talent, any cyberinfrastructure plan must include mechanisms that recognize and reward professional staff and ensure they have career opportunities that retain their talent within the academic community. One

important contribution to retaining and rewarding this skilled workforce is stability in funding for centers, recognizing that developing an expert staff is a long-term process that can be wasted with even a short-term gap in staff funding.

Programs are also needed to train future computational science and data analytics experts. The report of the NSF Task Force on Cyberlearning and Workforce Development[1] addressed this issue in depth and includes, more broadly, the use of computer-based approaches in learning and recognizes the need to train both the workforce that supports advanced computing and the practicing scientists who make use of advanced computing. Note that the effective use of advanced computing systems requires specialized and advanced training. NSF computing centers and other centers of advanced computing expertise (academic departments involved in advanced computing, national laboratories, and private industry) have leveraged their in-house expertise to offer such training. Examples include training programs for users offered by XSEDE and Blue Waters and the Argonne Training Program in Extreme Scale Computing. Such programs could benefit from a more formal approach and, in particular, long-term support for training materials and resources.

The pervasive NSF-wide and nationwide nature of advanced computing presents a perhaps unique opportunity, and responsibility, to pursue NSF's diversity and inclusion goals.[2] This includes ensuring the broadest possible benefit from and access to NSF's cyberinfrastructure, as well as translating this participation into creating and sustaining a computationally skilled workforce that reflects our nation. XSEDE has made significant progress in increasing the number of underrepresented minority and women users and, more notably, principal investigators (PIs) with allocations. The successful XSEDE campus champions program is a human network, which, while pursuing its primary mission of "empowering campus researchers, educators, and students to advance scientific discovery,"[3] also serves other missions including advancing diversity through increased awareness, training, and education. Increased access to statistics and metrics, concerning not just PIs and users but also those accessing online materials or participating in events or using other services, could better inform and guide actions by NSF, XSEDE, and

[1] National Science Foundation, Advisory Committee for Cyberinfrastructure, *Task Force on Cyberlearning and Workforce Development Final Report*, March 2011, https://www.nsf.gov/cise/aci/taskforces/FrontCyberLearning.pdf.

[2] National Science Foundation, *Diversity and Inclusion Strategic Plan 2012-2016*, http://www.nsf.gov/od/odi/reports/StrategicPlan.pdf, accessed January 29, 2016.

[3] XSEDE, "Campus Champions—Overview," https://www.xsede.org/campus-champions, accessed January 29, 2016.

the community, and XSEDE is already working toward increased public access to data.

6.2 ORGANIZATIONAL CHALLENGES AND COMMUNITY NEEDS

Although NSF's current mix of centralized and distributed cyberinfrastructure has had many notable successes, it is not without problems, both for infrastructure providers and for the research community. Some of these problems are rooted in history, some are embedded in the NSF culture, and some are consequences of NSF's organizational structure.

6.2.1 Competitive Challenges

From its origins, NSF's advanced computing programs—the original 1980s supercomputer centers program, the 1990s Partnership for Advanced Computational Infrastructure (PACI) program, the 2000s Distributed and Extensible Terascale Facilities, and now XSEDE—have all been based on a repeated cycle of competitions to host and operate large-scale cyberinfrastructure. This cycle continues to pit putative operators—universities and national laboratories—against one another in irregularly scheduled "winner take all" competitive battles. In each case, competitors build ad hoc hardware and software vendor alliances to mount proposals. To compete, they also leverage institutional funds to cover facility, hardware, and operations costs (which are capped in the competitions as a percentage of hardware costs). Much of this difficulty is rooted in the lack of distinction between research and infrastructure funding. Each has widely differing timescales and success metrics.

Not only does repeated infrastructure competition on 2- to 5-year cycles create strong disincentives for national collaboration, it convolves performance review, recompetition, and strategic planning in ways that are challenging for all. In addition, it leads to proposals designed to win a competition rather than maximize community scientific returns. For example, it places a premium on sometimes unproven, next-generation technology that can serve as a vendor-marketing showpiece, rather than on proven, production-quality infrastructure, and researchers have little input into vendor selection, configuration options, or service models. (There is a role for facilities to test novel and risky computing technologies, but it is not in production systems.)

Researchers whose work depends on access to shared facilities also face a form of "double jeopardy." The scientific merit of their proposed work is assessed via the standard peer review process. However, if funded, they are still not assured of access to the computing and storage resources

they need to conduct their research. A separate proposal for shared cyberinfrastructure access is conducted by either the XSEDE Resource Allocation Committee (XRAC) or the Petascale Computing Resource Allocations Committee (PRAC) to assess the competence of the researcher and his/her team to use the cyberinfrastructure resources efficiently. However, there is little operational follow-up to ensure the resources are in fact used wisely and efficiently. This is especially problematic because the monetary value of computing resource awards continues to increase.

Finally, as discussed earlier, the current model is structured largely in support of individual investigator and small team projects, with a nominal 3-year lifetime. Larger disciplinary projects and major scientific instruments (e.g., NSF MREFC projects or cross-agency partnerships) with longer production cycles have no mechanism to plan for and request cyberinfrastructure for a 10- or 20-year horizon, because there is no guarantee that any of the extant cyberinfrastructure facilities will still be operational. This adversely affects data preservation activities in particular, because, by definition, they target long-term access.

6.2.2 Structural Challenges

Since the beginning of the NSF supercomputing centers program in the 1980s, NSF ACI and its predecessor organizations have supported computational science research across NSF and provided services to a user base that spans all federal research agencies. Despite the clear recognition that computational science and data analytics are true peers with theory and experiment in the scientific process, NSF-wide coordination and support remain somewhat informal and ad hoc, with directorate participation often a secondary responsibility of the designees.

Although researchers in all NSF directorates are critically dependent on cyberinfrastructure, at present there are no formal mechanisms for coordinated strategic planning, nor are there ready ways to pool and disburse shared resources. Concretely, there are no shared negotiations for discounted infrastructure or services, nor an accepted strategy for prioritizing the balance of individual investigator, campus, and shared infrastructure. NSF would benefit from a formal roadmapping committee for cyberinfrastructure with representatives drawn from all directorates and shared responsibility for cross-directorate resource investment and strategy. In addition, it is crucial that advanced computing be treated as an NSF asset and funded accordingly, regardless of its organizational location. The need is too great and current resources are too limited for loosely coordinated action and reactive processes.

One corollary to the need for strategic coordination is scaling and scoping to match available resources. As a decentralized organization,

with frequent rotation of program officers, NSF regularly launches new programs and initiatives. For research, this is the distinguishing characteristic of NSF; it is community driven and adaptive. For infrastructure, this is often debilitating, because it leads to a proliferation of small efforts and projects that consume critical resources. When building and operating infrastructure, it is critical to do a small number of things extremely well. Successful infrastructure is derived from a sustained strategy and driven by relentless focus. The implication for NSF is clear. Given limited cyberinfrastructure resources, it must do a very small number of things extremely well, avoiding mission creep and resource dilution at all costs.

A second and equally important corollary is an integrated strategy for high-performance computing and big data analytics and a concomitant rebalancing of investments. Big data requires strongly coordinated big infrastructure, just as leading-edge computational science requires advanced computing systems. The lessons of commercial cloud computing are clear; centralization and scale create unprecedented opportunities for innovation and discovery. Clear and unambiguous requirements for data deposit and access are also needed. Only via such a mechanism, developed in broad community consultation, can the true benefits of data analytics be realized.

6.3 POTENTIAL SUSTAINABILITY APPROACHES

As the scale and scope of advanced computing demands and associated facilities and services have grown, the irregular, winner-take-all process described above has become more problematic. First, the scale and cost of high-end or leadership-class facilities needed to meet researcher demands is a large fraction of the total currently available in the NSF budget, whether within the ACI division budget or the budgets of other directorates. (Whether NSF needs a leadership-class or high-end system should be determined by the analysis of science requirements.) NSF could afford to purchase a significantly larger system than it is currently acquiring, but only by focusing on that investment rather than a larger number of much smaller investments.

Second, uncertainty regarding the timing and capability of infrastructure upgrades makes community planning difficult, and the timing is often not well matched to vendor hardware and software upgrade cycles. Third, the timescales are incompatible with the planning and life cycle of other scientific infrastructure, making use of centrally funded cyberinfrastructure difficult at best and often impossible.

Current models of funding for advanced computing (based on periodic recompetition) and service block allocations (via committee) create substantial uncertainty regarding service continuity and research access.

There are several ways to address these shortcomings while retaining the best elements of the current approach. These include approaches as varied as public-private partnerships for access to cloud services, federally funded research and development centers (FFRDCs) for organizational sustainability, and MREFC projects for facility construction. Many of these are not mutually exclusive and could be combined to address limitations of the current model.

6.3.1 A Regular Cadence of Infrastructure Investments

The cost of leading-edge advanced computing facilities and user support, whether for computational modeling or data analytics, is no longer measured in tens of millions of dollars. Rather, the costs are now denominated in hundreds of millions of dollars. Indeed, large-scale commercial data centers operated by cloud providers now cost over $1 billion each. The MREFC process may be a useful point of departure. Although there are some aspects of MREFC projects that match the needs of advanced computing infrastructure, the current MREFC mechanisms may need to be modified and adapted to the unique needs of advanced computing infrastructure, including the general nature of computing and the need for regular refresh of computing equipment.

To establish a regular cadence of infrastructure investments, NSF would plan and budget an upgrade every 3 to 5 years, with planning and construction of each generation overlapping the operation of the previous generation. This would clarify and systematize the technology upgrade and refresh process, provide a community mechanism to plan and shape infrastructure transitions, elevate budget planning and prioritization to NSF-wide discussion and approval, and provide the level of funding needed to maintain leading-edge capability.

As with MREFC projects, NSF would be able to request new funds as a line item in its annual budget request, explicitly acknowledging that that current, internal funding is inadequate to meet burgeoning need and scientific priorities. Finally, it would provide an operational instantiation for an NSF-wide advanced computing roadmap.

6.3.2 Leased Infrastructure

Historically, NSF cyberinfrastructure facilities have been operated by academic institutions on NSF's behalf, typically via cooperative agreements. In turn, the academic institutions have purchased computing, storage, and networking hardware from computing vendors at the start of the cooperative agreement to deliver the committed services. This hardware then depreciates over its nominal 3- to 5-year lifetime until its

residual economic value is minimal and its performance and capability are no longer competitive. At that point, only another infusion of capital will ensure service continuity.

Rather than purchasing hardware at the time of an award, NSF or its awardees might choose to lease the desired hardware from a vendor or a system integrator. In the simplest variation of this model, the hardware remains the property of the vendor but is located at the operator's facility. From an operational perspective, a simple leasing model is indistinguishable from outright purchase. Alternatively, the hardware could be hosted and maintained at a vendor facility, with a division of hardware service and user support between the partners.

Annual lease payments would smooth the punctuated budget shock of capital acquisitions, allowing amortization across multiple budget years. Lease terms at a higher level might also include periodic hardware upgrades to maintain leading-edge capability (e.g., equipment could be upgraded during the life of a cooperative agreement without competition to meet a series of performance targets) as well as quality of service and/or performance guarantees. Leases could also include exit clauses for termination, either with or without cause.

This is not a new idea. For example, the Department of Energy (DOE) has used this strategy successfully for its leading-edge computing deployments. University supercomputing centers in Japan also use leasing, which permits a regular and stable annual funding for each center.

6.3.3 Commercial Cloud Service Purchases

The explosive growth of commercial cloud services and their widespread adoption by both large corporations and small start-ups offers another alternative for provisioning advanced computing but is not a panacea (Boxes 6.1 and 6.2). Cloud computing now allows large organizations to outsource the provisioning, maintenance, and operation of computing infrastructure and commodity services, allowing them to focus resources and expertise on their core competence and differential value proposition. For smaller companies, the ability to offer services on a pay-as-you-go basis has reduced capital start-up requirements and lowered the barrier to market entry. The same could be true of individual laboratory users where computing use is highly episodic, with periods of low and high utilization.

The ability to scale services rapidly and dynamically across a wide range of demand is a consequence of the massive scale of cloud service deployment. All of the major cloud service vendors are investing billions of dollars annually to offer advanced computing and data analytics services. In addition, market competition is driving rapid declines in

BOX 6.1
The Role of Commercial Cloud Computing

Cloud computing has recently emerged as an effective way to provide computing to diverse communities. By taking advantage of economies of scale and easy network access, clouds can provide large amounts of computing power as well as convenient access to shared data. A natural question is whether cloud computing can meet the advanced computing needs of segments of the science community. This box considers some of the advantages and disadvantages of commercial cloud services today. The role of clouds for the National Science Foundation (NSF) will need to be re-evaluated frequently because the technology and ecosystem around clouds continues to change rapidly.

What Is Cloud Computing?

The term "cloud computing" has many parts and multiple definitions. The following aspects of cloud computing are relevant to the discussion here:

- Clouds are typically large clusters of computers that exploit economies of scale, providing both computing and data capabilities.
- The cloud is a shared resource. Many users can make use of it; the amount of resource is flexible (e.g., not a specific number of cores or nodes). The resource is not just hardware; it includes software and, often, data. It is easy to access cloud services, typically over the Internet.
- The resources available to a single job can vary from a single virtual CPU to a substantial faction of the entire cloud. This characteristic is sometimes described as "elastic." From the user's perspective, the cloud allows an application to use as much computing power as desired.
- Clouds may provide access to shared data, permitting a diverse user community to share the data and data products.
- Clouds provide a very flexible service model, permitting rapid access to resources with (usually) no long-term commitment.

Advantages of Cloud Computing

Cloud computing provides a number of advantages, particularly for single investigators or small research groups. Perhaps the most obvious advantage is that a cloud provides quick and easy access to computing power, and it is just as easy to get 10,000 cores as 1 core. For many users, the fact that access is available on demand within minutes of making the initial request is a major advantage. For others, the availability, if only for a short time, of more resources than they could otherwise afford is the key advantage.

For many users, the cost of cloud computing is much lower than the cost of buying and operating a system that is capable of meeting peak needs. The fixed costs, including the often-neglected cost of maintaining cybersecurity as well as data backups against both user errors (e.g., recovering a deleted file) and facility disaster (e.g., fire in the computer room), can be high. If there is no existing software base, then software must also be developed, sometimes at high cost.

Another advantage is the ability to share non-compute resources, such as data or networking. The easy-access model for clouds makes it simple to share data between users and communities at different institutions, and even different countries. Cloud computing also allows researchers to leverage rapid developments in data analytics that are being driven by the private sector and offered by commercial cloud computing providers. For those whose needs are not met by this software, it will be necessary, as with traditional high-performance computing (HPC), for communities to develop custom software.

Similar services can be offered by NSF centers, although the different allocation and resource model imposes some constraints. In particular, for very large data repositories, it may be impractical for each user to have a copy of the data, and also impractical to move the data to the user's site, even with high-performance, wide-area networking.

Note, however, that some of these advantages are or could be provided by NSF-operated advanced computing resources, which are already elastic. For example, an allocation on the Blue Waters system can be used for any number of nodes, permitting the use of as little as 32 cores (1 node) and as many as nearly 800,000 cores.

Clouds and Time and Space Sharing

The idea of sharing a computing resource to exploit economies of scale is not new. Computing centers of all types, including the supercomputing centers operated by NSF, the Department of Energy, and the Department of Defense, have done this almost from the beginning of computing.

However, there are important differences between cloud computing and conventional time- and space-sharing systems, although some of these are a matter of degree rather than being qualitatively different. First, clouds are accessed through a convenient network interface. This network connectivity makes it much easier to provide the resource to anyone on the planet, rather than those with access to the facility. NSF's advanced computing facilities are also conveniently accessible but less so than commercial cloud services, which require only a credit card for access. (For example, NSF's Blue Waters system requires two-factor authentication for access.) Second, virtualization support has made it much easier to securely run the customer's software environment, including the operating system. Third, standardized APIs for web access make it easy to provide *interactive* access to a computing resource on demand, including access to data repositories.

Cloud Cost Realities

Clouds provide many advantages, but it is important to separate cloud myths from realities. Clouds are not free. Some researchers have been given free cloud resources for small or high-profile research projects, and that initiative is to be applauded, but it is not realistic to expect 5 billion CPU-core hours as a gift from a commercial cloud provider (that is a small fraction of just the CPU time NSF consumes in a year and does not include data or network costs).

Commercial clouds, such as those operated by Amazon, Google, or Microsoft, are often very large in scale, with aggregate compute capacity larger than

continued

BOX 6.1 Continued

leadership-class systems. The very size of these systems gives these vendors substantial economies of scale as well as the ability to influence the hardware and software that goes into these systems. While this gives them some cost advantage, it does not necessarily mean they are cheaper than federally supported HPC centers.

Costs also include more than a charge for CPU time; in any cost comparison, it is important to include *all* costs. Costs for data handling, such as to/from disk, and for network access, are often a significant additional charge and might exceed the cost of CPU time. To further complicate the issue, commercial clouds rarely give enough details to make cost comparisons; for example, a cloud vendor may charge for a virtual CPU, but the specifics of the hardware (including details of cache size and speeds, specific processor model, and input/output [I/O] characteristics) are not provided.

Some analyses of the use of clouds for scientific computing found that commercial clouds are more expensive than a traditional supercomputing center,[1] although direct comparisons are difficult. Sustained system-performance benchmarking would help inform an understanding of true costs. To provide an updated rough-cost comparison, the committee estimated the cost of providing a leadership-class system using the Amazon Elastic Cloud (Box 6.2).

Convenience is not free. Demand for NSF's advanced computing services exceeds supply (see Figure 2.5); whenever that is the case, the supply must be rationed by some mechanism. NSF currently does this through the allocations process (which introduces various delays). The commercial market does this by adjusting price (not cost). There is no free lunch: on-demand access is, on average, more expensive than scheduled bulk access (although spot markets also offer an opportunity to get lower costs on occasions when demand is lower than supply). Cloud availability and cost savings depend, in large part, on uncorrelated use by the different customers. In the end, the only way to address the long queue times is to provide enough capacity. Using external clouds, at a higher unit cost, will decrease the available capacity, not increase it, given a fixed expenditure on computing.

Cloud service providers were not the first to seek to greatly improve their cost per unit of performance by exploiting commodity computing and the declining cost/performance ratio of its technologies. HPC has a long tradition of doing this. The best known is the Beowulf cluster. Although not the first effort to exploit commodity processors and networking, beginning in 1994, many groups built effective HPC systems from commodity parts. Many believe that this led to broader use of HPC by making systems more widely available. Today, many HPC systems are 100 percent commodity hardware, making use of high-end, but still commodity, interconnects such as InfiniBand, along with high-end ("server") processors and I/O systems. Today, commercial cloud systems employ a mix of commodity and custom hardware. For example, servers used by leading vendors include custom accelerators.

Both commercial cloud operators and government-funded HPC centers ex-

ploit significant economies of scale. Commercial cloud systems are very large, but not larger than other large HPC systems. For example, currently AWS has two systems on TOP500, but the highest is ranked only 180th on the November 2015 TOP500 list. Windows Azure reached number 165 on the 2012 list (but is not currently on the list).

Clouds also may not match the computing needs of large-scale, tightly coupled parallel science applications. Most clouds are designed to provide single "CPUs," possibly in large numbers, to the user. High-end HPC applications can use tens of thousands of nodes and require frequent and efficient communication among the nodes (e.g., communication every 100 microseconds with a communication overhead of 1-2 microseconds). This requires (1) a fast interconnect, (2) co-scheduling of all (not just most) of the processes in the program, and (3) efficient mapping of the program's processes onto the specific compute nodes to avoid communication interference with other jobs. (Clouds may in fact be distributed across the entire planet, adding significant speed-of-light latencies.) Clouds *can* be built to provide these capabilities, but only at additional cost.

In short, as a past study[2] has shown and as the discussion above further suggests, supercomputing centers already exploit many of the cost advantages of clouds and can be significantly cheaper than commercial cloud providers for some science applications.

Software and Expertise

Researchers will need more than access to the cloud services themselves if they are to make effective and efficient use of cloud services. Although some research communities have developed cloud-based applications and software stacks for certain applications, many disciplines lack common tools that reduce the development and management burden on researchers. Researchers will also need assistance selecting the appropriate services from a growing range of commercially offered options, including among multiple hardware configurations.

Some communities have been looking into taking advantage of clouds and seeing how to take advantage of improving software stacks. A few communities have developed "point-and-click" solutions, but these do not exist for the vast majority of scientific workflows. Just as NSF has invested in expertise to accompany its hardware acquisition programs, it seems natural to extend the model by helping support researchers who want to further explore using cloud services. Indeed, given that cloud services may be of the greatest immediate value in serving the long tail of users, who are less likely to have expertise and experience than larger users, provisioning expertise may be especially important.

---

[1] Department of Energy (DOE), *The Magellan Report on Cloud Computing for Science*, 2011, http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/Magellan_Final_Report.pdf.

[2] See, for example, DOE, *The Magellan Report on Cloud Computing for Science*, 2011, page ii, or Finding 7, p. iv.

BOX 6.2
The Price for a Leadership-Class Machine Implemented Using Amazon Elastic Cloud

It is natural to ask whether one could replace a large high-performance computing system with cloud services, especially given that clouds are often viewed as providing very-low-cost computing. The 2011 *Magellan Report on Cloud Computing for Science*,[1] prepared for the Department of Energy (DOE), asked just this question. Chapter 12 of the report contains an analysis of the cost of using a cloud to provide the compute and storage capability roughly in line with that at two DOE supercomputing centers, National Energy Research Scientific Computing Center and Argonne Leadership Computing Facility. The report's analysis considers the different costs, both on the cloud and at a center. Center costs include staffing for operation, building, and power, as well as the computing equipment. This analysis found that Amazon's commercial cloud offering was roughly three to seven times more expensive at providing compute cores and file storage than the two DOE centers. Section 12.6, "Late Update," noted a significant drop in price for the Amazon cloud as well as the introduction of more types of nodes, optimized for different types of computational needs. The authors are also careful to note that the analysis does not take into account the sustained performance on the sort of parallel science application that is common for DOE (and the National Science Foundation [NSF]) supercomputers, nor does it include the performance of the input/output (I/O) system and the cost of I/O operations. In addition, these analyses look solely at the cost of the computing resource and do not take into account the expertise in using these systems and working with computational scientists. The intent was to estimate a lower bound for the cost of a cloud; it is likely that the true cost will be higher.

However, 2011 was a long time ago, and cloud technologies and businesses have advanced. Are these conclusions still relevant? A check of Amazon Web Services pricing for computing and storage[2] suggests that they are. There are now many different tiers of nodes and I/O services, including nodes that provide GPUs

service costs and frequent service expansions (e.g., in software tools and packages).

NSF could make cloud services available to its researchers in one of several ways. All would likely involve NSF negotiating a bulk purchase agreement for data analytics and computing services.

- Individual investigators could request cloud services as part of a standard NSF proposal. The PIs of funded proposals could spend awarded funds with the cloud service provider of their choice. This is possible today, although cloud services incur indirect costs that may be more than 50 percent at many institutions, making them significantly less attractive than they otherwise would be compared to the purchase

and large memory nodes. In addition, substantial discounts are available by purchasing longer term (1- and 3-year) reserved instances. The committee compared the cost of simply providing the cores and the file storage for a large supercomputer to an estimate of the cost to NSF of the Blue Waters supercomputer, using data on January 12, 2016. Note that this cost only includes the processors, memory, and file space and does not include I/O operations (charged for separately by Amazon); the Blue Waters high-performance, low-latency interconnect; the Blue Waters tape library that can hold 320 PB of data; or an HPC-optimized software stack. Using 3-year reserved instances (which provide the greatest discount) and assuming 100 percent utilization of the Amazon resource and an estimate of about 75 percent utilization for Blue Waters, the cloud was still two to three times more expensive, depending on the exact choice of node type. Using 1-year reserved instances increases the cloud cost by about 50 percent.

This analysis does not mean that clouds must be more expensive; for example, NSF could negotiate a better (lower price) deal with a cloud provider. Rather, the point of this analysis is twofold. First, clouds are not necessarily cheaper than public supercomputing centers. Second, the costs must be very carefully analyzed to include all costs (which the committee did not do here) and to reflect the sustainable rather than peak performance available to the applications on the respective systems. For this reason, it will be important for NSF and the science community to continue to monitor the opportunities in cloud computing and to take advantage of them where it makes sense, but to also be aware that clouds are not necessarily cheaper than supercomputing centers and to be very careful in comparing costs.

---

[1] Department of Energy, *The Magellan Report on Cloud Computing for Science*, 2011, http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/Magellan_Final_Report.pdf.

[2] K. Asanovic, R. Bodik, B.C. Catanzaro, J.J. Gebis, P. Husbands, K. Keutzer, D.A. Patterson, W.L. Plishker, J. Shalf, S.W. Williams, and K.A. Yelick, *The Landscape of Parallel Computing Research: A View from Berkeley*, Technical Report No. UCB/EECS-2006-183, December 18, 2006, http://www.eecs.berkeley.edu/Pubs/TechRpts/ 2006/EECS-2006-183.pdf.

of computing hardware, which presently seems inequitable because the cost to an institution for purchasing cloud services is more akin to that of a recurring credit card charge or a subcontract. By bulk purchasing, NSF could eliminate this additional cost as well, potentially receiving more favorable rates than single investigators could obtain. Alternatively, mechanisms to reduce the indirect cost rate charged on cloud services can be explored.

• The current computing allocation review process could be expanded to include award of cloud services. Approved users would receive a budget to be spent with their chosen cloud provider. This would ensure centralized assessment of the appropriateness and likely efficiency

of the request, albeit with the double jeopardy of separate research and computing reviews.

- NSF could negotiate an agreement with one or more commercial cloud service providers (e.g., Amazon, Google, or Microsoft) and then operate a virtual facility on behalf of its users. In this model, user and application support would still rest with a noncommercial entity (e.g., via a cooperative agreement with an academic institution), and the cloud vendor would provide computing and storage services. NSF could leverage the Internet2 organization's NET+ initiative, which has selected commercial cloud services for its members and negotiated pricing and other terms.

All of these approaches would help take advantage of the rapid evolution of cloud services, the vibrant software ecosystem for cloud data analytics, the ability to use resources at massive scale, and the presence of large, shared data sets.

To address the structural disparity in the cost of cloud services compared to hardware acquisition, NSF would need to address the facilities and administrative (F&A) costs now charged for purchase of cloud services. Today, researchers can include cloud services as direct costs in research proposals, but these services are not excluded from the modified total direct cost (MTDC) on which F&A is computed. In contrast, capital equipment costs (e.g., computing equipment exceeding $5,000) are excluded from MTDC. The result is that $1 of cloud service costs $1.XX, where XX is the F&A rate at the researcher's institution. In contrast, the equivalent service on computing equipment purchased by an investigator on a research award costs only $1. In addition, power, cooling, and space for equipment are included in F&A, further skewing the incentive toward equipment purchase rather than service purchase. Removing this inequity would allow a more direct comparison and researcher selection based on perceived research value.

6.3.4 Cooperative Agreement Extension

Any funding and organizational structure must balance organizational stability and sustainability against responsiveness to technological change and customer needs. As noted earlier, NSF has long supported leading-edge cyberinfrastructure via a series of solicitations and open competitions. Although this has stimulated intellectual competition and increased NSF's financial leverage, it has also made deep and sustainable collaboration difficult among frequent competitors. Individual awardees quite rationally often focus more on maximizing their long-term prob-

ability of continued funding, rather than adapting and responding to community needs.

Frequent competitions have also made it more difficult for NSF-funded service providers to recruit and retain talented staff when the horizon for funding is only 2 to 5 years. This is especially true when the competition for IT and computational science expertise with industry is so great. Periodic review and rigorous performance assessment need not be coupled with "life or death" proposal competition and cooperative agreement funding.

Other federal agencies regularly review the performance of their service facilities, providing strategic and tactical guidance, without coupling those reviews to a facility termination decision. For example, DOE operates its National Energy Research Scientific Computing Center (NERSC) in this model. Hardware acquisition decisions, management reviews, and service priorities are subject to stringent reviews, but NERSC itself is not subject to termination review each time a new system is acquired. This also allows more honest and forthright discussion of problems, without existential fears.

NSF could consider designating one or more cyberinfrastructure centers as a core facility with a nominal lifetime of a decade—for example, as part of an extended cooperative agreement. Working with NSF and under regular review, the center would deploy and operate cyberinfrastructure on NSF's behalf. This would ensure organizational lifetime and planning horizons more similar to those of other NSF MREFC projects, which often last 10 to 20 years. In addition, longer horizons would also let NSF and its service providers evolve services and staffing in response to changing community needs and business partnerships. As extant examples, NSF's National Radio Astronomy Observatory and National Optical Astronomy Observatory play these roles in the astronomy community.

6.3.5 Federally Funded Research and Development Centers

As noted above, continuity is crucial to strategic planning, staff retention, and cross-domain partnerships. Cooperative agreements, whether for MREFC projects or other initiatives, provide one mechanism for collaborative planning and management. Implicit in all such approaches is a presumption that the project has a bounded lifetime. In turn, that presumption profoundly and adversely affects strategic planning and a commitment to sustainability within NSF and the community.

The centrality of advanced computing to research suggests that NSF treat it as a long-term, indefinite commitment that more clearly delineates the distinction between performance review and accountability and organizational continuity and service capabilities. Such separation

would allow service providers to work more collaboratively with NSF on responses to community needs and would encourage interorganizational collaboration.

An FFRDC is an excellent example of this balance. FFRDCs are independent nonprofit entities sponsored and funded by the U.S. government to meet specific long-term technical needs in areas of national interest. They operate as long-term strategic partners with their sponsoring government agencies. Many FFRDCs, such as DOE laboratories, include multiple programs spanning many areas of science and engineering research. NSF already uses an FFRDC, NCAR, as an integral part of NSF's cyberinfrastructure service strategy for the geoscience community; it can budget and plan new equipment acquisitions, and it offers staff career paths and continuity.

NSF could consider establishing one or more FFRDCs to support national cyberinfrastructure for research. Working with NSF, industry, and academia, such cyberinfrastructure FFRDCs could develop a strategic plan for cyberinfrastructure that meets evolving community needs, tracks technology developments, and provides a roadmap for NSF's directorates. The FFRDCs would also deploy and operate general or domain-specific cyberinfrastructure for the national community.

6.3.6 Partnerships with Other Agencies

NSF could explore partnerships with other federal agencies. For example, NSF could coordinate complementary leadership-class system configurations with DOE, especially with DOE systems that are used to support the DOE Innovative and Novel Computational Impact on Theory and Experiment program. The purpose of this partnership is not to shift the responsibility for providing cycles from NSF to DOE; rather, it is in recognition of the fact that there is not a simple one-dimensional configuration space for advanced cyberinfrastructure. Such a partnership would develop a way to fairly serve special needs from the population supported by each agency. For example, today NSF operates a system with more memory than any DOE system; conversely, DOE operates a system with more GPUs and peak floating-point operations per second (FLOP/s) than any NSF system. Currently, computational scientists request time on a variety of resources, taking advantage of DOE, NSF, and other providers of advanced computing infrastructure to the science community. But there is no formal coordination between agencies of the systems that they acquire, and trade-offs are made independently. Partnerships with other agencies could help ensure that the full spectrum of advanced cyberinfrastructure is available to the science community.

6.3.7 Strategic Public-Private Partnerships

As the demand for cyberinfrastructure continues to rise, the costs for deployment and operation rise commensurately. This is true for both aggregate demand—laboratory and institutional capabilities—and leading-edge computing and data storage systems. Superficially, this may seem paradoxical, given the dramatic increases in computing capability and storage capability regularly delivered by the computing industry. However, those same computing advances have birthed new sensors and scientific instruments and a torrent of new digital data, as well as new simulation models and expectations for ever-larger computing capability.[4]

Rising demands for computing and storage (end-to-end capabilities, not just hardware) now challenge the finances and social processes of both NSF and its academic grantees. Simply put, the rising cost of leading-edge facilities (NSF Track 1 and Track 2 systems) is not sustainable under the current partnership model and may not be sustainable under any government-funded model. Put another way, the perceived return on investment for a facility costing hundreds of millions of dollars must be substantial, particularly when the equipment has a useful lifetime of only 3 to 5 years.

NSF might consider alternative public-private partnership models that create financial incentives for private-sector partners to operate large-scale cyberinfrastructure facilities on the research community's behalf. These necessarily require more flexible approaches than traditional fee-for-service models and might include such options as access to university intellectual property in exchange for cyberinfrastructure services. Precisely how such arrangements might work would depend on the willingness of the academic community to agree on, for example, vendor exclusivity and intellectual property sharing.

6.3.8 User-Driven Acquisition and Allocation

All of the operational strategies described above are based on some variant of central planning and resource management. Alternatively, NSF could decentralize cyberinfrastructure acquisition and support and rely on social and economic forces to define and optimize community cyberinfrastructure. One first step in this process would be denominating all services in dollars, rather than the abstract, normalized service units (SUs) or storage allocations used today. SUs play an important role by attempting to enable the comparison of allocations on computers that may differ widely in both architecture (e.g., conventional processors or

[4] The end of Dennard scaling and limits of future microprocessor performance increases mean the "free lunch" of performance doubling will bring new and sobering economic constraints. Larger capability will require larger capital infusions.

graphical processing units) and time of deployment. For instance, the use of SUs makes more quantitative the assessment in Figure 2.5 of resources over the past decade. However, despite their merit, SUs obscure from users the actual costs associated with requests and allocations, and the use of SUs also distances the NSF programs and the user community from the prioritization processes about how the underlying funding is allocated. Moreover, the conversion factor between actual wall time on a computational resource and SUs is established by each site based on High-Performance Linpack benchmark results, which is just a single and outdated metric that does not capture the diversity of factors controlling the capability (which is more than just performance) of individual applications mapped to different architectures. Recently, XSEDE has started notifying both users and associated NSF program managers of the actual dollar value associated with an allocation, and there seem to be multiple significant potential benefits in making users even more cognizant of and ultimately responsible for the actual costs and effective use of resources.

Realizing these benefits can certainly start with increasing user awareness of costs and engaging users in resource planning and acquisition. In a more extensive realization of this model, however, individual researchers or research teams would be allowed to spend awarded cyberinfrastructure dollars at their discretion. This cyberinfrastructure marketplace might include the following options:

- Purchasing local computing infrastructure, services, or staff support for use within the individual researcher's laboratory;
- Contributing dollars to a university pool that operates a campus facility under a "campus condominium" model;[5]
- Pooling research dollars to purchase and operate shared regional or national facilities; and
- Purchasing commercial cloud services, exploiting the properties of elasticity and on-demand access.

All of these variants allow individual researchers and research teams to make separate decisions on how best to advance their research. They also remove researchers from double jeopardy, where they must compete separately for research funding and for computing resources. In addition, the options expose the costs of each option in a common currency. How-

[5] Under a condominium model, a university purchases a baseline computing and storage infrastructure and allows individual researchers to purchase and contribute nodes and storage to the shared pool. Researchers receive access priority in proportion to their financial contribution.

ever, the risk is that the sum of the local research optimizations may not be globally optimal for the national community.

Moreover, some form of such a model may provide an effective mechanism to encourage and formalize investments and responsibilities of researchers, institutions, and regions in private and shared local or national infrastructure. NSF already recognizes that there are significant computing resources "at the edges" (meaning within campuses and states) and that there is a clear need to coordinate and leverage investments. Programs such as Campus Cyberinfrastructure—Data, Networking, and Innovation Program (CC*DNI) and Major Research Instrumentation help develop this infrastructure, and elements of XSEDE, such as campus champions, are directed toward tying both communities and cyberinfrastructure together. However, the same economic and technological forces driving the decisions on national computing infrastructure are eroding the abilities of campuses to purchase and operate their own cyberinfrastructure, and especially challenging are the cost and complexity of managing research data. Thus, smaller institutions are now choosing to invest in infrastructure operated by larger neighbors or at national centers, which can provide both cost and other advantages compared to attempting to use the commercial cloud. However, in the absence of a scalable national model, such partnerships are presently ad hoc. The NSF Big Data Regional Innovation Hubs (BD Hubs) program is potentially a powerful catalyst to drive regional synergy, but this still needs to be tied to a national narrative that includes all aspects of advanced cyberinfrastructure.

Variations of this economic model have been explored in the past. Then called the "green stamps" model of resource allocation, it was analyzed in the 1995 *Report of the Task Force on the Future of the NSF Supercomputer Centers Program*.[6] The report noted

> The key concept in a green stamp mechanism is the use of the stamps to represent both the total allocation of dollars to the Centers and the allocation of those resources to individual PI's. NSF could decide a funding level for the Centers, which based on the ability of the Centers to provide resources, would lead to a certain number of stamps, representing those resources, being available. Individual directorates could disperse the stamps to their PI's, which could then be used by the researchers to purchase cycles. Multiple stamp colors could be used to represent different sorts of resources that could be allocated.
>
> The major advantages raised for this proposal are the ability of the di-

[6] National Science Foundation, *Report of the Task Force on the Future of the NSF Supercomputer Centers Program*, NSF9646, September 15, 1995, https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf9646.

rectorates to have some control over the size of the program by expressing interest in a certain number of stamps, improvement in efficiency gained by having the Centers compete for stamps, and improvements in the allocation process, which could be made by program managers making normal awards that included a stamp allocation.

Other than the mechanics of overall management, most of the disadvantages of such a scheme have been raised in the previous sections. In particular, such a mechanism (especially when reduced to cash rather than stamps) makes it very difficult to have a centralized high-end computing infrastructure that aggregates resources and can make long-term investments in large-scale resources.

NSF could conduct a pilot project to evaluate the power of market forces in allocating limited cyberinfrastructure support. Among the issues to evaluate is whether such an approach would exacerbate the problem of buying resources by the hour (see Section 5.5) without recognizing the fixed costs, such as the cost of retaining staff and supporting the use of new architectures.

Independently of any pilot projects, NSF will benefit by expressing in dollars the true cost of large cyberinfrastructure resource allocations (i.e., those now made by the XSEDE Resource Allocation Committee [XRAC] and Petascale Computing Resource Allocation Committees [PRAC]). First, it would allow researchers to identify the value of cyberinfrastructure awards to their institutions. Second, and equally important, it would make clear that such large allocations have true costs, encouraging wise and efficient use.

Appendixes

A

List of Individuals, Research Groups, and Organizations That Submitted Comments

Jay Alameda, National Center for Supercomputing Applications, University of Illinois, Urbana-Champaign
Richard B. Arthur, GE Global Research
Dan Atkins, University of Michigan
Nadine Aubry, Northeastern University
Troy Baer, University of Tennessee
Jim Belak, Lawrence Livermore National Laboratory
Francine Berman, Rensselaer Polytechnic Institute
Prentice Bisbal, Rutgers University
Alan Blatecky, RTI International
Adam Bowser, on behalf of The University Corporation for Advanced Internet Development (Internet2)
Robert F. Brammer, Brammer Technology, LLC
James G. Brasseur, Pennsylvania State University
Danielle Chandler, University of Illinois, Urbana-Champaign, on behalf of the Theoretical and Computational Biophysics Group
Coalition for Academic Scientific Computation, comments collected from senior U.S. cyberinfrastructure facility directors
Coalition for Academic Scientific Computation, comments from members (individual opinions)
Ronald Cohen, Carnegie Institution
Computing Community Consortium, Computing Research Association
George W. Crabtree, Argonne National Laboratory
Alan Crosswalk, Columbia University
Timothy Alden Davis, Texas A&M University

Carleton DeTar, University of Utah
Thom Dunning, Pacific Northwest National Laboratory and University of Washington
Rodolfo Barniol Duran, Purdue University
Bruce G. Elmegreen, IBM T.J. Watson Research Center
Ian Foster, University of Chicago
Lars Grabow, University of Houston
Victor Hazlewood, University of Tennessee
Hendrik Heinz, University of Akron on behalf of faculty of the College of Polymer Science and Engineering
Tony Hey, University of Washington eScience Institute
Alvin Kennedy, Morgan State University
Rubin H. Landau, Oregon State University
Randall LeVeque, University of Washington
Zachary H. Levine, National Institute of Standards and Technology
David A. Lifka, Cornell University
Yangzheng Lin, Carnegie Institution
Glenn K. Lockwood, 10X Genomics
Paul B. Mackenzie, Fermi National Accelerator Laboratory, on behalf of the U.S. Lattice Quantum Chromodynamics Collaboration
Jan Mandel, University of Colorado, Denver
Thomas A. Manz, New Mexico State University
J. Andrew McCammon, University of California, San Diego
Jonathan C. McKinney, University of Maryland
Charles Meneveau, Johns Hopkins University
Blake Mertz, West Virginia University
Rajat Mittal, Johns Hopkins University
Colin Morningstar, Carnegie Mellon University
Lawrence Murakami, University of Alaska
Annick Pouquet, University of Colorado, Boulder
Jeff F. Pummel, University of Arkansas
Ralph Roskies, Pittsburgh Supercomputing Center, University of Pittsburgh
Barry I. Schneider, National Institute of Standards and Technology
Bill Schultz, University of Michigan
Jerome Soller, CogniTech Corporation
James M. Stone, Princeton University
Alexander Tchekhovskoy, University of California, Berkeley
Greg van Anders, University of Michigan
Chris Van de Walle, University of California, Santa Barbara
Nancy Wilkins-Diehr, University of California, San Diego
Walt Wright, Check Twelve Leadership
P.K. Yeung, Georgia Institute of Technology
Peijun Zhang, Carnegie Institution

B

Information-Gathering Meetings

April 15, 2014, by telephone

Briefings from Irene Qualters, National Science Foundation (NSF), and Peter Arzberger, NSF

May 16, 2014, Washington, D.C.

Briefings from Michael Norman, San Diego Supercomputer Center; Michael Vogelius, NSF; Bogdan Mihaila, NSF; Jeryl Mumpower, NSF; and Eva Zanzerkia, NSF

November 19, 2014, Birds-of-a-feather session at SC-14, New Orleans, Louisiana

December 16-17, 2014, Workshop in Mountain View, California

Participants: Christian Ott, Caltech; Thomas Cheatham, University of Utah; Tom Jordan, University of Southern California; Steven Gottlieb, Indiana University; Tony Hey, Microsoft, by telephone; Ilkay Altintas, San Diego Supercomputer Center; Jacek Becla, SLAC National Accelerator Laboratory; Victoria Stodden, University of Illinois, Urbana-Champaign; and Ed Lazowska, University of Washington

February 19, 2015

Briefings from Jim Kurose, NSF; Irene Qualters, NSF; Rudi Eigenmann, NSF; and Steven Binkley, Department of Energy Office of Science

C

Biosketches of Committee Members

WILLIAM D. GROPP, *Co-Chair*, is the Thomas M. Siebel Chair in Computer Science at the University of Illinois, Urbana-Champaign, where he is also founding director of the Parallel Computing Institute. He held the positions of assistant (1982-1988) and associate (1988-1990) professor in the Computer Science Department at Yale University. In 1990, he joined the Numerical Analysis group at Argonne National Laboratory (ANL), where he was a senior computer scientist in the Mathematics and Computer Science Division, a senior scientist in the Department of Computer Science at the University of Chicago, and a senior fellow in the Argonne-Chicago Computation Institute. From 2000 through 2006, he was deputy director of the Mathematics and Computer Science Division at ANL. In 2007, he joined the University of Illinois, Urbana-Champaign, as the Paul and Cynthia Saylor Professor in the Department of Computer Science. In 2008, he was appointed deputy director for research for the Institute of Advanced Computing Applications and Technologies at the University of Illinois. His research interests are in parallel computing, software for scientific computing, and numerical methods for partial differential equations. He has played a major role in the development of the MPI message-passing standard, is one of the designers of the PETSc parallel numerical library, and has developed efficient and scalable parallel algorithms for the solution of linear and non-linear equations. Dr. Gropp is a fellow of the Association for Computing Machinery (ACM), the Institute of Electrical and Electronics Engineers (IEEE), and the Society for Industrial and Applied Mathematics (SIAM) and a member of the National Academy

of Engineering. He received the Sidney Fernbach Award from the IEEE Computer Society in 2008 and the Technical Committee on Scalable Computing Award for Excellence in Scalable Computing in 2010. Dr. Gropp received his B.S. in mathematics from Case Western Reserve University, an M.S. in physics from the University of Washington, and a Ph.D. in computer science from Stanford University.

ROBERT J. HARRISON, *Co-Chair*, is director, Institute of Advanced Scientific Computing, Stony Brook University, and director, Computational Science Center, Brookhaven National Laboratory. The core mission of the new Stony Brook institute is to advance the science of computing and its applications to solving complex problems in the physical sciences, the life sciences, medicine, sociology, industry, and finance. The institute works closely with the Brookhaven center, which specializes in data-intensive computing. Dr. Harrison's research interests are focused on scientific computing and the development of computational chemistry methods for the world's most technologically advanced supercomputers. From 2002 to 2012, he was director of the Joint Institute of Computational Science and professor of chemistry and corporate fellow at the University of Tennessee and Oak Ridge National Laboratory. Prior positions were at the Environmental Molecular Sciences Laboratory, Pacific Northwest Laboratory, and ANL. He has a prolific career in high-performance computing, with more than 100 publications on the subject, as well as extensive service on national advisory committees. He received his B.A. from Churchill College, University of Cambridge, and his Ph.D. in organic and theoretical chemistry from the University of Cambridge.

MARK R. ABBOTT is president and director of the Woods Hole Oceanographic Institution. He was previously dean of the College of Earth, Ocean, and Atmospheric Sciences at Oregon State University (OSU). Prior to his appointment at OSU, he served as a member of the technical staff at the Jet Propulsion Laboratory (JPL) and as a research oceanographer at Scripps Institution of Oceanography. Dr. Abbott's research focuses on the interaction of biological and physical processes in the upper ocean and relies on both remote sensing and field observations. He is a pioneer in the use of satellite ocean color data to study coupled physical/biological processes. As part of a NASA Earth Observing System interdisciplinary science team, Dr. Abbott led an effort to link remotely sensed data of the Southern Ocean with coupled ocean circulation/ecosystem models. His field research included the first deployment of an array of bio-optical moorings in the Southern Ocean as part of the U.S. Joint Global Ocean Flux Study. Dr. Abbott was a member of the National Science Board from 2006 to 2012 and served as a consultant to the board until 2013. He is the

vice chair of the Oregon Global Warming Commission. He is currently a member of the board of trustees for the Consortium for Ocean Leadership and the board of trustees of NEON, Inc. His past advisory posts include chairing the Coastal Ocean Applications and Science Team for NOAA and chairing the U.S. Joint Global Flux Study Science Steering Committee. He has also been a member of the Director's Advisory Council for JPL and NASA's MODIS and SeaWiFS science teams and the Earth Observing System Investigators Working Group. He was the 2011 recipient of the Jim Gray eScience Award, presented by Microsoft Research. Dr. Abbott is a national associate member of the National Academies of Sciences, Engineering, and Medicine and is currently a member of the Space Studies Board, chair of the Committee on Earth Science and Applications from Space, a member of the Committee to Advise the U.S. Global Change Research Program, and a member of the Panel on the Review of the Draft 2013 National Climate Assessment (NCA) Report. As part of his prolific service to the Academies, Dr. Abbott served on the Committee on Evaluating NASA's Strategic Direction, the Committee on the Assessment of NASA's Earth Science Programs, the Committee on the Role and Scope of Mission-Enabling Activities in NASA's Space and Earth Science Missions, and the Panel on Land-Use Change, Ecosystem Dynamics and Biodiversity for the 2007 Earth science and applications from space decadal survey. Dr. Abbott received his B.S. in conservation of natural resources from the University of California, Berkeley, and his Ph.D. in ecology from the University of California.

ROBERT L. GROSSMAN is a faculty member at the University of Chicago. He is the director of the Center for Data Intensive Science, a senior fellow and core faculty in the Computation Institute and the Institute for Genomics and Systems Biology, and a professor of medicine in the Section of Genetic Medicine. He also serves as the chief research informatics officer for the Biological Sciences Division. His research group focuses on data-intensive computing, data science, and bioinformatics. He is the founder and a partner of Open Data Group, which provides analytic services to help companies build predictive models over big data, and is the director of the not-for-profit Open Cloud Consortium, which provides cloud computing infrastructure to support the research community. He was elected a fellow of the American Association for the Advancement of Science in 2013. Dr. Grossman earned his Ph.D. in applied mathematics at Princeton University and an A.B. in mathematics from Harvard University.

PETER M. KOGGE is a professor of computer science and engineering and concurrent professor of electrical engineering at the University of Notre Dame. Dr. Kogge was with IBM, Federal Systems Division, from

1968 until 1994, and was appointed an IEEE fellow in 1990 and an IBM fellow in 1993. In 1977, he was a visiting professor in the ECE Department at the University of Massachusetts, Amherst. From 1977 through 1994, he was also an adjunct professor in the Computer Science Department of the State University of New York at Binghamton. In 1994, he joined the University of Notre Dame as first holder of the endowed McCourtney Chair in Computer Science and Engineering (CSE). Starting in the summer of 1997, he has been a distinguished visiting scientist at the Center for Integrated Space Microsystems at JPL. He is also the research thrust leader for architecture in Notre Dame's Center for Nano Science and Technology. For the 2000-2001 academic year, he was the Interim Schubmehl-Prein Chairman of the CSE Department at Notre Dame. From August 2001 until December 2008, he was the associate dean for research, College of Engineering; since fall 2003, he has been a concurrent professor of electrical engineering. His current research areas include massively parallel processing architectures, advanced VLSI and nanotechnologies and their relationship to computing systems architectures, non von Neumann models of programming and execution, parallel algorithms and applications, and their impact on computer architecture. While at IBM, one of his groups designed the first multi-processor PIM device with significant DRAM memory that may also be the world's first multicore chip. A paper on its architecture received the Daniel Slotnick Award at the 1994 International Conference on Parallel Processing. Dr. Kogge also designed and built the RTAIS parallel processor. Prior parallel machines included the IBM 3838 Array Processor and the space shuttle input/output processor (IOP), which probably represents the first true parallel processor to fly in space and is one of the earliest examples of multi-threaded architectures. Dr. Kogge received the IEEE Seymour Cray Award in 2012 and the IEEE Charles Babbage Award in 2014. He received his B.S. in electrical engineering from the University of Notre Dame, his M.S. in systems and engineering from Syracuse University, and his Ph.D. in electrical engineering from Stanford University.

PADMA RAGHAVAN is the associate vice president for research and director of strategic initiatives at the Pennsylvania State University, where she is also a distinguished professor of computer science and engineering. Dr. Raghavan is the founding director of the Penn State Institute for CyberScience, the coordinating unit on campus for developing interdisciplinary computation and data-enabled science and engineering. Prior to joining Penn State in 2000, she served as an associate professor in the Department of Computer Science at the University of Tennessee. Her research is in the area of high-performance computing and computational science and engineering. She has more than 95 peer-reviewed publications in three major areas, including scalable parallel computing;

energy-aware supercomputing (i.e., performance and power scalability of advanced computer systems); and computational modeling, simulation, and knowledge extraction. Dr. Raghavan currently serves on the editorial boards of the SIAM book series *Computational Science and Engineering* and *Software, Environments and Tools*, the *Journal of Parallel and Distributed Computing*, the *Journal of Computational Science*, and *IEEE Transactions on Parallel and Distributed Systems*. She serves on the program committees of major conferences sponsored by ACM, IEEE, and SIAM, and she co-chaired Technical Papers for Supercomputing 2012 and the 2011 SIAM Conference on Computational Science and Engineering. Dr. Raghavan also serves on various advisory and review boards, including the Academies' Panel on Digitization and Communication Science, the Network for Earthquake Engineering Simulation, and the Computer Research Association's (CRA's) Committee on the Status of Women in Computing Research. She is a fellow of the IEEE, and she received an NSF CAREER Award and the Maria Goeppert-Mayer Distinguished Scholar Award from the University of Chicago and ANL for her research on parallel sparse matrix computations. Dr. Raghavan received her Ph.D. in computer science from Penn State.

DANIEL A. REED is currently vice president for research and economic development, as well as a professor of computer science, electrical and computer engineering, and medicine at the University of Iowa. He also holds the University Computational Science and Bioinformatics Chair at Iowa. Dr. Reed was a corporate vice president at Microsoft from 2009 to 2012, responsible for global technology policy and extreme computing, and director of scalable and multicore computing at Microsoft from 2007 until 2009. Prior to Microsoft, he was the founding director of the Renaissance Computing Institute at the University of North Carolina, Chapel Hill, where he also served as Chancellor's Eminent Professor and vice chancellor for information technology. Before joining the University of North Carolina, Chapel Hill, in 2003, Dr. Reed was director of the National Center for Supercomputing Applications (NCSA), and Gutgsell Professor and head of the Department of Computer Science at the University of Illinois, Urbana-Champaign. He was appointed to the President's Council of Advisors on Science and Technology (PCAST) by President Bush in 2006 and served on the President's Information Technology Advisory Committee (PITAC) from 2003 to 2005. As chair of PITAC's computational science subcommittee, he was lead author of the report *Computational Science: Ensuring America's Competitiveness*. On PCAST, he co-chaired the Networking and Information Technology subcommittee (with George Scalise of the Semiconductor Industry Association) and coauthored a report on the Networking and Information Technology Research and

Development (NITRD) program called *Leadership Under Challenge: Information Technology R&D in Competitive World*. He is past chair of the board of directors of CRA and currently serves on its Government Affairs Committee. CRA represents the research interests of the university, national laboratory, and industrial research laboratory communities in computing across North America. Dr. Reed received his B.S. from the University of Missouri, Rolla, and his M.S. and Ph.D. degrees from Purdue University, all in computer science.

VALERIE TAYLOR is the senior associate dean of academic affairs in the Dwight Look College of Engineering and the Regents Professor and Royce E. Wisenbaker Professor in the Department of Computer Science and Engineering at Texas A&M University. In 2003, she joined Texas A&M as the department head of Computer Science and Engineering, where she remained in that position until 2011. Prior to joining Texas A&M, Dr. Taylor was a member of the faculty in the Electrical Engineering and Computer Sciences Department at Northwestern University for 11 years. She has authored or coauthored more than 100 papers in the area of high-performance computing. She is also the executive director of the Center for Minorities and People with Disabilities in IT. Dr. Taylor is an IEEE fellow and has received numerous awards for distinguished research and leadership, including the 2001 IEEE Harriet B. Rigas Award for a woman with significant contributions in engineering education, the 2002 Outstanding Young Engineering Alumni Award from the University of California, Berkeley, the 2002 CRA Nico Habermann Award for increasing the diversity in computing, and the 2005 Tapia Achievement Award for Scientific Scholarship, Civic Science, and Diversifying Computing. Dr. Taylor is a member of ACM. She earned her B.S. in electrical and computer engineering and M.S. in computer engineering from Purdue University and a Ph.D. in electrical engineering and computer sciences from the University of California, Berkeley.

KATHERINE A. YELICK is a professor of electrical engineering and computer sciences at the University of California, Berkeley, and the associate laboratory director for computing sciences at Lawrence Berkeley National Laboratory. Dr. Yelick is known for her research in parallel languages, compilers, algorithms, and libraries. She coinvented the UPC and Titanium languages and developed analyses, optimizations, and runtime systems for their implementation. She has also done research on memory hierarchy optimizations, communication-avoiding algorithms, and automatic performance tuning, including developing the first autotuned sparse matrix library. In her current role as associate laboratory director, she manages an organization that includes the National Energy Research

Scientific Computing Center (NERSC), the Energy Science Network (ESNet), and the Computational Research Division. She was the director of NERSC from 2008 to 2012. Dr. Yelick has received multiple research and teaching awards, including the Athena award, and she is an ACM fellow and an IEEE senior member. She is a member of the California Council on Science and Technology, the Academies' Computer Science and Telecommunications Board, and the Science and Technology Committee overseeing research at Los Alamos and Lawrence Livermore National Laboratories. She earned her Ph.D. in electrical engineering and computer science from the Massachusetts Institute of Technology.

D

Acronyms and Abbreviations

3D	three-dimensional
ACI	Division of Advanced Cyberinfrastructure (NSF)
aLIGO	Advanced Laser Interferometer Gravitational Wave Observatories
BBH	binary black hole
BD Hub	Big Data Regional Innovation Hub
BLAS	Basic Linear Algebra Subprograms
CC*DNI	Campus Cyberinfrastructure—Data, Networking, and Innovation Program
CCAM	Commonwealth Center for Advanced Manufacturing
CCP	Collaborative Computational Project
CIF21	Cyberinfrastructure Framework for 21st Century Science and Engineering
CISE	Directorate for Computer and Information Science and Engineering
CMOS	complementary metal-oxide semiconductor
CPU	central processing unit
CSIA	Cyber Security and Information Assurance
DOD	Department of Defense
DOE	Department of Energy

DRAM	dynamic random-access memory
EB	exabyte
ECSS	Extended Collaborative Support Service
EU	European Union
F&A	facilities and administrative
FFRDC	federally funded research and development center
FFTPACK	Fastest Fourier Transform in the West Package
FFTW	Fastest Fourier Transform in the West
FLOP/s	floating-point operations per second
FPGA	field-programmable gate array
GB	gigabyte
GENI	Global Environment for Network Innovations
GPU	graphical processing unit
GSL	GNU Scientific Library
GTEPS	giga-traversed edges per second
HCSS	High Confidence Software and Systems
HDD	hard disk drive
HECIA	high-end computing infrastructure and applications
HECRD	High-End Computing Research and Development
HPC	high-performance computing
HPCG	High-Performance Conjugate Gradient
I/O	input/output
IOPS	input/output operations per second
IT	information technology
LAPACK	Linear Algebra Package
LSN	Large Scale Networking
LSST	Large Synoptic Survey Telescope
MB	megabyte
MKL	Math Kernel Library
MPI	message passing interface
MREFC	Major Research Equipment and Facilities Construction
MTDC	modified total direct cost
MUMPS	Massachusetts General Hospital Utility Multi-Programing Systems

NASA	National Aeronautics and Space Administration
NCAR	National Center for Atmospheric Research
NCSA	National Center for Supercomputing Applications
NERSC	National Energy Research Scientific Computing Center
NITRD	Networking and Information Technology Research and Development
NSCI	National Strategic Computing Initiative
NSF	National Science Foundation
NTT	Nippon Telegraph and Telephone
OCI	Office of Cyberinfrastructure
PACI	Partnership for Advanced Computational Infrastructure
PAPI	precision approach path indicator
pARPACK	Parallel Arnoldi Package
PB	petabyte
PETSc	Portable, Extensible Toolkit for Scientific Computation
PGAS	partitioned global address space
PITAC	President's Information Technology Advisory Committee
PRAC	Petascale Computing Resource Allocations Committee
PRACE	Partnership for Advanced Computing in Europe
PSC	Pittsburgh Supercomputing Center
PSPLINE	Princeton Spline and Hermite Cubic Interpolation Routines
R&D	research and development
ScaLAPACK	Scalable Linear Algebra PACKage
SCEC	Southern California Earthquake Center
SciDAC	Scientific Discovery through Advanced Computing
SCOREC	Scientific Computation Research Center
SDP	Software Design and Productivity
SEW	Social, Economic, and Workforce Implications of IT and IT Workforce Development
SISI	Software Infrastructure for Sustained Innovation program
SLEPc	Scalable Library for Eigenvalue Problem Computations
SSD	solid-state disk
SU	service unit
TACC	Texas Advanced Computing Center
TB	terabyte

UPC	Universal Product Code
WAN	wide area network
XDMoD	XD Metrics on Demand
XRAC	XSEDE Resource Allocation Committee
XSEDE	Extreme Science and Engineering Discovery Environment