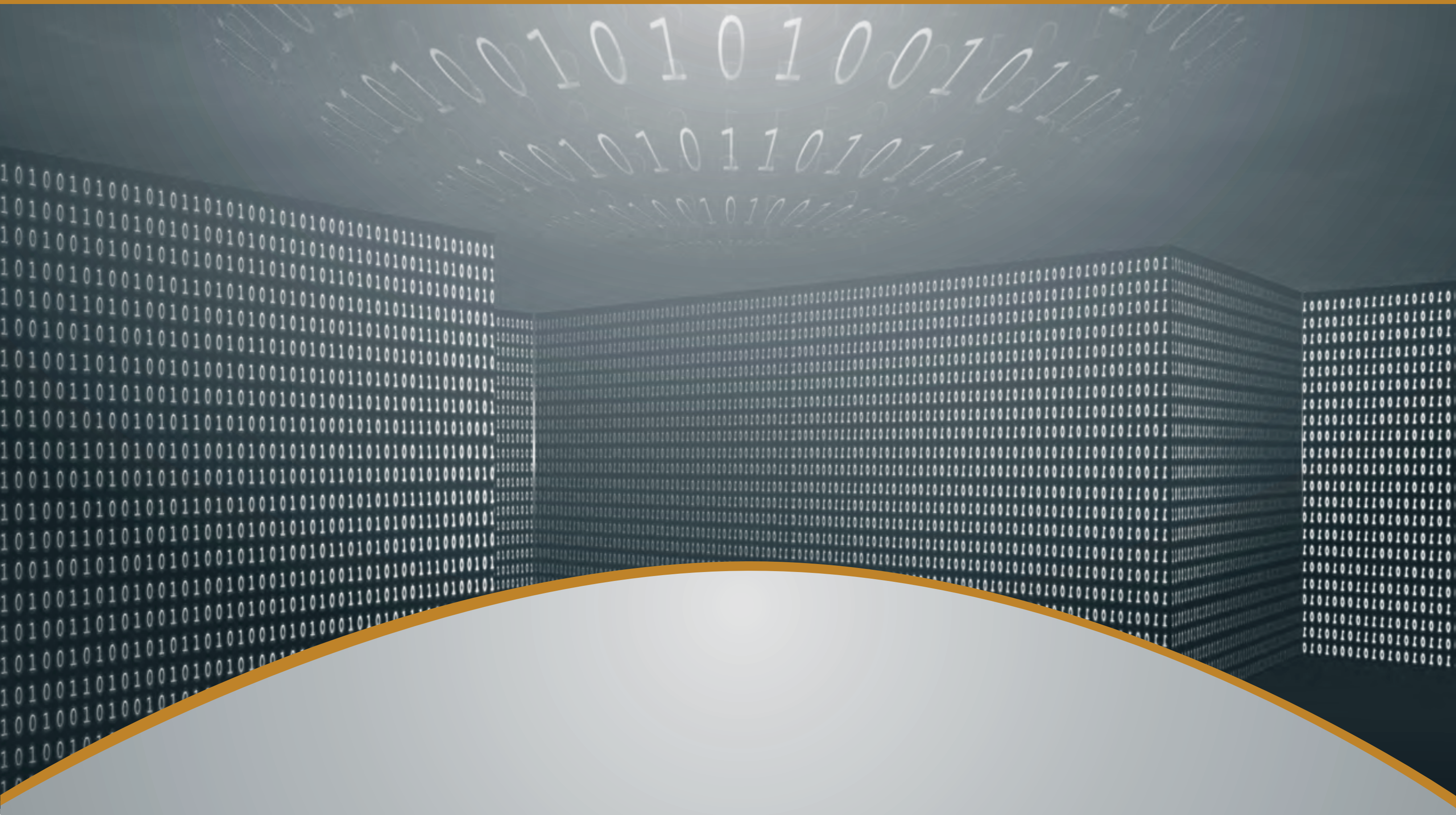

National Science Foundation
Advisory Committee for Cyberinfrastructure
Task Force on High Performance Computing

Final Report, March 2011

Report of the High Performance Computing Task Force

Overview

The High Performance Computing (HPC) Task Force is one of six task forces commissioned in April 2009 by the National Science Foundation (NSF) within the Advisory Committee on Cyberinfrastructure (ACCI), a panel of distinguished scientists and engineers from academia, government, and industry assembled to advise NSF on all cyberinfrastructure (CI) activities. Members of the HPC Task Force are listed in Table ES-1.

Table ES-1. HPC Task Force membership

Thomas Zacharia, Chair, University of Tennessee/Oak Ridge National Laboratory
Jim Kinter, Co-Chair, Center for Ocean-Land-Atmosphere Studies
Rob Pennington, NSF Liaison
Ravi Arimilli, IBM, Inc.
Ron Cohen, Carnegie Institution of Washington
Larry Davis, IDA
Tiziana Di Matteo, Carnegie Mellon University
Bill Harrod, Defense Advanced Research Projects Agency
George Karniadakis, Brown University
Rubin Landau, Oregon State University
Rich Loft, National Center for Atmospheric Research
Michael Macy, Cornell University
Dick McCombie, Cold Spring Harbor Laboratory
Dave Randall, Colorado State University
Steve Scott, Cray, Inc.
Horst Simon, Lawrence Berkeley National Laboratory
Thomas Sterling, Louisiana State University
Theresa Windus, Iowa State University
NSF Liaison – Rob Pennington

The HPC Task Force was charged with providing specific advice on the broad portfolio of HPC investments that NSF could consider to best advance science and engineering during the next 5–10 years, including not only the CI necessary to support breakthrough research in science and engineering, but also research on CI (hardware and software) and training to enable the effective use of these investments. The task force was directed to consider how to advance HPC and its use in the context of ensuring access, applications development, computer science and engineering, education, training, and policy implementation. In particular, the task force was asked to address six general questions:

1. **Ensuring access:** What process and strategy should NSF use to deliver access to advanced computing resources from 2011 to 2015?

2. **Applications development and support:** How should the software and algorithms necessary to support current and future applications of HPC be developed and maintained?
3. **Computer science and engineering:** What areas of computer science and engineering are ripe for innovations that can advance the development and use of HPC?
4. **Integration of research and education:** How can the ways in which HPC is used in scientific discovery be integrated into pre-college, undergraduate, and post-graduate education?
5. **Training:** What should be the scope and focus of training efforts to prepare the scientific workforce to make effective use of HPC and to continue to advance the development of HPC?
6. **Policy implementation:** What policies and programmatic activities can broaden the scientific and engineering exploitation of HPC approaches?

The HPC Task Force used a series of three workshops on sustainability for the national computational CI; application drivers for exascale computing and data CI; and on broader engagement to gather input from the community on these questions. Participants in these workshops, held in Arlington, VA, were drawn from universities, industry, and government. Presentations, reports, and position papers collected from all workshops are available on line, http://www.nics.tennessee.edu/workshop. The findings from the workshops were considered and refined by the Task Force and are presented in this document as the recommendations of the Task Force.

Workshop 1: Sustainability for the National Computational Cyberinfrastructure

The first workshop, held on 4 December 2009, focused on sustainability and user requirements for the national computational CI. The 38 participants were asked to address the following issues:

- Sustainability and user requirements in HPC
  - Should the current acquisition model be revised?
  - What is the proper balance between production and experimental systems?
- Building on TeraGrid and XD for integration and advanced services
- Alternate models of computing (clouds, grids, etc.; the use of commercial providers; and the potential of pay-for-service models)
- Exascale and beyond: Advancing the applications community

Workshop 2: Application Drivers for Exascale Computing and Data CI

The second workshop, held 29 July 2010, focused on applications at the exascale. The 44 participants were asked to consider the following issues:

- What is exascale computing and data CI essential for?
- What new science will be enabled by exascale computing and data CI?
- What are the challenges to reach exascale?
- What scientific areas have not yet considered exascale, and how do we help them get there?

Workshop 3: Broader Engagement

At the third workshop, held on 3 December 2010, the 41 participants were asked to consider the topic of broader engagement, focusing on the following issues:

- What is broader engagement and why is it important?
- What strategies should we use to engage new communities?
- How can we train and sustain the next generation?
- What are the challenges and how can we overcome them?

Summary recommendations are presented here, followed by descriptions of the findings and recommendations resulting from in each of the three areas.

Strategic Recommendations to NSF

In consideration of the input received from the three workshops and after due deliberation, the HPC Task Force makes the following overarching strategic recommendations to the NSF:

1. Develop a sustainable model to provide the academic research community with access, by 2015–2016, to a rich mix of HPC systems that:
   – deliver sustained performance of 20–100 petaflops on a broad range of science and engineering applications;
   – are integrated into a comprehensive national CI environment; and
   – are supported at national, regional, and/or campus levels.

2. Invest now to prepare for exascale systems that will be available by 2018–2020. NSF should consider the use of co-design partnerships to provide the HPC systems and data CI needed to enable data-driven science.

3. Broaden outreach to improve the preparation of HPC researchers and to engage industry, decision-makers, and new user communities in the use of HPC as a valuable tool.

4. Establish a continuing process for soliciting community input on plans for HPC investments.

The HPC Task Force considers the issues confronting the NSF in the area of high-performance computing to fall into the following 3 categories:

1. Cyberinfrastructure Sustainability
2. Exascale Computing
3. Broader Engagement

The remainder of the report provides detailed findings and recommendations in each of these areas.

Cyberinfrastructure Sustainability

Findings:

The U.S. science and engineering research and education (R&E) community is critically dependent on NSF CI. However, the current model for HPC centers creates uncertainty for both the centers and the users by diverting the attention of center staff from services to competition. The impact of competition, uncertainty due to instability, on the centers' ability to attract and retain high-quality staff impairs the quality and quantity of services. In addition, the model creates barriers to long-term relationships with vendors, thus limiting joint research and imposing financial risk.

The scale of available HPC and the appetite for it are growing rapidly. The rapid growth of HPC systems offers potential performance improvements (more processors of slowly increasing speed) and increased problem scale (more memory), but the development of science and engineering applications that can make effective use of leading-edge computer systems with hundreds of thousands of compute cores, hundreds of terabytes of memory, and tens of petabytes of disk storage is becoming increasingly complex. Additional funding is needed to support the intensive and sustained research on software and algorithms that will be required to enable effective use of these systems by the computational science and engineering R&E community. In particular, new utility and math codes are needed both for today's petascale systems and for the exascale systems that are expected to be available within the next decade. In addition, the NSF HPC centers need additional funding for user support, training, outreach, and education programs to promote the understanding of HPC and its application in scientific research.

Data-driven science is insufficiently supported at the NSF HPC centers. As research has become increasingly data-intensive, the need for data management tools and software, mass storage, and archival hardware and services has not been adequately addressed.

Finally, the heavy use of the NSF HPC centers by researchers supported by other federal agencies contributes to the oversubscription of NSF resources and raises additional concerns about long-term sustainability. This situation can be taken as an indication of the success of the centers, and it shows that NSF is filling an important gap for agencies that lack the expertise needed to manage and operate their own HPC systems. However, competition for scarce resources is frustrating for researchers in all disciplines, and consideration should be given to the development of a process that provides for interagency cooperation in distributing costs.

Recommendations for NSF:

1. By 2015–2016, NSF should ensure that academic researchers have access to a rich mix of HPC systems that:

- deliver a sufficient number of systems with sustained performance of 20–100 petaflops on a broad range of science and engineering codes;
- are integrated into a comprehensive, national cyberinfrastructure environment; and

- are supported at national, regional, and/or campus levels.

This mix of systems is expected to include NSF investment in a few systems delivering sustained performance in the 10–50 petaflops range across a broad range of science and engineering research applications and at least one system capable of exceeding 250 petaflops of sustained performance on the most computationally challenging research codes, both in terms of actual throughput rate achieved and strong scalability (time to solution). In addition, because the scale, resolution, and fidelity of scientific simulations are limited by the configuration and balance of HPC systems, it is crucial for NSF to engage the scientific and engineering community in shaping the solicitations for these next-generation systems.

2. To sustain and promote the stability of resources, NSF should direct the evolution of its supercomputing program in a sustainable way, allowing researchers and HPC centers to select the best value in computational and data platforms and enabling centers to offer continuous service to the community.

The recommended investments in trans-petascale hardware will be effective only if NSF simultaneously provides (1) a stable, experienced pool of expertise in the management and operation of leading-edge resources, (2) support services for researchers using these resources, and (3) support for the development and deployment of the new tools needed to make optimal use of these resources.

In acting on these recommendations, NSF should:
- Commit to stable and sustained funding for HPC centers to allow them to recruit and develop the expertise needed to maximize the potential offered by NSF's hardware investments. Rigorous review and oversight processes, with "institutional memory" that can monitor progress over time, can be developed and implemented to provide assurance that centers meet NSF expectations for performance.
- Encourage HPC centers to build long-term relationships with vendors, thus providing researchers with the benefits of a planned road map for multiple generations of chip technology upgrades and with continuity in architecture and software environments. Results-oriented acquisition strategies can be applied to ensure that vendor performance meets center and NSF needs.

Exascale Computing

Findings:

Extensive reviews by the scientific community have identified a number of scientific "grand challenges" that can exploit computing at extreme scales to bring about dramatic progress toward their resolution. Past experience with the application of HPC to challenges in biology, chemistry, climate, materials science, nuclear fission, nuclear fusion, physics, and other fields provides confidence that vastly expanded compute power and data capacity will enable continuing advances as simulation fidelity improves and/or more time steps can be computed. In addition, progress in uncertainty quantification will enhance acceptance of the results of increasingly sophisticated models of complex physical systems.

As computational resources, sensor networks, and other large-scale instruments and experiments expand to the exascale, the amount of data generated by these sources will also grow commensurately in many cases. Opportunities exist for new discoveries in data-rich fields—not only in areas such as biology and medicine, chemistry, climate science, earth science, and astrophysics, but also in public health, economics, and social science. New exascale CI will be essential, not only to manage and extract useful information from these massive data sets, but also to ensure the preservation of persistent and definitive data repositories and to facilitate and encourage sharing of data from high-cost experiments and simulations across communities.

Reaching the exascale presents a number of substantial challenges that can be overcome only through a broad effort. The dramatic changes in hardware and architecture needed to attain this level of performance create the need for new systems software, tools, and programming models. Research and development in mathematical models, applied mathematics, and computer science is required to meet the needs of a spectrum of applications at the exascale. Research and development is also needed in algorithm design, notably to address the issue of resiliency and fault-tolerance or fault-obliviousness, but also for other issues that will arise approaching the exascale. Without further research and development the range of benefits may prove too narrow in range of applications and the deployment and operational costs prohibitive except for single national showcase machines.

In particular, application development is critical both in broadening the use of petascale computing and in advancing to the exascale. Participants noted that "co-design" of architectures and algorithms (in which scientific problem requirements influence architecture design and technology constraints inform the formulation and design of algorithms and software) offers the opportunity to improve the effectiveness of both petascale and exascale systems. In addition, training programs are needed both to encourage use of resources at the lower levels of the "Branscomb pyramid" and to address the new operating models and different memory hierarchies expected for exascale systems.

There is widespread uncertainty about the architectures and programming models that will be dominant over the next five years. The wide range of experimental architectures, exploring varying degrees of heterogeneity, that are currently being considered gives rise to uncertainty about whether or not MPI will be the programming model of choice and how fast compilers will keep pace with the rapidly evolving architecture.

Recommendations for NSF:

1. Given the major challenges involved in the transition to HPC at the exascale, NSF should consider new models for partnerships, such as expanded collaborations with

industry, academia, and other agencies, to accelerate the development of exascale systems and applications.

2. Given the promise of exascale computing and data CI to accelerate innovation and enhance competitiveness, NSF should expand its efforts to engage new user communities in HPC, with particular emphasis on broadening industry use of HPC and on applying new and expanded capabilities to data-intensive fields of research.

3. Given the opportunities and challenges presented by the generation of exabytes of digital data, NSF should provide funding for a digital data framework designed to address the issues of knowledge discovery in the exascale ecosystem, including the co-location of archives and community data resources with compute and visualization resources as appropriate.

4. Given the rapidly changing environment for HPC, NSF should establish a continuing process for soliciting community input on its plans for HPC investments to facilitate the growth of an ecosystem for exascale science. The NSF should also invest in research to determine if a new execution model is needed, and, if so, what it should look like to drive us towards exascale capability (including strong scaling).

Broader Engagement

Findings:

Traditionally, broader engagement has been defined as increasing the involvement in HPC of individuals from underrepresented groups (e.g., women, minorities, people with disabilities). Today, this definition is expanding to include efforts to engage new research communities. Participants noted that broader engagement is a competitiveness issue: HPC is widely recognized as a critical tool for accelerating scientific discovery and innovation in many fields of research, but few researchers have the skills necessary to apply this tool. As is the case in many technical fields, women and U.S.-born minorities remain underrepresented in HPC.

A variety of strategies will be needed to engage new communities. These include mentoring programs to encourage members of underrepresented groups to pursue careers in HPC; focused HPC outreach programs at the campus level, such as the NSF Campus Champions; and the development of tools and training to facilitate the use of HPC and computational techniques for R&E in additional disciplines. It should be noted that one community that previously had a large presence in HPC has diminished in activity: computational engineering.

Mechanisms needed to train and sustain the next generation include programs for broader engagement and workforce development at all levels, starting with K–12; realignment of current computing and computational science education programs; and continuing education to support lifelong learning about HPC.

It is essential that the NSF recognize the gap in appreciation and respect among young people for the sciences, engineering, and computational science and undertake initiatives to motivate and inspire them. The NSF should fill this gap and support programs to motivate young people to learn more about the computational sciences and consider adopting careers in these fields, which can be highly rewarding. Students must be offered career choices independent of the socio-economic and demographic groups in which they find themselves.

Culture change will be needed to facilitate the development of innovative inter-disciplinary curricula and appropriate reward systems, both to train computational specialists and to provide scientists and engineers in other fields with the knowledge and skills needed to use HPC effectively. The development and growth of the necessary programs in a financially constrained environment was also cited as a major challenge.

Recommendations to NSF:

1. NSF should continue and grow a variety of education, outreach, and training programs to expand awareness and encourage the use of high-end modeling and simulation capabilities for a variety of purposes: accelerating innovation, increasing industrial competitiveness, supporting decision-makers, and enabling advances in fields that have not yet embraced HPC as a key research tool. Improvements in the preparation of HPC researchers (e.g., through curriculum development to integrate HPC into computational science) should be encouraged. Outreach to industry and to new user communities is particularly important in demonstrating the value of next-generation HPC for accelerating innovation, while expanded workforce development programs, starting with K–12, are essential to educating not only a diverse group of computer scientists and engineers, but also domain scientists with the mathematical and computational skills needed to take full advantage of HPC for discovery.

2. NSF should consider providing support for large-scale simulations of practical, large-scale engineering systems (e.g. a nuclear reactor, an internal combustion engine or the entire human arterial tree). Scientists at the NSF HPC centers should be encouraged to collaborate more directly with academic groups. More generally, the NSF should consider sponsoring postdoctoral positions and sabbatical leaves for faculty who want to spend some time at the HPC centers learning about new algorithms and new tools, which can then disseminate to their campuses. This can complement the Campus Champions program to further foster meaningful and long-lasting interactions.

3. NSF should develop mechanisms for measuring the impacts of education, outreach, and training programs (e.g., expansion in use of HPC or delivery of science). Such mechanisms will make it possible to identify and foster the most productive programs.

http://www.nsf.gov/od/oci/taskforces