Testimony of Daniel A. Reed
Director, National Center for Supercomputing Applications (NCSA)
University of Illinois at Urbana-Champaign

July 16, 2003

Good morning, Mr. Chairman and Members of the Committee. Thank you very much for granting me this opportunity to comment on appropriate paths for scientific computing. I am Daniel Reed, Director of the National Center for Supercomputing Applications (NCSA), one of three NSF-funded high-end computing centers. I am also a researcher in high-performance computing and a former head of the Department of Computer Science at the University of Illinois.

In response to your questions, I would like to make three points today regarding the status and future of advanced scientific computing.

1. Success and Scientific Opportunities

First, the National Science Foundation has played a critical role in providing high-end computational infrastructure for the broad university-based science and engineering community, a critical role that must continue. NCSA began with NSF funding for the supercomputer centers program in the 1980s. Via this program and its successors, the NSF supercomputing centers have provided community access to high-end computers that previously had been available only to researchers at national laboratories. These supercomputing investments have not only enabled scientific discovery, they have also catalyzed development of new technologies and economic growth.

The Internet sprang from early DARPA investments and from funding for NSFNet, which first connected NSF's supercomputing centers and provided open access to high-end computing facilities. NCSA Mosaic, the first graphical web browser, which spawned the web revolution, grew from development of tools to support collaboration among distributed scientific groups. Research by NCSA and the other supercomputing centers was instrumental in the birth of scientific visualization – the use of graphical imagery to provide insight into complex scientific phenomena. Via ONR-funded security center, NCSA is creating new cybersecurity technologies to safeguard the information infrastructure of our nation and our military forces.

Today, the NSF supercomputing centers and their Partnerships for Advanced Computational Infrastructure (PACI) program and Extended Terascale Facility (ETF) partners are developing new tools for analyzing and processing the prodigious volumes of experimental data being produced by new scientific instruments. They are also developing new Grid technologies that couple distributed instruments, data archives and high-end computing facilities, allowing national research collaborations on an unprecedented scale.

Access to high-end computing facilities at NCSA, the San Diego Supercomputing Center (SDSC) and the Pittsburgh Supercomputing Center (PSC) is provided by national peer review, with computational science researchers reviewing the proposals for supercomputing time and support. This process awards access to researchers without regard to the source of their research funds. The three centers support researchers with DOE, NIH and NASA awards, among others. This ensures that the most promising proposals from across the entire science and engineering community gain access to high-end computing hardware, support staff and software. Examples include

- *Simulation of cosmological evolution to test basic theories of the large-scale structure of the universe*
- *Quantum chromodynamics simulations to test the Standard Model of particle physics*
- *Numerical simulation of weather and severe storms for accurate prediction of severe storm tracks*

- *Climate modeling to understand the effects of climate change and assess global warming*
- *Studying the dynamics and energetics of complex macromolecules for drug design*
- *Nanomaterials design and assessment*
- *Fluid flow studies to design more fuel-efficient aircraft engines*

From these, let me highlight just two examples of scientific exploration enabled by high-end computing, and the limitations on computing power we currently face.

A revolution is underway in both astronomy and high-energy physics, powered in no small part by high-end scientific computing. New data, taken from high-resolution sky surveys, have exposed profound questions about the nature and structure of the universe. We now know that the overwhelming majority of the matter in the universe is of an unknown type, dubbed "dark matter," that is not predicted by the Standard Model of physics. Observations also suggest that an unknown "dark energy" is causing the universe to expand at an ever-increasing rate. Both of these discoveries are profound and unexpected.

Large-scale simulations at NCSA and other high-end computing centers are being used to investigate models of the evolution of the universe, providing insight in the viability of competing theories. The goal of these simulations is to create a computational "universe in a box" that evolves from first principles to conditions similar to that seen today. **These cosmological studies also highlight one of the unique capabilities of large-scale scientific simulation – the ability to model phenomena where experiments are otherwise not possible.**

As a second example, the intersecting frontiers of biology and high-end computing are illuminating biological processes and medical treatments for disease. We have had many successes and others are near. For example, given large numbers of small DNA fragments, high-end computing enabled "shotgun sequencing" approaches for assembling the first maps of the human genome. High-end computing has also enabled us to understand how water moves through cell walls, how blood flow in the heart can disrupt the plaque that causes embolisms, and how new drugs behave.

Unraveling the DNA code for humans and organisms has enabled biologists and biomedical researchers to ask new questions, such as how genes control protein formation and cell regulatory pathways and how different genes increase the susceptibility to disease. Biophysics and molecular dynamics simulations are now being used to study the structure and behavior of biological membranes, drug receptors and proteins. In particular, understanding how proteins form three-dimensional structures is central to designing better drugs and combating deadly diseases such as HIV and SARS. **Today's most powerful high-end computing systems can only simulate microseconds of the protein folding process; complete folding takes milliseconds or more. Such large-scale biological simulations will require vast increases in computing capability, perhaps as much as 1000 times today's capability.**

Simply put, we are on the threshold of a new era of scientific discovery, enabled by computational models of complex phenomena. From astronomy to zoology, high-end computing has become a peer to theory and experiment for exploring the frontiers of science and engineering.

This brings me to my second point: the challenges before us.

2. Challenges

Although large-scale computational simulation has assumed a role equal to experiment and theory in the scientific community, within that community, we face critical challenges. There is a large and unmet demand for access to high-end computing in support of basic scientific and engineering research. There are neither enough high-end computing systems available nor are their capabilities adequate to address fully the research challenges and opportunities. This view is supported by recent workshops, reports and surveys, including the NSF report on high-end computing and cyberinfrastructure, the DOD integrated high-end computing report, and the DOE study on a science case for large-scale simulation.

On behalf of the interagency High-End Computing Revitalization Task Force (HECRTF), I recently chaired a workshop to gain community input on high-end computing opportunities and needs. At the workshop, researchers from multiple disciplines made compelling cases for sustained computing performance of 50-100X beyond that currently available Moreover, researchers in every discipline at the HECRTF workshop cited the difficulty in achieving high, sustained performance (relative to peak) on complex applications. to reach new, important scientific thresholds. Let me cite just a few examples to illustrate both the need and our current shortfall.

In high-energy physics, lattice quantum chromodynamics (QCD) calculations, which compute the masses of fundamental particles from first principles, require *a sustained performance* of 20-50 teraflops/second (one teraflop is 10^{12} arithmetic operations per second). This would enable predicative calculations for ongoing and planned experiments. In magnetic fusion research, sustained performance of 20 teraflops/second would allow full-scale tokamak simulations, providing insights into the design and behavior of proposed fusion reactor experiments such as the international ITER project. HECRTF workshop participants also estimated that a sustained performance of 50 teraflops/second would be needed to develop realistic models of complex mineral surfaces for environmental remediation, and to develop new catalysts that are more energy efficient.

Note that each of these requirements is for sustained performance, rather than peak hardware performance. This is notable for two reasons. First, the aggregate peak performance of the high-end computing systems now available at NSF's three supercomputing centers (NCSA, SDSC and PSC) and DOE's National Energy Research Scientific Computing Center (NERSC) is roughly 25 teraflops.[1] Second, researchers in every discipline at the HECRTF workshop cited the difficulty in achieving high, sustained performance (relative to peak) on complex applications.

Simply put, the nation's aggregate, open capability in high-end computing is at best equal to the scientific community's estimate of that needed for a single, breakthrough scientific application study. This is an optimistic estimate, as it assumes one could couple all these systems and achieve 100 percent efficiency. Instead, these open systems are shared by a large number of users, and the achieved application performance is often a small fraction of the peak hardware performance. This is not an agency-specific issue, but rather a shortfall in high-end computing capability that must be addressed by all agencies to serve their community's needs.

Achieving high-performance for complex applications requires a judicious match of computer architecture, system software and software development tools. Most researchers in high-end computing believe the key reasons for our current difficulties in achieving high performance on complex scientific applications can be traced to (a) inadequate research investment in software and (b) use of processor and memory architectures that are not well matched to scientific applications. Today, scientific applications are developed with software tools that are crude compared to those used in the commercial sector. Low-level programming, based on message-passing libraries, means that application developers must provide deep knowledge of application software behavior and its interaction with the underlying computing hardware. This is a tremendous intellectual burden that, unless rectified, will continue to limit the usability of high-end computing systems, restricting effective access to a small cadre of researchers.

[1] According to the most recent "Top 500" list (www.top500.org), 6 teraflops at PSC, 6 teraflops at NCSA, 3.7 teraflops at SDSC and 10 teraflops at NERSC. Upcoming deployments at NCSA, SDSC and PSC will raise this number, but the aggregate will still be far less than user requirements.

Developing effective software (programming languages and tools, compilers, debuggers and performance tools) requires time and experience. Roughly twenty years elapsed from the time vector systems such as the Cray-1 first appeared in the 1970s until researchers and vendors developed compilers that could automatically generate software that operated as efficiently as that written by a human. This required multiple iterations of research, testing, product deployment and feedback before success was achieved.

In the 1990s, the U.S. HPCC program supported the development of several new computer systems. In retrospect, we did not learn the critical lesson of vector computing, namely the need for long-term, sustained and balanced investment in both hardware and software. We under invested in software and expected innovative research approaches to high-level programming to yield robust, mature software in only 2-3 years. One need only look at the development history of Microsoft Windows™ to recognize the importance of an iterated cycle of development, deployment and feedback to develop an effective, widely used product. High quality research software is not cheap, it is labor intensive, and its successful creation requires the opportunity to incorporate the lessons learned from previous versions.

The second challenge for high-end computing is dependence on products derived too narrowly from the commercial computing market. Although this provides enormous financial leverage and rapid increases in peak processor performance, commercial and scientific computing workloads differ in one important and critical way – access to memory. Most commercial computer systems are designed to support applications that access a small fraction of a system's total memory during a given interval.

For commercial workloads, caches – small, high-speed memories attached to the processor – can hold the critical data for rapid access. In contrast, many, though not all, scientific applications (and several of those critical to signals intelligence and cryptanalysis) have irregular patterns of access to a large fraction of a system's memory. This is not a criticism of vendors, but rather a marketplace reality we must recognize and leverage. New high-end computing designs are needed to support these characteristics, both for fully custom high-end computer designs and more appropriate designs based on commodity components.

The dramatic growth of the U.S. computing industry, with its concomitant economic benefits, has shifted the balance of influence on computing system design away from the government to the private sector. As the relative size of the high-end computing market has shrunk, we have not sustained the requisite levels of innovation and investment in high-end architecture and software needed for long-term U.S. competitiveness. Alternative strategies will be required.

This leads me to my third and final point: appropriate models of cooperation and support for high-end computing.

3. Actions

We must change the model for development, acquisition and deployment of high-end computing systems if the U.S. is to sustain the leadership needed for scientific discovery and national security in the long term. The Japanese Earth System Simulator is a wakeup call, as it highlights the critical importance of both industry-government collaboration and long-term sustained investment. Reflecting the lessons of long-term investment I discussed earlier, the Earth System Simulator builds on twenty years of continued investment in a particular hardware and software model, and the lessons of six product generations. To sustain U.S. leadership in computational science, we must pursue two concurrent and mutually supporting paths, one short to medium term and the second long term.

In the short to medium term, we must acquire and continue to deploy additional high-end systems at larger scale if we are to satisfy the unmet demand of the science and engineering research community.

NSF's recent cyberinfrastructure report,[2] DOD's integrated high-end computing report, and DOE's ultrascale simulation studies have all made such recommendations. As one example, the cyberinfrastructure report noted "The United States academic research community should have access to the most powerful computers that can be built and operated in production mode at any point in time, rather than an order of magnitude less powerful, as has often been the case in the past decade." The cyberinfrastructure report estimated this deployment as costing roughly $75M/year per facility, with $50M/year per facility allocated to high-end computing hardware..

Given the interdependence between application characteristics and hardware architecture, this will require deployment of high-end systems based on diverse architectures, including large-scale message-based clusters, shared memory systems (SMPs) and vector systems.[3] Moreover, these systems must not be deployed in a vacuum, but rather must leverage another critical element of sustainable infrastructure – the experienced support staff members who work with application scientists to use high-end systems effectively. These high-end centers must also interoperate with a broad infrastructure of data archives, high-speed networks and scientific instruments.

High-end computing system deployments should not be viewed as an interagency competition, but rather as an unmet need that requires aggressive responses from multiple agencies. NSF and its academic supercomputing centers have successfully served the open academic research community for seventeen years; NSF should build on this success by deploying larger systems for open community access. Similarly, DOE has well served the high-end computing needs of laboratory researchers; it too should build on its successes. NIH, DOD, NASA, NSA and other agencies also require high-end capabilities in support of their missions, both for research and for national needs. The need is so large, and the shortfall is so great, that broader investment is needed by all agencies.

Concurrent with these deployments, we must begin a coordinated research and development effort to create high-end systems that are better matched to the characteristics of scientific applications. To be successful, these efforts must be coordinated across agencies in a much deeper and tighter way than in the past. This will require a broad, interagency program of basic research into computer architectures, system software, programming models, software tools and algorithms.

In addition, we must fund the design and construction of large-scale prototypes of next-generation high-end systems that includes balanced exploration of new hardware and software models, driven by scientific application requirements. Multiple, concurrent efforts will be required to reduce risk and to explore a sufficiently broad range of ideas; six efforts, each federally funded at a minimum level of $5M-$10M/year for five years, is the appropriate scale. At smaller scale, one will not be able to gain the requisite insights into the interplay of application needs, hardware capabilities, system software and programming models.

Such large-scale prototyping efforts will require the deep involvement and coordinated collaboration of vendors, national laboratories and centers, and academic researchers, with coordinated, multi-agency investment. After experimental assessment and community feedback, the most promising efforts should then transition to even larger scaling testing and vendor productization, and new prototyping efforts should be launched. It is also important to remember the lesson of the Earth System Simulator – the

---

[2] "*Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*," January 2003, www.cise.nsf.gov/evnt/reports/toc.htm

[3] This is the approach we have adopted at NCSA, deploying multiple platforms, each targeting a distinct set of application needs.

critical cycle of prototyping, assessment, and commercialization must be a long-term, sustaining investment, not a one time, crash program.

I believe we face both great opportunities and great challenges in high-end computing. Scientific discovery via computational science truly is the "endless frontier" of which Vannevar Bush spoke so eloquently in 1945. The challenges are for us to sustain the research, development and deployment of the high-end computing infrastructure needed to enable those discoveries.

In conclusion, Mr. Chairman, let me thank you for this committee's longstanding support for scientific discovery and innovation. Thank you very much for your time and attention. I would be pleased to answer any questions you might have.