Testimony of Daniel A. Reed
Chair, Computing Research Association (CRA)

May 8, 2008

Good afternoon, Mr. Chairman and Members of the Committee. Thank you for granting me this opportunity to comment on current U.S. computational capabilities and the research and infrastructure needs to support climate modeling. I am Daniel Reed, Chair of the Board of Directors for the Computing Research Association (CRA). I am a researcher in high-performance computing; a member of the President's Council of Advisors on Science and Technology (PCAST); the former Director of the National Center for Supercomputing Applications (NCSA), one of NSF's high-performance computing centers; and Director of Scalable and Multicore Computing Strategy at Microsoft.

I would like to make five points today regarding the status and future of high-performance computing (HPC) for climate change modeling, beginning with the relationship between HPC and climate change models.

1. High-end Computational Science: Enabling Climate Change Studies

We know the Earth's climate has changed during the planet's history, due to the complex interplay of the oceans, land masses and atmosphere, the solar flux and the biosphere. Recently, the U.S. Climate Change Science Program and the Intergovernmental Panel on Climate Change (IPCC)[1] concluded that climate change will accelerate rapidly during the 21st century unless there are dramatic reductions in greenhouse emissions. **We now face true life and death questions – the potential effects of human activities and natural processes on our planet's ecosystem. I believe HPC tools and technologies provide one of our best options for gaining that understanding.**

In 2005, I was privileged to chair the computational science subcommittee of the President's Information Technology Advisory Committee (PITAC), which examined the competitive position of the U.S. in computing-enabled science. In our report, *Computational Science: Ensuring America's Competitiveness,*[2] we noted that

> *Computational science is now indispensable to the solution of complex problems in every sector, from traditional science and engineering domains to such key areas as national security, homeland security, and public health. Advances in computing and connectivity make it possible to develop computational models and capture and analyze unprecedented amounts of experimental and observational data to address problems previously deemed intractable or beyond imagination.*

Computational science now constitutes the third pillar of the scientific enterprise, a peer alongside theory and physical experimentation. This is especially important in a field such as climate change studies, where the models are complex – multidisciplinary and multivariate – and one cannot conduct parametric experiments at planetary scale.

Why then, is HPC especially critical to climate change studies? First, one must simulate hundreds to thousands of Earth years to validate models and to assess long-term consequences. This is practical only if one can simulate a year of climate in at most a few hours of elapsed time. Each of these simulations must of sufficient fidelity (i.e., temporal and spatial resolution) to capture salient features. Today, for example, most climate models that are run for several hundred to several thousand

[1] R. Alley *et al*, *Climate Change 2007: The Physical Science Basis*, IPCC, Working Group 1 for the Fourth Assessment, WMO.

[2] *Computational Science: Ensuring America's Competitiveness* President's Information Technology Advisory Committee (PITAC), June 2005, http://www.nitrd.gov/pitac/reports/20050609_computational/computational.pdf

simulated years do not explicitly resolve important regional features like hurricanes. These are large-scale, capability computing problems (i.e., ones requiring the most powerful computing systems).

Second, to understand the effects of environmental changes and to validate climate models, one must conduct parameter studies (e.g., to assess sensitivity to different conditions such as the rate of CO_2 emissions or changes in the planet's albedo). Each of these studies involves hundreds to thousands of individual simulations. This is only practical if each simulation in the ensemble takes a modest amount of time. These are large-scale, capacity computing problems (i.e., ones requiring ongoing access to multiple, large-scale computing systems).

Third, understanding the sensitivity of physical and biogeochemical processes to social, behavioral and economic policies requires evaluation of statistical ensembles and many model variants. These are hypothesis-driven computational scenarios that are only possible after the physical and biogeochemical processes are understood, requiring additional capacity and capability computing.

This is a daunting problem – developing, validating and evaluating multidisciplinary climate models in time to provide the necessary answers to critical questions:

- *How many simulation scenarios are necessary (minimally and optimally)*
- *What model elements are needed for each scenario?*
- *What temporal and spatial resolution, along with physical models, is affordable?*
- *What are the errors and uncertainties in model predictions?*
- *When must research end and production simulation begin to produce policy guidance?*

Underlying these questions is the need for powerful computers to model climate change at regional and fine scales, and to support the sophisticated and computationally expensive algorithms needed to represent the complexities of both natural and human effects. **We must also manage the tsunami of observational data now being captured via a new generation of environmental sensors, integrating high-resolution Earth system models with assimilated satellite and other data, supported by large data archives and intelligent data mining and management systems.**

Finally, we must develop the multiphysics algorithms and models needed to represent the complex interactions of biological, geophysical, chemical and human activities. New scientific and mathematical advances will also be required to quantify model uncertainty for such complex systems. This fusion of sensor data with complex models is large-scale computational science in its clearest and most compelling form. Equally importantly, those HPC systems must be available for researcher use.

2. High-Performance Computing Resource Availability

In the early 1980s, HPC facilities were accessible only by a handful of U.S. researchers. Most access required both a national security clearance and partnership with one of the U.S. weapons laboratories or international travel – for access to computing research facilities outside the U.S. **The rising importance of computing to science and the dearth of HPC facilities for open scientific research stimulated creation of the National Science Foundation (NSF) supercomputing centers and similar facility investments by the Department of Energy's (DOE) Office of Science.** Although other agencies also support HPC facilities, NSF and DOE now provide the overwhelming fraction of the unclassified resources for computational science, including climate change.

This NSF program and its descendents, the Partnerships for Advanced Computational Infrastructure (PACI) and the TeraGrid, continue to support academic researchers via consulting, HPC systems and archival storage. All of the NSF-supported resources, with the exception of the majority at the National Center for Atmospheric Research (NCAR), are allocated by peer review across all disciplines. The

computing facilities at NCAR include peer-reviewed resources allocated for weather and climate research and the Climate Simulation laboratory (CSL) resources dedicated to climate change research. Historically, all NSF computing resources have been substantially over-subscribed, with unmet demand from academic researchers. Recently, however, NSF has funded a series of competitive hardware acquisitions to help address this shortfall, with the largest slated to sustain one petaflop[3] on selected applications.[4]

The DOE Office of Science also maintains a set of unclassified computing facilities, anchored by the National Energy Research Scientific Computing Center (NERSC), two leadership-class computing systems at Oak Ridge and Argonne National Laboratories, and a smaller facility at the Pacific Northwest Research Laboratory. The majority of DOE's NERSC resources are also allocated by peer review, with the requirement that the proposed use be relevant to the DOE Office of Science mission. Finally, the DOE leadership-class facilities target focused projects that could benefit from access to the largest-scale facilities in the country, including the climate change modeling program. Most of these resources are allocated by the INCITE initiative.[5]

Our computational science infrastructure is enormously greater than twenty years ago. However, so are our expectations and needs – science and computing are now synonymous. Equally tellingly, because almost all of our NSF and DOE HPC resources are shared across disciplines, only a modest fraction of these systems is dedicated to climate change studies. Rather, researchers rely on a combination of proposal peer review and programmatic resource allocation to conduct climate change studies on a diverse array of HPC systems.

At present, there is no truly large scale U.S. climate change computing research facility, architected, configured and dedicated to multidisciplinary climate change studies that can deliver timely and accurate predictions. A recent DOE study estimated that climate and environmental modeling could use an exascale system effectively (i.e., one thousand times faster than any extant computer system). Simply put, change modeling is a deep and challenging scientific problem that requires computing infrastructure at the largest scale.

3. Computing Evolution: Lessons and Challenges

In the late 1970s and the 1980s, HPC was defined by vector processors, as exemplified by the eponymously named systems designed by the legendary Seymour Cray. These systems combined high-speed, custom processor design with fast memories and innovative packaging. Researchers and software developers were able to tune selected portions of their codes to the vector hardware, achieving unprecedented performance with modest effort.

With the birth of the PC, a new approach to HPC began to emerge in the 1980s. The increasing performance and low cost of commodity microprocessors – the "Attack of the Killer Micros" – transformed HPC. This new model of massive parallelism partitions computations across large numbers of processors. Via this approach, one can increase peak hardware performance to levels limited only my economics and reliability. However, achieving high performance on complex applications is more problematic and challenging, particularly for multidisciplinary applications. The climate change community expressed great concern about this disruptive technology transition during the 1990s, with concomitant political controversy.

---

[3] One teraflop is 10^{12} floating point operations/second; one petaflop is one thousand teraflops, or 10^{15} floating point operations/second; one exaflop is one thousand petaflops, or 10^{18} floating point operations/second.

[4] See the NSF Office of Cyberinfrastructure, http://www.nsf.gov/dir/index.jsp?org=OCI for details on the NSF acquisition program.

[5] Department of Energy Innovative and Novel Computational Impact on Theory and Experiment (INCITE) initiative, http://hpc.science.doe.gov/

Recognizing this technological shift, the associated challenges and the opportunities, the Defense Advanced Research Projects Agency (DARPA) launched an aggressive research and development program that engaged academia, industry and national laboratories. Other federal agencies, notably the National Science Foundation (NSF), the Department of Energy's (DOE) Office of Science and the National Aeronautics and Space Administration (NASA), joined in the High-Performance Computing and Communications (HPCC) program.[6]

In the 1990s, research flourished in computer architecture, system software, programming models, algorithms and applications. Computer vendors launched new initiatives, and parallel computing startup companies were born. Planning began for petascale systems, based on integrated hardware, architecture, software and algorithms research. **After a promising start, much of the initiative faded and attention shifted elsewhere.** The most notable exception was DOE's National Nuclear Security Administration (NNSA). Needing to certify the weapons stockpile without testing, NNSA embraced HPC to verify and validate weapon safety and readiness. The complex physics drove new algorithm and software development and acquisition of some of the world's most power computing systems, all based on massive parallelism and commodity microprocessors.

While the U.S. computing industry largely abandoned purpose-built supercomputers in favor of commodity designs, Japanese vendors, notably Hitachi and Fujitsu, continued to develop and evolve vector supercomputers. In 2002, Japan announced the Earth Simulator – then the world's fastest computer. The Earth Simulator was designed specifically for large-scale climate and weather studies and drew on many years of vector computing research and development.

Although the Japanese plan had long been public, it precipitated considerable concern. The interagency High-End Computing Revitalization Task Force (HECRTF) was chartered to assess the competitive position of the United States. I was privileged to chair the 2003 HECRTF community workshop and edited the associated community report.[7] The federal agencies produced a complementary report and a proposed action plan. Several agencies launched new programs, of which the largest and most visible were the NSF OCI petascale initiative and the DOE Office of Science's Scientific Discovery through Scientific Computing (SciDAC)[8] and INCITE programs.

Today, the majority of the world's largest HPC systems, dominated by U.S. laboratory and academic holdings, remain based on commodity building blocks and community-developed software. In this high-performance "monoculture," vendor profit margins are small, and competition for sales is intense, with limited vendor opportunity to recover research and development investments in alternative architectures. Equally worrisome, the pool of academic researchers in HPC and computational science is small, and research funding is limited.

Without doubt, the explosive growth of scientific computing based on clusters of commodity microprocessors has reshaped the HPC market. The U.S. remains the undisputed world leader in this space. Petascale systems are being deployed by NSF and DOE for academic and laboratory research, and feasibility assessments of exascale systems[9] are underway. Although this democratization of HPC has had

---

[6] The High-Performance Computing and Communications (HPCC) program became the Networking and IT Research and Development (NITRD) program, http://www.nitrd.gov/about/about_NITRD.html

[7] The documents for the High-End Computing Revitalization Task Force (HECRTF), including the community workshop report, can be found at http://www.nitrd.gov/subcommittee/hec/hecrtf-outreach

[8] Department of Energy, Scientific Discovery through Scientific Computing (SciDAC), http://www.scidac.gov/

[9] *Modeling and Simulation at the Exascale for Energy and the Environment*, Summer 2007, http://www.sc.doe.gov/ascr/ProgramDocuments/TownHall.pdf

many salutatory effects, including broad access to commodity clusters across laboratories and universities, it is not without its negatives.

Not all aspects of climate change models map efficiently to the cluster programming model of loosely coupled, message-based communication. It is also unclear if we have the resources needed to address the climate change problem at appropriate scale and in a timely manner, particularly given dramatic changes now underway in computing technology.

4. The Brave New World: Multicore and Massive Data

Over the past twenty years, computational science and HPC have exploited the ever-increasing performance of commodity microprocessors. Each new processor generation combined greater transistor density, new architectural techniques and higher chip power to deliver greater single processor performance. This tripartite evolution is now over. Although transistor densities on chip will continue to rise, physics and power constraints make it impractical to increase clock frequencies further. Future chip performance increases will depend on explicit parallelism and architectural innovations. No longer will current software execute faster in the future without change. Parallelism is now required, even at the chip level, to deliver greater performance.

This multicore revolution – the placement of multiple, slower processors on each chip – poses major new challenges for the computing industry. It is just as disruptive as the transition from vector to parallel computing was fifteen years ago. Today's quad-core chips will soon be replaced by chips containing tens, then hundreds and perhaps thousands of cores (processors). **The technical challenges are daunting, and we have no straightforward technical solutions that will hide this complexity from software developers.**[10] This will profoundly affect the software industry and scientific researchers.

For multicore chips, new programming models and tools are needed to develop parallel applications, and existing software must be retrofitted. New chip architectures are needed to exploit rising transistor densities, support parallel execution and enable heterogeneous processing. New memory technologies and interconnects are needed to support chips with tens to hundreds of cores. Equally importantly, new algorithms are needed that map efficiently to these new architectures. All of these changes will affect parallel climate models now being developed and executed on clustered commodity systems. **Today, we are suffering some of the delayed consequences of limited research investment in parallel computing – architecture, system software, programming tools, data management and algorithms.**

In addition to dramatic changes in processors and computation, our models of data capture and management are in flux. We can now generate, transmit, and store data at rates and scales unprecedented in human history. Many of our new environmental instruments can routinely produce many tens to hundreds of petabytes of data annually. **The scientific data deluge threatens to overwhelm the capacity of our federal institutions to manage, preserve and process and of our climate modeling researchers to access and integrate the data with multidisciplinary models.** This data integration is critical to climate model validation.

Although industry is developing massive data centers to host Internet search, social networks and software as a service, our research data infrastructure has not kept pace. Climate researchers need better data management tools, including provenance tracking, translation, mining, fusion, visualization, and analysis. We must not focus exclusively on computing, but on the fusion of sensors and data management with computing hardware and rich climate models.

[10] This realization recently motivated Microsoft and Intel to invest $20M in academic multicore research at the University of Illinois at Urbana-Champaign and the University of California at Berkeley, http://www.microsoft.com/presspass/press/2008/mar08/03-18UPCRCPR.mspx

5. Actions: A Sustainable, Integrated Approach

One can and must draw several important, salutary lessons from the changing nature of computing technology. The U.S. HPC industry is now driven by business and consumer technology economics, with concomitant advantages and disadvantages. Large product volumes and amortized research and development costs lead to rapid innovation and technological change. However, those same consumer economics mean that today's HPC systems are built from commodity hardware and software components, and they are often ill-suited to the numerically and communication intensive nature of climate change models. In consequence, they rarely deliver a large fraction of their advertized peak performance.

Given their unique attributes, the highest capability computing systems have a very limited commercial market. The high non-recurring engineering costs to design HPC systems matched to scientific and government needs are not repaid by sales in the commercial marketplace. Hence, we must rethink our models for research, development, procurement and operation of high-end systems. We must target exploration of new systems that better support the needs of scientific and national defense applications and sustain the federal investment needed to design, develop and procure those systems. Today's approach is unlikely to provide the necessary resources to address the climate change model problem fully.

New programming models and tools are also needed that simplify application development and maintenance and that target emerging multicore processors. Today, almost all parallel scientific applications are developed using low-level message-passing libraries. Climate modeling teams must have deep knowledge of application software behavior and its interaction with the underlying computing hardware, and they often spend inordinate amounts of time tailoring algorithms and software to hardware and software idiosyncrasies, time more profitably spent on science and engineering research.

Climate change analysis requires large-scale data archives, connections to scientific instruments and collaboration infrastructure to couple distributed scientific groups. **Any investment in HPC facilities must be balanced with appropriate investments in hardware, software, storage, algorithms and collaboration environments.** Simply put, climate change modeling, as with all scientific discovery, requires a judicious match of computer architecture, system software, algorithms and software development tools.

These facts illustrate the importance of a long-term, integrated research and development program that considers the entire computational science ecosystem, something I advocated as chair and co-chair of two recent PITAC and PCAST subcommittees, respectively. **Both the 2005 President's IT Advisory Committee (PITAC) report on computational science and the 2007 President's Council of Advisors on Science and Technology (PCAST) review of the Networking and Information Technology Research and Development (NITRD) program recommended creation of an interagency strategic roadmap for computational science and computing research.** In particular, the 2005 PITAC report found that

> *The continued health of this dynamic computational science "ecosystem" demands long-term planning, participation, and collaboration by Federal R&D agencies and computational scientists in academia and industry. Instead, today's Federal investments remain short-term in scope, with limited strategic planning and a paucity of cooperation across disciplines and agencies.*

The report also recommended creation of a long-term, interagency roadmap to

> *… address not only computing system hardware, networking, software, data acquisition and storage, and visualization, but also science, engineering, and humanities algorithms*

and applications. The roadmap must identify and prioritize the difficult technical problems and establish a timeline and milestones for successfully addressing them.

In that same spirit, the 2007 PCAST review of the NITRD program, *Leadership Under Challenge: Information Technology R&D in a Competitive World,*[11] which I co-chaired, reiterated the need for a strategic plan and roadmap for high-performance computing and noted that

The Federal NIT R&D portfolio is currently imbalanced in favor of low-risk projects; too many are small-scale and short-term efforts. The number of large-scale, multidisciplinary activities with long time horizons is limited and visionary projects are few.

Based on these studies, I believe we face both great opportunities and great challenges in high-end computing for climate change. Computational science truly is the "third pillar" of the scientific process. The challenges are for us to sustain the research, development and deployment of the high-end computing infrastructure needed to enable discoveries and to ensure the health of our planet.

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

Biographical Sketch

Daniel A. Reed is Director of Scalable and Multicore Computing Strategy at Microsoft. Previously, he was the Chancellor's Eminent Professor at the University of North Carolina at Chapel Hill, as well as the Director of the Renaissance Computing Institute (RENCI), which explored the interactions of computing technology with the sciences, arts and humanities. He is a former Director of the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign, where he also led National Computational Science Alliance, a consortium of roughly fifty academic institutions and national laboratories to develop next-generation software infrastructure of scientific computing. He was also one of the principal investigators and chief architect for the NSF TeraGrid.

Dr. Reed is a member of President Bush's Council of Advisors on Science and Technology (PCAST) and a former member of the President's Information Technology Advisory Committee (PITAC). He recently chaired a review of the federal networking and IT research (NITRD) portfolio, and he is chair of the board of directors of the Computing Research Association (CRA), which represents the research interests of universities, government laboratories and industry. He received his PhD in computer science in 1983 from Purdue University.

[11] *Leadership Under Challenge: Information Technology R&D in a Competitive World*, President's Council of Advisors on Science and Technology (PCAST), August 2007, http://www.ostp.gov/pdf/nitrd_review.pdf