Title

How Are We Doing? A Self-Assessment of the Quality of Services and Systems at NERSC, 2005-2006

Authors

Kramer, William T.C.
Hules, John

Publication Date

2007-03-13

LBNL-62117

How Are We Doing?

A Self-Assessment of the Quality of Services and Systems at NERSC, 2005–2006

William T.C. Kramer and John Hules

National Energy Research Scientific Computing Center Division
Ernest Orlando Lawrence Berkeley National Laboratory
Berkeley, CA 94720

March 13, 2007

This work was supported by the U.S. Department of Energy's Office of Science, Office of Advanced Scientific Computing Research, Mathematical, Information and Computational Sciences Division under Contract No. DE-AC02-05CH11231.

CONTENTS

EXECUTIVE SUMMARY 1

INTRODUCTION — HIGHLIGHTS OF 2005–2006..2

1. RELIABLE AND TIMELY SERVICE

   For the systems NERSC provides, service will be assessed regarding availability, mean time between interruptions and mean time to repair for computational and storage systems after six months of a system going into full service. In addition, NERSC must be timely in responding to user problem reports and issues 7

2. CLIENT SUPPORT GOALS

   The end measure of a facility is how much productive scientific work users accomplish. Sites must assist users in being as productive as possible by providing systems, tools, information, consulting services and training. The objective is to understand codes and how they are used, and target bottlenecks for elimination or minimization..10

3. NEVER BE A BOTTLENECK TO MOVING NEW TECHNOLOGY INTO SERVICE

   NERSC is a primary vehicle for achieving the SC goal of making leading-edge technology available to its scientists. To do this, NERSC continually evaluates, tests, integrates and supports early systems and software. Therefore, NERSC must help ensure future high-performance technologies are available to Office of Science computational scientists in a timely way 16

4. ENSURE ALL NEW TECHNOLOGY AND CHANGES IMPROVE (OR AT LEAST DO NOT DIMINISH) SERVICE TO OUR CLIENTS.

   In striving to provide users with the latest systems for computational sciences, NERSC has the responsibility to ensure system changes have a maximum benefit and minimal detrimental impact on the clients' ability to do work 21

5. DEVELOP INNOVATIVE APPROACHES TO HELP THE CLIENT COMMUNITY EFFECTIVELY USE NERSC SYSTEMS

   NERSC must assist our clients in being as productive as possible by providing systems, enhancements, tools, information, training, consulting and other assistance. In addition to the traditional approaches that are effective, NERSC will constantly try new approaches to help make our clients effective in an ever-more-changing environment. NERSC will help design strategies and integrate and develop technology to enable our clients to improve their use of our systems and to more effectively accomplish their science 24

6. DEVELOP AND IMPLEMENT WAYS TO TRANSFER RESEARCH PRODUCTS AND KNOWLEDGE INTO PRODUCTION SYSTEMS AT NERSC AND ELSEWHERE.

   NERSC is uniquely placed to establish methods and procedures that enable research products and knowledge, particularly those developed at LBNL/UC, to smoothly flow into production..31

7. IMPROVE METHODS OF MANAGING SYSTEMS WITHIN NERSC AND LBNL AND BE A LEADER IN LARGE-SCALE SYSTEMS MANAGEMENT AND SERVICES.

As the Department of Energy's flagship unclassified scientific computing facility, NERSC continually provides leadership and helps shape the field of high performance computing. As HPC technology evolves at an increasing rate, it is crucial that NERSC and LBNL remain at the forefront of getting the most out of these systems........................32

8. EXPORT KNOWLEDGE, EXPERIENCE AND TECHNOLOGY DEVELOPED AT NERSC, PARTICULARLY TO AND WITHIN NERSC CLIENT SITES.

In order for NERSC to be a leader in large-scale computing, NERSC must export experience, knowledge, and technology. Transfers must be made to other client sites, supercomputer sites, and industry..35

9. NERSC WILL BE ABLE TO THRIVE AND IMPROVE IN AN ENVIRONMENT WHERE CHANGE IS THE NORM.

High-performance organizations that deal with advanced technology must be able to adapt and embrace change as a way of life. HPC centers that are not growing and changing are dying (or have died). Providing reliable cycles is not enough to serve the NERSC users in a time of constant change. Research is needed to ensure that tomorrow's systems are accessible and productive to our users 44

10. IMPROVE THE EFFECTIVENESS OF NERSC STAFF BY IMPROVING INFRASTRUCTURE, CARING FOR STAFF, ENCOURAGING PROFESSIONALISM AND PROFESSIONAL IMPROVEMENT

Every employee has a stake in the success of NERSC, and management encourages staff to contribute their ideas for helping the organization succeed. NERSC looks for and implements new ways to increase staff effectiveness. This leads to NERSC being able to support more activities and innovations 52

CONCLUSION 54

EXECUTIVE SUMMARY

This is the sixth self-assessment of the systems and services provided by the U.S. Department of Energy's National Energy Research Scientific Computing Center, describing many of the efforts of the NERSC staff to support advanced computing for scientific discovery. The report is organized along the 10 goals set by our staff and outlines how we are working to meet those goals. Our staff applies experience and expertise to provide world-class systems and unparalleled services for NERSC users. At the same time, members of our organization are leading contributors to advancing the field of high-performance computing through conference presentations, published papers, collaborations with scientific researchers and through regular meetings with members of similar institutions. In the fast-moving realm of high-performance computing, adopting the latest technology while reliably delivering critical resources can be a challenge, but we believe that this self-assessment demonstrates that NERSC continues to excel on both counts.

INTRODUCTION — HIGHLIGHTS OF 2005–2006

In 2005, NERSC set the stage for the next five years of its development through a series of important planning activities. Through the well-established Greenbook process, our active user community provided their input to the planning process. NERSC management then developed a new five-year plan for 2006 to 2010, which was then thoroughly reviewed in a programmatic review by DOE. The reviewers fully endorsed our plans, stating in part: “NERSC is a strong, productive, and responsive science-driven center that possesses the potential to significantly and positively impact scientific progress.… NERSC is extremely well run with a lean and knowledgeable staff.”

NERSC added significant capacity in 2005 by introducing two new clusters, named “Jacquard” and “Bassi.” This increase in capacity was highly welcomed by the NERSC community. Our users make the most progress in computational science with exactly this type of system that provides high performance in a reliable, predictable environment. One of our users, Robert Duke of the University of North Carolina, commented on the two systems as follows: *“I have to say that both of these machines are really nothing short of fabulous. While Jacquard is perhaps the best-performing commodity cluster I have seen, Bassi is the best machine I have seen, period.”* By delivering this quality, NERSC makes it possible for its users to concentrate on their science. Not surprisingly, our user community enjoys continued and unparalleled scientific productivity. In 2005 we were able to list more than 1200 scientific publications that were written on the basis of simulations carried out at NERSC.

Another milestone was reached in 2006 with the announcement that Cray had won a $52 million contract to deliver a supercomputer that will deliver over 16 teraflop/s of sustained performance. A successor to the Cray XT3 supercomputer, this new XT4 system, which NERSC has named “Franklin,” will be among the world’s fastest general-purpose systems. The system uses thousands of AMD Opteron processors running tuned, lightweight operating system kernels and interfaced to Cray’s unique SeaStar network. Installation of Franklin is expected to be completed in the first half of 2007, with acceptance in mid-2007.

We also saw a long-term goal accomplished that will significantly set apart the NERSC production environment from other centers. In 2005, NERSC deployed the NERSC Global Filesystem (NGF) into production, providing seamless data access from all of the Center’s computational and analysis resources. NGF is intended to facilitate sharing of data between users and between machines. NGF’s single unified namespace makes it easier for users to manage their data across multiple systems. Users no longer need to keep track of multiple copies of programs and data, and they no longer need to copy data between NERSC systems for pre- and post-processing. NGF provides several other benefits as well: storage utilization is more efficient because of decreased fragmentation, and computational resource utilization is more efficient because users can more easily run jobs on an appropriate resource. In early 2006, NERSC upgraded NGF to over 70 TB of user-accessible storage.

Statistical Snapshots of NERSC Users

Meeting the computational science needs of the DOE Office of Science encompasses a broad range of researchers in terms of scientific disciplines, geographic location or home institution. Here are some statistics on the NERSC user community.

NERSC served 2,677 scientists throughout the United States in 2005 and 2,978 scientists in 2006. These researchers work in DOE laboratories, universities, industry, and other Federal agencies. Figures 1 and 2 shows the proportion of NERSC usage on massively parallel processing (MPP) systems by each type of institution for 2005 and 2006. Figures 3, 4, 5, and 6 show laboratory, university, and other organizations that used large allocations of computer time in both years. Computational science conducted at NERSC covers the entire range of scientific disciplines, but is focused on research that supports the DOE's mission and scientific goals, as shown in Figures 7 and 8.

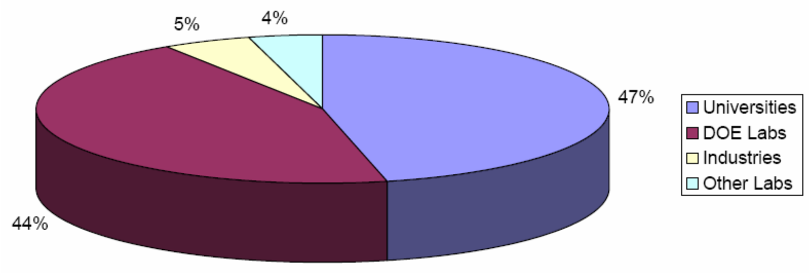

Figure 1. NERSC MPP usage by institution type, 2005.

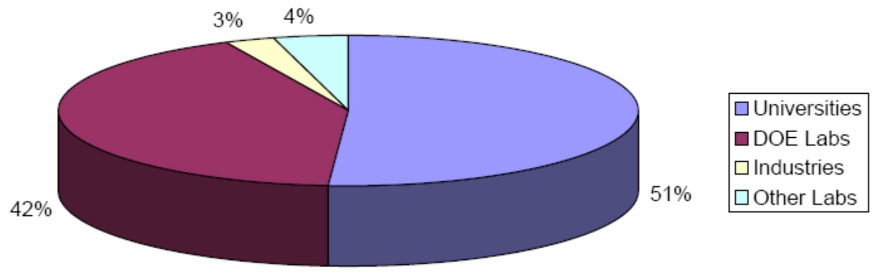

Figure 2. NERSC MPP usage by institution type, 2006.

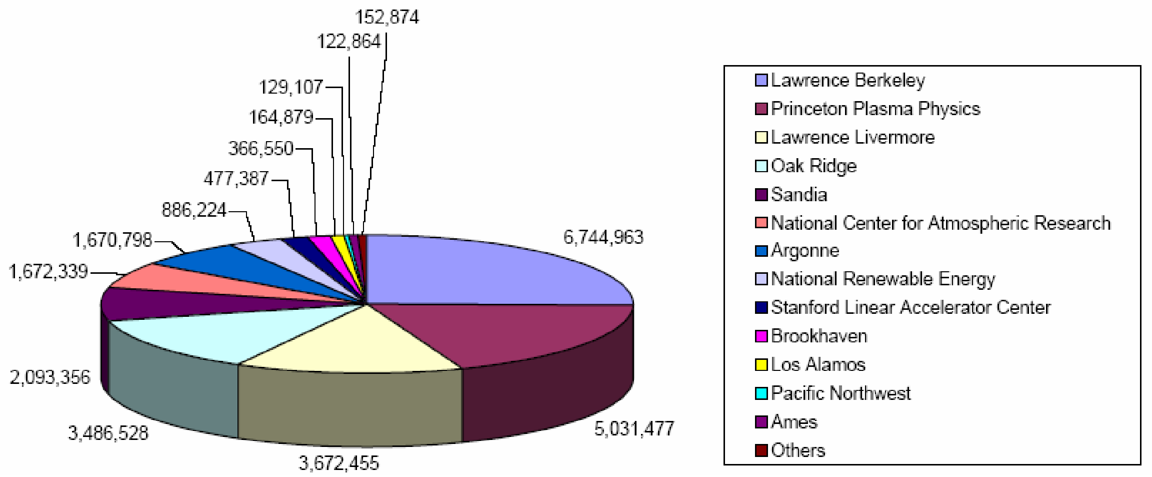

Figure 3. DOE and other Federal laboratory usage at NERSC, 2005 (MPP hours).

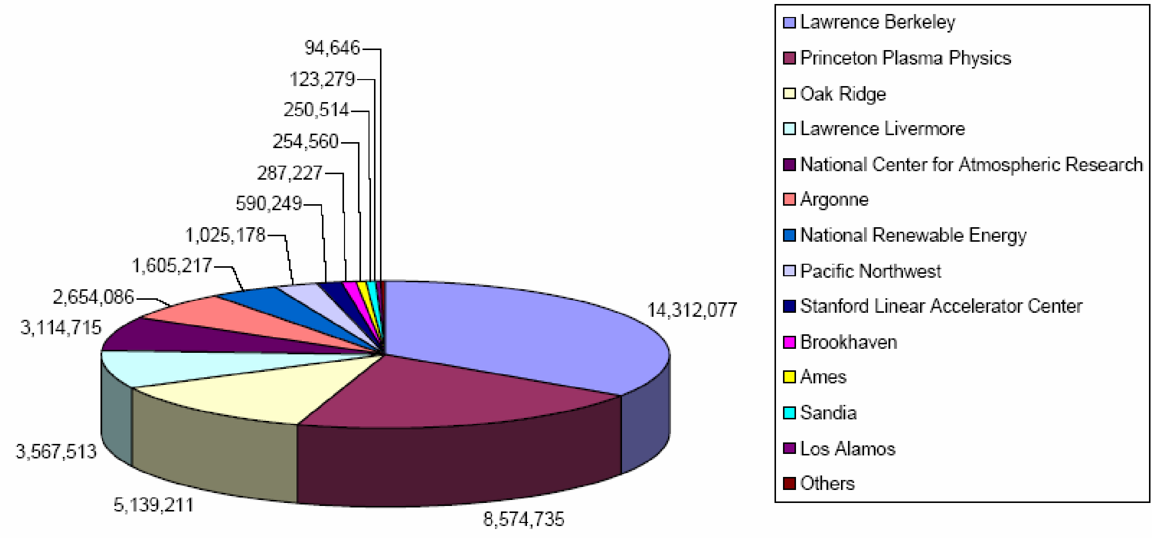

Figure 4. DOE and other Federal laboratory usage at NERSC, 2006 (MPP hours).

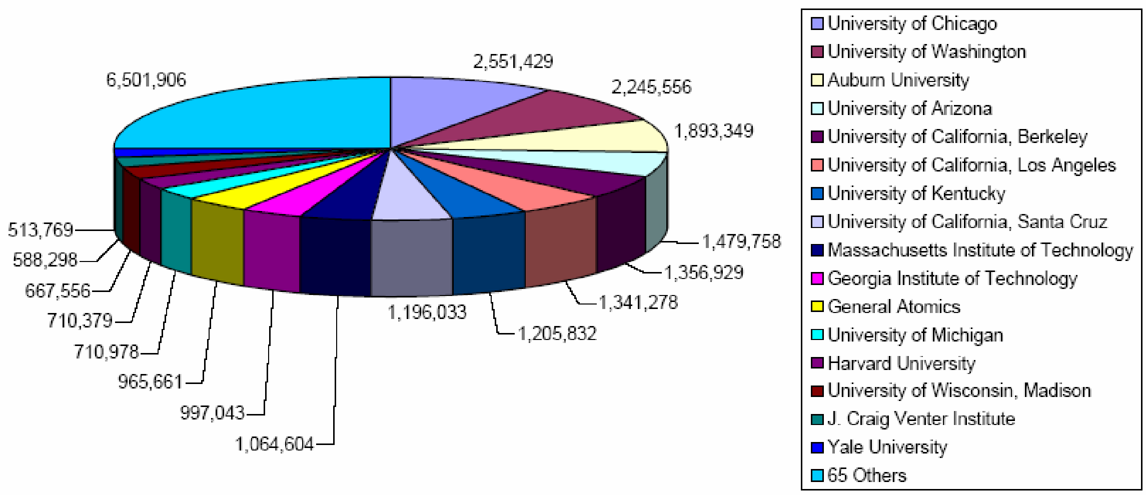

Figure 5. Academic and private laboratory usage at NERSC, 2005 (MPP hours).

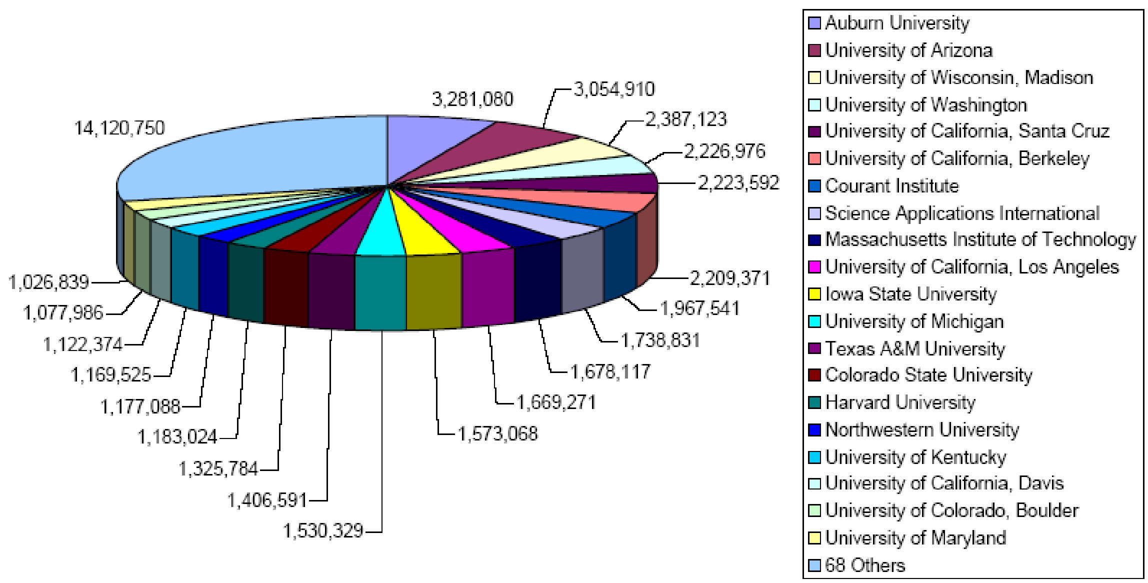

Figure 6. Academic and private laboratory usage at NERSC, 2006 (MPP hours).

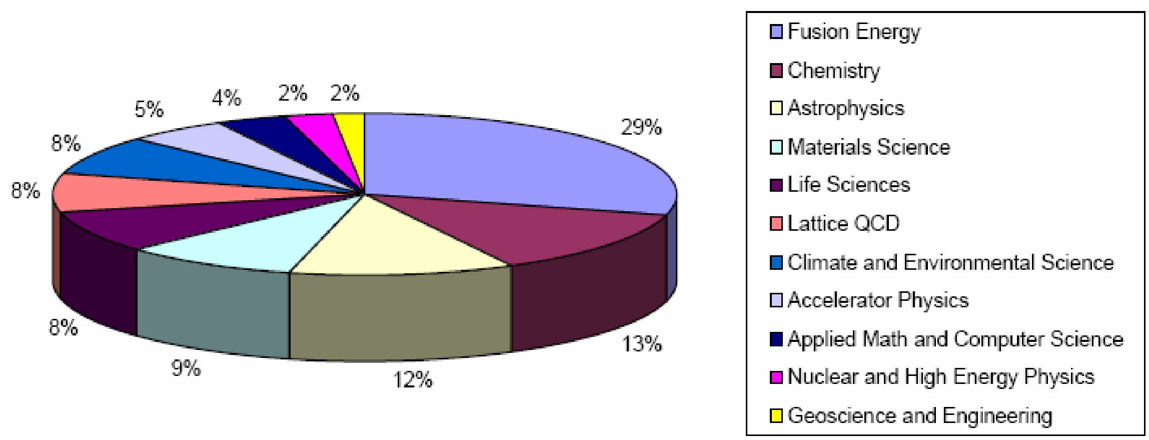

Figure 7. NERSC usage by scientific discipline, 2005.

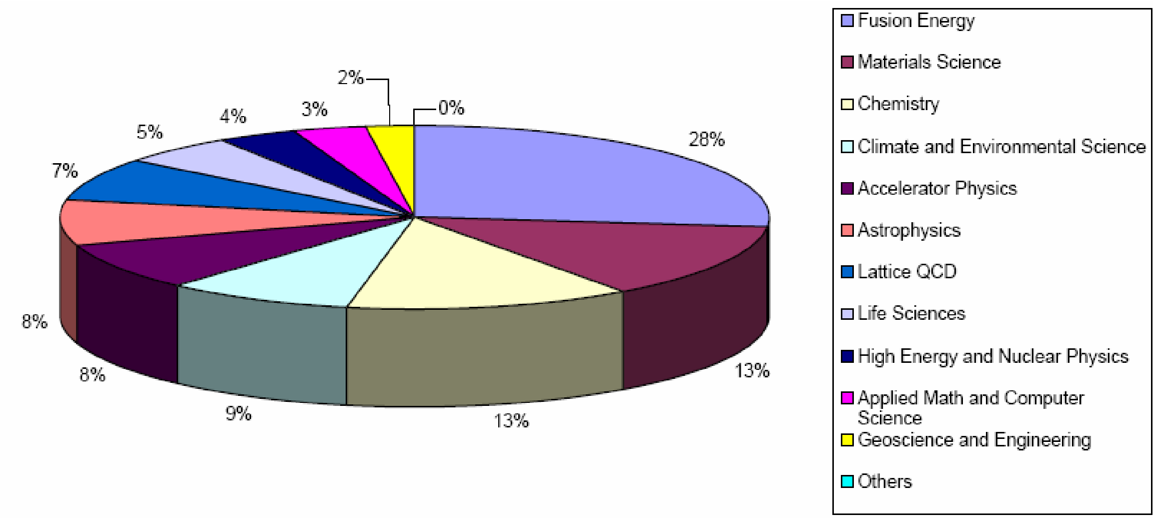

Figure 8. NERSC usage by scientific discipline, 2006.

1. Reliable and Timely Service

For the systems NERSC provides, service will be assessed regarding availability, mean time between interruptions and mean time to repair for computational and storage systems after six months of a system going into full service. In addition, NERSC must be timely in responding to user problem reports and issues.

For the systems NERSC provides, service is assessed regarding availability, mean time between interruptions and mean time to repair computational and storage systems after six months of a system going into full service.

NERSC strives to provide reliable service to all of our clients. Our efforts address two general areas:

- How reliably our systems operate (i.e., availability to clients); and
- How responsive we are to clients when they have a problem.

To meet our goals, various groups within NERSC organization must work together to provide users with both the high-performance computing systems and the expert services for achieving research goals. To achieve this, NERSC takes a two-pronged approach. First, the NERSC staff is continually seeking out new techniques and technologies to anticipate and meet users' needs. Second, when a problem arises, we respond promptly to acknowledge, address and correct it. One user who responded to our 2005 user survey summed up the results of our efforts by saying, "I run at NERSC's Seaborg because the performance is the most consistent and reliable of any of the facilities at which I compute. I never have trouble running on even 1000+ processors."

NERSC strives to provide users with the maximum availability of our resources, not just in terms of scheduled availability, but in terms of overall availability. After all, if a system isn't available and a job can't run, it doesn't matter to the user whether it's a scheduled outage or unanticipated downtime. Here are the system metrics definitions we use in evaluating our performance.

- Scheduled availability is the percentage of time a system is available for users, accounting for any scheduled downtime for maintenance and upgrades.

$$\frac{\Sigma \text{ scheduled hours} - \Sigma \text{ outages during scheduled time}}{\Sigma \text{ scheduled hours}}$$

- Overall availability is the percentage of time a system is running.

$$\frac{\Sigma \text{ available hours} - \Sigma \text{ (unscheduled outages + scheduled downtime)}}{\Sigma \text{ available hours}}$$

- A failure is any event (hardware, software, human, environment) that disrupts full service to the client base.
- Any partial degradation of committed services levels (e.g., dropping below the promised number of compute nodes on a system) is treated, for the sake of these goals, as a complete failure.
- Any shutdown that has less than 24 hours notice is treated as an unscheduled interruption.

- A service outage is the time from when computational processing halts to the restoration of computation (e.g., not when the system was booted, but rather when user jobs are recovered and restarted).
- If an outage occurs within two hours of the system that does not have checkpoint/restart being restored to service, it is treated as one continuous outage.

Tables 1 and 2 show how NERSC is achieving its system availability goals.

Table 1
System Availability Metrics for FY05

Systems	Scheduled Availability	Overall Availability	Mean Time between Failures (Day:Hr:Min)	Mean Time to Restoration (Hours)
MPP	98.54%	97.27%	10:18:12	7.6
Storage	99.55%	98.15%	8:12:35	3.0
File Servers	100.00%	100.00%		
Math/Vis Servers	99.88%	99.77%	29:16:36	3.0

Table 2
System Availability Metrics for FY06

Systems	Scheduled Availability	Overall Availability	Mean Time between Failures (Day:Hr:Min)	Mean Time to Restoration (Hours)
MPP	98.70%	98.30%	16:02:03	7.1
Storage	99.85%	98.71%	6:08:33	2.2
File Servers	No longer reported because redundant servers provide 100% uptime.			
Math/Vis Servers	99.07%	98.35%	11:19:41	5.1

Tables 3 and 4 show how quickly users' problems are resolved, specifically, the number of days between when a trouble ticket is opened and closed. (Note that trouble tickets include internal issues that may be left open simply to track a problem.)

Table 3
Problem Resolution Metrics for FY05
(3,590 total incidents)

Number of days to resolve problem	Number of problems resolved	Percentage of problems resolved	Cumulative percentage of problems resolved
1	2342	65.24	65.24
2	231	6.43	71.67
3	151	4.21	75.88
4	79	2.20	78.08
5	82	2.28	80.36
6	67	1.87	82.23
7	60	1.67	83.90
8	53	1.48	85.38
9	29	0.81	86.18
10	28	0.78	86.96
...			
15	62	1.73	90.58
...			
30	3	0.08	95.15
...			
180	47	1.31	100.00

Table 4
Problem Resolution Metrics for FY06
(3,971 total incidents)

Number of days to resolve problem	Number of problems resolved	Percentage of problems resolved	Cumulative percentage of problems resolved
1	2540	63.96	63.96
2	233	5.87	69.83
3	164	4.13	73.96
4	62	1.56	75.52
5	66	1.66	77.18
6	62	1.56	78.75
7	42	1.06	79.80
8	50	1.26	81.06
9	28	0.71	81.77
10	30	0.76	82.52
...			
15	106	2.67	87.53
...			
30	7	0.18	93.55
...			
180	25	0.63	100.00

2. Client Support Goals

The end measure of a facility is how much productive scientific work users accomplish. Sites must assist users in being as productive as possible by providing systems, tools, information, consulting services and training. The objective is to understand codes and how they are used, and target bottlenecks for elimination or minimization.

Support for Large-Scale Projects

NERSC works directly with scientists on major projects that require extensive scientific computing capabilities, such as the SciDAC and INCITE collaborations. These projects are often characterized by large collaborations, the development of community codes, and the involvement of computer scientists and applied mathematicians. In addition to high-end computing, these large projects handle issues in data management, data analysis, and data visualization, as well as automation features for resource management.

NERSC provides its highest level of support to these researchers, including special service coordination for queues, throughput, increased limits, etc.; and specialized consulting support, which may include algorithmic code restructuring to increase performance, I/O optimization, visualization support—whatever it takes to make the computation scientifically productive. The three INCITE projects for 2005 are good examples of this kind of support.

The INCITE project “Magneto-Rotational Instability and Turbulent Angular Momentum Transport,” led by Fausto Cattaneo of the University of Chicago, had the goal of understanding the forces that help newly born stars and black holes increase in size by simulating laboratory experiments that study magnetically caused instability. “With the help of NERSC staff, we were able to tune our software for Seaborg’s hardware and realize performance improvements that made additional simulations possible,” Catteneo said. NERSC also provided crucial help in creating animated visualizations of the simulation results, which involve the formation of complex, three-dimensional structures that need to be seen to be understood.

For the INCITE project “Direct Numerical Simulation of Turbulent Nonpremixed Combustion,” Jacqueline Chen, Evatt Hawkes, and Ramanan Sankaran of Sandia National Laboratories have performed the first 3D direct numerical simulations of a turbulent nonpremixed flame with detailed chemistry. After analyzing and optimizing the code’s performance with the help of NERSC staff, the researchers improved the code’s efficiency by 45%. The simulations generated 10 TB of raw data, and NERSC consultants helped the researchers figure out the best strategy for efficiently transferring all that data from NERSC systems to the researchers’ local cluster. “The assistance we received from the NERSC computing staff in optimizing our code and with terascale data movement has been invaluable,” Chen said. “The INCITE award has enabled us to extend our computations to three dimensions so that we may investigate interactions between turbulence, mixing, and finite-rate detailed chemistry in combustion.”

The “Molecular Dynameomics” INCITE project, led by Valerie Daggett of the University of Washington, is an ambitious attempt to use molecular dynamics simulations to characterize and catalog the folding/unfolding pathways of representative proteins from all known protein folds.

David Beck, a graduate student in the Daggett lab, worked with NERSC consultants to optimize the performance of the group's code on Seaborg. "The INCITE award gave us a unique opportunity to improve the software, as well as do good science," Beck said. Improvements included load balancing, which sped up the code by 20%, and parallel efficiency, which reached 85%. The INCITE award enabled the team to do five times as many simulations as they had previously completed using other computing resources. "We are quite satisfied with our experience at NERSC," Daggett commented.

Archiving Strategies for Genome Researchers

When researchers at the Production Genome Facility of DOE's Joint Genome Institute (JGI) found they were generating data faster than they could find somewhere to store the files, a collaboration with NERSC's Mass Storage Group developed strategies for improving the reliability of data storage while also making retrieval easier.

JGI is one of the world's leading facilities in the scientific quest to unravel the genetic data that make up living things. With advances in automatic sequencing of genomic information, scientists at the JGI's Production Genome Facility (PGF) found themselves overrun with sequence data, as their production capacity had grown so rapidly that data had overflowed their existing storage capacity. Since the resulting data are used by researchers around the world, PGF has to ensure that the data are reliably archived as well as easily retrievable.

As one of the world's largest public DNA sequencing facilities, the PGF produces 2 million files per month of trace data (25 to 100 KB each), 100 assembled projects per month (50 MB to 250 MB), and several very large assembled projects per year (~50 GB). In aggregate, this averages about 2 TB per month.

In addition to the amount of data, a major challenge is the way the data are produced. Data from the sequencing of many different organisms are produced in parallel each day, resulting in a daily archive that spreads the data for a particular organism over many tapes.

DNA sequences are considered the fundamental building blocks for the rapidly expanding field of genomics. Constructing a genomic sequence is an iterative process. The trace fragments are assembled, and then the sequence is refined by comparing it with other sequences to confirm the assembly. Once the sequence is assembled, information about its function is gleaned by comparing and contrasting the sequence with other sequences from both the same organism and other organisms. Current sequencing methods generate a large volume of trace files that have to be managed—typically 100,000 files or more. And to check for errors in the sequence or make detailed comparisons with other sequences, researchers often need to refer back to these traces. Unfortunately, these traces are usually provided as a group of files with no information as to where the traces occur in the sequence, making the researchers' job more difficult.

This problem was compounded by the PGF's lack of sufficient online storage, which made organization (and subsequent retrieval) of the data difficult and led to unnecessary replication of files. This situation required significant staff time to move files and reorganize filesystems to find sufficient space for ongoing production needs; and it required auxiliary tape storage that was not particularly reliable.

Staff from NERSC's Mass Storage Group and the PGF worked together to address two key issues facing the genome researchers. The most immediate goal was for NERSC's HPSS to become the archive for the JGI data, replacing the less-reliable local tape operation and freeing up disk space at the PGF for more immediate production needs. The second goal was to collaborate with JGI to improve the data handling capabilities of the genome sequencing and data distribution processes.

NERSC storage systems are robust and available 24 hours a day, seven days a week, as well as highly scalable and configurable. NERSC has high-quality, high-bandwidth connectivity to the other DOE laboratories and major universities provided by ESnet.

Most of the low-level data produced by the PGF are now routinely archived at NERSC, with ~50 GB of raw trace data being transferred from JGI to NERSC each night.

The techniques used in developing the archiving system allow it to be scaled up over time as the amount of data continues to increase—up to billions of files can be handled with these techniques. The data have been aggregated into larger collections which hold tens of thousands of files in a single file in the NERSC storage system. This data can now be accessed as one large file, or each individual file can be accessed without retrieving the whole aggregate, greatly improving the speed with which a researcher can retrieve the data once it is stored.

Not only will the new techniques be able to handle future data, they also helped when the PGF staff discovered raw data that had been previously processed by software that had an undetected bug. The staff were able to retrieve the raw data from NERSC and reprocess it in about a month and a half, rather than go back to the sequencing machines and produce the data all over again—which would have taken about six months. In addition to saving time, this also saved money—a rough estimate is that the original data collection comprised up to 100,000 files per day at a cost of $1 per file, which added up to $1.2 million for processing six months' worth of data. Comparing this figure to the cost of a month and a half of staff time, the estimated savings are about $1 million—and the end result is a more reliable archive.

A National Priority: Analyzing Hurricane Coastal Surges

"NERSC … has a well-earned reputation for providing highly reliable systems, fast turnaround on critical projects, and dedicated support for users," said Secretary of Energy Samuel Bodman when announcing an allocation of NERSC computer resources to the U.S. Army Corps of Engineers. In 2006, the Office of Science made allocated 800,000 processor hours of supercomputing time at NERSC to the Corps of Engineers for studying ways to improve hurricane defenses along the Gulf Coast. "Because these simulations could literally affect the lives of millions of Americans, we want to ensure that our colleagues in the Corps of Engineers have access to supercomputers which are up to the task," the Secretary stated, giving NERSC credit for its proven record of delivering highly reliable production supercomputing services.

As hurricanes move from the ocean toward land, the force of the storm causes the seawater to rise as it surges inland. The Corps of Engineers used its DOE supercomputer allocations to create revised models for predicting the effects of 100-year storm-surges—the worst-case scenario based on 100 years of hurricane data—along the Gulf Coast (Figures 9 and 10). In particular, simulations were generated for the critical five-parish area of Louisiana surrounding New Orleans and the Lower Mississippi River.

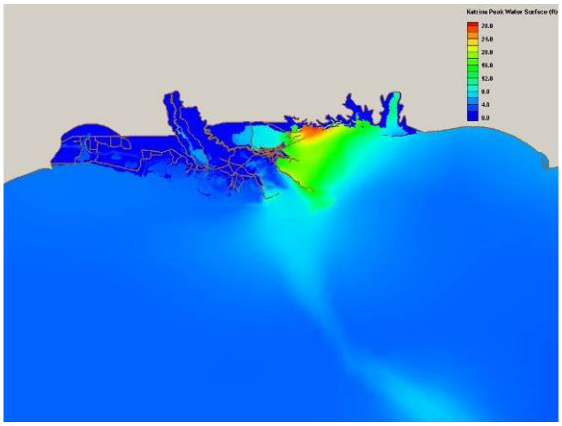

Figure 9. Overview simulation showing elevated storm surges along the Gulf Coast.

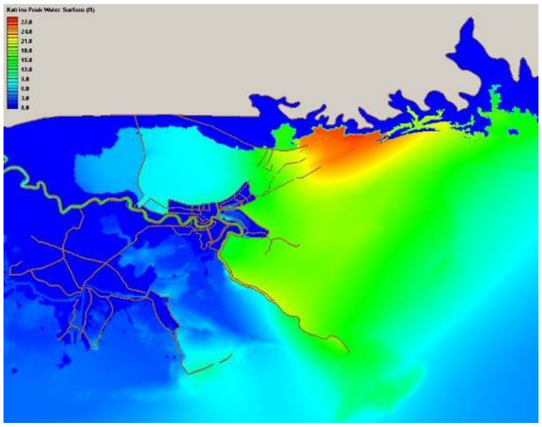

Figure 10. Simulation detail showing highest surge elevation (in red) striking Biloxi, Miss. New Orleans is the dark blue crescent to the lower left of Biloxi.

These revised models of the effects known as “storm-surge elevations” are serving as the basis of design for levee repairs and improvements currently being designed and constructed by the Corps of Engineers in the wake of Hurricane Katrina’s destruction in the New Orleans Metro Area.

Additionally, Gulf Coast Recovery Maps were generated for Southern Louisiana based on FEMA’s revised analysis of the frequency of hurricanes and estimates of the resulting waves. These maps are being used on an advisory basis by communities currently rebuilding from the 2005 storms.

Having access to the NERSC supercomputer allowed the Corps of Engineers to create more detailed models of the effects of Hurricane Rita and other storms along the Texas-Louisiana coasts. Increased detail gave the Corps of Engineers and FEMA more information about the local effects of such storms.

For example, storm surge elevations are greatly influenced by local features such as roads and elevated railroads. Representing these details in the model greatly improves the degree to which computed elevations match observed storm surge high-water marks and allows the Corps to make better recommendations to protect against such surges.

As a result of the runs, the Corps determined that the applications produced incorrect results at topographic boundaries in some instances, and codes were modified to improve the accuracy of the results. For example, the runs at NERSC have improved the Corps’ ability to model the effects of vegetation and land use on storm surges which propagate far inland, as Hurricane Rita did on Sept. 24, 2005.

Changes Resulting from User Survey Feedback

The results from the 2005 and 2006 user surveys show generally high satisfaction with NERSC's systems and support. Areas with the highest user satisfaction in 2005 included account support services, the reliability and uptime of the HPSS mass storage system, and HPC consulting. The largest increases in satisfaction over the 2004 survey included the NERSC CVS server, the Seaborg batch queue structure, PDSF compilers, Seaborg uptime, available computing hardware, and network connectivity.

Areas with the lowest user satisfaction in 2005 included batch wait times on both Seaborg and Jacquard, Seaborg's queue structure, PDSF disk stability, and Jacquard's performance and debugging tools. Only three areas were rated significantly lower in 2005: PDSF overall satisfaction and uptime, and the amount of time taken to resolve consulting issues. The introduction of three major systems during the year combined with a reduction in consulting staff explain the latter.

Eighty-two users in 2005 answered the question "What does NERSC do well?" Forty-seven respondents stated that NERSC gives them access to powerful computing resources without which they could not do their science; 32 mentioned excellent support services and NERSC's responsive staff; 30 pointed to very reliable and well managed hardware; and 11 said "Everything."

Sixty-five users responded to "What should NERSC do differently?" The areas of greatest concern are the interrelated issues of queue turnaround times (24 comments), job scheduling and resource allocation policies (22 comments), and the need for more or different computational resources (17 comments). Users also voiced concerns about data management, software, group accounts, staffing, and allocations.

As in the past, comments from the previous survey led to changes in 2005, including a restructuring of Seaborg's queuing polices, the addition of the new Jacquard and Bassi clusters, the upgrade of ESnet's connectivity to NERSC to 10 gigabits per second, and the installation of additional visualization software.

Areas with the highest user satisfaction in 2006 included the HPSS mass storage system, account and consulting services, DaVinci C/C++ compilers, Jacquard uptime, network performance within the NERSC center, and Bassi Fortran compilers. The largest increases in satisfaction over the 2005 survey were for the Jacquard Linux cluster; Seaborg batch wait times and queue structure; NERSC's available computing hardware; and the NERSC Information Management (NIM) system. Areas with the lowest user satisfaction in 2006 included Seaborg batch wait times; PDSF disk storage, interactive services and performance tools; Bassi and Seaborg visualization software; and analytics facilities.

In 2006, 113 users answered the question “What does NERSC do well?” Eighty-seven respondents stated that NERSC gives them access to powerful computing resources without which they could not do their science; 47 mentioned excellent support services and NERSC’s responsive staff; 27 highlighted good software support or an easy-to-use user environment; and 24 pointed to hardware stability and reliability.

In previous years, the greatest areas of concern were dominated by queue turnaround and job scheduling issues. In 2004, 45 users reported dissatisfaction with queue turnaround times. In 2005 this number dropped to 24, and in 2006 only five users made such comments. NERSC has made many efforts to acquire new hardware, to implement equitable queuing policies across the NERSC machines, and to address queue turnaround times by allocating fewer of the total available cycles. These efforts have clearly paid off. The top three areas of concern in 2006 were job scheduling, more compute cycles, and software issues. Other user survey comments were addressed by improvements to Jacquard’s computing infrastructure and by the deployment of the NERSC Global Filesystem.

The complete survey results can be found at http://www.nersc.gov/news/survey/2005/ and http://www.nersc.gov/news/survey/2006/.

3. Never Be a Bottleneck to Moving New Technology into Service

NERSC is a primary vehicle for achieving the SC goal of making leading-edge technology available to its scientists. To do this, NERSC continually evaluates, tests, integrates and supports early systems and software. Therefore, NERSC must help ensure future high-performance technologies are available to Office of Science computational scientists in a timely way.

Two New Clusters: Jacquard and Bassi

In August 2005 NERSC accepted a 722-processor Linux Networx Evolocity cluster system named “Jacquard” for full production use (Figure 11). The acceptance test included a 14-day availability test, during which a select group of NERSC users were given full access to the Jacquard cluster to thoroughly test the entire system in production operation. Jacquard had a 99 percent availability during the testing while users and scientists ran a variety of codes and jobs on the system.

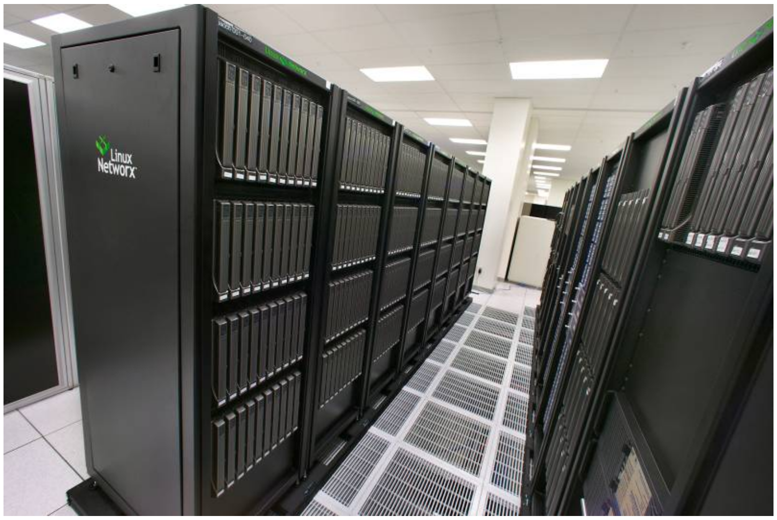

Figure 11. Jacquard is a 722-processor Linux Networx Evolocity cluster system with a theoretical peak performance of 2.8 teraflop/s.

The Jacquard system is one of the largest production InfiniBand-based Linux cluster systems and met rigorous acceptance criteria for performance, reliability, and functionality that are unprecedented for an InfiniBand cluster. Jacquard is the first large system to deploy Mellanox 12x InfiniBand uplinks in its fat-tree interconnect, reducing network hot spots and improving reliability by dramatically reducing the number of cables required.

The system has 712 AMD 2.2 GHz Opteron processors devoted to computation, with the rest used for I/O, interactive work, testing, and interconnect management. Jacquard has a peak performance of 3.1 teraflop/s. Storage from DataDirect Networks provides 30 TB of globally available formatted storage.

Following the tradition at NERSC, the system was named for someone who has had an impact on science or computing. In 1801, Joseph-Marie Jacquard invented the Jacquard loom, which was the first programmable machine. The Jacquard loom used punched cards and a control unit that allowed a skilled user to program detailed patterns on the loom.

In January 2006, NERSC launched a 976-processor IBM cluster named "Bassi" into production use (Figure 12). Earlier, during the acceptance testing, users reported that codes ran from 3 to 10 times faster on Bassi than on NERSC's other IBM supercomputer, Seaborg, leading one tester to call the system the "best machine I have seen."

Figure 12. Bassi is an 888-processor IBM p575 POWER5 system with a theoretical peak performance of 6.7 teraflop/s.

Bassi is an IBM p575 POWER5 system, and each processor has a theoretical peak performance of 7.6 gigaflop/s. The processors are distributed among 111 compute nodes with eight processors per node. Processors on each node have a shared memory pool of 32 GB.

The compute nodes are connected to each other with a high-bandwidth, low-latency switching network. Each node runs its own full instance of the standard AIX operating system. The disk storage system is a distributed, parallel I/O system called GPFS (IBM's General Parallel File System). Additional nodes serve exclusively as GPFS servers. Bassi's network switch is the IBM "Federation" HPS switch, which is connected to a two-link network adapter on each node.

One of the test users for NERSC's two new clusters was Robert Duke of the University of North Carolina, Chapel Hill, the author of the PMEMD code, which is the parallel workhorse in modern versions of the popular chemistry code AMBER. PMEMD is widely used for molecular dynamics simulations and is also part of NERSC's benchmark applications suite. Duke has worked with NERSC's David Skinner to port and improve the performance of PMEMD on NERSC systems.

"I have to say that both of these machines are really nothing short of fabulous," Duke wrote to Skinner. "While Jacquard is perhaps the best-performing commodity cluster I have seen, Bassi is the best machine I have seen, period."

Other early users during the acceptance testing included the INCITE project team "Direct Numerical Simulation of Turbulent Nonpremixed Combustion." "Our project required a very long stretch of using a large fraction of Bassi processors—512 processors for essentially an entire month," recounted Evatt Hawkes. "During this period we experienced only a few minor problems, which is exceptional for a pre-production machine, and enabled us to complete our project against a tight deadline. We were very impressed with the reliability of the machine."

Hawkes noted that their code also ported quickly to Bassi, starting with a code already ported to Seaborg's architecture. "Bassi performs very well for our code. With Bassi's faster processors we were able to run on far fewer processors (512 on Bassi as opposed to 4,096 on Seaborg) and still complete the simulations more rapidly," Hawkes added. "Based on scalar tests, it is approximately 7 times faster than Seaborg and 1½ times faster than a 2.0 GHz Opteron processor. Also, the parallel efficiency is very good. In a weak scaling test, we obtained approximately 78 percent parallel efficiency using 768 processors, compared with about 70 percent on Seaborg."

The machine is named in honor of Laura Bassi, a noted Newtonian physicist of the eighteenth century. Appointed a professor at the University of Bologna in 1731, Bassi was the first woman to officially teach at a European university.

New Visual Analytics Server: DaVinci

In mid-August of 2005, NERSC put into production a new server specifically tailored to data-intensive visualization and analysis. The 32-processor SGI Altix, called DaVinci (Figure 13), offers interactive access to large amounts of large memory and high performance I/O capabilities well suited for analyzing large-scale data produced by the NERSC high performance computing systems (Bassi, Jacquard, and Seaborg).

Figure 13. DaVinci is a 32-processor SGI Altix with 6 GB of memory per processor and 25 TB of disk memory, a configuration designed for data-intensive analysis and interactive visualization.

With its 192 gigabytes (GB) of RAM and 25 terabytes (TB) of disk, DaVinci's system balance is biased toward memory and I/O, which is different from the other systems at NERSC. This balance favors data-intensive analysis and interactive visualization. DaVinci has 6 GB of memory per processor, compared to 2 GB per processor on Jacquard, 4 GB on Bassi, and 1 to 4 GB on Seaborg.

Users can obtain interactive access to 80 GB of memory from a single application (or all 192 GB of memory by prior arrangement), whereas the interactive limits on production NERSC supercomputing systems restrict interactive tasks to a smaller amount of memory (256 MB on login nodes). While DaVinci is available primarily for interactive use, the system is also configured to run batch jobs, especially those jobs that are data intensive.

The new server runs a number of visualization, statistics, and mathematics applications, including IDL, AVS/Express, CEI Ensight, VisIT (a parallel visualization application from Lawrence

Livermore National Laboratory), Maple, Mathematica, and MatLab. Many users depend on IDL and MatLab to process or reorganize data in preparation for visualization. The large memory is particularly beneficial for these types of jobs.

DaVinci is connected to the NERSC Global Filesystem (see below), High Performance Storage System (HPSS), and ESnet networks by two independent 10 gigabit Ethernet connections.

With DaVinci now in production, NERSC has retired the previous visualization server, Escher, and the math server, Newton.

Cray Provides the Next Major NERSC System

On August 10, 2006, Cray Inc. and the DOE Office of Science announced that Cray has won the contract to install a next-generation supercomputer at NERSC. The systems and multi-year services contract includes delivery of a Cray massively parallel processor supercomputer, code-named “Hood.”

The contract also provides options for future upgrades that would quadruple the size of the system and eventually boost performance to one petaflops (1,000 trillion floating point operations per second) and beyond.

A successor to the massively parallel Cray XT3 supercomputer, the XT4 system installed at NERSC will be among the world’s fastest general-purpose systems and will be the largest XT4 system in the world. It will deliver sustained performance of at least 16 trillion calculations per second when running a suite of diverse scientific applications at scale. The system uses thousands of AMD Opteron processors running tuned, lightweight operating system kernels and interfaced to Cray’s unique SeaStar network.

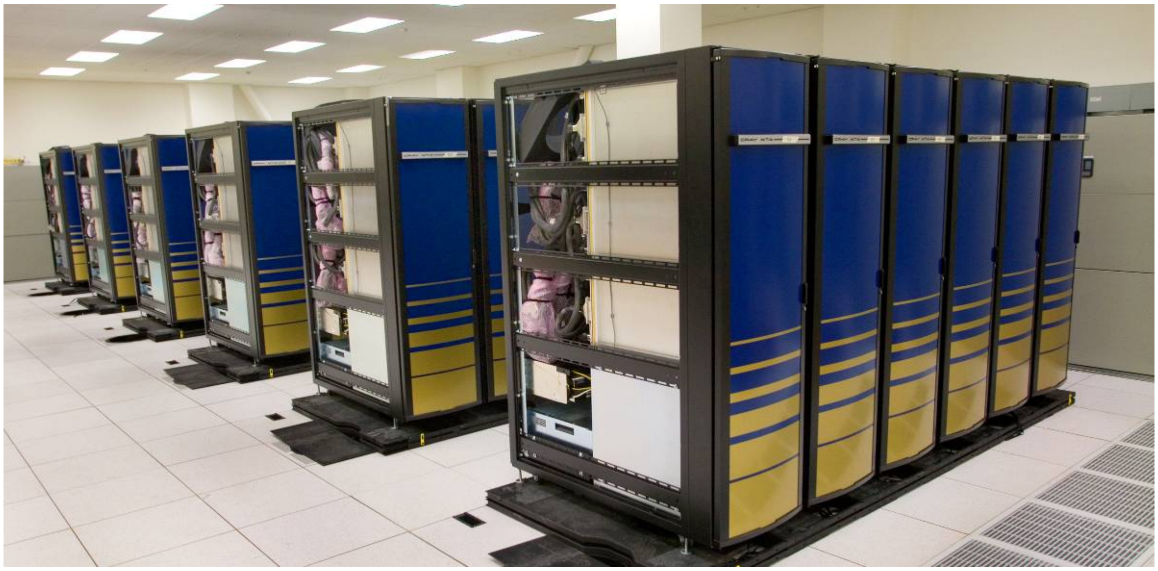

Figure 14. The first 36 cabinets of NERSC’s Cray XT4 supercomputer, which, when complete, will deliver sustained performance of at least 16 teraflop/s.

Cray began building the new supercomputer at the manufacturing facility in late 2006 and delivered it in early 2007 (Figure 14), with completion of the installation and acceptance scheduled for mid-2007.

As part of a competitive procurement process, NERSC evaluated systems from a number of vendors using the NERSC Sustained System Performance (SSP) metric. The SSP metric, developed by NERSC, measures sustained performance on a set of codes designed to accurately represent the challenging computing environment at the Center.

“While the theoretical peak speed of supercomputers may be good for bragging rights, it’s not an accurate indicator of how the machine will perform when running actual research codes,” said Horst Simon, director of the NERSC Division at Berkeley Lab. “To better gauge how well a system will meet the needs of our 2,500 users, we developed SSP. According to this test, the new system will deliver over 16 teraflop/s on a sustained basis.”

“The Cray proposal was selected because its price/performance was substantially better than other proposals we received, as determined by NERSC’s comprehensive evaluation criteria of more than 40 measures,” said Bill Kramer, general manager of the NERSC Center.

The XT4 supercomputer at NERSC will consist of almost 20,000 AMD Opteron 2.6-gigahertz processor cores (19,344 compute CPUs), with two cores per socket making up one node. Each node has 4 gigabytes (4 billion bytes) of memory and a dedicated SeaStar connection to the internal network. The full system will consist of over 100 cabinets with 39 terabytes (39 trillion bytes) of aggregate memory capacity. When completely installed, the system will increase NERSC's sustained computational capability by almost a factor of 10, with an SSP of 16.09 teraflop/s (as a reference, Seaborg’s SSP is 0.89 Tflop/s, and Bassi’s SSP is 0.8 Tflop/s). The system will have a bisection bandwidth of 6.3 terabytes per second and 402 terabytes of usable disk.

In keeping with NERSC’s tradition of naming supercomputers after world-class scientists, the new system will be called “Franklin” in honor of Benjamin Franklin, America’s first scientist. This year is the 300th anniversary of Franklin’s birth.

“Ben Franklin’s scientific achievements included fundamental advances in electricity, thermodynamics, energy efficiency, material science, geophysics, climate, ocean currents, weather, materials science, population growth, medicine and health, and many other areas,” said NERSC’s Bill Kramer. “In the tradition of Franklin, we expect this system to make contributions to science of the same high order.”

4. Ensure All New Technology and Changes Improve (or at Least Do Not Diminish) Service to Our Clients.

In striving to provide users with the latest systems for computational sciences, NERSC has the responsibility to ensure system changes have a maximum benefit and minimal detrimental impact on the clients' ability to do work.

Bay Area Metropolitan Area Network Inaugurated

On August 23, 2005, the NERSC Center became the first of six DOE research sites to go into full production on the Energy Science Network's (ESnet's) new San Francisco Bay Area Metropolitan Area Network (MAN). The new MAN provides dual connectivity at 20 to 30 gigabits per second (10 to 50 times the previous site bandwidths, depending on the site using the ring) while significantly reducing the overall cost. The Bay Area MAN is the first implementation of several MANs ESnet has planned over the next several years.

The connection to NERSC consists of two 10-gigabit Ethernet links. One link is used for production scientific computing traffic, while the second is dedicated to special networking needs, such as moving terabyte-scale datasets between research sites or transferring large datasets which are not TCP-friendly.

"What this means is that NERSC is now connected to ESnet at the same speed as ESnet's backbone network," said ESnet engineer Eli Dart.

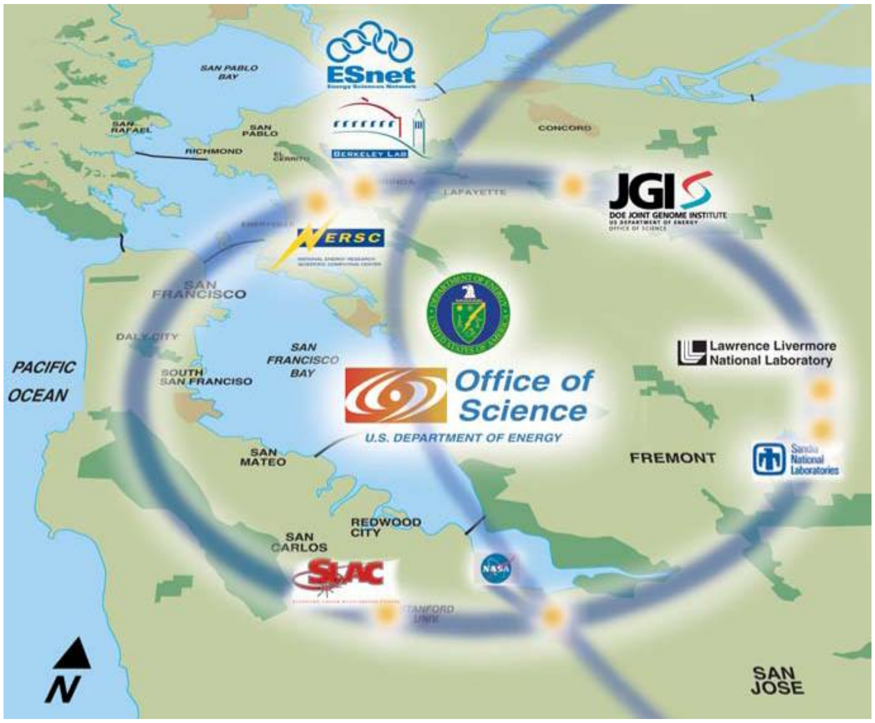

Figure 15. ESnet's new San Francisco Bay Area Metropolitan Area Network provides dual connectivity at 20 to 30 gigabits per second to six DOE sites and NASA Ames Research Center.

The new architecture is designed to meet the increasing demand for network bandwidth and advanced network services as next-generation scientific instruments and supercomputers come on line. Through a contract with Qwest Communications, the San Francisco Bay Area MAN provides dual connectivity to six DOE sites—the Stanford Linear Accelerator Center, Lawrence Berkeley National Laboratory, the Joint Genome Institute, NERSC, Lawrence Livermore National Laboratory, and Sandia National Laboratories/California (Figure 15). The MAN also provides high-speed access to California's higher education network (CENIC), NASA's Ames Research Center, and DOE's R&D network, Ultra Science Net. The Bay Area MAN connects to both the existing ESnet production backbone and the first segments of the new Science Data Network backbone.

Due to coordination and hard work by both ESnet and NERSC staff, the network upgrade was transparent to NERSC users.

OSF Power Supply Is Upgraded

A planned power outage at the Oakland Scientific Facility (OSF) during the week of October 30, 2006, allowed the NERSC computer room to be safely upgraded to accommodate a new uninterruptible power supply (UPS) and future computing systems, including Franklin, NERSC's soon-to-be-installed new Cray supercomputer. Several carefully timed email notices during the previous month had informed all NERSC users about the outage that began on Monday morning, October 30, and was scheduled to last for two days.

The electrical substations in the OSF basement were built to deliver up to 6 megawatts (MW) of power, but until now, only 2 MW were actually used in the machine room. Soon, however, NERSC will need 4 MW to power the increased computing capability and cooling requirements of Franklin and future machines.

To meet these needs, PG&E upgraded its connection to the building, and new 480V feeds were connected between the basement and the machine room to deliver the increased power. The chilled water piping under the machine room floor was also rearranged to improve the air flow, since each of Franklin's 102 racks will need 2300 cubic feet of cooled air per minute.

NERSC staff began shutting down the computing, storage, and network systems at 4 a.m. on Monday, and the OSF power was shut off at 9:30 a.m. so the work could proceed safely. The power upgrade was completed a little ahead of schedule, with the OSF power restored and the computer room stabilized around 8 p.m. on Tuesday. NERSC staff then returned NERSC systems to production, with most systems restored by 10 a.m. Wednesday, Nov. 1. A few more hours of unscheduled hardware and software maintenance were required on Seaborg on Wednesday and Thursday evenings due to a new kernel bug that caused nodes to crash, but the NERSC web site kept users informed with system status updates.

In February 2007, NERSC completed the power upgrade by installing its first uninterruptible power supply (UPS) to protect critical data in the NERSC Global Filesystem (NGF) and HPSS. If an unscheduled power outage were to crash NGF—which is mounted on all NERSC production systems and holds up to 70 TB of data—new data that had not yet been backed up might be lost, and previously backed up data could take a week to restore. With the UPS in operation, if an unscheduled power outage does happen, the UPS will allow a graceful shutdown of NERSC's critical storage disks and databases. And that added margin of safety will benefit NERSC staff and users with increased reliability and short times to recover from power failures.

5. Develop Innovative Approaches to Help the Client Community Effectively Use NERSC Systems

NERSC must assist our clients in being as productive as possible by providing systems, enhancements, tools, information, training, consulting and other assistance. In addition to the traditional approaches that are effective, NERSC will constantly try new approaches to help make our clients effective in an ever-more-changing environment. NERSC will help design strategies and integrate and develop technology to enable our clients to improve their use of our systems and to more effectively accomplish their science.

NERSC Global Filesystem

In late 2005, NERSC deployed the NERSC Global Filesystem (NGF) into production, providing seamless data access from all of the Center's computational and analysis resources. NGF is intended to facilitate sharing of data between users and/or machines. For example, if a project has multiple users who must all access a common set of data files, NGF provides a common area for those files. Alternatively, when sharing data between machines, NGF eliminates the need to copy large datasets from one machine to another. For example, because NGF has a single unified namespace, a user can run a highly parallel simulation on Seaborg, followed by a serial or modestly parallel post-processing step on Jacquard, and then perform a data analysis or visualization step on DaVinci—all without having to explicitly move a single data file.

NGF's single unified namespace makes it easier for users to manage their data across multiple systems. Users no longer need to keep track of multiple copies of programs and data, and they no longer need to copy data between NERSC systems for pre- and post-processing. NGF provides several other benefits as well: storage utilization is more efficient because of decreased fragmentation; computational resource utilization is more efficient because users can more easily run jobs on an appropriate resource; NGF provides improved methods of backing up user data; and NGF improves system security by eliminating the need for collaborators to use "group" or "world" permissions.

"NGF stitches all of our systems together," said Greg Butler, leader of the NGF project. "When you go from system to system, your data is just there. Users don't have to manually move their data or keep track of it. They can now see their data simultaneously and access the data simultaneously."

NERSC staff began adding NGF to computing systems in October 2005, starting with the DaVinci visualization cluster and finishing with the Seaborg system in December. To help test the system before it entered production, a number of NERSC users were given preproduction access to NGF. Early users helped identify problems with NGF so they could be addressed before the filesystem was made available to the general user community.

"I have been using the NGF for some time now, and it's made my work a lot easier on the NERSC systems," said Martin White, a physicist at Berkeley Lab. "I have at times accessed files on NGF from all three compute platforms (Seaborg, Jacquard, and Bassi) semi-simultaneously."

NGF also makes it easier for members of collaborative groups to access data, as well as ensure data consistency by eliminating multiple copies of critical data. Christian Ott, a Ph.D. student and member of a team studying core-collapse supernovae, wrote that "the project directories make our collaboration much more efficient. We can now easily look at the output of the runs managed by other team members and monitor their progress. We are also sharing standard input data for our simulations."

NERSC General Manager Bill Kramer said that as far as he knows, NGF is the first production global filesystem spanning five platforms (Seaborg, Bassi, Jacquard, DaVinci, and PDSF), three architectures, and four different vendors. While other centers and distributed computing projects such as the National Science Foundation's TeraGrid may also have shared filesystems, Butler said he thinks NGF is unique in its heterogeneity.

The heterogeneous approach of NGF is a key component of NERSC's five-year plan (see section 9). This approach is important because NERSC typically procures a major new computational system every three years, then operates it for five years to support DOE research. Consequently, NERSC operates in a heterogeneous environment with systems from multiple vendors, multiple platforms, different system architectures, and multiple operating systems. The deployed filesystem must operate in the same heterogeneous client environment throughout its lifetime.

Butler noted that the project, which is currently based on IBM's proven GPFS technology (in which NERSC was a research partner), started about five years ago. While the computing systems, storage, and interconnects were mostly in place, deploying a shared filesystem among all the resources was a major step beyond a parallel filesystem. In addition to the different system architectures, there were also different operating systems to contend with. However, the last servers and storage have now been deployed. To keep everything running and ensure a graceful shutdown in the event of a power outage, a large uninterruptible power supply has been installed in the basement of the Oakland Scientific Facility.

While NGF is a significant change for NERSC users, it also "fundamentally changes the Center in terms of our perspective," Butler said. For example, when the staff needs to do maintenance on the filesystem, the various groups need to coordinate their efforts and take all the systems down at once.

Storage servers, accessing the consolidated storage using the shared-disk filesystems, provide hierarchical storage management, backup, and archival services. The first phase of NGF is focused on function and not raw performance, but in order to be effective, NGF has to have performance comparable to native cluster filesystems. The current capacity of NGF is approximately 70 TB of user-accessible storage and 50 million inodes (the data structures for individual files). Default project quotas are 1 TB and 250,000 inodes. The system has a sustainable bandwidth of 3 GB/sec bandwidth for streaming I/O, although actual performance for user applications will depend on a variety of factors. Because NGF is a distributed network filesystem, performance will be equal to or slightly less than that of filesystems that are local to NERSC compute platforms. This should only be an issue for applications whose performance is I/O bound.

NGF will grow in both capacity and bandwidth over the next several years, eventually replacing or dwarfing the amount of local storage on systems. NERSC is also working to seamlessly integrate NGF with the HPSS data archive to create much larger "virtual" data storage for projects. Once NGF is completely operational within the NERSC facility, Butler said, users at other centers, such as the National Center for Atmospheric Research and NASA Ames Research Center, could be allowed to remotely access the NERSC filesystem, allowing users to read and visualize data without having to execute file transfers. Eventually, the same capability could be extended to experimental research sites, such as accelerator labs.

Integrated Performance Monitoring Simplifies Code Assessment

As the HPC center of choice for the DOE research community, NERSC consistently receives requests for more computing resources than are available. Because computing time is so valuable, making the most of every allocated processor-hour is a paramount concern. Evaluating the performance of application codes in the diverse NERSC workload is an important and challenging endeavor. As NERSC moves toward running more large-scale jobs, finding ways to improve performance of large-scale codes takes on even greater importance.

For this reason identifying bottlenecks to scalable performance of parallel codes has been an area of intense focus for NERSC staff. To identify and remove these scaling bottlenecks, NERSC's David Skinner has developed Integrated Performance Monitoring, or IPM. IPM is a portable profiling infrastructure that provides a performance summary of the computation and communication in a parallel program. IPM has extremely low overhead, is scalable to thousands of processors, and was designed with a focus on ease of use, requiring no source code modification. These characteristics are the right recipe for measuring application performance in a production environment like NERSC's, which consists of hundreds of projects and parallelism ranging from 1 to 6,000 processors.

Skinner points to the lightweight overhead and fixed memory footprint of IPM as one of its biggest innovations. Unlike performance monitoring based on traces, which consume more resources the longer the code runs, IPM enforces strict boundaries on the resources devoted to profiling. By using a fixed memory hash table, IPM achieves a compromise between providing a detailed profile and avoiding impact on the profiled code.

IPM was also designed to be portable. It runs on the IBM SP, Linux clusters, Altix, Cray X series, NEC SX6, and the Earth Simulator. Portability is a key to enabling cross-platform performance studies. Portability, combined with IPM's availability under an open source software license, will hopefully lead to other centers adopting and adding to the IPM software.

Skinner characterizes IPM as a "profiling layer" rather than a performance tool. "The idea is that IPM can provide a high-level performance summary which feeds both user and center efforts to improve performance," Skinner said. "IPM finds 'hot spots' and bottlenecks in parallel codes. It also identifies the overall characteristics of codes and determines which compute resources are being used by a code. It really provides a performance inventory. Armed with that information, users can improve their codes and NERSC can better provide compute resources aligned to meet users' computational needs."

IPM automates a number of monitoring tasks that NERSC consultants used to perform manually. By running a code with IPM, NERSC staff can quickly generate a comprehensive performance picture of a code, with the information presented both graphically (Figure 16) and numerically.

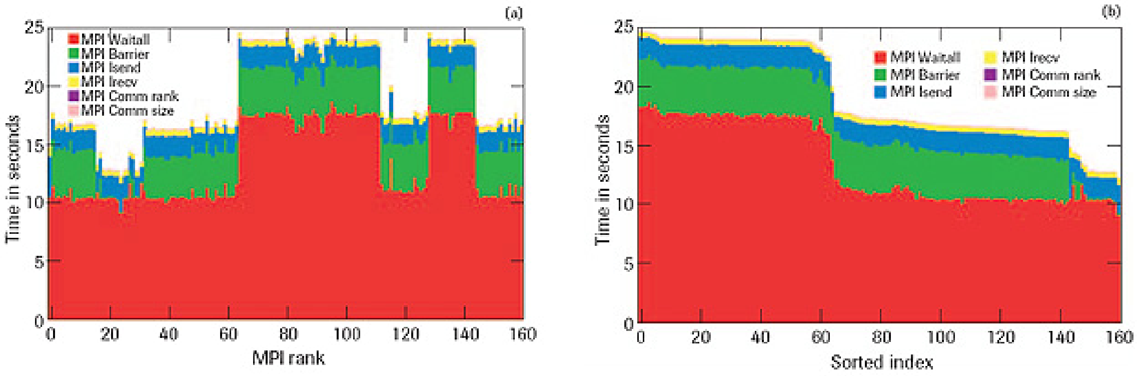

Figure 16. IPM can graphically present a wide range of data, including communication balance by task, sorted by (a) MPI rank or (b) MPI time.

The monitors that IPM currently integrates include a wide range of MPI communication statistics; HPM (Hardware Performance Monitor) counters for things like flop rates, application memory usage, and process topology; and system statistics such as switch traffic.

The integration in IPM is multi-faceted, including binding the above information sources together through a common interface, and also integrating the records from all the parallel tasks into a single report. On many platforms IPM can be integrated into the execution environment of a parallel computer. In this way, an IPM profile is available either automatically or with minor effort. The final level of integration is the collection of individual performance profiles into a database that synthesizes the performance reports via a Web interface. This Web interface can be used by all those concerned with parallel code performance: users, HPC consultants, and HPC center managers. As different codes are characterized, the results are posted to protected Web pages. Users can access only the pages for the codes they are running.

One of the first uses for IPM was to help the INCITE projects make the most effective use of their large allocations. Subsequently it has been expanded to other projects. Even a small improvement — say 5 percent — in a code that runs on a thousand processors for millions of processor-hours is a significant gain for the center. "Our primary goal is to help projects get the most out of their allocated time," Skinner said.

But the same information is also interesting to the center itself. Obtaining a center-wide picture of how computational resources are used is important to knowing that the right resources are being presented and in the right way. It also guides choices about what future NERSC computational resources should look like. For example, IPM shows which parts of MPI are widely used by NERSC customers and to what extent. "It's good to know which parts of MPI our customers are using," Skinner said. "As an HPC center this tells us volumes about not only what we can do to make codes work better with existing resources as well as what future CPUs and interconnects should look like."

"We are looking for other programmers to contribute to IPM," Skinner added. "IPM complements existing platform-specific performance tools by providing an easy-to-use profiling layer that can motivate and guide the use of more detailed, in-depth performance analysis."

More information about IPM is available at http://www.nersc.gov/projects/ipm/.

Science-Driven Analytics

Simulations and experiments are generating data faster than it can be analyzed and understood. Addressing this bottleneck in the scientific discovery process is the emerging discipline of *analytics,* which has the simple goal of understanding data.

The term *analytics* refers to a set of interrelated technologies and intellectual disciplines that combine to produce insight and understanding from large, complex, disparate, and sometimes conflicting datasets. These technologies and disciplines include data management, visualization, analysis, and discourse aimed at producing specific types of understanding. These in turn rely on the computational infrastructure, expertise in using that infrastructure, and close cooperation between domain scientists, computational scientists, and computer scientists.

More specifically, the term *visual analytics* is the science of analytic reasoning facilitated by interactive visual interfaces. Its objective is to enable analysis of overwhelming amounts of information, and it requires human judgment to make the best possible evaluation of incomplete, inconsistent, and potentially erroneous information.

NERSC's analytics strategy builds on two of the Center's existing strengths: (1) proven expertise in effectively managing large, complex computing, infrastructure, and data storage systems to solve scientific problems of scale; and (2) exemplary user services, consulting, and domain scientific knowledge that help the NERSC user community effectively employ the Center's resources to solve challenging scientific problems. On this foundation, NERSC's analytics strategy adds an increased emphasis on facilities, infrastructure, expertise, and alliances that can be used to realize analytics solutions.

With the establishment of its new Analytics Team (see section 9), NERSC is realigning its resources to support analytics activities. The NERSC Center's infrastructure is being broadened to include elements such as database deployment and support, with an increased focus on data analysis and scientific data management to support analytics. The existing visualization program is being expanded to include information visualization and integrated data management, analysis, and distributed computing. The goal is a well-rounded service and technology portfolio that is responsive to the analytics needs of NERSC's user community.

NERSC's analytics strategy includes five elements:

1. *Taking a proactive role in deploying emerging technologies.* NERSC will increasingly become a conduit for prototype technologies that emerge from the DOE computer science research community. Analytics will require adapting and deploying technologies from several different areas—data management, analysis, visualization, dissemination—into a unified workflow that functions effectively in a time-critical production environment. The role of NERSC staff will include deploying new system and support software, helping applications software engineers effectively use NERSC resources, and playing a proactive role in providing feedback to the original computer science researchers and developers to address security or performance concerns.

2. *Enhancing NERSC's data management infrastructure.* The NERSC Global Filesystem offers increased performance for all applications, including data-intensive analytics tasks. It also helps streamline distributed workflows and provides high I/O rates, which are important for large datasets. NERSC also plans to increase its archival storage to nearly 40 PB over the next five years. In the near term, NERSC will evaluate and deploy software that provides distributed, file-level data management.

3. *Expanding NERSC's visualization and analysis capabilities.* One of the most significant activities performed by the NERSC visualization staff is in-depth, one-on-one consulting services, such as those provided to INCITE and other large projects. These activities typically involve finding or engineering solutions where none exist off the shelf. In addition, visualization staff will evaluate new visualization hardware and software technologies to determine which are beneficial to the user community. These technologies may include *information visualization,* which differs from the better-known *scientific visualization* in that the underlying data does not readily lend itself to spatial mapping—for example, comparing the results of genome alignment across multiple species. As data size and complexity grow, it will become increasingly crucial to use analysis technologies to reduce the processing load through the computational and visualization pipelines, as well as to reduce the "scientific processing load" on the humans who must interpret and understand the results. A portfolio of commercial, production, open-source, and research-grade technologies is expected to be most effective in meeting users' scientific needs.

4. *Enhancing NERSC's distributed computing infrastructure.* NERSC's strategy for supporting distributed computing will be tailored to provide services that have the broadest possible benefit and that conform to security requirements. In addition to providing low-level infrastructure such as the Open Grid Services Architecture (OGSA) and similar technologies that provide authentication and secure data movement across the network, NERSC will investigate and deploy higher-level applications and services that emerge from research and applications communities like the Open Science Grid which rely on standard services for brokering access to data and tools that serve large, distributed user communities. NERSC will work closely with the user community to provide the documentation and assistance they need to construct analytics workflows.

5. *Understanding the analytics needs of the user community.* To be effective, NERSC's new program focus on Science-Driven Analytics will require additional information from the user

community. To that end, the entire user community was surveyed in early 2006 to identify their most pressing analytics needs. The findings from this survey have been instrumental in shaping and prioritizing the emerging analytics effort. NERSC will continue soliciting input from users as well as tracking analytics trends in the larger scientific community.

6. Develop and Implement Ways to Transfer Research Products and Knowledge into Production Systems at NERSC and Elsewhere

NERSC is uniquely placed to establish methods and procedures that enable research products and knowledge, particularly those developed at LBNL/UC, to smoothly flow into production.

Another Checkpoint/ Restart Milestone

On the weekend of June 11 and 12, 2005, IBM personnel used NERSC's Seaborg supercomputer for dedicated testing of IBM's latest HPC Software Stack, a set of tools for high performance computing. To maximize system utilization for NERSC users, instead of "draining" the system (letting running jobs continue to completion) before starting this dedicated testing, NERSC staff checkpointed all running jobs at the start of the testing period. "Checkpointing" means stopping a program in progress and saving the current state of the program and its data—in effect, "bookmarking" where the program left off so it can start up later in exactly the same place.

This is believed to be the first full-scale use of the checkpoint/restart software with an actual production workload on an IBM SP, as well as the first checkpoint/restart on a system with more than 2,000 processors. It is the culmination of a collaborative effort between NERSC and IBM that began in 1999. Of the 44 jobs that were checkpointed and restarted, approximately 65% checkpointed successfully. Of the 15 jobs that did not checkpoint successfully, only 7 jobs were deleted from the queuing system, while the rest were requeued to run again at a later time. This test enabled NERSC and IBM staff to identify and fix some previously undetected problems with the checkpoint/restart software.

In 1997 NERSC made history by being the first computing center to achieve successful checkpoint/restart on a massively parallel system, the Cray T3E. NERSC is now working with Berkeley Lab's Future Technologies Group and Cray to implement Berkeley Lab Checkpoint/ Restart (BLCR) on the XT4.

Integrating HPSS into Grids

NERSC's Mass Storage Group is currently involved in a development collaboration with Argonne National Laboratory and IBM to integrate High Performance Storage System (HPSS) accessibility into the Globus Toolkit for Grid applications.

At Argonne, researchers are adding functionality to the Grid file transfer daemon2 so that the appropriate class of service can be requested from HPSS. IBM is contributing the development of an easy-to-call library of parallel I/O routines that work with HPSS structures and are also easy to integrate into the file transfer deamon. This library will ensure that Grid file transfer requests to HPSS movers are handled correctly.

NERSC is providing the HPSS platform and testbed system for IBM and Argonne to do their respective development projects. As pieces are completed, NERSC tests the components and works with the developers to help identify and resolve problems. The public release of this capability is scheduled with HPSS 6.2, as well as future releases of the Globus Toolkit.

7. Improve Methods of Managing Systems Within NERSC and LBNL and Be a Leader in Large-Scale Systems Management and Services

As the Department of Energy's flagship unclassified scientific computing facility, NERSC continually provides leadership and helps shape the field of high performance computing. As HPC technology evolves at an increasing rate, it is crucial that NERSC and LBNL remain at the forefront of getting the most out of these systems.

NERSC-5 Procurement Involves Interagency Collaboration

As part of NERSC's regular computational system acquisition cycle, the NERSC-5 procurement team was formed in October 2004 to develop an acquisition plan, select and test benchmarks, and prepare a request for proposals (RFP). The RFP was released in September 2005; proposals were submitted in November; and the resulting award to Cray was announced in August 2006 (see section 3). The RFP set the following general goals for the NERSC-5 system:

- Support the entire NERSC workload, specifically addressing the DOE Greenbook recommendations.
- Integrate with the NERSC environment, including the NERSC Global Filesystem, HPSS, Grid software, security and networking systems, and the user environment (software tools).
- Provide the optimal balance of the following system components:
  - computational: CPU speed, memory bandwidth, and latency
  - memory: aggregate and per parallel task
  - global disk storage: capacity and bandwidth
  - interconnect: bandwidth, latency, and scaling
  - external network bandwidth.

The RFP also stated specific goals for performance (as measured by NERSC's Sustained System Performance [SSP] metric), disk storage, space and power requirements, software, etc. "The Cray proposal was selected because its price/performance was substantially better than other proposals we received, as determined by NERSC's comprehensive evaluation criteria of more than 40 measures," said Bill Kramer, general manager of the NERSC Center.

Two recent reports[1,2] on high-end computing recommended interagency collaboration on system procurements. The National Research Council report stated, "Joint planning and coordination of acquisitions will increase the efficiency of the procurement processes from the government viewpoint and will decrease variability and uncertainty from the vendor viewpoint."[3] NERSC-5 is

[1] *Federal Plan for High-End Computing: Report of the High-End Computing Revitalization Task Force (HECRTF).* Washington, D.C.: National Coordination Office for Information Technology Research and Development, May 10, 2004.

[2] Susan L. Graham, Marc Snir, and Cynthia A. Patterson, eds., *Getting Up to Speed: The Future of Supercomputing,* Committee on the Future of Supercomputing, National Research Council. Washington, D.C.: The National Academies Press, 2005.

[3] Ibid., p. 171.

possibly the first procurement involving collaboration with other government agencies. This collaboration includes the sharing of benchmarks with DOD and NSF (as described in section 8). In addition, four organizations—the DOD HPC Modernization Program, the National Center for Supercomputing Applications, the Pittsburgh Supercomputing Center, and Louisiana State University—sent representatives to observe NERSC's Best Value Source Selection process. The NSF centers have adopted several of NERSC's procurement practices.

NERSC Helps IBM Refine System Software Testing

Nick Cardo, NERSC's IBM SP project lead, was invited in 2005 to give a customer perspective to staff at IBM's test lab in Poughkeepsie, NY. Cardo spent two days at the facility, demonstrating how he runs various systems tests regularly on Seaborg, NERSC's IBM supercomputer.

For two days, Cardo worked side by side with IBM staff on their test SP, showing them how he runs tests on a daily basis. The result was that the IBM staff were able to see what a user encounters.

"By sitting down with the testers at their internal test machine, I was able to give them a customer's perspective of a production environment, running the checkouts I would normally run during the course of the day," Cardo said. "This effort, which was unique, is a reflection of the working partnership we have developed with IBM over the years."

Curtis Vinson, Cardo's contact at IBM, summarized the results of the testing at the SP-XXL meeting held a short time later in Edinburgh, Scotland. The SP-XXL user group focuses on large-scale scientific and technical computing on IBM hardware.

For his part, Cardo produced a seven-page report describing some of the problems he encountered during the March 29–30 testing stint and outlining ways to fix them. The overall objective, Cardo said, was to help IBM find ways to prevent "field escapes," the term for software bugs that make it out of the testing lab and into the user community.

"Our concern is that sometimes when we do system updates, we hit problems that should have been caught in the test lab," Cardo said. "By showing IBM how we use the system, we were able to help them refine their testing procedures and take steps to eliminate the bugs before they become field escapes."

As part of his responsibilities at NERSC, Cardo runs certain tests twice a day on Seaborg. This helps the Computational Systems Group find and fix problems quickly, before they become major hindrances to running users' jobs.

What Cardo and the IBM testers realized is that while each software component may have been well tested at the lab individually, the components were not always tested together for overall compatibility.

"The benefit of all this is that the software upgrades produced by IBM will be more stable right out of the box," Cardo said. "Users of all IBM systems will benefit from this work."

Improving the Software Side of HPC Support

In September 2005, Mike Stewart of NERSC was invited to give a talk on customer service issues at the Linux Networx (LNXI) user group meeting. Stewart and Francesca Verdier prepared a presentation that described NERSC users' and staff members' experiences with Jacquard and with the LNXI service organization, focusing in particular on software support.

Stewart's talk described the challenge of the NERSC user environment: A diverse user base runs hundreds of constantly changing codes that use the entire gamut of scientific algorithms. These codes exercise many of the compiler and library features and thus expose many bugs, which NERSC must rapidly respond to. NERSC also needs to run multiple versions of software and to test, install, and potentially back out of new software releases.

NERSC requires an integrated software environment for building and running parallel scientific applications, including compilers, scientific libraries, and a variety of tools and utilities. While previous vendors produced and supported all of these elements directly, LNXI acquires these products from third-party vendors and integrates them into the NERSC environment, which complicates the process of reporting and fixing bugs and testing new versions of software.

NERSC made three general recommendations on how LNXI can improve software support:

- LNXI should develop expertise on each product they provide at least sufficient to write test cases for bug reports the customer submits.
- For each customer LNXI should have a test suite of every bug in every product submitted by that customer as well as other interesting codes, and this test suite should be run against new releases of any product.
- LNXI should insist that every fix and new feature for a customer go into a standard release version of that product and not a customer-specific version.

Stewart's talk had been taped and was later shown to every LNXI employee. After the user group meeting, Steven C. Caruso, LNXI's Program Manager for Western DOE Labs, wrote:

> I believe that I am representing a large number of LNXI employees, including executives, when I thank you for your participation in our first user's group conference and especially for your presentation regarding software support with respect to Linux clusters. I can tell you that it has caught the attention of many employees and we have discussed the ideas you presented very enthusiastically.
>
> We appreciate your balanced, constructive criticisms and also appreciate your citing the positive experiences as well. We are certainly aware of the shortcomings of the current method of supporting an open source/third party software stack, and your presentation summing up the experiences we both have experienced with Jacquard has been able to succinctly elucidate the associated problems and possible solutions.
>
> Concurrently with the Jacquard experience, and especially motivated by your presentation, LNXI has actively embarked on a process of devising a solution to the problems you have cited. You will be hearing more about our plans for improving our software support in the very near future….

The company subsequently reorganized their software service division.

8. Export Knowledge, Experience and Technology Developed at NERSC, Particularly to and within NERSC Client Sites

In order for NERSC to be a leader in large-scale computing, NERSC must export experience, knowledge, and technology. Transfers must be made to other client sites, supercomputer sites, and industry.

Promoting Cross-Platform Filesystems

Using GPFS for the NERSC Global Filesystem (NGF, see section 5) was made possible by IBM’s decision to make its GPFS software available across mixed-vendor supercomputing systems. This strategy was a direct result of IBM’s collaboration with NERSC. “Thank you for driving us in this direction,” wrote IBM Federal Client Executive Mike Henesy to NERSC General Manager Bill Kramer when IBM announced the project in December 2005. “It’s quite clear we would never have reached this point without your leadership!”

NERSC’s Mass Storage Group collaborated with IBM and the San Diego Supercomputer Center to develop a Hierarchical Storage Manager (HSM) that can be used with IBM’s GPFS. The HSM capability with GPFS provides a recoverable GPFS filesystem that is transparent to users and fully backed up and recoverable from NERSC’s multi-petabyte archive on HPSS. GPFS and HPSS are both cluster storage software: GPFS is a shared disk filesystem, while HPSS supports both disk and tape, moving less-used data to tape while keeping current data on disk.

One of the key capabilities of the GPFS/HPSS HSM is that users’ files are automatically backed up on HPSS as they are created. Additionally, files on the GPFS which have not been accessed for a specified period of time are automatically migrated from online resources as space is needed by users for files currently in use. Since the purged files are already backed up on HPSS, they can easily be automatically retrieved by users when needed, and the users do not need to know where the files are stored to access them. “This gives the user the appearance of almost unlimited disk storage space without the cost,” said NERSC’s former Mass Storage Group Leader, Nancy Meyer.

This capability was demonstrated in the Berkeley Lab and IBM booths at the SC05 and SC06 conferences. Bob Coyne of IBM, the industry co-chair of the HPSS executive committee, said, “There are at least ten institutions at SC05 who are both HPSS and GPFS users, many with over a petabyte of data, who have expressed interest in this capability. HPSS/GPFS will not only serve these existing users but will be an important step in simplifying the storage tools of the largest supercomputer centers and making them available to research institutions, universities, and commercial users.”

“Globally accessible data is becoming the most important part of Grid computing,” said Phil Andrews of the San Diego Supercomputer Center. “The immense quantity of information demands full vertical integration from a transparent user interface via a high performance filesystem to an enormously capable archival manager. The integration of HPSS and GPFS closes the gap between the long-term archival storage and the ultra high performance user access mechanisms.” The GPFS/HPSS HSM was included in the release of HPSS 6.2 in spring 2006.

Benchmarking and Performance Monitoring

An important responsibility of NERSC's Science-Driven System Architecture (SDSA) Team (see section 9) is sharing performance and workload data, along with benchmarking and performance monitoring codes, with others in the HPC community. Benchmarking suites, containing application codes or their algorithmic kernels, are widely used for system assessment and procurement. NERSC has recently shared its SSP benchmarking suite with National Science Foundation (NSF) computer centers. With the Defense Department's HPC Modernization Program, NERSC has shared benchmarks and jointly developed a new one. Furthermore, NERSC now has a web site for all the SSP benchmarks, at which other sites can download tests and report their own results (http://www.nersc.gov/projects/SDSA/software/?benchmark=ssp).

Software Roadmap to Plug and Play Petaflop/s

In the next five years, the DOE expects to field systems that reach a petaflop of computing power in scale. In the near term (two years), DOE will have several "near-petaflops" systems that are 10% to 25% of a peraflop-scale system. A common feature of these precursors to petaflop systems (such as the Cray XT3 or the IBM BlueGene/L) is that they rely on an unprecedented degree of concurrency, which puts stress on every aspect of HPC system design. Such complex systems will likely break current "best practices" for fault resilience, I/O scaling, and debugging, and even raise fundamental questions about programming languages and application models. It is important that potential problems are anticipated far enough in advance that they can be addressed in time to prepare the way for petaflop-scale systems.

DOE asked the NERSC and Computational Research Divisions at Lawrence Berkeley National Laboratory to address these issues by considering the following four questions:

1. What software is on a critical path to make the systems work?
2. What are the strengths/weaknesses of the vendors and of existing vendor solutions?
3. What are the local strengths at the labs?
4. Who are other key players who will play a role and can help?

Berkeley Lab responded to these questions in the report "Software Roadmap to Plug and Play Petaflop/s" (https://www.nersc.gov/ news/reports/LBNL-59999.pdf). The report is organized as follows.

Section 1 provides a high-level answer to question #1, "What software is on the critical path to make the systems work?"

We broadened the response to include both hardware and software issues because the two are so intricately entwined on systems of this scale. We also differentiate near-term (2007) challenges from those that we anticipate in the long term (2008 and beyond).

Section 2 addresses question #2, "Describe the strengths and weaknesses of the vendors and existing vendor solutions," using data collected from the NERSC-5 procurement.

Section 3 addresses question #3, “What are the local strengths at the labs?” by describing the local strengths at NERSC and LBNL for responding to the challenges of petascale computing described in the earlier sections.

Section 4 addresses question #4, “Who are other key players who will play a role and can help?” by identifying key players at other institutions who can be considered key partners for addressing the problems posed in earlier sections.

Section 5 provides supplemental information regarding NERSC’s effort to use non-invasive workload profiling to identify application requirements for future systems (see discussion of IPM in section 5). The data collected by NERSC may be valuable for proactively identifying bottlenecks in current systems and anticipating future application requirements.

Section 6 describes a set of codes that provide good representation of the application requirements of the broader DOE scientific community. The success of these codes is a bellwether for the overall success of these computing platforms for the range of DOE scientific applications.

Section 7 is a comprehensive production software requirements checklist that was derived from the experience of the NERSC-3, NERSC-4, and NERSC-5 procurement teams. It presents an extremely detailed view of the software requirements for a fully functional petaflop-scale system environment. It also includes an assessment of how emerging near-petaflop systems (XT3, BG/L, Power SP) conform or fail to conform to these requirements.

ASCR Metrics Effort

In spring of 2006, Dr. Raymond Orbach, the Department of Energy Under Secretary for Science, asked the Advanced Scientific Computing Research Advisory Committee (ASCAC) “to weigh and review the approach to performance measurement and assessment at [ALCF, NERSC, and NLCF], the appropriateness and comprehensiveness of the measures, and the [computational science component] of the science accomplishments and their effects on the Office of Science’s science programs.” The Advisory Committee formed a subcommittee to respond to the charge, which was chaired by Gordon Bell.

NERSC has long used goals and metrics to assure what we do is meeting the needs of DOE and its scientists. Hence, it was natural for NERSC to take the lead, working with representatives from the other sites to formulate a joint plan for metrics. NERSC then worked with the other sites and the subcommittee to review all the information and suggestions.

Throughout the summer, NERSC worked with the other sites and the committee to develop a robust approach to metrics. The committee report, accepted in February 2007, identified two classes of metrics — *control metrics* and *observed metrics. Control metrics* have specific goals which must be met, and *observed metrics* are used for monitoring and assessing activities. The subcommittee felt that there should be free and open access to the many observed metrics computing centers collect and utilize but “it would be counter-productive to introduce a large number of spurious ‘control’ metrics beyond the few we recommend below.”

The committee report pointed out, "It should be noted that NERSC pioneered the concept of 'project specific services' which it continues to provide as part of SciDAC and INCITE projects." Another panel recommendation is that the all centers "use a 'standard' survey based on the NERSC survey that has been used for several years in measuring and improving service."

The final committee report is available at http://www.sc.doe.gov/ascr/ASCAC/ DOE_ASCAC_Petascale_Metrics_Panel_Interim_Report_AND_Exec_Summary_061016.pdf.

Presentations, Workshops, and Tutorials

Wes Bethel, "Finding the Unknown in a Sea of Data: Leveraging Human Intuition with Scientific Visual Data Analysis," LBNL Projects Office, Washington, D.C., February 17, 2005.

Wes Bethel, "Query-Driven Visualization," 12th SIAM Conference on Parallel Processing for Scientific Computing, San Francisco, February 21–24, 2006.

Wes Bethel, "The SciDAC2 Visualization and Analytics Center for Enabling Technologies: Overview and Objectives," SC06, Tampa, Florida, November 11–17, 2006.

Jonathan Carter and Lenny Oliker, "Leading Computational Methods on the Earth Simulator," 12th SIAM Conference on Parallel Processing for Scientific Computing, San Francisco, February 21–24, 2006.

Jonathan Carter, "Introducing the SciDAC Outreach Center," SC06, Tampa, Florida, November 11–17, 2006.

Phillip Finck, David Keyes, and Rick Stevens (eds.; Horst Simon, William T.C. Kramer, Wes Bethel, participants/co-authors), "Workshop on Simulation and Modeling for Advanced Nuclear Energy Systems," Office of Nuclear Energy and Office of Advanced Scientific Computing Research, U.S. Department of Energy, August 15–17, 2006.

William T.C. Kramer, "Creating Science Driven Architectures," Report for the Workshop on the Use of High Performance Computing in Meteorology, Reading, England, October 2004, workshop proceedings 2005.

William T.C. Kramer, "A Brief Overview of Performance Investigations at NERSC," invited presentation to the DOD High Performance Computing Modernization Program, Virginia, March 2005, and NSF Terascale Application Benchmarking workshop, Sept. 19–21, 2005.

William T.C. Kramer, "NERSC: Where Data and Simulation Meet," presentation at the International IEEE-Computer Society Symposium on Mass Storage Systems and Technologies, Sardinia, Italy, June 19–24, 2005.

William T.C. Kramer, "Future Trends in HPC Hardware and Software," keynote presentation at the DOD High Performance Modernization Program User's Meeting, Nashville, TN, July, 2005 (http://www.hpcmo.hpc.mil/Htdocs/UGC/UGC05/tuesday.html).

William T.C. Kramer, "Ten Years of Clusters: Where They Have Been and Where They May Go," keynote presentation at the 3rd Annual Symposium on the Use of Commodity Clusters for Large Scale Scientific Applications, Greenbelt, Maryland, July 26–28, 2005.

William T.C. Kramer, “A Perspective on Future Trends and Directions in Hardware and Software,” presentation at the HPC Expo Conference, September 2005.

William T.C. Kramer, “Data — Who Needs It?” keynote speaker at the Data Intensive Computing Environment (DICE) Workshop, Springfield, OH, March 2006.

William T.C. Kramer, “Petascale Systems and Services: What a National Facility Needs to Do to Help Science Use Petascale Computing,” invited presentation for the Oak Ridge Seminar Series, Oak Ridge, TN, June 6, 2006.

William T.C. Kramer, “Acquisition and Operation of an HPC System” (panel discussion co-chair), International Conference on Supercomputing (ISC2006), Dresden, Germany, June 28, 2006.

William T.C. Kramer and Horst Simon, “The NERSC Global File System,” (peer reviewed poster paper), International Conference on Supercomputing (ISC2006), Dresden, Germany, June 28, 2006.

William T.C. Kramer, “Make the Most of What You Buy by Assessing Cluster Performance,” presentation at the HPC Expo Conference, September 14, 2006.

William T.C. Kramer, “Science Driven Supercomputing,” presentation at the IDC HPC User Forum, Oak Ridge, TN, September 15, 2006.

William T.C. Kramer, “NERSC Experience Implementing a Facility Wide Global File System,” 12th ECMWF (European Center for Mid-range Metrological Forecasting) Workshop on the Use of High Performance Computing in Meteorology, Reading, England, October 2006 (http://www.ecmwf.int/newsevents/meetings/workshops/2004/high_performance_computing-12th/presentations.html). Workshop proceeding to be published in Spring 2007.

William T. Kramer, “NERSC Experience and Plans for Petascale Data,” the Petascale Data Storage Workshop at SC06, Tampa, Florida, November 18, 2006.

William T.C. Kramer, Michael Resch, “Best Practice in HPC Procurements,” (workshop), SC06, Tampa, Florida, November 11–17, 2006.

Stephen Q. Lau, Scott Campbell, William T.C. Kramer, “Cybersecurity at Open Scientific Facilities,” (tutorial), SC05, Seattle, WA, November 2005.

Stephen Q. Lau, Scott Campbell, William T. Kramer, Brian L. Tierney, “Computing Protection in Open HPC Environments,” (tutorial), SC06, Tampa, Florida, November 11–17, 2006.

Horst Simon, “Petascale Computing for Science,” The Salishan Conference on High-Speed Computing, April 18–21, 2005, Gleneden Beach, Oregon.

Horst Simon, “Petascale Computing for Science,” (Invited Speaker), ICCSA2005 Conference,” May 7-11, 2005, Singapore.

Horst Simon, “Toward Petascale Computing for Science,” (Invited Speaker), 1st Erlangen International High-End Computing Symposium, June 16, 2005, Erlangen, Germany.

Horst Simon, “What Supercomputers Still Can’t Do,” (Invited Speaker), Paderborn Opening Ceremony for the pc2,” Paderborn University, June 21, 2005, Paderborn, Germany.

Horst Simon, “Progress in Supercomputing: The Top Three Breakthroughs of the Last 20 and the Top Three Challenges for the Next 20 Years,” (Invited Speaker), June 20–24, 2005, ISC2005 Conference, Heidelberg, Germany.

Horst Simon, “Petascale Computing for Science,” (Invited Speaker), ICCSE 2005, June 27–30, 2005, Istanbul, Turkey.

Horst Simon, “Does Architecture Matter?” (Invited Speaker), NSF CyberInfrastructure Council, July 27, 2005, Arlington, VA.

Horst Simon, “Progress in Supercomputing: The Top Three Breakthroughs of the Last 20 and the Top Three Challenges for the Next 20 Years,” (Invited Speaker), September 30, 2005, Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, Utah.

Horst Simon, “Cyberinfrastructure Direction at NSF and Implications for UC Research,” UC Research Cyberinfrastructure Meeting, October 10–11, 2005, La Jolla, CA.

Horst Simon, “International Review of Research Using HPC in the UK,” (Panel Chair), HPC Users Meeting, London, UK, December 12, 2005.

Horst Simon, “High Performance Computing,” Panel Discussion, SIAM Conference on Computational Science & Engineering, Orlando, Florida, February 12–15, 2005.

Horst Simon, “Capability Computing,” Panel Discussion at SOS10, Maui, March 6–9, 2006.

Horst Simon, “Let's Design Our Own Petaflops System!” Workshop on Algorithms and Architectures for Petascale Computing, Schloss Dagstuhl, Wadern, Germany, February 13, 2006.

Horst Simon, “Petascale Computing in the U.S.” (presentation) and “The Asian Attack” (panel discussion chair), ISC2006, Dresden, Germany, June 27–30, 2006.

David Skinner, “Integrated Performance Monitoring of Highly Parallel HPC Workloads,” 12th SIAM Conference on Parallel Processing for Scientific Computing, San Francisco, February 21–24, 2006.

David Skinner and William T.C. Kramer, “Understanding the Causes of Performance Variability in HPC Workloads,” 12th SIAM Conference on Parallel Processing for Scientific Computing, San Francisco, February 21–24, 2006.

Wei Xu, Joseph Hellerstein, William T.C. Kramer, and David Patterson, “Control Considerations for Scaling Event Correlation,” 16th IFIP/IEEE Distributed Systems Operations and Management (DSOM 05), Universitat Politècnica de Catalunya, Barcelona, Spain, October 24–26, 2005.

Published Papers and Articles

Krste Asanovíc, Rastislav Bodik, Bryan Catanzaro, Joseph Gebis, Parry Husbands, Kurt Keutzer, David Patterson, William Plishker, John Shalf, Samuel Williams, and Katherine Yelick, *The Landscape of Parallel Computing Research: A View from Berkeley,* University of California at Berkeley Technical Report No. UCB/EECS-2006-183, December 18, 2006 (http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-183.pdf).

Horst Simon, Michael Heroux and Padma Raghavan, eds., *Frontiers of Parallel Processing for Scientific Computing* (SIAM, Philadelphia, 2006).

Horst Simon, William T.C. Kramer, William Saphir, John Shalf, David Bailey, Leonid Oliker, Michael Banda, C. William McCurdy, John Hules, Andrew Canning, Marc Day, Philip Colella, David Serafini, Michael Wehner, and Peter Nugent, Science-Driven System Architecture: A New Process for Leadership Class Computing, Journal of the Earth Simulator, Vol. 2, January 2005.

Kurt Stockinger, E. Wes Bethel, Scott Campbell, Eli Dart, and Kesheng Wu, "Detecting Distributed Scans Using High-Performance Query-Driven Visualization," Proceedings of SC06, Tampa, Florida, November 11–17, 2006.

LBNL/PUB-5503. D. Skinner (2005). Performance monitoring of parallel scientific applications. http://www-library.lbl.gov/docs/PUB/5503/PDF/PUB-5503.pdf | http://www-library.lbl.gov/docs/PUB/5503/PDF/PUB-5503.doc

LBNL/PUB-904 (2006). D. Skinner, F. Verdier, H. Anand, J. Carter, M. Durst, and R. Gerber (2005). Parallel Scaling Characteristics of Selected NERSC User Project Codes. http://www-library.lbl.gov/docs/PUB/904/PDF/PUB-904_2006.pdf | http://www-library.lbl.gov/docs/PUB/904//SRC/PUB-904_2006.doc

LBNL/PUB-960. J. Srinivasan (2006). An Evaluation of the EverGrid Deja Vu Checkpoint/Restart Software.

LBNL-53656. D. H. Bailey and M. Misiurewicz (2005). A Strong Hot Spot Theorem. Proceedings of the American Mathematical Society 134, 2495-2501. http://www-library.lbl.gov/docs/LBNL/536/56/PDF/LBNL-53656_Journal.pdf

LBNL-57369. J. Hules, J. Bashor, L. Yarris, J. McCullough, P. Preuss, and W. Bethel (2005). NERSC Annual Report 2004. http://www-library.lbl.gov/docs/LBNL/573/69/PDF/LBNL-57369.pdf

LBNL-57488. D. H. Bailey and A. Snavely (2005). Performance Modeling: Understanding the Present and Predicting the Future. In Euro-Par 2005, Lisbon, Portugal. http://www-library.lbl.gov/docs/LBNL/574/88/PDF/LBNL-57488.pdf

LBNL-57490. D. H. Bailey and J. M. Borwein (2005). Experimental Mathemataics: Examples, Methods and Implications. Notices of the American Mathematical Society 52 (5), 502-514. http://www-library.lbl.gov/docs/LBNL/574/90/PDF/LBNL-57490.pdf

LBNL-57491. D. H. Bailey and J. M. Borwein (2005). Highly Parallel, High-Precision Numerical Integration. In SC2005, Seattle, WA. http://www-library.lbl.gov/docs/LBNL/574/91/PDF/LBNL-57491.pdf

LBNL-57493. P. Luszczek, J. J. Dongarra, D. Koester, R. Rabenseifner, B. Lucas, J. Kepner, J. McCalpin, D. Bailey, and D. Takahashi (2005). Introduction to the HPC Challenge Benchmark Suite. In SC2005, Seattle, WA. http://www-library.lbl.gov/docs/LBNL/574/93/PDF/LBNL-57493.pdf

LBNL-57500. J. A. Greenough, B. R. de Supinski, R. K. Yates, C. A. Rendleman, D. Skinner, V. Beckner, M. Lijewski, J. Bell, and J. C. Sexton (2005). Performance of a Block Structured, Hierarchical Adaptive Mesh Refinement Code on the 64k Node IBM BlueGene/L Computer. In SC|05 International Conference for High Performance Computing Networking and Storage, Seattle, WA. http://www-library.lbl.gov/docs/LBNL/575/00/PDF/LBNL-57500.pdf

LBNL-57581. D. H. Bailey, J. M. Borwein, and D. M. Bradley (2005). Experimental Determination of Apery-Like Identities for Zeta(2n+2). Experimental Mathematics.

LBNL-57582. H. D. Simon, W. T. C. Kramer, D. H. Bailey, M. J. Banda, E. W. Bethel, J. M. Craw, W. J. Fortney, J. A. Hules, N. L. Meyer, J. C. Meza, E. G. Ng, L. E. Rippe, W. C. Saphir, F. Verdier, H. A. Walter, and K. A. Yelick (2005). Science-Driven Computing: NERSC's Plan for 2006-2010. http://www-library.lbl.gov/docs/LBNL/575/82/PDF/LBNL-57582.pdf

LBNL-57582. Horst D. Simon, William T. C. Kramer, David H. Bailey, Michael J. Banda, E. Wes Bethel, Jonathon T. Carter, James M. Craw, William J. Fortney, John A. Hules, Nancy L. Meyer, Juan C. Meza, Esmond G. Ng, Lynn E. Rippe, William C. Saphir, Francesca Verdier, Howard A. Walter, Katherine A. Yelick (2005). Science-Driven Computing: NERSC's Plan for 2006-2010. http://www-library.lbl.gov/docs/LBNL/575/82/PDF/LBNL-57582.pdf

LBNL-58001. J. Borrill, J. Carter, L. Oliker, D. Skinner, and R. Biswas (2005). Integrated Performance Monitoring of a Cosmology Application on Leading HEC Platforms. In THE 2005 INTERNATIONAL CONFERENCE ON PARALLEL PROCESSING (ICPP-05), edited by L. N. Olav Lynse, IEEE Computer Society, Univ. of Oslo, Norway, 00. http://www-library.lbl.gov/docs/LBNL/580/01/PDF/LBNL-58001.pdf | http://www-library.lbl.gov/

LBNL-58002. J. Carter, M. Soe, L. Oliker, Y. Tsuda, G. Vahala, L. Vahala, and A. Macnab (2005). Magnetohydrodynamic Turbulence Simulations on the Earth Simulator Using the Lattice Boltzmann Method. In Conference on High Performance Computing, Networking and Storage, edited by B. B. Jeff Kuehn, IEEE Computer Society Press, Seattle, WA. http://www-library.lbl.gov/docs/LBNL/580/02/PDF/LBNL-58002.pdf | http://www-library.lbl.gov/docs/LBNL/580/02/SRC/LBNL-58002.doc

LBNL-58004. J. Carter, L. Oliker, and J. Shalf (2005). Performance Evaluation of Plasma Physics and Astrophysics Applications on Modern Parallel Vector Systems. In VECPAR'06 7th International Conference on High Performance Computing for Computational Science, edited by J. D. M. Daydé, V. Hernandez and J.M.L.M. Palma, Springer, Rio de Janeiro, Brazil.

LBNL-58243. R. Ryne, D. Abell, A. Adelmann, J. Amundson, C. Bohn, J. Cary, P. Colella, D. Dechow, V. Decyk, A. Dragt, R. Gerber, S. Habib, D. Higdon, T. Katsouleas, K.-L. Ma, P. McCorquodale, D. Mihalcea, C. Mitchell, W. Mori, C. T. Mottershead, F. Neri, I. Pogorelov, J. Qiang, R. Samulyak, D. Serafini, J. Shalf, C. Siegerist, P. Spentzouris, P. Stoltz, B. Terzic, M. Venturini, and P. Walstrom (2005). SciDAC Advances and Applications in Computational Beam Dynamics. In SciDAC 2005, edited by L. C. Graham Douglas, IOPP, San Francisco, CA.

LBNL-58246. D. H. Bailey and J. M. Borwein (2005). Effective Error Bounds in Euler-Maclaurin-Based Quadrature Schemes. Mathematics of Computation.

LBNL-58247. D. H. Bailey and A. M. Frolov (2005). Positron Annihilation in the Bipositronium Ps2. Physical Review A 72 (014501), 014501(1-4). http://www-library.lbl.gov/docs/LBNL/582/47/PDF/LBNL-58247.pdf

LBNL-58768. K. Stockinger, K. Wu, S. Campbell, S. Lau, M. Fisk, E. Gavrilov, A. Kent, C. E. Davis, R. Olinger, R. Young, J. Prewett, P. Weber, T. P. Caudell, E. W. Bethel, and S. Smith (2005). Network Traffic Analysis With Query Driven VisualizationSC 2005 HPC Analytics Results. In Supercomputing 2005, ACM - Association for Computing Machinery, Seattle, Washingoton. http://www-library.lbl.gov/docs/LBNL/587/68/PDF/LBNL-58768.pdf

LBNL-58868. W. Kramer, J. Shalf, and E. Strohmaier (2005). The NERSC Sustained System Performance (SSP) Metric. http://www-library.lbl.gov/docs/LBNL/588/68/PDF/LBNL-58868.pdf

LBNL-58914. K. Yelick, S. Kamil, W. T. Kramer, L. Oliker, J. Shalf, H. Shan, and E. Strohmaier (2005). Science Driven Supercomputing Architectures: Analyzing Architectural Bottlenecks with Applications and Benchmark Probes. http://www-library.lbl.gov/docs/LBNL/589/14/PDF/LBNL-58914.pdf

LBNL-58927. A. Aspuru-Guzik, V. Batista, E. W. Bethel, J. D. Borrill, J. Chen, P. Colella, D. J. Dean, T. DeBoni, D. Donzis, S. Ethier, A. Friedman, A. Harrison, E. Hawkes, M. Head-Gordon, B. E. Hingerty, S. C. Jardin, Y. Jung, D. Keyes, W. A. J. Lester, F. Lui, M. Mavrikakis, A. Mezzacappa, D. Olson, K. Refson, C. Ren, D. Rotman, R. Ryne, V. Sankaran, C. Sovinec, D. Spong, G. Sposito, G. M. Stocks, R. Sugar, D. Swesty, H. Wang, V. Wayland, and P. K. Yeung (2005). DOE Greenbook: Needs and Directions in High Performance Computing for the Office of Science. A Report from the NERSC User Group.

LBNL-58935. J. Gabler (2006). Better Bonded Ethernet Load Balancing. http://www-library.lbl.gov/docs/LBNL/589/35/PDF/LBNL-58935.pdf | http://www-library.lbl.gov/docs/LBNL/589/35/SRC/LBNL-58935.doc

LBNL-59304. S. Campbell (2005). How to Think About Security Failures. Communications of the ACM 49 (1), 37-39.

LBNL-59340. J. Carter and L. Oliker (2006). Performance Evaluation of Lattice-Boltzmann Magnetohydrodynamics Simulations on Modern Parallel Vector Systems. In 2nd Teraflop Workshop, edited by M. Resch, Springer, Stuttgart, Germany. http://www-library.lbl.gov/docs/LBNL/593/40/PDF/LBNL-59340.pdf

LBNL-59999. W. Kramer, J. Carter, D. Skinner, L. Oliker, P. Husbands, P. Hargrove, J. Shalf, O. Marques, E. Ng, T. Drummond, K. Yelick (2006). Software Roadmap to Plug and Play Petaflop/s.

LBNL-60060. S. Kamil, A. Pinar, D. Gunter, M. Lijewski, L. Oliker, J. Shalf, and D. Skinner (2006). Reconfigurable Hybrid Interconnection for Static and Dynamic Scientific Applications. In International Conference for High Performance Computing, Networking, Storage and Analysis , Tampa, FL. http://www-library.lbl.gov/docs/LBNL/600/60/PDF/LBNL-60060.pdf

LBNL-60166. T. L. Killeen and H. D. Simon (2006). Supporting National User Communities at NERSC and NCAR. Cyberinfrastructure Technology Watch 2 (2). http://www-library.lbl.gov/docs/LBNL/601/66/PDF/LBNL-60166.pdf

9. NERSC Will Be Able to Thrive and Improve in an Environment Where Change Is the Norm.

High-performance organizations that deal with advanced technology must be able to adapt and embrace change as a way of life. HPC centers that are not growing and changing are dying (or have died). Providing reliable cycles is not enough to serve the NERSC users in a time of constant change. Research is needed to ensure that tomorrow's systems are accessible and productive to our users.

Science-Driven Computing

NERSC is continuously reassessing its approach to supporting high-end scientific computing, and from time to time undertakes a major reevaluation and realignment. In 2005 this involved several activities: the NERSC Users' Group writing and publishing the latest DOE Greenbook, the development of NERSC's five-year plan for 2006–2010, and a programmatic peer review conducted for the DOE. The overall theme of this evaluation and planning effort was *Science-Driven Computing.*

DOE Greenbook Published

The *DOE Greenbook: Needs and Directions in High-Performance Computing for the Office of Science* was compiled by Steve Jardin for the NERSC Users Group and published in June 2005 (Figure 17). With contributions from 37 scientists from a variety of disciplines and organizations, this report documents the computational science being done at NERSC and other DOE computing centers and provides examples of computational challenges and opportunities that will guide the evolution of these centers over the next few years.

According to the Greenbook, researchers in all of the disciplines supported by the Office of Science are finding that large-scale computational capabilities are now essential for the advancement of their research. Today's most powerful computers and scientific application codes are being used to produce new and more precise scientific results at the cutting edge of each discipline, and this trend is destined to continue for years to come. The Greenbook presents many examples of the impact of large-scale computations on the sciences.

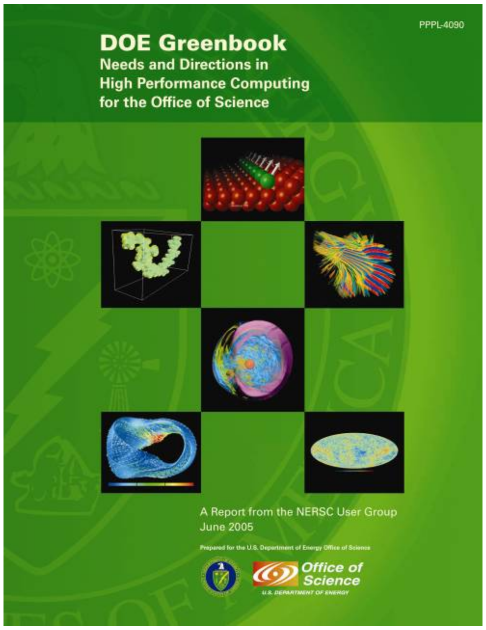
PPPL-4090
DOE Greenbook
Needs and Directions in
High Performance Computing
for the Office of Science
A Report from the NERSC User Group
June 2005
Prepared for the U.S. Department of Energy Office of Science
Office of Science
U.S. DEPARTMENT OF ENERGY

Figure 17. The DOE Greenbook, prepared by the NERSC Users Group, is available online at http://www.nersc.gov/news/greenbook/2005greenbook.pdf.

However, the Greenbook points out that the current computational resources available through NERSC are saturated, and the lack of additional computing resources is becoming "a major bottleneck in the scientific research and discovery process.” The report advises, “A large increase in computer power is needed in the near future to take the understanding of the science to the next level and to help secure the U.S. DOE SC leadership role in these fundamental research areas.”

The Greenbook’s specific recommendations include:

- Expand the high performance computing resources available at NERSC, maintaining an appropriate system balance to support the wide range of large-scale applications involving production computing and development activities in the DOE Office of Science.
- Configure the computing hardware and queuing systems to minimize the time-to-completion of large jobs, as well as to maximize the overall efficiency of the hardware.
- Actively support the continued improvement of algorithms, software, and database technology for improved performance on parallel platforms.
- Significantly strengthen the computational science infrastructure at NERSC that will enable the optimal use of current and future NERSC supercomputers.
- Carefully evaluate the requirements of data- or I/O-intensive scientific applications in order to support as wide a range of science as possible.

NERSC’s Five-Year Plan

With guidance from the DOE Greenbook and other interactions with the NERSC user community, as well as monitoring of technology trends, NERSC developed a five-year plan focusing on three components: Science-Driven Systems, Science-Driven Services, and Science-Driven Analytics (Figure 18). NERSC management and staff identified three trends that need to be addressed over the next several years:

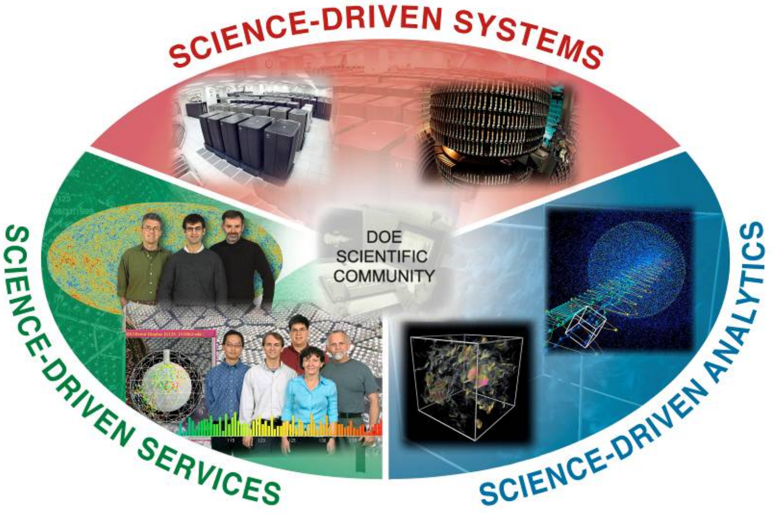

Figure 18. Conceptual diagram of NERSC’s plan for 2006–2010.

- the widening gap between application performance and peak performance of high-end computing systems
- the recent emergence of large, multidisciplinary computational science teams in the DOE research community
- the flood of scientific data from both simulations and experiments, and the convergence of computational simulation with experimental data collection and analysis in complex workflows.

NERSC's responses to these trends are the three components of the science-driven strategy that NERSC will implement and realize in the next five years:

- *Science-Driven Systems:* Balanced introduction of the best new technologies for complete computational systems—computing, storage, networking, visualization and analysis—coupled with the activities necessary to engage vendors in addressing the DOE computational science requirements in their future roadmaps.
- *Science-Driven Services:* The entire range of support activities, from high-quality operations and user services to direct scientific support, that enable a broad range of scientists to effectively use NERSC systems in their research. NERSC will concentrate on resources needed to realize the promise of the new highly scalable architectures for scientific discovery in multidisciplinary computational science projects.
- *Science-Driven Analytics:* The architectural and systems enhancements and services required to integrate NERSC's powerful computational and storage resources to provide scientists with new tools to effectively manipulate, visualize, and analyze the huge data sets derived from simulations and experiments.

This balanced set of objectives will be critical for the future of the NERSC Center and its ability to serve the DOE scientific community. Elements of this strategy that are currently being implemented are discussed in the following pages. The full five-year plan can be read at http://www.nersc.gov/news/reports/LBNL-57582.pdf.

DOE Review of NERSC

On May 17–19, 2005, a Programmatic Review of NERSC was conducted for the Department of Energy. The peer review committee was chaired by Frank Williams of the Arctic Region Supercomputing Center at the University of Alaska, Fairbanks. Other members were Walter F. Brooks of the NASA Advanced Supercomputing Division at NASA Ames Research Center, Lawrence Buja of the National Center for Atmospheric Research, Cray Henry of the Defense Department's High Performance Computing Modernization Program, Robert Meisner of the National Nuclear Security Administration, José L. Muñoz of the National Science Foundation, and Tomasz Plewa of the Center for Astrophysical Thermonuclear Flashes at the University of Chicago.

In addition to reviewing the DOE Greenbook and NERSC's five-year plan, the review panel heard presentations and engaged in conversations covering all aspects of NERSC's operations. The DOE had requested that the panel address a number of specific topics, but they were also given the freedom to look into any aspect of NERSC and to comment accordingly. The panel responded by presenting a detailed list of findings and recommendations to DOE and NERSC managers, who are now using those findings to improve NERSC's operations.

The overall conclusions of the review committee included a strong endorsement of NERSC's approach to enabling computational science:

> NERSC is a strong, productive, and responsive science-driven center that possesses the potential to significantly and positively impact scientific progress by providing users with

> access to high performance computing systems, services, and analytics beneficial to the support and advancement of their science….
>
> Members of the review panel each report that NERSC is extremely well run with a lean and knowledgeable staff. The panel members saw evidence of strong and committed leadership, and staff who are capable and responsive to users' needs and requirements. Widespread, high regard for the center's performance, reflected in such metrics as the high number of publications supported by NERSC, and its potential to positively impact future advancement of computational science, warrants continued support.

Organizational Changes

In order to implement the new initiatives and the changes in emphasis derived from the planning and review process, in November 2005 NERSC announced several organizational changes, including two new associate general managers, two new teams, and a new group.

"In order to efficiently carry out our plan and meet the expectations of our users and sponsors, we are modifying the NERSC Center organization," General Manager Bill Kramer wrote in announcing the changes. "In addition to the Division, Department and Group components of the organization, we will have two other components: Functional Areas and Teams."

NERSC has created two functional areas—Science-Driven Systems and Science-Driven Services. The majority of the NERSC staff will work in these two areas (Figure 19). The functional areas are responsible for carrying out the responsibilities and tasks discussed in the respective sections of NERSC's five-year plan. Functional areas will be led by Associate General Managers (AGMs), who are responsible for coordinating activities across the groups and teams in their areas. Francesca Verdier is associate general manager for Science-Driven Services, and Howard Walter is associate general manager for Science-Driven Systems.

The Accounts and Allocations Team, the Analytics Team, the Open Software and Programming Group, and the User Services Group report to the Science-Driven Services AGM. The Computational Systems Group, the Computer, Operations and ESnet Support Group, the Mass Storage Group, and the Networking, Security and Servers Group report to the Science-Driven Systems AGM.

The reorganization included the creation of one new group and two new teams. They are:

- *Analytics Team:* Analytics is the intersection of visualization, analysis, scientific data management, human-computer interfaces, cognitive science, statistical analysis, and reasoning. The primary focus of the Analytics Team is to provide visualization and scientific data management solutions to the NERSC user community to better understand complex phenomena hidden in scientific data. The responsibilities of the team span the range from applying off-the-shelf commercial software to advanced development to realizing new solutions where none previously existed. The Analytics Team is a natural expansion of the visualization efforts that have been part of NERSC since it moved to Berkeley Lab. Wes Bethel is the team leader. (NERSC's analytics strategy is discussed in more detail in section 5.)

- *Open Software and Programming (OSP) Group:* The growing use of open-source software and partially supported software requires a change of approach to NERSC's needs for the future. These areas now are a key component of NERSC's ability to provide high-quality systems and services. This group is responsible for the support and improvement of open-source and other partially supported software, particularly the software that NERSC uses for infrastructure, operations, and delivery of services. Key efforts include open-source engineering, development and support of middleware (Grid and Web tools), and NERSC's software infrastructure. David Skinner is the leader of the OSP Group.

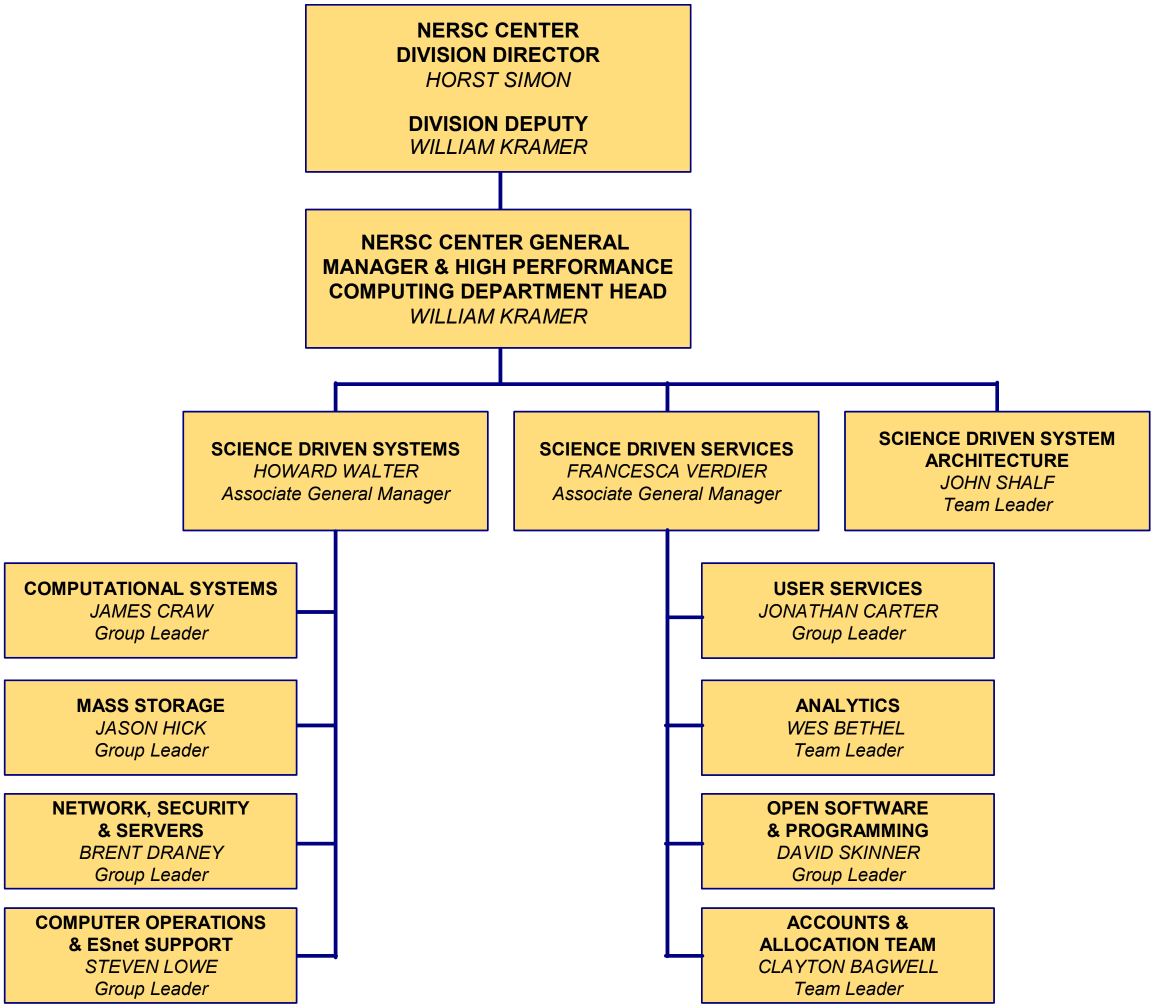

Figure 19. NERSC's new organization reflects new priorities and promotes coordination across groups and teams.

- *Science-Driven System Architecture (SDSA) Team:* This team performs ongoing evaluation and assessment of technology for scientific computing. The SDSA Team has expertise in benchmarking, system performance evaluations, workload monitoring, use of application modeling tools, and future algorithm scaling and technology assessment. Using scientific methods, the team will develop methods for analyzing possible technical alternatives and will create a clear understanding of current and future NERSC workloads. The SDSA Team will engage with vendors and the general research community to advocate technological features that will enhance the effectiveness of systems for NERSC scientists. The team is responsible for ongoing management of a suite of benchmarks that NERSC and Berkeley Lab use for architectural evaluation and procurement. This includes composite benchmarks and metrics such as SSP, ESP, variation, reliability, and usability. The team will matrix staff from both NERSC and Berkeley Lab's Computational Research Division for specific areas such as algorithm tracking and scaling, which are designed to develop and document future algorithmic requirements. The scientific focus for this effort will change periodically and will start with applied mathematics and astrophysics. The SDSA Team leader is John Shalf.

Completing the reorganization, Jonathan Carter succeeded Francesca Verdier as leader of the User Services Group, and Brent Draney succeeded Howard Walter as leader of the Networking, Security and Servers Group.

Science-Driven System Architecture

The creation of NERSC's Science-Driven System Architecture (SDSA) Team formalizes an ongoing effort to monitor and influence the direction of technology development for the benefit of computational science. NERSC staff collaborate with scientists and computer vendors to refine computer systems under current or future development so that they will provide excellent sustained performance per dollar for the broadest possible range of large-scale scientific applications.

While the goal of SDSA may seem ambitious, the actual work that promotes that goal deals with the nitty-gritty of scientific computing—for example, why does a particular algorithm perform well on one system but poorly on another—at a level of detail that some people might find tedious or overwhelming, but which the SDSA team finds fascinating and challenging.

"All of our architectural problems would be solvable if money were no object," said SDSA Team Leader John Shalf, "but that's never the case, so we have to collaborate with the vendors in a continuous, iterative fashion to work towards more efficient and cost-effective solutions. We're not improving performance for its own sake, but we are improving user effectiveness."

Much of the SDSA work involves performance analysis: how fast do various scientific codes run on different systems, how well do they scale to hundreds or thousands of processors, what kinds of bottlenecks can slow them down, and how can performance be improved through hardware development. A solid base of performance data is particularly useful when combined with workload analysis, which considers what codes and algorithms are common to NERSC's diverse scientific workload. These two sets of data lay a foundation for assessing how that workload would perform on alternative system architectures. Current architectures may be directly analyzed, while future architectures may be tested through simulations or predictive models.

The SDSA Team is investigating a number of different performance modeling frameworks, such as the San Diego Supercomputer Center's Memory Access Pattern Signature (MAPS), in order to assess their accuracy in predicting performance for the NERSC workload. SDSA team members are working closely with San Diego's Performance Modeling and Characterization Laboratory to model the performance of the NERSC-5 SSP benchmarks and compare the performance predictions to the benchmark results collected on existing and proposed HPC systems.

Seemingly mundane activities like these can have an important cumulative impact: as more research institutions set specific goals for application performance in their system procurement specifications, HPC vendors have to respond by offering systems that are specifically designed and tuned to meet the needs of scientists and engineers, rather than proposing strictly off-the-shelf systems. By working together and sharing performance data with NERSC and other computer centers, the vendors can improve their competitive position in future HPC procurements, refining their system designs to redress any architectural bottlenecks discovered through the iterative process of benchmarking and performance modeling. The end result is systems better suited for scientific applications and a better-defined niche market for scientific computing that is distinct from the business and commercial HPC market.

The SDSA Team also collaborates on research projects in HPC architecture. One key project, in which NERSC is collaborating with Berkeley Lab's Computational Research Division and computer vendors, is ViVA, or Virtual Vector Architecture. The ViVA concept involves hardware and software enhancements that would coordinate a set of commodity scalar processors to function like a single, more powerful vector processor. ViVA would enable much faster performance for certain types of widely used scientific algorithms, but without the high cost of specialized processors. The research is proceeding in phases. ViVA-1 is focused on a fast synchronization register to coordinate processors on a node or multicore chip. ViVA-2 is investigating a vector register set that hides latency to memory using vector-like semantics. Benchmark scientific kernels are being run on an architectural simulator with ViVA enhancements to assess the effectiveness of those enhancements.

Another research collaboration is the RAMP Project (Research Accelerator for Multiple Processors), which focuses on how to build low cost, highly scalable hardware/software prototypes, given the increasing difficulty and expense of building hardware. RAMP is exploring emulation of parallel systems via field programmable gate arrays (FPGAs). Although FPGAs are slower than real hardware, they are much faster than simulators, and thus can be used to evaluate novel ideas in parallel architecture, languages, libraries, and so on.

SDSA Team members John Shalf and Kathy Yelick are two of the co-authors of a white paper called "The Landscape of Parallel Computing Research: A View from Berkeley." Based on two years of discussions among a multidisciplinary group of researchers, this paper addresses the challenge of finding ways to make it easy to write programs that run efficiently on manycore systems. The creation of manycore architectures — hundreds to thousands of cores per processor — demands that a new parallel computing ecosystem be developed, one that is very different from the environment that supports the current sequential and multicore processing systems. Since real-world applications are naturally parallel and hardware is naturally parallel, what is needed is a programming model, system software, and a supporting architecture that are naturally parallel. Researchers have the rare opportunity to re-invent these cornerstones of computing, provided they simplify the efficient programming of highly parallel systems. The paper provides strategic suggestions on how to accomplish this (see http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-183.pdf).

Perhaps the most ambitious HPC research project currently under way is the Defense Advanced Research Projects Agency’s (DARPA’s) High Productivity Computer Systems (HPCS) program. HPCS aims to develop a new generation of hardware and software technologies that will take supercomputing to the petascale level and increase overall system productivity ten-fold by the end of this decade. NERSC is one of several “mission partners” participating in the review of proposals and milestones for this project.

10. Improve the Effectiveness of NERSC Staff by Improving Infrastructure, Caring for Staff, Encouraging Professionalism and Professional Improvement

Every employee has a stake in the success of NERSC, and management encourages staff to contribute their ideas for helping the organization succeed. NERSC looks for and implements new ways to increase staff effectiveness. This leads to NERSC being able to support more activities and innovations.

Individual and Team Recognition

To recognize individual and group contributions to the success of our organization, NERSC honors employees with both "Spot Awards" and Outstanding Performance Awards. The Spot Awards program was developed by the Laboratory to provide "on the spot" recognition with a cash award and certificate. The Outstanding Performance Awards are typically presented to employees for exemplary performance outside the scope of their usual responsibilities.

During 2005 and 2006, NERSC presented Spot Awards to Harsh Anand, Will Baird, Nick Balthaser Elizabeth Bautista, Scott Campbell, Shane Canon, Nicholas Cardo, Eli Dart, Brent Draney, Aaron Garrett, Damian Hazen, Rusty Huie, John Hules, Wayne Hurlbert, Bill Iles, Stephen Lau, Bob Neylan, Ken Okikawa, David Paul, Tony Quan, David Skinner, Gary Smith, Jay Srinivasan, Tavia Stone, David Turner, and Steve Warner.

Outstanding Performance Awards were given to the following employees for their work on special projects: David Paul for technical leadership at the SC2004 conference; David Skinner for consulting support to the Sandia INCITE project; the NERSC strategic review and planning team (Francesca Verdier, Howard Walter, James Craw, Jonathan Carter, Nancy Meyer); the NCS-b procurement selection team (Tina Butler, Richard Gerber); the NSF Terascale Facilities proposal team (John Shalf, David Skinner, Nicholas Cardo, Jonathan Carter, and Howard Walter).

Staff Web Site

To help facilitate the professional exchange of ideas and information, NERSC has created a password-protected Web site for staff only, where they can share information and expertise; create and manage documents; develop, manage, and collaborate on projects; and review Berkeley Lab and NERSC policies and procedures. The Web site was created using TWiki, a flexible, powerful, secure, yet simple collaboration platform. The NERSC TWiki expanded rapidly and has become an essential part of the organization's operations, improving communications and project management.

Berkeley Lab Citizenship

Although NERSC is a national user facility, the center is also integrated into the fabric of the Laboratory. The NERSC staff complies with all Environmental, Health and Safety programs of the Lab and actively participates in the Computing Sciences Safety Committee. NERSC staff also are members of Lab-wide committees, such as Jonathan Carter serving on the Information Technology Advisory Committee.

CONCLUSION

Today NERSC users have the benefit of scalable high-end capability computing in Seaborg and the soon-to-be-added Franklin, along with reliable capacity machines, in an integrated environment that offers a global filesystem, analytics support, and a seemingly infinite mass storage system (now close to 40 petabytes). Our five-year plan will move this integrated system environment to the petaflop/s performance level, which we will reach in 2010 with the planned introduction of NERSC-6. Scalability to tens of thousands of processors, both for applications and systems software, managing petabytes of data, and at the same time continuing the excellent support, reliability, and quality of service, will be the big challenges ahead. Thanks to the ongoing support from our program management at the Office of Advanced Scientific Computing Research at DOE, the NERSC budget plan has been set at a level that makes these ambitious plans feasible. Thus we are confidently looking forward to continuing scientific and computing accomplishments at NERSC.

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.