LBNL-51725 (2001)

# How Are We Doing?

## A Self-Assessment of the Quality of Services and Systems at NERSC (2001)

William T. Kramer

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

September 2002

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-AC03-76SF00098.

This report was compiled by John Hules (JAHules@lbl.gov) and Jon Bashor (JBashor@lbl.gov). To learn more about the National Energy Research Scientific Computing Center, visit our web site at: http://www.nersc.gov.

## CONTENTS

## EXECUTIVE SUMMARY

This is the fifth annual 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 for 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 THE PAST YEAR

In 2001, the senior management of NERSC responded to a request from DOE's Office of Science to prepare a strategic proposal outlining directions and goals for the fiscal years 2002-2006. The proposal was endorsed by DOE and NERSC is now implementing these strategies to continue its leadership role among the world's high-performance computing centers. Taking stock of current practices to make improvements for the future has been a hallmark of NERSC. Since moving to Berkeley Lab, NERSC has annually produced this self-assessment to provide statistical and anecdotal evidence of our efforts. Now in its fifth year, "How Are We Doing?" again assesses where the NERSC center is heading and how its staff is striving to achieve the goals guiding the center's operation and evolution.

NERSC began the year anticipating the arrival of a 2,568-processor IBM RS/6000 SP supercomputer, a system that would constitute the world's most powerful unclassified supercomputer, offering a theoretical peak performance level of 3.8 teraflops. In fact, when the SP went on line mid-year, the system had been expanded to 3,328 processors with a peak speed of 5 teraflops. While this performance level is important, even more compelling is the quality of science the system enables. As part of the acceptance testing for the new IBM SP, NERSC invited a small group of key users to put the system through its pacing by running very demanding codes scaled to take maximum advantage of the system.

While these grueling test runs were being processed, Berkeley Lab officially dedicated the Oakland Scientific Facility housing the IBM and NERSC's other high-performance computing and data storage systems. This dedication of the facility and the acceptance of the SP, followed by the strategic proposal, puts NERSC on firm foundation for building a high-performance computing center for 21$^{st}$ century science.

As with most construction projects, it's the skills, knowledge and abilities of the workforce that determine the quality of the end result. NERSC's dedicated staff, with their depth and expertise, are critical to the success of the center and its users. Unlike many building projects, however, there is no "end" to the efforts at NERSC. The center is constantly being improved, with new ideas and technologies put in place on a regular basis.

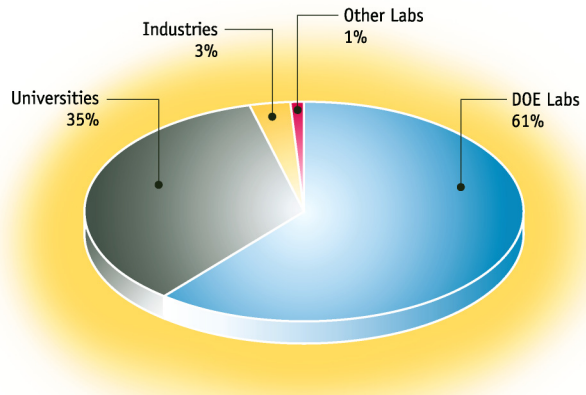
While making these advances, the NERSC staff continues to provide our user community with highly rated support services and with an unmatched availability of computing resources. Every year, an annual survey of users shows that the quality of services provided by NERSC gets better and better. In addition to this survey, NERSC has also established a series of related goals and annually assess our performance against them to ensure that our staff remains focused on meeting the needs of NERSC and advancing computational science in supporting DOE's mission areas. This report, the fifth in a series, describes how the NERSC staff is working to achieve these goals and the overall objective of providing unparalleled systems and services to the scientific community.

## A Statistical Snapshot 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.

### By Type of Institution

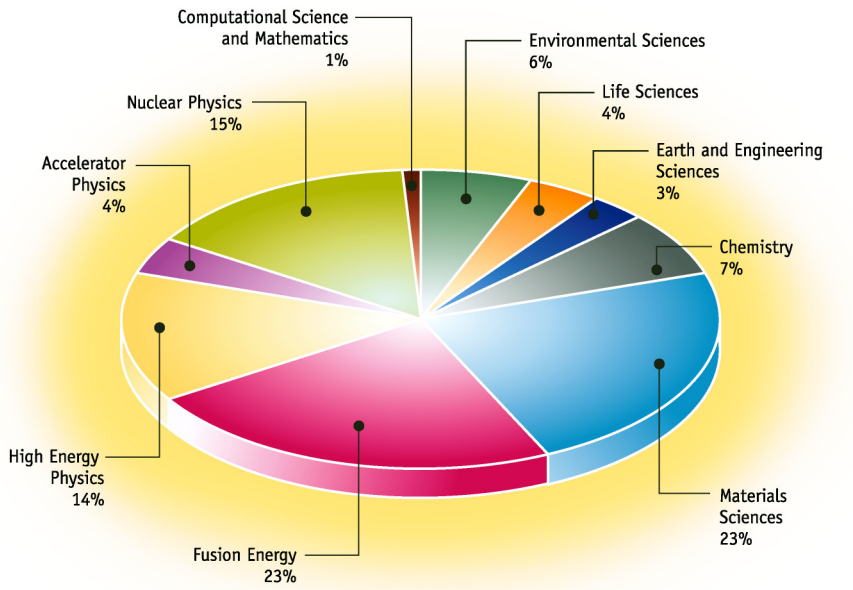
NERSC predominantly serves users at DOE national laboratories, which account for 61 percent of the center's MPP use. Universities account for 35 percent of MPP use, other labs claim 1 percent and use by industry is about 3 percent. These figures are based on FY01 use.




***NERSC MPP usage by institution type, FY01.***

### By Scientific Discipline

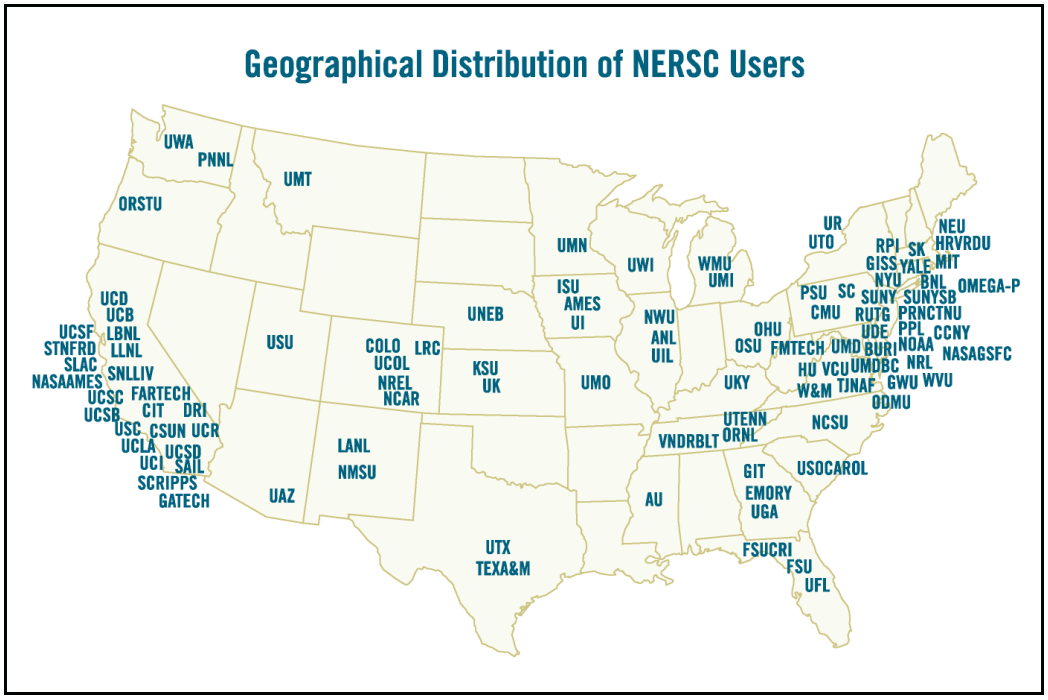
NERSC supports research across the scientific spectrum of programs in DOE's Office of Science. The breakdown of the total usage of NERSC's MPP systems by scientific disciplines shows Computational Science and Mathematics, 1 percent; Environmental Sciences, 6 percent; Life Sciences, 4 percent; Earth and Engineering Sciences, 3 percent; Chemistry, 7 percent; Materials Sciences, 23 percent; Fusion Energy, 23 percent; High Energy Physics, 14 percent; Accelerator Physics, 4 percent; and Nuclear Physics, 15 percent.




***NERSC MPP usage by scientific dicipline, FY01.***

## Truly a National Resource

The user community served by NERSC goes well beyond the national laboratories operated by DOE. NERSC now serves users in 36 states and the District of Columbia, as shown in the map below.



## A Leader by Any Measure

We believe that when evaluated either as an organization supporting the Office of Science or compared with our counterparts supporting researchers funded by other federal and state agencies, NERSC will clearly stand out as a leading force in the world of high-performance computing.

## 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 computational and storage systems within six months of a system going into full service.*

For the systems NERSC provides, service is assessed regarding availability, mean time between interruptions and mean time to repair computational and storage systems within 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 2000 user survey summed up the results of our efforts by saying the NERSC "Provides excellent computing resources with high reliability and ease of use." Table 1 shows how we're achieving that goal.

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. To ensure our systems are available, NERSC has set a goal of high availability on both a scheduled and overall basis, and as the chart above shows, we are exceeding our goal in both

**Table 1**
**System Metrics for FY01 FY01 Goal (as measured in FY01)**

| System | % Availability | | * MTBI* Hours | MTTR** Hours |
|---|---|---|---|---|
| | Scheduled | Overall | | |
| Vector | 98.0 (97.21) | 97 (96.57) | 281 | 9.00 |
| Parallel (T3E/SP) | 96.0 (97.49) | 95.0 (96.97) | 241 | 7.90 |
| Storage | 98.0 (98.21) | 97.0 (97.45) | 170 | 4.20 |
| File Servers | 99.5 (99.97) | 97 (99.97) | 8760 | N/A |
| Math/Vis | 99.92 | 99.73 | 296 | 4.03 |

* Mean Time Between Interruptions: Total wall clock hours/total number of downtime periods
** Mean Time to Restoral: Total downtime hours/total number of downtime periods

areas. Here are the system metrics definitions we use in setting our goals and 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. In NERSC's 24 × 7 environment, 100 percent availability for FY01 would be 8,760 hours.

$$\frac{\text{available hours} - \Sigma \text{ unscheduled outages and scheduled downtime}}{\text{available hours}}$$

- A service interruption is any event or failure (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.
- If an outage occurs within two hours of the system being restored to service, it is treated as one continuous outage.

### NERSC Achieves 95 Percent Utilization on Cray T3E

In February 2001, staffers from NERSC's Computational Systems and User Services groups and Cray Inc. achieved an elusive milestone—exceeding 95 percent utilization (averaged over 30 days) of the Cray T3E supercomputer. The highest percentage reached was 95.34 and would likely have been higher, had not the demolition of a nearby building in downtown Oakland gone awry and brought down power lines with the damaged structure. Those contributing to the successful effort were Tina Butler of the Computational Systems Group, Jonathan Carter and Therese Enright of the User Services Group, and Bryan Hardy, Terence Brewer and Bill Contento of Cray Inc.

## 2. Client Support Goals

*The end measure of a site 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.*

### Balancing System Utilization and Response Time

Over the years, NERSC has been a pioneer in achieving high utilization on massively parallel systems—in simple terms, keeping as many processors as possible working for as long as possible. High utilization, made possible by customized scheduling and load balancing software, gives researchers more available processor hours and maximizes the value of the computing resource. NERSC's Cray T3E reached 95% utilization (averaged over 30 days) in February 2001; and taking advantage of our experience with the T3E, we were able to get 85% utilization on the IBM SP, better than the T3E's first year.

An occasional side effect of high utilization—and an undesirable one from the user's viewpoint—is a long wait for a job to run, if that particular job is not the right size for the system's current load balance. Because response time is an important factor in researchers' productivity—and scientific productivity is the ultimate measure of NERSC's success—we are now working with a user committee to find the optimum balance of utilization and response time. The committee is researching solutions and developing guidelines to be implemented on NERSC systems. Issues that have been resolved to date include:

- Guaranteed throughput for a selected group of projects—for example, providing a mechanism to allow periods of nearly 24 × 7 run time to allow a climate modeling project to utilize their large allocation.
- Improving debugging and interactive turnaround during prime time. Now 5% of the compute pool is set aside for interactive and debugging jobs weekdays from 5 a.m. to 6 p.m. (Pacific time).
- Implementing "priority aging" on Regular class jobs. Now, after 36 hours of being queued, Regular jobs cannot be preempted by newer Premium class jobs. A new queue priority algorithm considers class priority, user priority, and the time the job entered the queue, multiplied by appropriate weighting factors, to achieve this result.

The following figures show IBM SP utilization statistics for the full IBMSP showing overall utilization (Figure 1) and on job size (Figure 2).

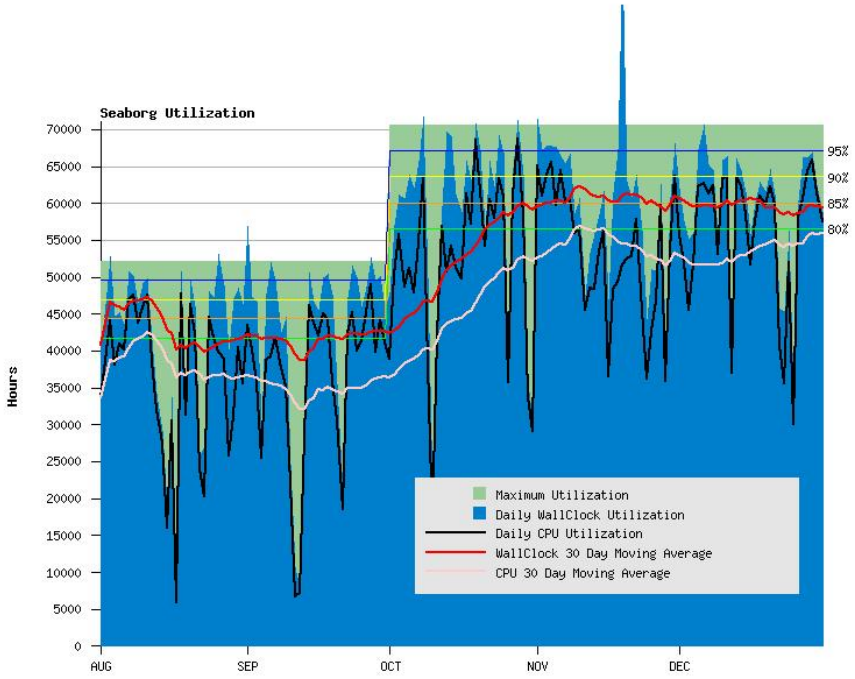



**Figure 1. IBM SP overall utilization, August through December 2001.**

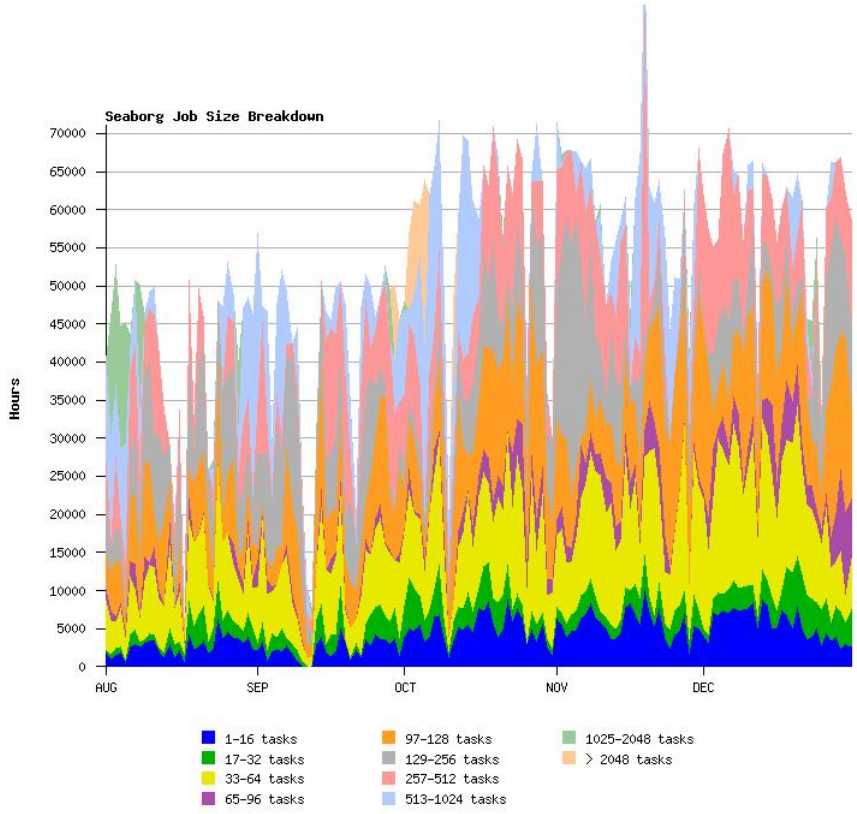



**Figure 2. IBM SP utilization by job size, August through December 2001.**

### Improving End-to-End Network Performance

NERSC's Networking and Security Group is concerned with day-to-day performance. The networking staff monitor traffic for signs of poor performance, and they proactively work with remote users to improve end-to-end rates. For example, they traced slow data transfers from Brookhaven National Laboratory and the Jet Propulsion Laboratory to a bug in the respective labs' firewalls and then worked with Cisco Systems to fix the bug. This kind of creative problem solving has resulted in 4 to 30 times faster data transfer rates from some user sites.

### Changes Resulting from User Survey Feedback

Every year NERSC institutes changes based on the user survey. Some of the changes resulting from the FY 2000 survey are:

- We increased the SP home inode and disk quotas as well as the SP scratch space. SP disk configuration satisfaction was higher this year and only one user requested more inodes on this year's survey.
- Last year one of the two top SP issues was that the "SP is hard to use." Based on comments we received in last year's survey we wrote more SP web documents and made changes to the user environment. This year only 12% (compared with 25% last year) of the comments reflected that the SP is hard to use.
- We added resources to the T3E pe512 queue and created a new long64 queue: satisfaction with T3E turnaround time improved this year.
- Last year we moved PVP interactive services from the J90 to the SV1 architecture and provided more disk resources. Overall PVP satisfaction was rated higher in this year's survey.

The FY2001 survey had 227 respondents, the most ever (more than 10% of users). The responses represented all five DOE Science Offices, nine national laboratories and 13 universities. There were significant increases in user satisfaction in the available computing hardware, the allocations process, and the PVP cluster. Other areas showing increased satisfaction were T3E and SP batch wait times, SP disk configuration, SP Fortran compilers, and HPSS. Areas with continuing high user satisfaction included HPSS reliability, performance, and uptime; consulting responsiveness, quality of technical advice, and follow-up; Cray programming environment; PVP uptime; and account support.

When asked what NERSC does well, some respondents pointed to our stable and well managed production environment, while others focused on NERSC's excellent support services. Other areas singled out include well done documentation, good software and tools, and the mass storage environment. When asked what NERSC should do differently, the most common responses were to provide more hardware resources and to enhance our software offerings. Other areas of concern were visualization services, batch wait times on all platforms, SP interactive services, training services, and SP performance and debugging tools.

There were many favorable comments about NERSC's client support in the user survey. For example:

- [NERSC] “listens to users, and tries to set up systems to satisfy users and not some managerial idea of how we should compute.”
- “I have also found the user support staff to be very helpful and responsive (I particularly appreciate how rapidly they respond to both e-mails and phone calls.)”
- “Provides computing resources in a manner that makes it easy for the user. NERSC is well run and makes the effort of putting the users first, in stark contrast to many other computer centers.”
- “They are always there to help you out.”
- “Providing first rate consulting services. Providing first rate training.”
- “NERSC responds well to users’ needs.”
- “Very good web site with well arranged information.”
- “The web page, hpcf.nersc.gov, is well structured and complete. Also, information about scheduled down times is reliable and useful.”
- “Training and information on web pages are excellent.”
- “Web management, especially NIM, is great advantage for the users.”
- “PDSF is as close to a perfectly run facility as I have ever experienced. Clear strategic planning, highly competent technical support, intelligent management. Don’t change it.”
- “Excellent software maintenance.”
- “Choice of libraries is very, very good.”

Of course, there is always room for improvement. The following improvements in client support were suggested:

- “More debugging and optimization support for MPP platforms like seaborg.”
- “Better maintained NERSC online consulting answers page — perhaps a more comprehensive FAQ type page.”
- “Batch queue structure improved or explained in more detail.”
- “Better indexing of the sprawling website. Finding, e.g. compiler options or queue limits takes some knowledge.”
- “Better organization of web based documentation and tutorials.”
- “Orient webpage more towards beginners to supercomputing with detailed discussion of issues such as optimization etc.”
- “Training for the users far from the site would be beneficial.”
- “I’d like to see better online training tools.”
- “More training classes so that I can effectively use my NERSC time. I’m constantly worried that I’m wasting MPP time with memory leaks, code inefficiencies and the like.”
- “How about having some of NERSC’s people involved in profiling and performance tuning of some of the major codes running on Seaborg?”

- “Sort out the mess of different home user space on the SP3 and mcurie.”
- “The NERSC allocation process needs to be streamlined. The ERCAP form asks for too much overlapping information.”
- “Sort out the mess where I need to remember 3 or 4 different passwords and change them at different times on different machines.”

When such surveys show high user satisfaction, there can sometimes be a tendency to ease up and rest on one’s laurels. NERSC views the results as meaningful input for making a good thing even better, and reports the resulting changes back to its users.

## 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.*

### IBM SP Acceptance

The goals of DOE computational science projects, and the system requirements that derive from those projects, compel NERSC to remain one of the top 10 most powerful computational facilities in the world. To satisfy this expectation, NERSC is an early implementer of advanced computational systems. Our Best Value acquisition strategy typically results in the selection of a system that is just coming into production, and NERSC is often the first site, or one of the first sites, to install the new technology.

Before NERSC accepts the system, it undergoes an extensive period of configuration, customization, fine-tuning, troubleshooting, debugging and testing to ensure that it meets the performance standards specified in the contract and, more importantly, that it will satisfy the needs of the computational scientists who use it.

The NERSC-3 IBM SP system, which incorporated IBM's newest processor and interconnect technology, was installed in two phases. Phase 1 was the first implementation of the 64-bit POWER3 microprocessor, with two processors per node. Phase 2, the first new system installed in the Oakland Scientific Facility, utilized 16-CPU SMP nodes with enhanced POWER3+ microprocessors and a unique 512-way double/single switch configuration. During acceptance testing, NERSC and IBM staff worked together to solve more than 40 bugs, ranging from hardware to microcode in the switches to compilers.

NERSC tested the new system's functionality with the Sustained System Performance (SSP) test suite, which estimates the amount of scientific computation that can really be delivered over a time period by using actual scientific applications in a realistic production mode. SSP is conservative, so most applications actually do better. The NERSC-3 contract specified more than 40 performance and 150 functional requirements that stem from the application workload that exists at NERSC. Almost every area of function and performance is covered by overlapping but independently relevant measures. For example, there are two disk I/O measures, one for large parallel files and another for small parallel and serial files. Both metrics are necessary for an accurate estimate of the true performance for scientists.

Because the NERSC-3 contract was based on specific, firm performance metrics and functional requirements rather than on peak performance or hardware specifications, the Phase 2 IBM SP system has 22% more processing power and 50% more memory than originally planned. Figure 3 shows the SSP performance results, and Tables 2 and 3 show how the system evolved from the original plan to its final configuration. The final, long-term result is 10 million additional processor hours per year available for the research community (Figure 4).

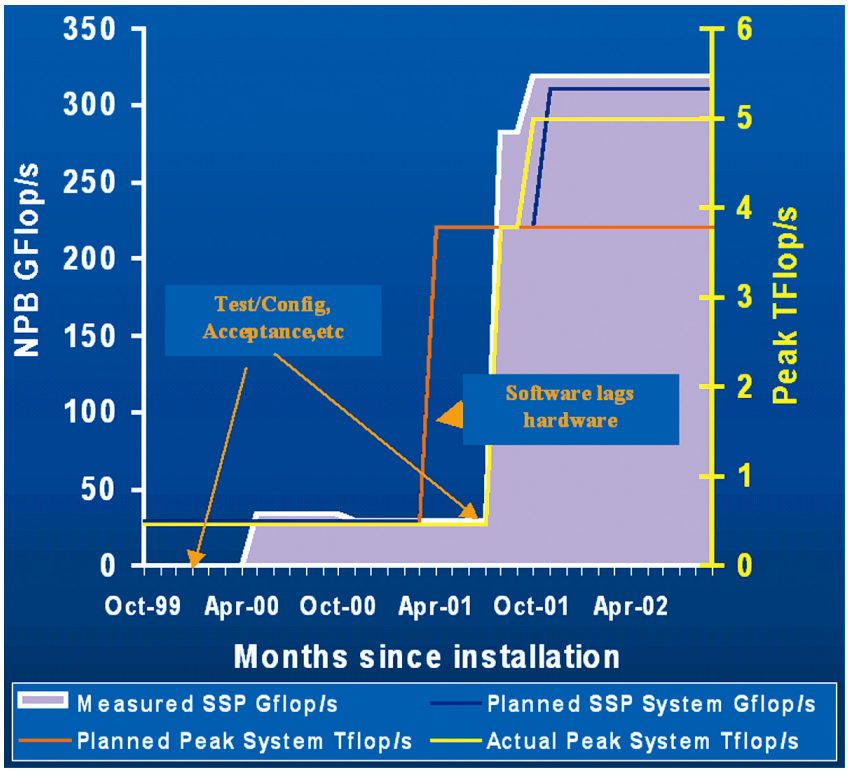



**Figure 3. Peak vs Sustained System Performance**
**(SSP = measured performance × time)**

**Table 2**
**NERSC-3 Original Plan**

| | Phase 1 | Phase 2a/b |
|---|---|---|
| **Compute Nodes Processors** | 256<br>256 × 2 = 512 | 134*<br>134 × 16 = 2144* |
| **Networking Nodes** | 8 | 2 |
| **Interactive Nodes** | 8 | 2 |
| **GPFS Nodes** | 16 | 16 |
| **Service Nodes** | 16 | 4 |
| **Total Nodes (CPUs)** | 304 (604) | 158 (2528) |
| **Total Memory (compute nodes)** | 256 GB | 1.6 TB |
| **Total Global Disk (user accessible)** | 10 TB | 20 TB |
| **Peak (compute nodes)** | 409.6 Gflop/s | 3.2 Tflop/s* |
| **Peak (all nodes)** | 486.4 Gflop/s | 3.8 Tflop/s* |
| **Sustained System Performance** | 33 Gflop/s | 235+ Gflop/s / 280+ Gflop/s |
| **Production Dates** | April 1999 | April 2001 / Oct 2001 |
| * Minimum — may increase due to Sustained System Performance measure | | |

**Table 3**
**NERSC-3 Evolution**

| | Initial Expectations | RFP Contract | Actual System |
|---|---|---|---|
| **Peak Performance** | 4 × T3E ~ 2 Tflop/s | 3.8 Tflop/s | 5 Tflop/s |
| **Computational CPUs** | Twice T3E ~ 1300 | 2048 | 188* × 16 = 3008 |
| **Memory** | Twice T3E ~ 360 GB<br>512 MB/CPU | 1.8 TB<br>758 MB/CPU | 4.5 TB*<br>1.4 GB/CPU |
| **Disk** | 4 × T3E = 10 TB | 32 TB | 35 TB |
| **Schedule** | | | |
| **Initial Service (Phase 1)** | FY 00 | FY 00 | FY 00 |
| **Final Service (Phase 2b)** | FY 01 | FY 01 | FY 01 |
| **Allocation Increase** | 4 × T3E ~ 20M | 35 M | 45 M |
| * 16 Nodes and 1.28 TB of memory purchased in addition to base contract | | | |

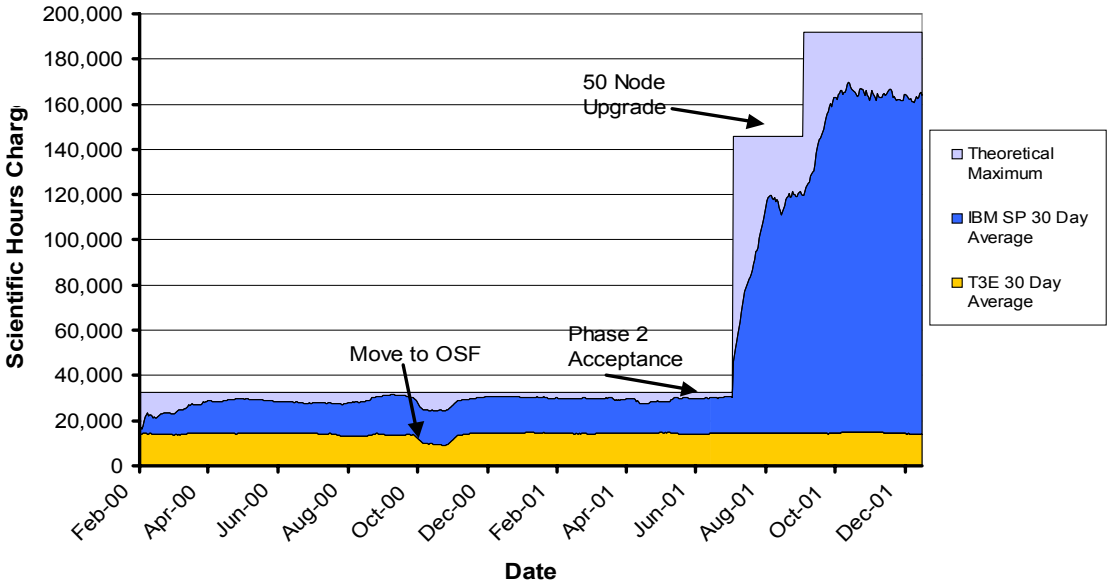



**Figure 4. NERSC MPP usage.**

At the end of acceptance testing on the expanded system, IBM had exceeded the performance and functional requirements on the specifications. NERSC also purchased some additional hardware in order to make the system even more suitable for certain applications. By the time the system was opened up for general production use, it was running so well that significant scientific results in several fields were produced in just the first few weeks (see http://www.nersc.gov/news/N3acceptance100801.html), and users expressed a high level of satisfaction with the system.

### Probing New Storage Technologies

Two years ago, NERSC and Oak Ridge National Laboratory established the Probe wide-area distributed-storage testbed to advance the performance and utility of storage systems by experimenting with, developing, and testing new storage technologies, equipment, and utilities. The goals of the Probe collaboration include providing higher-speed access to storage, making data more widely available on both local and wide area networks (including Grids), and catching the flood of data from theoretical studies and experiments.

In 2001, NERSC used Probe to test a new, faster version of HSI (the HPSS interface utility) that enables access to multiple HPSS systems, making distributed mass storage simpler and more convenient. A Grid-capable FTP daemon was also tested as another way to bring Grid connectivity to HPSS. In addition, NERSC experimented with remote WAN network movers as a way to bypass conventional file transfer methods and stream data quickly between sites. All of these efforts, together with the GUPFS project (described below under Goal 4), will help bring mass storage and Grid technologies together.

See <http://hpcf.nersc.gov/storage/hpss/probe/> for more information.

## 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.*

### Move to the Oakland Scientific Facility

In October 2000, Berkeley Lab took possession of its new state-of-the-art Oakland Scientific Facility (OSF). This site is capable of housing the world's most powerful unclassified computing and data storage systems. Between October 26 and November 3, 2000, NERSC relocated all of its hardware systems (except the IBM SP Phase 1 system, which remained in Berkeley) to the new facility without a complete interruption of service and with little, if any, impact on our clients' work. In fact, every system went online ahead of schedule (Table 4).

The new 16,000-square-foot OSF computer room was designed for flexibility and expandability. The computer room can easily be expanded to 20,000 square feet by removing a non-load-bearing wall, resulting in a facility able to accommodate several new generations of computing systems. Berkeley Lab also has an option for an additional 20,000-square-foot computer room.

### Providing Fast Network Connections

To connect our new facilities and systems to the Grid, NERSC has implemented major upgrades to our networking infrastructure. At the end of May, the Oakland Scientific Facility upgraded its ESnet connection to OC-12 (622 Mb/s) to accommodate the growing demands of data-intensive computing. NERSC's sustained usage rate of the ESnet connection is about one-third, with significant peaks over one-half of the capacity. Gigabit Ethernet utilizing Jumbo Frames (9 KB packets) is the new standard for our internal network connecting the IBM SP and HPSS systems; when our current Cray systems are replaced, older HIPPI and FDDI connections will also be phased out. Ongoing network upgrades are being planned for the future to anticipate the ever-increasing bandwidth needs of our users. The Networking and Security Group works with clients, vendors, and ESnet staff to troubleshoot end-to-end problems and provide point-to-point network tuning, often resulting in dramatic improvements in data transfer rates:

**Table 4**
**System Online Times After Move to OSF**

| System | Scheduled | Actual |
|---|---|---|
| SP | 10/27 – 9 am | no outage |
| T3E | 11/3 – 10 am | 11/3 – 3 am |
| SV1's | 11/3 – 10 am | 11/2 – 3 pm |
| HPSS | 11/3 – 10 am | 10/31 – 9:30 am |
| PDSF | 11/6 – 10 am | 11/2 – 11 am |
| Other Systems | 11/3 – 10 am | 11/1 – 8 am |

- from 1 MB/s to 12 MB/s between NERSC and Oak Ridge National Laboratory
- from 600 KB/sec to 5 MB/s between NERSC and the University of Washington
- from less than 250 KB/s to 1 MB/s between NERSC and the Jet Propulsion Laboratory
- from less than 200 KB/s to 1 MB/s between NERSC and Brookhaven National Laboratory

### Ongoing Storage Upgrades

The amount of data stored at NERSC has been doubling every year for the last three years (Figures 5 and 6 show statistics for 1998–2001). To stay ahead of the growing needs of our clients and the growing capacity of our computational systems, NERSC works constantly to increase the capacity, bandwidth, and functionality of our HPSS mass storage system. This year NERSC began increasing our disk cache to 20 terabytes (TB) by adding more Fibre Channel disks. Internal data transfer rates were improved by the replacement of our HIPPI internal storage network with Jumbo Gigabit Ethernet. Doubling the number of high-capacity Fibre tape drives and adding 33% more tape slots expanded our archive from 1.3 to 2.5 petabytes (PB). And AFS was upgraded with a bigger disk cache and faster processors.

The numbers tell the story of mushrooming data:

- There are 12 million files in the storage systems, compared with 3 million 42 months ago.
- There are 331 TB of active data in the storage systems, compared with 33 TB 42 months ago.
- An average of 30,000 files totaling 1.5 TB of data are transferred to or from NERSC storage every day.
- There are 1,500 tape mounts per day.

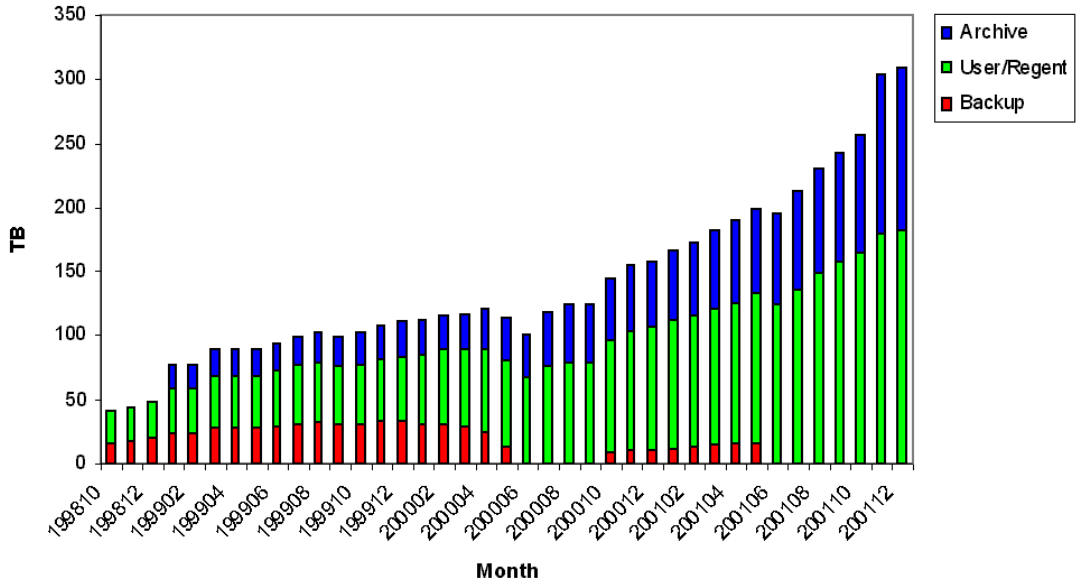



**Figure 5. Cumulative storage by month and system**

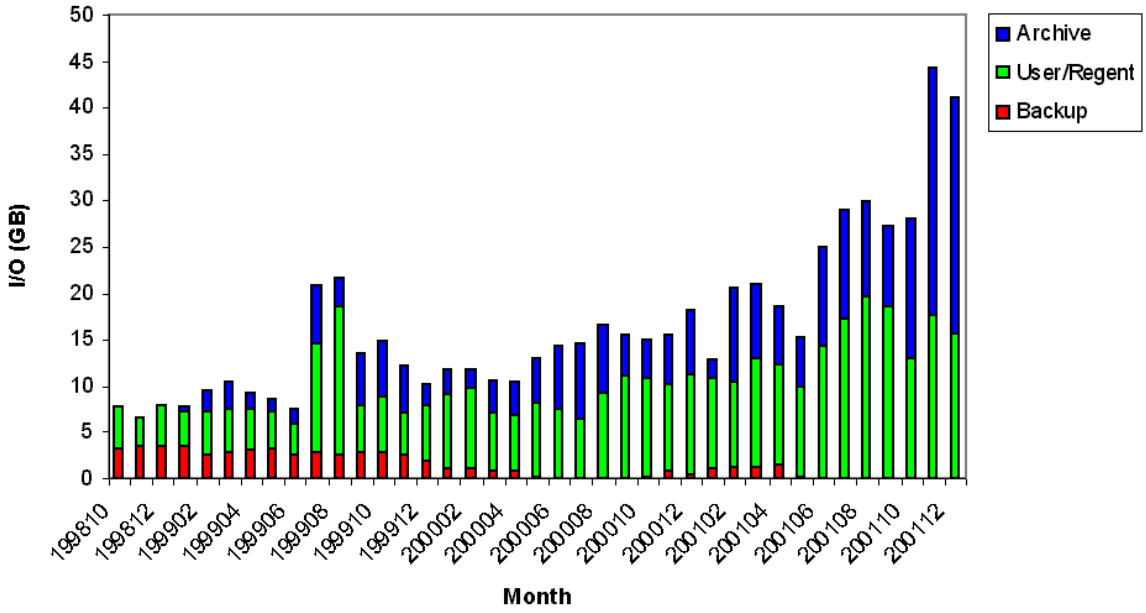



**Figure 6. Monthly I/O by month and system**

## Alvarez

In November 2000, as the result of a joint agreement between the Department of Energy, NERSC, and LBNL, Berkeley Lab procured a moderately sized cluster computer system. The goal of the cluster, named "Alvarez" after LBNL Nobel Laureate Luis Alvarez, is to serve as a platform on which to find out whether a cluster can be a high performance production environment for the scientifically diverse workload typical of NERSC and LBNL. Alvarez serves as a system testbed for the NERSC Center to explore the applicability of such a system to support a highly parallel, numerically intensive workload while at the same time providing a moderately parallel computational resource for strategic LBNL projects.

The system consists of 87 dual-processor compute nodes with 866 MHz Intel Pentium III processors. Each node has 1 GB of main memory, and the cluster has 1.5 TB of shared disk space. The nodes are connected with a high performance Myrinet-2000 interconnect. The peak performance of this system is estimated to be 150 Gflops.

Alvarez was the second IBM Netfinity cluster shipped, and the internal interconnect, the Myrinet 2000 switch, was one of the first five switches shipped by Myricom. Neither the integrated system nor its hardware and software components underwent extensive testing before Alvarez was delivered. The early result, as one might expect, was a less than robust system. As is typical with early products, a number of defective hardware components had to be identified and replaced, and software and microcode bugs had to be identified and fixed. Despite these hurdles, the system successfully passed acceptance testing in June 2001, exceeding benchmark expectations, and achieved 99.7% availability under user load. Subsequently, several periods of instability occurred due to hardware and software problems. Many of these problems have been corrected, and the system is now providing a usable environment. Whether this environment can be as scientifically productive as a high-end system is a topic of ongoing study.

## 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.*

### Support for SciDAC Projects

Soon after the DOE announced the awards in its "Scientific Discovery through Advanced Computing" (SciDAC) program in August 2001, NERSC was getting ready to provide specialized consulting and algorithmic support for these SciDAC projects:

- Advanced Computing for 21st Century Accelerator Science and Technology
- The Summit Framework
- Center for Extended MHD Modeling
- Electromagnetic Wave-Plasma Interactions in Multi-Dimensional Systems
- Computational Atomic Physics for Fusion
- Magnetic Reconnection
- The Supernova Science Center
- TeraScale Supernova Initiative
- Explicitly Correlated Methods for Computations of Properties to Chemical Accuracy
- Terascale Optimal PDE Simulations
- High-End Computer System Performance: Science and Engineering
- DOE Science Grid

A consulting project facilitator from the User Services Group has been assigned to each SciDAC scientific discipline to help define project requirements, get resources, and tune and optimize codes, as well as coordinate special services such as special queues and throughput, increased limits, etc. Algorithmic project facilitators from the Scientific Computing Group have been assigned to develop and improve algorithms, enhance performance, and coordinate software development with the various integrated software infrastructure centers.

A variety of other services are also being offered to SciDAC researchers. For example, they can request support for software that is particular to their work and not as applicable to the general community. NERSC will also help develop and improve visualization methods specific to the projects. And NERSC staff are available to provide presentations and custom training at SciDAC conferences and workshops.

NERSC hosts a Web site for these SciDAC projects at <http://scidac.nersc.gov/>. Web services for interested projects include password-protected areas, safe "sandbox" areas for dynamic script

development, Web infrastructure tools, an archive for mailing lists, and consulting support to help projects organize and manage Web content.

In addition, NERSC hosts a CVS server that is available to SciDAC projects and other selected groups. CVS is an increasingly popular source code versioning system which allows managed contributions and edits to programming projects from many programmers. CVS can help users save space in their home directory by storing large software trees either locally or on a remote CVS server. NERSC helps projects set up and manage code repositories, and provides backup, administration and access control.

## Special Services for "Big Splash" and Data-Intensive Projects

To promote the productivity of cutting-edge science, NERSC adapts its systems to provide the resources needed by data-intensive applications and provides "Red Carpet" support to strategic projects so that they can make rapid progress. These projects typically need support such as:

- very large scratch disk (and often hundreds of inodes)—Seaborg has 12.5 TB of scratch disk for users
- terabytes of usable memory and support for 64-bit computing—Seaborg has 4.3 TB of memory for users
- very large computing allocations
- large-scale visualization
- consulting support to make effective use of these resources
- good bandwidth between the resources
- large archival storage.

The four case studies below show how the unique resources of the IBM SP combined with excellent consulting support have produced scientific breakthroughs. They also demonstrate the size of the resources needed for rapidly growing, data-intensive projects. Today, three of the Big Splash projects could each use all of NERSC's resources just.

### *Black Hole Merger Simulations*

One of the earliest results of NERSC's special services was the first successful simulation of the spiraling merger of two black holes, performed by researchers at the Max Planck Institute for Gravitational Physics in Germany, led by Ed Seidel, with visualization assistance from NERSC's John Shalf (Figure 7). This simulation is particularly important for interpreting the gravitational wave signatures that will soon be seen by new laser interferometric detectors (such as LIGO and VIRGO) around the world. Detection of the first gravitational waves (or failure to do so) will strongly test Einstein's Theory of General Relativity, the results of which will have ramifications that extend throughout the world of physics.

The Cactus code used to simulate the black hole merger performs a direct evolution of Einstein's equations, which are a system of coupled nonlinear elliptic hyperbolic equations that contain thousands of terms if fully expanded. Consequently, the computational resource requirements are

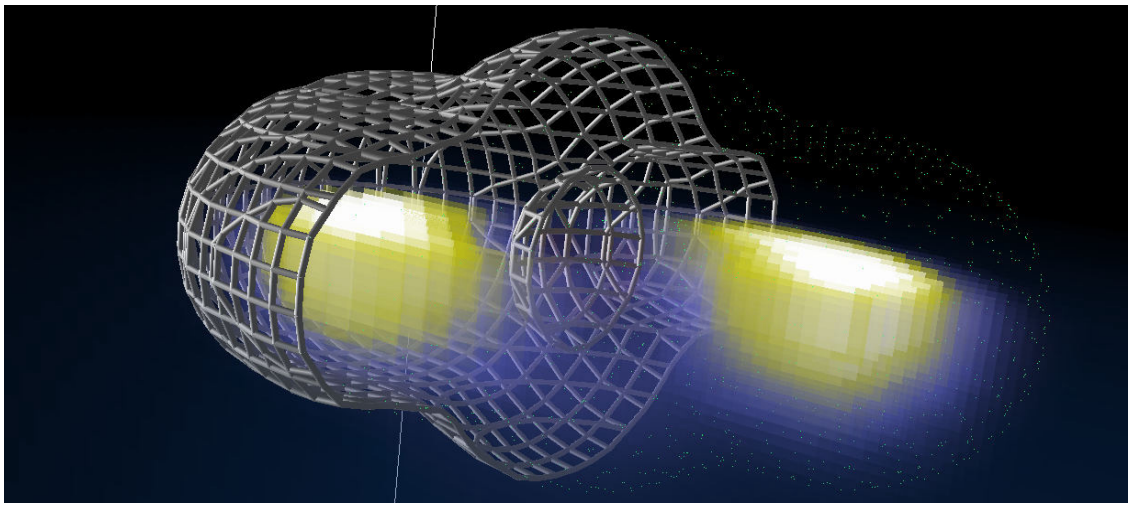


**Figure 7. Results related to this work, including visualizations of binary black hole inspiral, appeared in the April 2002 edition of Scientific American, on the Discovery Channel in June 2002, in the June 2002 IEEE Computer Magazine, and will also appear in Nature in the near future.**

enormous just to do the most basic simulation. The simulation has previously been limited by both the memory and CPU performance of supercomputers as they attempt to move from calibrating against analytic black hole solutions to non-analytic astrophysically relevant cases in full 3-D. The spiraling merger is just such a non-analytic case.

This simulation must use more than one-third of the NERSC IBM SP's available aggregate memory of 4.3 TB in order to achieve the resolution required to accurately simulate these phenomena. This simulation uses 1.5 terabytes of memory and more than 2 terabytes of disk space (250,000 inodes) for each run on the NERSC IBM SP system. These runs typically consume 64 of the large-memory nodes of the SP (a total of 1024 processors) for 48 wall-clock hours at a stretch. (The simulation can use all 184 nodes, but this would only allow simulations that are fractionally larger than using the large-memory nodes due to memory/load-balancing issues.)

NERSC provided access to a special queue to improve turnaround, opened ports to allow remote steering and Grid access, and provided consulting support for 64-bit integration and code debugging. In the space of two months, this simulation consumed 700,000 of the allocated 760,000 CPU hours, simulating three-fourths of a full orbit before coalescence. In the near future, this project could use 10 TB of disk for each run, 5 TB of uniform, user-available memory, and 15 million MPP hours.

The results so far indicate that the Meudon model for black hole coalescence seems to match the simulation data more accurately than the competing Cook-Baumgarte model.

***Supernova Explosions and Cosmology***

This collaboration, led by Adam Burroughs of Arizona State University and Peter Nugent of NERSC, brings together the SciDAC Supernova Science Center and the members of the PHOENIX/SYNPOL collaboration. The goal is a better understanding of supernovae of all types through simulation and model validation. Specific objectives are to clarify the physics of supernova explosions, to improve the reliability of such explosions as calibrated standard candles, and to measure fundamental cosmological parameters. Despite decades of research and modeling, no one understands in detail how supernovae work. The problem persists largely because, until recently, computer resources have been inadequate to carry out credible multi-dimensional calculations.

On June 4, 2002, at the American Astronomical Society meeting in Albuquerque, N.M., Michael Warren and Chris Fryer from Los Alamos National Laboratory presented the results of one of several projects in this collaboration, the first 3-D supernova explosion simulation, based on computation at NERSC (Figure 8). This research eliminates some of the doubts about earlier 2-D modeling and paves the way for rapid advances on other questions about supernovae.

Earlier one-dimensional simulations of core-collapse supernovae almost always failed to explode. Two-dimensional simulations were qualitatively different from 1-D, leading to a robust explosion without fine-tuning of the star's physical properties. They showed that the explosion process is critically dependent on convection, the mixing of the matter surrounding the iron core of the collapsing star. It was believed that the results could again be changed radically by adding a third dimension, but the 3-D simulations turned out to be similar to the 2-D results. The explosion energy, explosion time scale, and remnant neutron star mass do not differ by more than 10 percent between the 2-D and 3-D models. With these 3-D results, researchers are ready to attack more exotic problems that involve rotation and non-symmetric accretion.

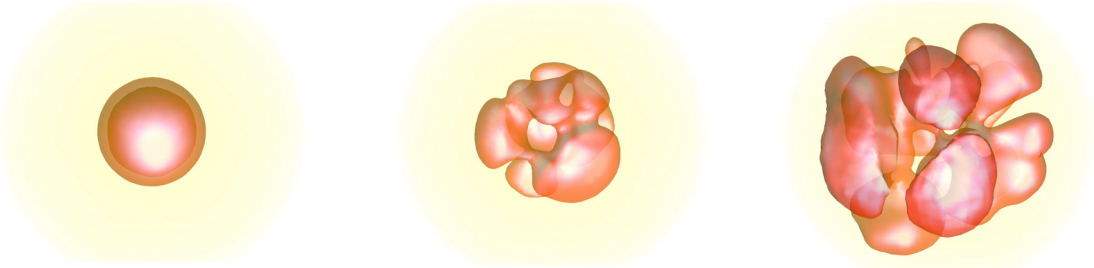


**Figure 8. Computer visualization shows (left to right) three stages of a simulated supernova explosion over a period of 50 milliseconds, starting about 400 milliseconds after the core begins to collapse. The surfaces show the material which is flowing outward at a speed of 1000 kilometers/second. Left is the initial spherical implosion. Center, as in-falling gas approaches the core, it is exposed to a higher and higher influx of neutrinos that heat the gas and make it buoyant. Right, as more cold gas sinks in, it is heated and rises, resulting in enough convective energy transfer to create an explosion. (Michael S. Warren, Los Alamos National Laboratory)**

The 3-D simulation used a parallel smooth particle hydrodynamics (SPH) code coupled with a flux-limited diffusion radiation transport. Supernova explosion calculations are computationally demanding because many processes, involving all four fundamental forces of physics, must be modeled and followed for more than 100,000 time steps. Typical simulations (1 million particles) took about three months on the IBM SP at NERSC.

In the next five years, the Supernova Cosmology Project and the Nearby Supernova Factory experiments will increase both the quality and quantity of observational supernova data at low and high redshift by several orders of magnitude. The purpose of these experiments is to improve the use of Type 1a supernovae as tools for cosmology by determining the underlying physics behind these catastrophic events and to utilize these tools to help us understand the dark energy that drives the acceleration of the universe. The only way to fully exploit the power of this amazing data set is to make a similar order-of-magnitude improvement in computational studies of supernovae, via spectrum synthesis and radiation hydrodynamics. The focus of the PHOENIX/SYNPOL collaboration's portion of this Big Splash project is to start the process of creating 3-D spectrum synthesis models of Type 1a supernovae in order to constrain the observations and place limits on the explosion models and progenitors of supernovae using the full-physics 1-D models as a guide (Figure 9).

Currently two sets of spectrum synthesis codes, PHOENIX and SYNPOL, are used at NERSC to study the model atmospheres of supernovae. PHOENIX models astrophysical plasmas in one dimension under a variety of conditions, including differential expansion at relativistic velocities found in supernovae. The current version solves the fully relativistic radiative transport equation for a variety of spatial boundary conditions in both spherical and plane-parallel geometries for both continuum and line radiation simultaneously and self-consistently using an operator splitting

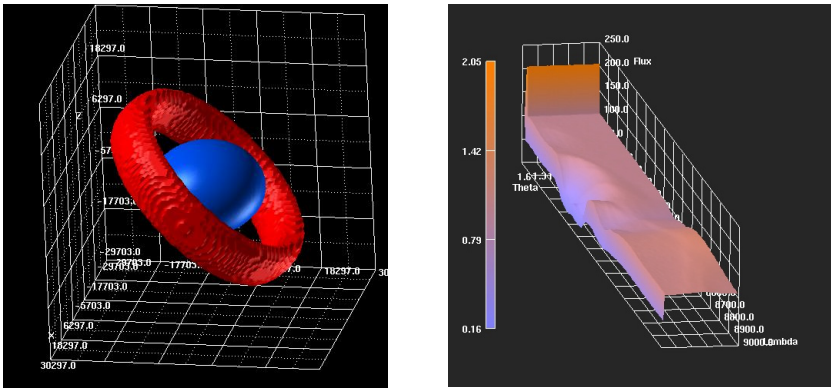



**Figure 9. A spectrum synthesis calculation of a supernova atmosphere surrounded by a toroid. The layout of the atmosphere is presented on the left, while at the right is a graph of the flux vs. wavelength vs. viewing angle. As the viewing angle shifts towards the toroid, the strength of the absorption increases dramatically. Data that confirm such a model would for the first time put strong constraints on the progenitors of Type Ia supernovae. Such flux features are seen in the spectrum of SN 2001el. (Peter Nugent and Daniel Kasen, Lawrence Berkeley National Laboratory)**

technique. PHOENIX also solves the full multi-level non-local thermodynamic equilibrium (NLTE) transfer and rate equations for a large number of atomic species (with a total of more than 10,000 energy levels and more than 100,000 primary NLTE lines), including non-thermal processes. PHOENIX accurately solves the fully relativistic radiation transport equation along with the non-LTE rate equations (currently for ~150 ions) while ensuring radiative equilibrium (energy conservation).

SYNPOL is a 3-D radiative transfer code developed at NERSC to study the spectropolarimetry of supernovae as part of the Big Splash initiative. It is based on a Monte Carlo treatment of line formation via the Sobolev approximation and includes electron scattering. Because SYNPOL does not solve rate equations and does not do continuum transfer, it is not used for quantitative abundance determinations or for absolute flux calculations. Rather its value lies in establishing line identifications (the intervals of ejection velocity within which the presence of particular ions is detected) and in probing the geometry of the supernova and its ejecta. For a full 3-D run, with signal-to-noise and resolution an order of magnitude greater than the observational data, approximately $10^{12}$ photons are generated within a Cartesian grid of 300 per side. Due to the size of the atomic data—over 42 million lines whose strengths can vary at each cube in the grid—the memory requirements and the time it takes to process the scattering of such a large number of photons are quite large: 1 million MPP hours for 3-D simulation with simplified physics, and 10 GB input and 1 GB output per iteration, with 20 iterations per star model for 20 to 30 models.

NERSC provided a new 24-hour run queue to accommodate this simulation. Within the next two or three years, 100 times more CPUs will be needed to run 3-D simulations with complex physics if there are no algorithmic improvements. In addition, the Supernova Factory will need to receive 50 GB of new data daily into HPSS and Seaborg, then retrieve 50 GB of archived data from HPSS for comparison with the new data, and store 25 GB of additional data back to HPSS.

### *JAZZ Genome Assembler*

Almost a thousand new human genes have been discovered in the human genome by scientists who have decoded the genome of a very distant relative, the fugu, or puffer fish, as reported in the July 26, 2002 issue of Science. As many as three-quarters of the fish's genes have direct human counterparts, despite the 450 million years of evolution since the two vertebrates shared an ancestor (Figure 10). Though the fugu fish has much the same number of genes as people, its genome is a mere eighth the size because it lacks much of the "junk DNA" that clutters the human genome. Dr. Daniel Rokhsar and colleagues at the DOE's Joint Genome Institute (JGI) used NERSC resources in 2001 to write and run the computer program, JAZZ, that assembled the fugu genome from 3.1 million DNA fragments produced by coding machines.

The fugu assembly required 30 GB for database files and 150 GB of scratch space. NERSC staff, led by Jonathan Carter, worked with Rokhsar to port the JAZZ assembler, the BLAST alignment tool, the cross match alignment tool, and the MySQL client to the IBM SP, and provided consulting support for parallelizing BLAST and cross match. NERSC also provided a dedicated MySQL server and resolved issues with installing a MySQL server on the SP.

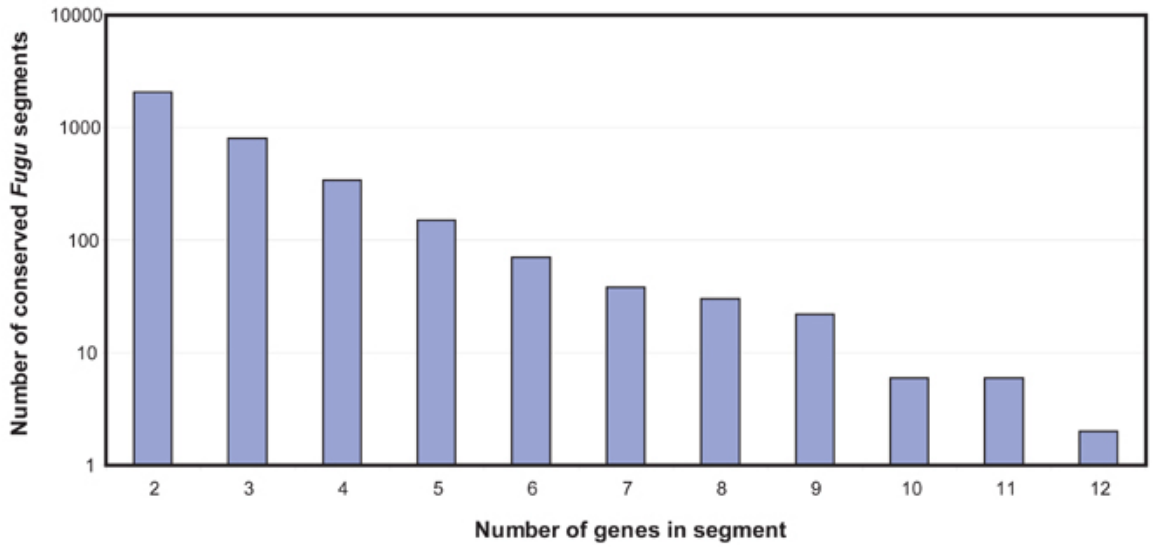



**Figure 10. Relationship between the abundance of conserved segments between fugu and human and the number of conserved genes per segment.**

The JGI team is currently working on assembling the mouse genome, which will require 75 GB for database files and 500 GB of intermediate data. As more raw data is added, these requirements could easily double.

### *Accelerator Science*

The Advanced Computing for 21st Century Accelerator Science and Technology SciDAC project, co-led by Robert Ryne of Berkeley Lab, is developing a new generation of accelerator simulation codes that will help use existing accelerators more efficiently and will strongly impact the design, technology, and cost of future accelerators. The project's current requirements include 1.6 million MPP hours, up to 2 TB of memory, and 64-bit MPI. NERSC provided 3 TB of scratch space, consulting support for large memory management and performance analysis, CVS support, and Web hosting. Within the next three years, this project will require 15 to 20 million MPP hours, more than 5 TB of scratch space, and continued consulting support.

These simulations are helping physicists understand beam heating for SLAC's PEP-II upgrade, helping design the Next Linear Collider accelerating structure, providing a better understanding of emittance growth in high-intensity beams, and exploring laser wakefield accelerator concepts for future accelerator design.

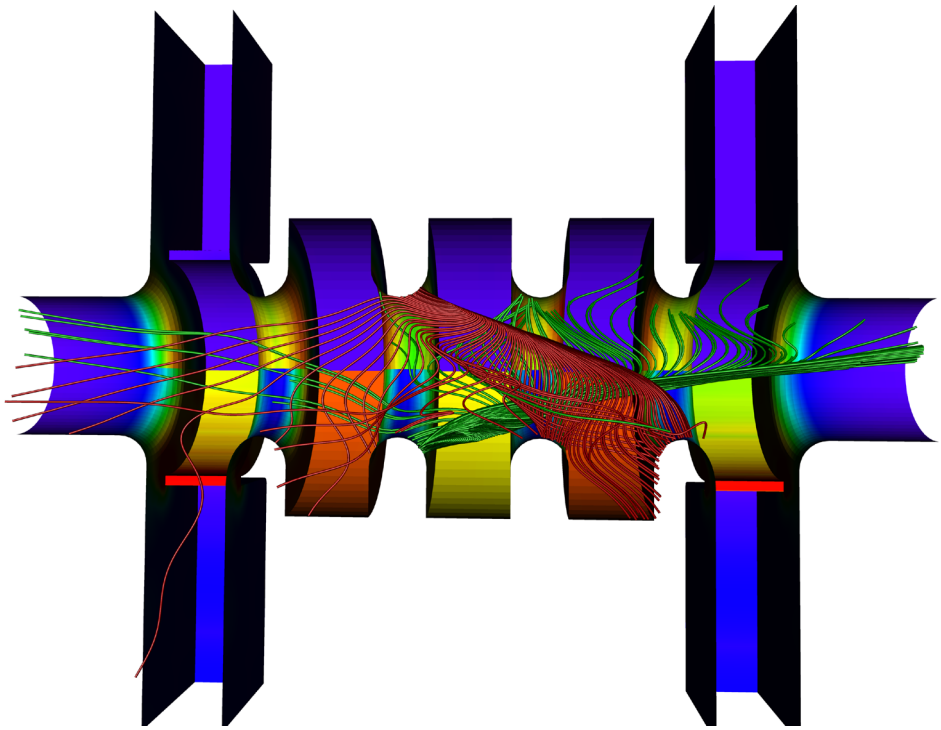


**Figure 11. Trajectories of field emitted electrons in a traveling wave accelerating structure originating from the high field area around the nose of the disks (red for field gradient of 50 MV/m and green for 120 MV/m) to study dark currents. Simulation was done using the particle tracking module (Ptrack) within the parallel time-domain electromagnetic code Tau3P. Peak surface electric field is shown in the upper half of the structure while the lower half displays the maximum surface magnetic field that corresponds to wall loss.**

### New Network Web Page Gives Current Traffic Reports

For several years NERSC has displayed current system status for the computational and storage systems on the <http://hpcf.nersc.gov> Web page so that users have easy access to this information, which is important to their productivity. This year NERSC added a new network statistics Web page <http://hpcf.nersc.gov/network/> which allows users to monitor network activity in real time. Network traffic statistics are collected every five minutes and can be displayed for each border router interface in daily, weekly, monthly and yearly graphs. Alternatively, a current traffic summary consisting of the daily graph for each interface can be displayed on one page. This service can help users decide the best time for a large data transfer, or it can show them why they may be experiencing slower than usual interactivity at a given point in time.

### Expanded PDSF Brings Cosmic Mysteries to Light

Eighty-two new servers and significant improvements in the overall computing and networking infrastructure were added this year to the PDSF (Parallel Distributed Systems Facility), a large Linux-based computer cluster that is currently operated as a partnership between NERSC and the Berkeley Lab Nuclear Science and Physics divisions.

Computing power was expanded to 390 processors, and the number of disk vaults grew to 49, with a total 35 terabytes of shared storage. Gigabit Ethernet networking for the high-bandwidth compute nodes enables the PDSF to run MPI jobs that require up to 50 nodes (100 processors). Gigabit Ethernet for the disk vaults makes it possible to take advantage of the server-side performance improvements of the 2.4 Linux kernel.

In addition to the new servers and disk vaults, the expansion included added memory for some compute nodes as well as upgrades to console servers, switches, networking cables, disk drives, software, and miscellaneous tools — all with the goal of continuing to provide a reliable, well-supported resource that meets its users' growing needs.

"The large computing and storage capabilities of the PDSF play an essential role in helping researchers extract the physics from the terabytes of data that they produce," said Lee Schroeder, Director of Berkeley Lab's Nuclear Science Division. "For example, the STAR experiment is producing descriptions of nuclear collisions at the highest heavy ion collider energies ever achieved, and SNO's first physics results are helping to solve the 'solar neutrino problem' and provide further evidence for neutrino oscillations. These great physics results were made possible by the dedication of the NERSC/PDSF staff and their close collaboration with the research teams."

The PDSF provides many benefits to its users:

- Professional administration of the cluster with high uptimes (>95%) and rapid response to problems
- System is kept stable while constantly evolving and improving the system
- System is kept secured, patched, and monitored
- Uses Linux and commodity hardware which decreases both acquisition and maintenance costs
- PDSF participates in evaluations and beta testing of new products to stay at the forefront of new technologies
- Frees scientists from dealing with hardware and system software issues and allows them to focus on their own applications
- PDSF users have access to user services and support

In addition, PDSF provides several special services to the HENP users:

- Installation of standard HENP libraries
- Porting and maintenance of user applications and software for the local environment
- Works with remote sites to improve network transfers
- Tutorials and user meetings
- Strong participant in the HENP computing community

Contributions of various projects to the PDSF and a sampling of their utilization of the system are depicted in the Figure 12.

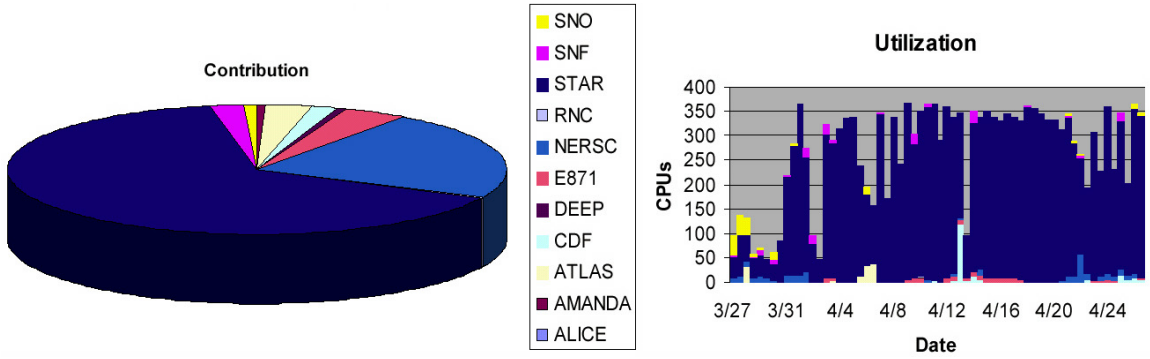



**Figure 12. User contributions to and utilization of the PDSF.**

## 6. Develop and Implement Ways to Transfer Research Products and Knowledge into Production Systems at NERSC and Elsewhere (New for 2001)

*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.*

### The Global Unified Parallel File System

In a typical high-performance computing (HPC) environment, each large computational system has its own local disk, and access to additional network-attached storage and archival storage servers. Such an environment prevents the consolidation of storage between systems, thus limiting the amount of working storage available on each system to its local disk capacity.

The result? An unnecessary replication of files on multiple systems, an increased workload on users to manage their files, and a burden on the infrastructure to support file transfers between the various systems.

NERSC is implementing a solution using existing and emerging technologies to overcome these inefficiencies. The Global Unified Parallel File System (GUPFS) Project aims to provide a scalable, high-performance, high-bandwidth, shared-disk file system for use by all of NERSC's high-performance production computational systems. GUPFS will provide unified file namespace for these systems and will be integrated with the High Performance Storage System (HPSS). Storage servers, accessing the consolidated storage through the GUPFS shared-disk file systems, will provide hierarchical storage management (HSM), backup and archival services. An additional goal is to distribute GUPFS-based file systems to geographically remote facilities as native file systems over the DOE Science Grid.

This environment will eliminate unnecessary data replication, simplify the user environment, provide better distribution of storage resources, and permit the management of storage as a separate entity while minimizing impacts on the computational systems.

The major enabling components of this envisioned environment (see Figure 13) are a high-performance shared-disk file system and a cost-effective, high-performance storage area network (SAN). These emerging technologies, while evolving rapidly, are not targeted towards the needs of high-performance scientific computing. The GUPFS project intends to encourage the development of these technologies to support HPC needs through collaborations with other institutions and vendors, while also aiding in their development.

The GUPFS project is expected to span five years. During the first three years, NERSC plans to test, evaluate and advance shared-disk file systems, SAN technology, and other components of the GUPFS environment. This investigation is expected to include open-source and commercial shared-disk file systems, new SAN fabric technologies as they become available, SAN and file

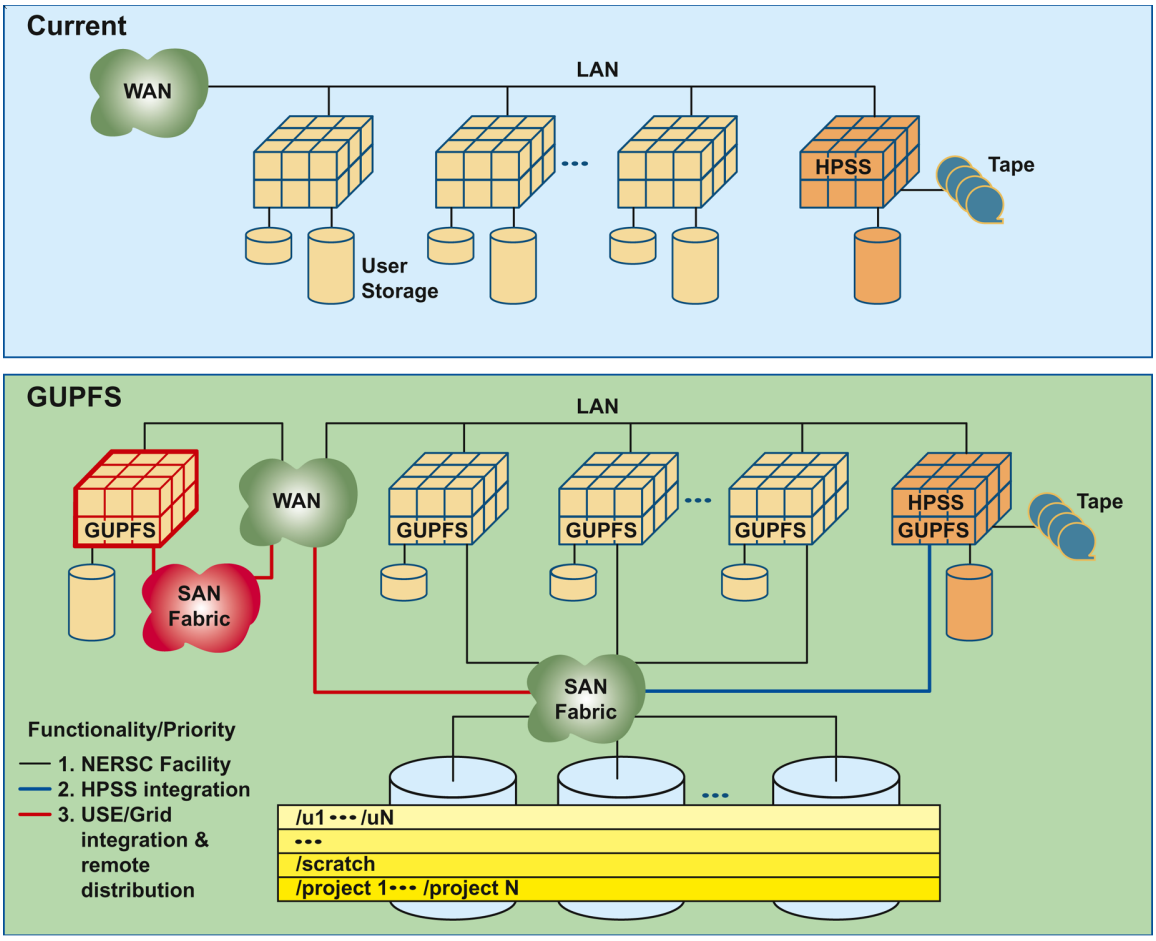



**Figure 13. Conceptual diagram of current storage architecture and GUPFS.**

system distribution over the WAN, HPSS integration, and file system performance and scaling. During this time NERSC also plans to form collaborations and become active in the shared-disk file system community. At the end of this period, NERSC will assess the feasibility of moving forward with a full implementation in the NERSC production environment.

Provided the assessment is favorable, the last two years of the GUPFS project will focus on implementation. The first step in implementation is to choose a file system based upon the previous evaluation and testing. The next step is building up the SAN infrastructure, while simultaneously starting the full development efforts required for Grid distribution and HSM integration. Subsequently, there will be a phased rollout of the GUPFS file system on the various production systems, including HPSS. Once adequate SAN infrastructure, production systems, and required software are in place, DOE Science Grid distribution of the shared file systems over the WAN will be initiated.

For those interested in additional information or participating in the GUPFS project see: http://www.nersc.gov/projects/gupfs

# 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 largest 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.*

## Proposals Sought for NERSC-4 System

NERSC acquires a new capability-focused computational system every three years. The three-year interval is based on the length of time it takes to introduce large systems, the length of time it takes for NERSC clients to become productive on new systems, and the types of funding and financial arrangements NERSC uses. At any given time, NERSC has two generations of computational systems in service, so that each system will have a lifetime of five to six years. This overlap provides time for NERSC clients to move from one generation to the next, and provides NERSC with the ability to fully test, integrate, and evolve the latest generation while maintaining service on the earlier generation.

NERSC uses the "Best Value" process for procuring its major systems. Rather than setting mandatory requirements and using a quantitative rating scheme, the Best Value method requests baseline and value-added characteristics. These characteristics are not meant to design a specific solution but rather to signify a range of parameters that will produce an excellent and cost-effective solution. Thus, Best Value does not limit a site to the lowest common denominator requirements, but rather allows NERSC to push the limits of what is possible in order to get the best solution. Vendors indicate they prefer this method as well, because it provides them more flexibility in crafting their best solution.

A request for proposals for the NERSC-4 system was issued in November 2001, with proposals due in February 2002. Like the NERSC-3 contract described above, the NERSC-4 contract will be based on performance metrics and functional requirements, especially NERSC's Sustained System Performance (SSP) benchmark suite and Effective System Performance (ESP) test. One new feature of this contract is that the SSP value will no longer be based on the NAS Parallel Benchmarks, but will be based on the NERSC Application Performance Suite. Delivery of the initial NERSC-4 system is expected in early 2003.

# 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. Transfer must be made to other client sites, supercomputer sites, and industry.*

## Distributing NERSC-Developed Software

At the SC01 conference, NERSC distributed several CDs containing the documentation, code and open source licenses for software developed at NERSC. These CDs included:

**Tools for Grid Computing:** Developing and effectively using application programs in distributed environments such as collaboratories or computational grids presents a number of challenges. The Berkeley Lab/NERSC Distributed Systems Department has had more than five years of experience enabling applications to function efficiently in such environments. In the process we have developed a variety of tools to help support the development and tuning of distributed applications and the collaboration of scientists. These tools include:

- Akenti Authorization Service: A PKI certificate based authorization service for distributed environments
- The InterGroup Protocols: Protocols for reliable multicast communication that scale to the Internet
- NetLogger: A library to facilitate diagnoses of performance problems in networks and in distributed systems code by combining network, host, and application-level monitoring
- Network Characterization Tools: Tools to diagnose and troubleshoot networks hop-by-hop in an easy and timely fashion
- Nettest Monitoring Framework: Nettest is a secure, real-time network monitoring utility
- pyGlobus: A Python interface to Globus
- The Grid Portal Development Kit: A collection of components and an SDK for the development of Grid enabled portals
- Remote Camera Control: A client-server system for remotely controlling serial devices used for videoconferencing

**M-VIA and MVICH:** This CD contains Virtual Interface Architecture (VIA) software for low-latency, high-bandwidth, inter-process communication. VIA is an industry standard high performance communication interface for system area networks (SANs). VIA provides protected user-level zero-copy data transfers, enabling low latency and high bandwidth. The communication model includes both cooperative communication (send/recv) and remote memory access (get/put).

M-VIA is a modular implementation of the VIA standard for Linux. It provides a software framework that eases the development of drivers for new VIA-aware hardware as well as support for legacy network devices. MVICH is an MPICH-based implementation of MPI for VIA. It

provides receive-side buffering for short messages and high performance zero-copy RDMA transfers for large messages.

**Berkeley Lab AMR:** Many everyday situations, such as running an internal combustion engine or predicting weather, involve complex physical processes that scientists are only now beginning to understand. Berkeley Lab/NERSC mathematicians are developing adaptive mesh refinement (AMR) tools and algorithms for computer modeling of problems like these. AMR serves as a "numerical microscope," allowing researchers to "zoom in" on the specific regions of a problem that are most important to its solution. Rather than requiring the whole calculation have the same spatial resolution, AMR allows for different resolution in different regions of the problem. Areas of interest are covered with a finer mesh than the surrounding regions. Not having to perform the entire calculation at the finest resolution allows scientists to make the most of their available computer resources, so that they can solve bigger, harder problems.

The Berkeley Lab AMR CD contains a gallery of research highlights and images produced using AMR, an extensive list of technical papers and publications relating to Berkeley Lab's development and application of AMR, and the source code for various applications in Berkeley Lab's AMR libraries. The information and applications can also be found on the Web at <http://seesar.lbl.gov/AMR/>.

NERSC staff members also share their expertise and experience by writing papers, giving presentations at conferences, and helping to organize conferences and workshops to facilitate the exchange of ideas. NERSC staff members are also regularly invited to speak at meetings and facilities. Here is a partial list of these activities for the year 2001:

### Invited talks

"Finding New Math Identities Using Computers," David Bailey, Harvey Mudd College, Claremont, Calif., January 2001.

"NERSC Introduction," Bill Kramer, Celera Genomics, Rockville, Md., January 2001.

"Future Challenges of High-End Computing," David Bailey, Johns Hopkins University, Baltimore, Md., February 2001.

"M-VIA: Virtual Interface Architecture for Linux Clusters," Paul H. Hargrove, Cluster Computing in the Sciences, Salt Lake City, Utah, February 2001.

"Parallelization of a Dynamic Unstructured Application using Three Leading Paradigms," Leonid Oliker, EPFL Swiss Federal Institute of Technology, Lausanne, Switzerland, February 2001.

"Efficient Parallelization of Irregularly Structured Computations," Leonid Oliker, CERN, Geneva, Switzerland, February 2001.

"A New View Of The Early Universe," Julian Borrill, American Association for the Advancement of Science, San Francisco, CA, February 2001.

"NERSC A Supercomputer Facility for the Next Millennium," Bill Kramer, Arctic Region Supercomputer Center Technology Panel, Fairbanks, Alaska, February 2001.

"NERSC Expansion Plans," Bill Kramer, SP-XXL (IBM's Large Scale Advisory Group), Maui, Hawaii, February 2001.

"The Ordering Problem for Sparse Nonsymmetric Factorizations," Esmond Ng, Department of Aerospace Engineering, Old Dominion University, Norfolk, Va., March, 2001.

"Large scale atomistic electronic structure calculations of nanostructures," Lin-Wang Wang, Material Research Society spring meeting, San Francisco, Calif. April 2001.

"Librerias para computacion de alto rendimiento: el proyecto ACTS" (Libraries for high-performace computing: The ACTS project), Tony Drummond, Departamento de Ciencia de la computacion e inteligencia artificial, Universidad de Alicante, Spain, June 2001.

"Entornos para la computacion de alto rendimiento: direcciones dela siguiente generacion de codigos cientificos" (Frameworks for High-Performance Computing: Directions of the next generation of scientific codes), Tony Drummond, Departamento de Sistemas Informaticos y Computacion. Grupo de Investigacion en Redes y Computacion de Altas Prestaciones., Universidad Politecnica de Valencia, Spain, June 2001

"Computational Issues in Large Scale Electromagnetic Simulations," Esmond Ng, High Performance Computing Working Group, Snowmass 2001 – The Future of Particle Physics, Snowmass, Colo., July, 2001.

"Performance of Incomplete Factorizations as Preconditioners," Esmond Ng, Minisymposium on Krylov Space Methods and Preconditioners, 2001 SIAM Annual Meeting, San Diego, Calif., July 2001."Future Directions in Scientific Supercomputing for Computational Physics," Horst Simon, (with Bill McCurdy, Bill Kramer, Bob Lucas and David Bailey), Conference on Computational Physics, Aachen, Germany, September 2001.

"Climate Research at NERSC," Bill Kramer, Computational Atmospheric Sciences Conference, Annecy, France, October–November, 2001.

"The National Energy Research Scientific Computing Center," Esmond Ng, the National Center for High-Performance Computing, Hsinchu, Taiwan, December, 2001.

### Conference Presentations and Proceedings

"Visualization of Adaptive Mesh Refinement Data," G. Weber, Bernd Hamann, K. Joy, Terry Ligocki, K. Ma, John Shalf, Visual Data Exploration and Analysis VIII, Proceedings of the SPIE, Photonics West - Electronic Imaging, January 2001.

"Ordering Schemes for Sparse Matrices using Modern Programming Paradigms," Leonid Oliker, Xiaoye "Sherry" Li, Parry Husbands and Rupak Biswas, IASTED International Conference on Applied Informatics (AI 2001), Innsbruck, Austria, February, 2001.

"Design Strategies for Irregularly Adapting Parallel Applications," Leonid Oliker, SIAM Conference on Parallel Processing, Portsmouth, Va., March 2001

"Ordering Sparse Matrices for Cache-Based Systems," Leonid Oliker, SIAM Conference on Parallel Processing, Portsmouth, Va., March 2001

"Communication Support for Adaptive Computation," Ali Pinar and B. Hendrickson, SIAM Conference on Parallel Processing for Scientific Computing, Portsmouth, Va., March 2001.

“The ACTS Toolkit project," New Vistas in Physics Conference, National Society of Black Physicists meeting, Stanford University, Palo Alto, Calif., March 2001.

“A Parallel Adaptive Projection Method for Low Mach Number Flows,” John Bell, Marcus Day, Ann. Almgren, Michael. Lijewski and Charles Rendleman, Proceedings of the ICFD Conference on Numerical Methods for Fluid Dynamics, University of Oxford, Oxford, England, March 2001.

“Unfavorable Strides in Cache Memory Systems", David Bailey, SIAM Parallel Processing Conference, Norfolk, Va., March 2001.

“Carrier localization in InGaN alloy,” Lin-Wang Wang, American Physical Society Annual March Meeting, Seattle, Wash., March 2001.

“Million atom electronic structure calculations of nanostructures,” Lin-Wang Wang, 2001 ACRS Joint Meeting (International Conference on Computational Nanoscience 2001), Hilton Head Island, S.C., March 2001.

“Graph Partitioning for Complex Objectives,” Ali Pinar and B. Hendrickson, International Parallel and Distributed Processing Symposium, San Francisco, Calif., April 2001.

“Message Passing vs. Shared Address Space on a Cluster of SMPs,” Leonid Oliker, International Parallel and Distributed Processing Symposium (IPDPS 2001), San Francisco, Calif., April 2001.

“Coupling Scientific Application using a Distributed Data Broker,” Tony Drummond, 2001 International Conference in Computing Science, San Francisco, Calif., April 2001.

“Optimization of Sparse Matrix Kernels for Data Mining,” Kathy Yelick, Proceedings of Text Mine Workshop ’01, Chicago, Ill., April 2001.

“Optimizing Sparse Matrix Computations for Register Reuse in Sparsity,” Kathy Yelick, Proceedings of the International Conference on Computational Science, San Francisco, Calif., May 2001.

“Extraction of Crack-Free Isosurfaces from Adaptive Mesh Refinement Data", G. Weber, Oliver Kreylos, Terry Ligocki, John Shalf, H. Hagen, Bernd Hamann, K. Joy, Data Visualization 2001, Proceedings of VisSym 2001, Springer-Verlag, Vienna, Austria, May 2001.

“Extraction of Crack-free Isosurfaces from Adaptive Mesh Refinement Data,” Gunther H. Weber, Oliver Kreylos, Terry J. Ligocki, John M. Shalf, Hans Hagen, Bernd Hamann, Kenneth I. Joy, Proceedings of IEEE Eurographics Symposium on Visualization, Ascona Switzerland, May 2001.

“Optimizing Sparse Matrix Computations for Register Reuse in Sparsity,” Kathy Yelick, International Conference on Computational Science, San Francisco, Calif., May 2001.

“The Data Broker: A decentralized mechanism for periodic exchange of fields between multiple ensembles of, parallel computations”, K. Sklower, Tony. Drummond, C.R. Mechoso, J. Spahr, E. Mesrobian, H. Robinson, paper and presentation, CERFACS, Toulouse, France, 2001

“Adaptive numerical simulation of turbulent premixed combustion,” John Bell, Marcus Day, Ann. Almgren, Michael. Lijewski and Charles Rendleman, Proceedings of the First MIT Conference on Computational Fluid and Solid Mechanics, Massachusetts Institute of Technology, Cambridge, Mass., June 2001.

“NERSC Information Management with PHP and PHPLIB,” John McCarthy and Mikhail Avrekh, OSCon2001 (Open Software Conference, 2001), San Diego, Calif., July 2001.

"Problem Solving Environments: The Cactus Toolkit Conference," John Shalf, Ed Seidel, Gabrielle Allen CASC/Components 2001, Livermore, Calif., July 2001.

"An Integrated Solution for Secure Group Communication in WAN," Olivier Chevassut, Deborah Agarwal, Mary. Thompson, G. Tsudik, 6th IEEE Symposium on Computers and Communications, Hammamet, Tunisia, July, 2001.

"The Astrophysics Simulation Collaboratory: Case Study of a Grid-Enabled Application Environment," John Shalf, Ian Foster, Gregor Von Laszewsk, Michael Russel, Jason Novotny, Gabrielle Allen, Greg Daues, Tom Goodale, Wai-Mo Suen, Ed Seidel, HPDC10, San Francisco, Calif., August 2001.

"Network Characterization Service (NCS)," G. Jin, G. Yang, B. Crowley, D. Agarwal, Proceedings of the 10th IEEE Symposium on High Performance Distributed Computing HPDC-10, San Francisco, Calif., August 2001.

"Enabling Network-Aware Applications," Brian Tierney, Dan Gunter, Jason Lee, Martin Stoufer, Proceedings of the 10th IEEE Symposium on High Performance Distributed Computing ( HPDC-10 ), San Francisco, Calif., August 2001.

"Applied Techniques for High Bandwidth Data Transfers across Wide Area Networks," Jason Lee, Dan Gunter, Brian Tierney, W. Allock, J. Bester, J. Bresnahan, S. Tuecke, Conference on High Energy Physics (CHEP 01), Beijing, China, September 2001.

"LSF Batch in Multi-Experiment Environment (The Art of Herding Cats)," Iwona Sakrejda, HEPiX/HEPNT Fall 2001 Meeting, Berkeley, Calif., October 2001.

"NERSC High Performance Storage System (HPSS) from a User Perspective," Thomas M. DeBoni, Nancy Meyer, Harvard Holmes, SCICOMP 4 - The Fourth Meeting of SCICOMP, the IBM System Scientific User Group, Knoxville, Tenn., October, 2001.

"COMBAT – Cosmic Microwave Background Analysis Tools," Julian Borrill, Advanced Information Systems Research Program, Baltimore, Md., October 2001.

"High-Performance Remote Access to Climate Simulation Data: A Challenge Problem for Data Grid Technologies," B. Allcock, Foster, I., Nefedova, V., Chervenak, A., Deelman, E., Kesselman, C ., Alex Sim, Arie Shoshani, Jason Lee, B. Drach, D. Williams, Proceedings of the IEEE Supercomputing 2001 Conference, Denver, Colo., November 2001.

"Introducing the RAGE Robot," Zach Radding, Deb Aggarwal, John Shalf, Marcia Perry, Josh Boverhof, Martin Stoufer, et al., SC-Global (SC-Global Showcase), Denver Colorado, November 2001.

"Cosmic Microwave Background Data Analysis With MADCAP," Julian Borrill, P.G. Ferreira, A.H. Jaffe and R. Stompor, in "Mining The Sky" ESO Astrophysics Symposia Series, 2001.

"An evaluation of search tree techniques in the presence of caches," Costin Iancu, ISPASS-2001: 2001 IEEE International Symposium on Performance Analysis of Systems and Software, Tucson, Ariz., November 2001

"High-quality Volume Rendering of Adaptive Mesh Refinement Data," Gunther H. Weber, Oliver Kreylos, Terry J. Ligocki, John M. Shalf, Hans Hagen, Bernd Hamann, Kenneth I Joy and Kwan-Liu Ma, Proceedings of Vision, Modeling and Visualization, Stuttgart, Germany, November 2001.

"Scalable Preconditioning Using Incomplete Factors," Esmond Ng, Keita Teranishi and Padma Raghavan, Proceedings of the Tenth SIAM Conference on Parallel Processing for Scientific Computing, SIAM, 2001.

"Bipartite Graph Partitioning and Data Clustering," Horst Simon, Hongyuan Zha, Xiaofeng He, Chris Ding and Ming Gu, ACM 10th International Conference on Information and Knowledge Management, Atlanta, Ga., November 2001.

"A Min-max Cut Algorithm for Graph Partitioning and Data Clustering," Horst Simon, Chris Ding, Xiaofeng He, Hongyuan Zha and Ming Gu, First IEEE International Conference on Data Mining, San Jose, Calif., November-December 2001.

"Automatic Topic Identification Using Webpage Clustering," Horst Simon, Xiaofeng He, Chris H.Q. Ding and Hongyuan Zha, First IEEE International Conference on Data Mining, San Jose, Calif., November-December 2001.

"Spectral Relaxation for K-means Clustering," Horst Simon, Hongyuan Zha, Chris Ding, Ming Gu, and Xiaofeng He), Neural Information Processing Systems, NIPS*2001. Vancouver, British Columbia, Canada. December 2001.

### Workshops and Tutorials

"Building and Measuring a High Performance Network Architecture," Bill Kramer, SC 2000 Wrap-Up Meeting, New Orleans, La., January 2001.

"Titanium: Language and Compiler Support for Grid-Based Computation," Kathy Yelick, NPACI All Hands Meeting, San Diego, Calif., January 2001.

"Support for Adaptive Computations Applied to Simulation of Fluids in Biological Systems, Kathy Yelick, NPACI All Hands Meeting, San Diego, Calif., February 2001.

"Internet X.509 Public Key Infrastructure Restricted Delegation Certificate," Mary Thompson, D. Engert, S.Tueke, Global Grid Forum 1, Amsterdam, the Netherlands, March 2001.

"Empirical Pseudopotential Calculations for Quantum Dots," Lin-Wang Wang, colloquium, Physics Department, University of California at Davis, April 2001.

"Solver Technology from the SciDAC Integrated Software Infrastructure Center for Terascale Optimal PDE Simulations," Esmond Ng, Advanced Computing for 21st Century Accelerator Science & Technology (SciDAC) Project Kickoff Meeting, SLAC, Stanford, Calif., August, 2001.ACTS Toolkit Workshop: Solving problems in science and engineering, organized by Tony Drummond and Osni Marques, Lawrence Berkeley National Laboratory, Berkeley, Calif., October 2001

"Cosmic Microwave Background Data Analysis," Julian Borrill, ACTS Toolkit Workshop, Berkeley, CA, October 2001.

The ACTS Toolkit: How can it work for you", Tony Drummond and Osni Marques, SC 2001, Denver, November 2001.

Report to the National Science Foundation Directorate for Computer and Information Science and Engineering (CISE), Advanced Networking Infrastructure and Research Division, NSF Grand

Challenges in e-Science Workshop, John Shalf and 30+ other participants, University of Illinois, Chicago, Ill., December 2001.

"Scalable Incomplete Factorizations," Esmond Ng, The 2001 Workshop on Scalable Solver Software: Multiscale Coupling and Computational Earth Science (SSS2001), University of Tokyo, Tokyo, Japan, December, 2001."An Implicit Algorithm for Solving a Sparse Linear System," Esmond Ng, Workshop on Scientific Computing, Chinese University of Hong Kong, Hong Kong, December, 2001.

### Published Papers and Articles

"TCP tuning Guide for Distributed Applications on Wide Area Networks," Brian Tierney, Usenix; login Journal, February 2001.

"A Monitoring Sensor Management System for Grid Environments," Brian Tierney, Brian Crowley, Dan Gunter, Jason Lee, Mary Thompson, Cluster Computing Journal, March 2001.

"Multi-class Protein Fold Recognition Using Support Vector Machines and Neural Networks," Chris H.Q. Ding and Inna Dubchak, Bioinformatics Journal, March 2001.

An Optimal Index Reshuffle Algorithm for Multidimensional Arrays and Its Applications for Parallel Architectures. C.H.Q. Ding. IEEE Transactions on Parallel and Distributed Systems, March 2001.

"Large scale LDA-band-gap-corrected GaAsN calculations," Lin-Wang Wang, Applied Physics Letters, 2001.

"Using Accurate Arithmetics to Improve Numerical Reproducibility and Stability in Parallel Applications," Yun He, Chris H. Q. Ding, Journal of Supercomputing , March 2001.

"Thick-restart Lanczos method for large symmetric eigenvalue problems," Horst Simon and Kesheng Wu, SIAM. Journal of Matrix Analysis Applications, Vol. 22, No. 2, 2001.

"High Performance Computations for Large Scale Simulations of Subsurface Multiphase Fluid and Heat Flow," Erik Elmroth, Chris Ding, Yu-Shu Wu, Journal of Supercomputing, March 2001.

"First Estimations of Cosmological Parameters From BOOMERANG," A.E. Lange, Julian Borrill, et al., Physical Review, 2001.

"Asymmetric Beams in Cosmic Microwave Background Anisotropy Experiments," J. H. P. Wu et al., Julian Borrill, Astrophysical Journal Supplement.

"Cosmology from Maxima-1, Boomerang and COBE/DMR CMB Observations," A. H. Jaffe et al., Julian Borrill, Physical Review Letters.

"Tests for Gaussianity of the MAXIMA-1 CMB Map," J. H. P. Wu, et al., Julian Borrill, Physical Review Letters.

"A High Spatial Resolution Analysis of the MAXIMA-1 Cosmic Microwave Background Anisotropy Data," A. T. Lee, et al., Julian Borrill, Astrophysical Journal Letters.

"Cosmological implications of the MAXIMA-I high resolution Cosmic Microwave Background anisotropy measurement," R. Stompor et al., Julian Borrill, Astrophysical Journal, 2001.

"Community Software Development with the Astrophysics Simulation Collaboratory," Gregor von Laszewski, Michael Russell, Ian Foster, John Shalf, Gabrielle Allen, Greg Daues, Jason Novotny, Edward Seidel, Concurrency and Computation: Practice and Experience, 2001

"The Cactus Worm: Experiments with Dynamic Resource Discovery and Allocation in a Grid Environment," Gabrielle Allen, David Angulo, Ian Foster, Gerd Lanfermann, Chang Liu, Thomas Radke, Ed Seidel, John Shalf, International Journal of High Performance Computing Applications, Volume 15, Number 4, 2001.

"Prediction of alloy precipitate shapes from first principle calculations," S. Muller, L.W. Wang, A. Zunger, and C. Wolverton, Europhys. Letters, 2001.

"Large scale LDA-band-gap-corrected GaAsN calculations," L.W. Wang, Appl. Phys. Lett., 2001.

"Calculating the influence of external charges on the photoluminescence of a CdSe quantum dot," L.W. Wang, J. Phys. Chem., 2001.

"Calculations of carrier localization in InGaN," L.W. Wang, Phys. Rev. B, 2001.

"Dependence of the band gap and valence band splitting on the order parameter for partially ordered Ga_xIn_(1–x)P alloys," Y. Zhang, A. Mascarenhas, L.W. Wang,, Phys. Rev. B, 2001.

"Thomas-Fermi charge mixing for obtaining self-consistency in density functional calculations," D. Raczkowski, A. Canning, L.W. Wang, Phys. Rev. B, 2001.

"Statistics aspects of electronic and structural properties in partially ordered semiconductor alloys," Y. Zhang A. Mascarenhas, L.W. Wang, Phys. Rev. B, 2001.

"Linearly Polarized Emission from Colloidal Semiconductor Quantum Rods," J. Hu, L. Li, W. Yang, L. Manna, L.W. Wang, A.P. Alivisatos, Science magazine, 2001.

"Mask function real space implementations of nonlocal pseudopotentials," L.W. Wang, Phys. Rev. B, 2001.

"The Astrophysics Simulation Collaboratory: A Science Portal Enabling Community Software Development," Michael Russell, Gabrielle Allen, Greg Daues, Ian Foster, Edward Seidel, Jason Novotny, John Shalf, Gregor Von Laszewski, Journal of Cluster Computing, Spring/Summer 2001.

"The persistence of regular reflection during strong shock diffraction over rigid ramps," L. F. Henderson, K. Takayama, William Y. Crutchfield, and S. Itabashi, Journal of Fluid Mechanics, 2001.

"Asymmetric Beams in Cosmic Microwave Background Anisotropy Experiments', J. H. P. Wu, Julian Borrill, et al., Ap J Supp, 2001.

"Cosmology from Maxima-1, Boomerang and COBE/DMR CMB Observations," A.H. Jaffe. Julian Borrill et al., Phys Rev Lett, 2001.

"A Cartesian grid embedded boundary method for the heat equation on irregular domains," Peter McCorquodale, Phillip Colella, H. Johansen, Journal of Computational Physics, 2001.

"A projection method for incompressible viscous flow on moving quadrilateral grids," D. Trebotich and Phillip Colella, Journal of Computational Physics, 2001.

"Robust and Efficient Surface Reconstruction from Contours," Ge Cong, Bahram Parvin, Visual Computers, May 2001.

"Detection of Vortices and Saddle Points in SST Data," Qing Yang, Bahram Parvin, A. Mariano, Geophysical Research Letters, 2001.

"Critical currents and vortex states at fractional matching fields in superconductors with periodic pinning," Charles Reichhardt and Niels Gronbech-Jensen, Accepted for publication in Physical Review B, 2001.

"Moving Wigner glasses and smectics: Dynamics of disordered Wigner crystals," Charles Reichhardt, Cynthia J. Olson, Niels Gronbech-Jensen, and Franco Nori, Physical Review Letters (2001).

"Mode-locking in ac-driven vortex lattices with random pinning," Alejandro B. Kolton, Daniel Dominguez, and Niels Gronbech-Jensen, Physical Review Letters, 2001.

"Cactus Tools for Grid Applications," Gabrielle Allen, Werner Benger, Tom Goodale, Hans-Christian Hege, Gerd Lanfermann, André Merzky, Thomas Radke, Edward Seidel, John Shalf, Cluster Computing 2001.

"A Monitoring Sensor Management System for Grid Environments," Brian Tierney, Brian Crowley, Dan Gunter, Jason Lee, Mary Thompson, Cluster Computing Journal, vol 4-1, 2001.

"Biannual TOP-500 Computer Lists Track Changing Environments for Scientific Computing," Horst Simon, Jack Dongarra, Hans Meuer and Erich Strohmaier, SIAM News, November 2001.

"On computing row and column counts for sparse QR factorization" Esmond Ng, John Gilbert, Xiaoye Li, and Barry Peyton, BIT, 41, 2001.

### Conference Leadership and Organization

In October, Carey Whitney of NERSC's Parallel Distributed Systems Facility (PDSF) team organized HEPiX/HEPNT, an international meeting of PDSF's high energy physics UNIX (HEPiX) and Windows NT (HEPNT) user groups. The meeting, held Oct. 15–18 at the Oakland Scientific Facility, drew 40 attendees, with 15 coming from overseas and a handful sitting in via videoconference. The agenda included discussions of the Grid, data storage, computer security and methods for improving system performance, as well as site reports from research institutions in the United States and Europe. Cary Whitney of the PDSF team took the lead in organizing the conference. NERSC presenters included Horst Simon, Steven Chan, Nick Cardo, Harvard Holmes, Shane Canon, Iwona Sajkreda, Stephen Lau and Brent Draney.

Here are other examples of NERSC staff playing key roles in scientific and technical meetings:

8th International Symposium on Solving Irregularly Structured Problems in Parallel, Esmond Ng, chair, San Francisco, Calif., April 2001.

2001 International Conference on Preconditioning Techniques for Large Sparse Matrices from Industrial Applications, Esmond Ng, co-chair, Tahoe City, Calif., April-May, 2001.

Numerical Linear Algebra Workshop, Esmond Ng, co-organizer, The Fields Institute for Research in Mathematical Sciences, Toronto, Canada, October-November, 2001.

"Vistas in Computational Sciences" session, Tony Drummond, chair and organizer, National Society of Black Physicists meeting, Stanford University, Palo Alto, Calif., March 2001.

2001 International Conference on Computational Nanoscience, Niels Gronbech-Jensen, Scientific Advisory Board member, Hilton Head Island, S.C., March 2001.

ACTS Workshop: High-Performance Libraries for Science and Engineering, organized by Tony Drummond and Osni Marques, during the LACSI Symposium. Drummond and Marques also presented talks on "ACTS Toolkit and its Applications" and "PETSc" and moderated two, "Promoting Reusability and Performance" and "Promoting Software Interoperability," Santa Fe, New Mexico, October 2001.

SCICOMP 4 – The Fourth Meeting of SCICOMP, the IBM System Scientific User Group, Tom DeBoni, secretary, Knoxville, Tenn., October 2001.

SC2001 conference, Bill Kramer, Bandwidth Challenge chair, Denver, Colo., November 2001.

HiPC 2001 (8th Int'l Conf on High Performance Computing), Chris Ding, Program Committee member.

SIAM 1st Data Mining Conf, Workshop on Text Mining, Chris Ding, 2001. Program Committee member.

## 9. NERSC Will Be Able to Thrive and Improve in an Environment Where Change Is the Norm. (New for 2001)

*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.*

### Anticipating Technology Changes

It is essential that NERSC continuously evaluate technologies to determine what can best solve the problems of its clients in the future. Key technologies relevant to NERSC include supercomputer architectures, data communications technologies, online disk storage, archival storage, and application development software.

NERSC's experiences in operating the IBM SP and the Alvarez cluster, both described above, give us a basis for comparing various types of SMP systems in the future. And NERSC's feedback to IBM regarding hardware and software for both systems provides the vendor with important information about the needs of a high performance production center.

NERSC tracks the progress of special architectures—such as Cray's MTA and SV2, IBM's Blue Gene and Blue Light, and academic research products such as UC Berkeley's iRAM—and evaluates their applicability to the DOE computational science challenges. For example, NERSC has a staff member serving on the Cray SV2 design review team and another on Cray's Corporate Advisory Committee. NERSC also works closely with Berkeley Lab computer science research staff to evaluate new architectures.

The Probe wide-area distributed-storage testbed, described above under Goal 3, has involved research in several new storage technologies. And NERSC's computational system procurement system, which is based on performance characteristics, is open-ended regarding which technology will best meet the requirements.

### Preparing for the Final PVP to SMP Transition

Since NERSC acquired the Cray T3E in 1996, we have been helping users make the transition from parallel vector (PVP) to massively parallel (MPP) and symmetric multiprocessor (SMP) architectures. With the Cray SV1s scheduled for decommissioning at the end of FY 2002, we began planning in 2001 to help the remaining PVP users make the transition to SMPs. We conducted a survey of PVP users to determine their reasons for using PVP systems and to identify any barriers to moving to a new architecture. From the survey results, we derived the following requirements summary:

- Modern SMP with similar amount of CPU, memory and disk
- High memory bandwidth
- Resources/limits no less than current values

- Interactive work should be well supported
- Good support should remain available
- Utilities should be provided for file conversion
- Replacements for common Cray routines and functions should be provided

We will continue exploring options to meet the needs of PVP users during FY 2002.

### Security

The Networking and Security Group monitors the network for intrusions and unauthorized use, and responds to security incidents in cooperation with Berkeley Lab's cybersecurity staff. Attempted intrusions have been rising steadily in the last few years, but security research is making progress as well. The NERSC BRO border intrusion detection system can take action to block certain attacks without human intervention. In addition, we have begun working with Juniper Networks, Inc. to test the BRO intrusion detection software at speeds of 2.4 Gb/s (OC-48).

## 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. To help facilitate the professional exchange of ideas and information, NERSC has adopted a series of guidelines and information. They are posted at* <http://www.nersc.gov/staff/#nersc>.

### 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 the year, NERSC presented Spot Awards to John Shalf, Alex Sim, Junmin Gu, Vijaya Natarajan, Francesca Verdier, Tom DeBoni, David Skinner, Howard Walter, Jed Donnelley, William Iles, Zaida McCunney, Zachary Radding, Cary Whitney, Steve Lowe, Robert Neylan, Matthew Andrews, Wayne Hurlbert, Nancy Johnston, and William Harris.

Oustanding Performance Awards were given to the following employees for their work on special projects:

William Fortney for his assistance in developing the NERSC Strategic Proposal.

Igor Gaponenkom, Akbar Mokhtarani, and Simon Patton for upgrading the database for the BaBar experiment at the Stanford Linear Accelerator Center.

Wayne Hurlbert, Harvard Holmes, Steve Lowe, James Lee, Del Black, Richard Beard, Russell Huie, and Robert Neylan for moving NERSC's archival data storage center from the main Lab site to the Oakland Scientific Facility as part of the center's relocation to its new site in Oakland.

### Distributed Facility Management

While NERSC has developed expertise in tools and techniques for managing distributed systems, the relocation of computing resources and a significant number of employees from the main Laboratory site to the new Oakland Scientific Facility – creating a "distributed facility" proved to be a new challenge.

In order to provide management with a connection of suitable "bandwidth," provisions were made for shared office space at both LBNL and OSF, allowing employees and managers to have space to park their laptops and conduct their work in either location as needed. Also, an Access Grid node was built at the OSF, providing an interactive videoconferencing link with the Lab and the rest of the world.

Additionally, NERSC allows for employees to telecommute as appropriate, allowing them to maximize their productivity and enjoy flexibility in commuting and scheduling.

### 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 William Harris serving on the Diversity Committee and Tammy Welcome on the Computing and Communications Services Advisory Committee.

## CONCLUSION

As NERSC nears the start of its fourth decade, the center is among the most senior centers in the world of high-performance computing. But, to paraphrase the tagline from an old TV commercial, we're not getting older — we're getting better. In fact, NERSC continues to set the pace in the realm of scientific computing, providing the latest reseources, assessing and developing new methods to improve the delivery of systems and services, and actively sharing our expertise with other members of the HPC community. As DOE's flagship computing center for unclassified research, however, the primary measure of NERSC's success is the computational science discoveries of our more than 2,000 users. These research achievements are highlighted each year in the NERSC Annual Report (the 2001 report is available on the Web at <http://www.nersc.gov/research/annrep01/>). Essential to this scientific success, however, are the contributions of our staff in providing critical systems and services for our users. It is these measures which we have described in this self-assessment.

But we also take our definition of success one step farther to include our interaction with HPC vendors, other DOE national laboratories, the scientific community and the HPC community. Successful leadership means working with vendors to improve the ability of computing and storage resources to meet the day-to-day production demands of our users. This experience helps vendors develop better products, while giving other centers the tools to assess how different systems perform in the real world of scientific computing.

NERSC staff members produce hundreds of scientific and technical papers each year, and after careful review, the articles are published for the benefit of other researchers. When organizers of conference need speakers on any number of HPC topics, NERSC staffers are usually on the invitation list. Putting our ideas and experiences out for scrutiny by our peers and having those ideas validated is another measure of our success – and a source of pride for us. If our work helps a scientist make a significant discovery, helps another facility improve its systems or service, or gives a student a new idea for research, then we've succeeded.