LBNL-41522

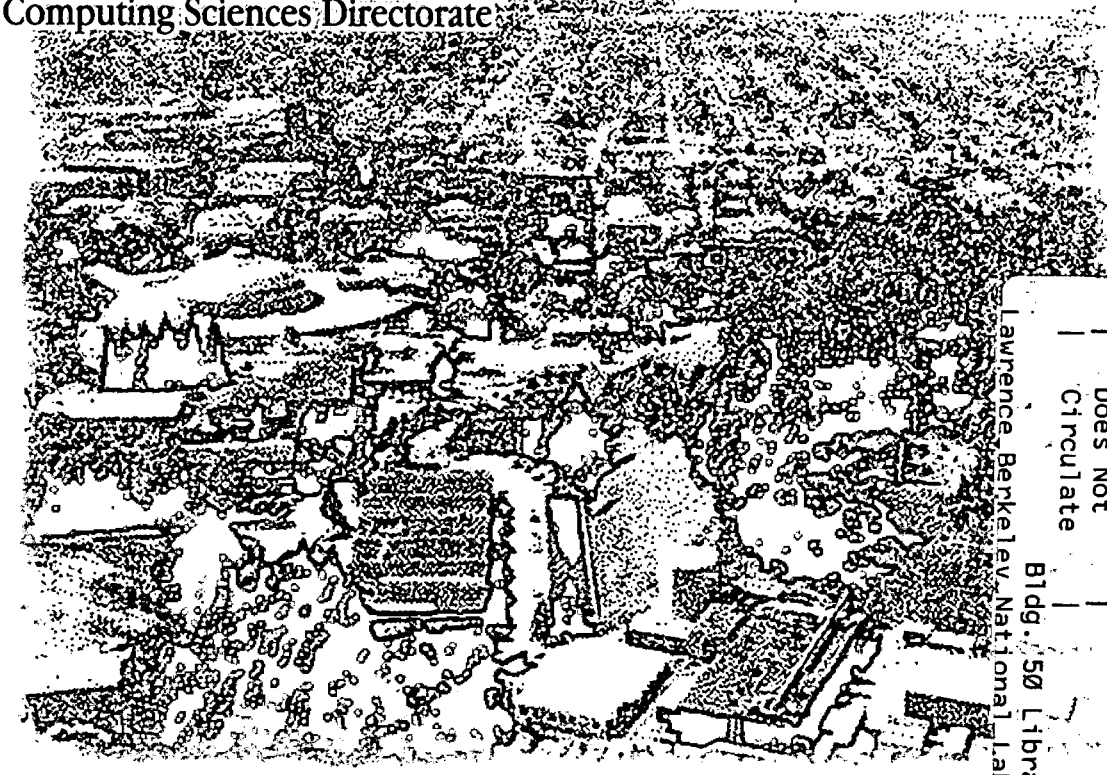
ERNEST ORLANDO LAWRENCE BERKELEY NATIONAL LABORATORY

How Are We Doing?

A Self-Assessment of the Quality of Services and Systems at NERSC (October 1, 1996 to September 30, 1997)

Issued January 1998 by the
High Performance Computing Department
William T. Kramer, Department Head

National Energy Research Scientific Computing Center
Computing Sciences Directorate



REFERENCE COPY
Does Not
Circulate
Bldg. 50 Library - Ref.
Lawrence Berkeley National Laboratory
LBNL-41522
Copy 1

DISCLAIMER

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

LBNL-41522

How Are We Doing?

A Self-Assessment of the Quality of Services and Systems at NERSC
(Oct. 1, 1996 to Sept. 30, 1997)

Issued January 1998 by the
High Performance Computing Department
William T. Kramer, Department Head

National Energy Research Scientific Computing Center
Computing Sciences Directorate

Ernest Orlando Lawrence Berkeley National Laboratory
Berkeley, California 94720

INTRODUCTION

Since its inception nearly 25 years ago, the National Energy Research Scientific Computing Center has provided its ever-expanding client base with the latest in scientific computing resources. A key element of NERSC's successful operation is its ability to anticipate and meet the diverse needs of clients. In order to further this strong working relationship, NERSC staff and clients meet periodically via ERSUG to share views, offer training and identify problems and solutions.

The decision to move NERSC from Livermore to Berkeley Lab in 1995 wasn't just to gain a change of scenery. The new, improved NERSC provides clients with state-of-the-art supercomputers as well as an extensive range of support services aimed at accelerating the pace of scientific discovery. When the center moved to Berkeley, the name was changed to the National Energy Research Scientific Computing Center. This signaled a new philosophy — one of enabling scientific computing, not just providing supercomputer cycles.

We have re-invented NERSC with a new architecture, a design that emphasizes both our high-performance computing systems and our commitment to excellent client service. Working in parallel, these two sides of our organization aim to give clients the necessary resources for conducting their research.

In fact, the success of NERSC is measured in large part by the quality of science produced by our clients. Our job is to give them the reliable tools they need — client support, software and access to computing resources.

To ensure that we are meeting those needs, we have established a set of 10 performance goals pertaining to our systems and service. These goals were developed in consultation with our staff, our client community and our stakeholders. Within the NERSC organization, we framed our goals through both top-down and a bottom-up processes. Each group within the organization came up with their own goals, and also reviewed those set by other groups. Senior managers then outlined overall management goals and reviewed the groups' goals to ensure that they supported the management goals and merged them into a single package. We then submitted this list to our DOE program sponsors for review and validation. The final step was to seek and gain endorsement from ERSUG.

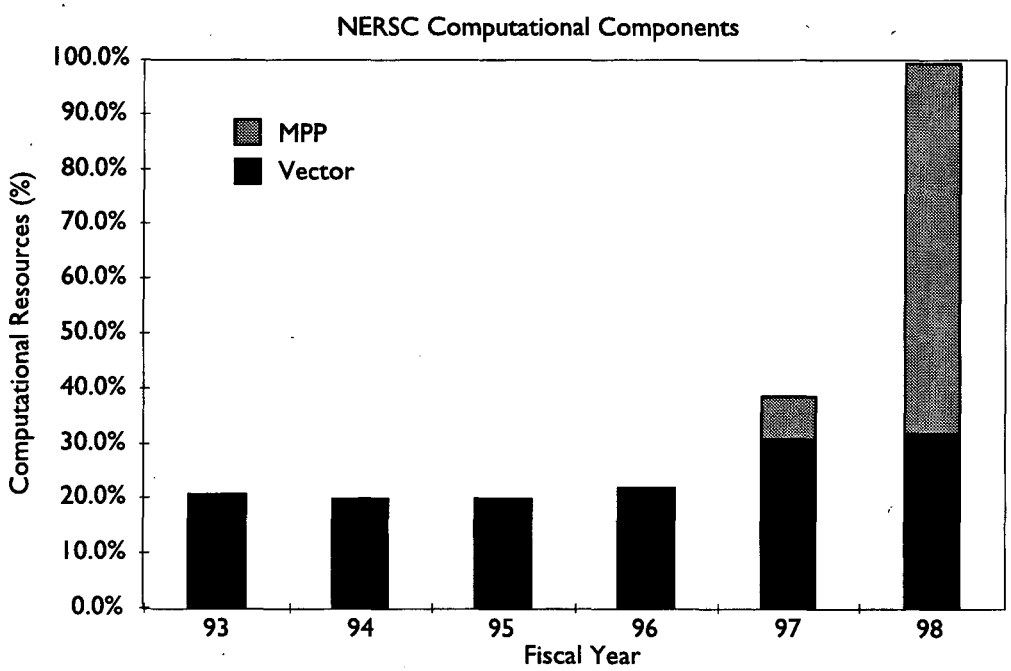
With our agreed-upon goals in place, we can now proactively gauge just how well we're doing in meeting those common expectations. We've tried to ensure that they reflect our efforts from a client's perspective, as opposed to an internal one. For example, a measurement of system availability needs to reflect the number of hours a machine is available to our clients, not how long it takes to identify a problem and initiate corrective action on our end. The goals we have set out cover the following areas:

- Reliable and Timely Service
- Innovative Assistance
- Timely and Accurate Information
- New Technologies
- Wise Technology Integration
- Progress Measurement
- High-Performance Computing Center Leadership
- Technology Transfer

- Staff Effectiveness
- Protected Infrastructure

To give our clients, our sponsors and our own staff a better idea of how we're performing, we've produced this report, covering our work from October 1996 through September 1997. We believe it shows we're constantly improving our efforts, yet also seeking other ways to do even better. As with any organization, there is always room for improvement. At the same time, improvements also raise expectations. Given this situation, NERSC intends to constantly assess, improve and set new standards for the systems and services we provide to clients.

At the same time, we will continue to provide our clients with the greatest amount of computer time we can allocate, and we are constantly striving to improve our facilities and procedures to optimize the number of cycles we can provide. The following chart demonstrates our past and anticipated future allocation of cycles.



RELIABLE AND TIMELY SERVICE

Goal: Have all systems and support functions provide reliable and timely service to their clients.

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 components of the NERSC organization must work hand in hand. When there is a problem, we respond promptly to acknowledge, address and correct it. As described previously, NERSC provides clients with both the high-performance computing systems and the expert services for achieving research goals. To measure how well we're doing, we have established metrics for system reliability and responsiveness to clients' problems.

The following chart shows various aspects of our systems' reliability since the NERSC facility was moved from Lawrence Livermore to Lawrence Berkeley National Laboratory. Figures in bold represent the measured time, while goals are shown in parentheses. Scheduled availability refers to the amount of time the systems are expected to be available (accounting for scheduled maintenance and upgrades), while gross availability refers to the overall availability. The "Mean Time to Repair" refers to the amount of time between a system failure and the point at which full service is restored to clients.

In most cases, our measured performance meets or exceeds the goals.

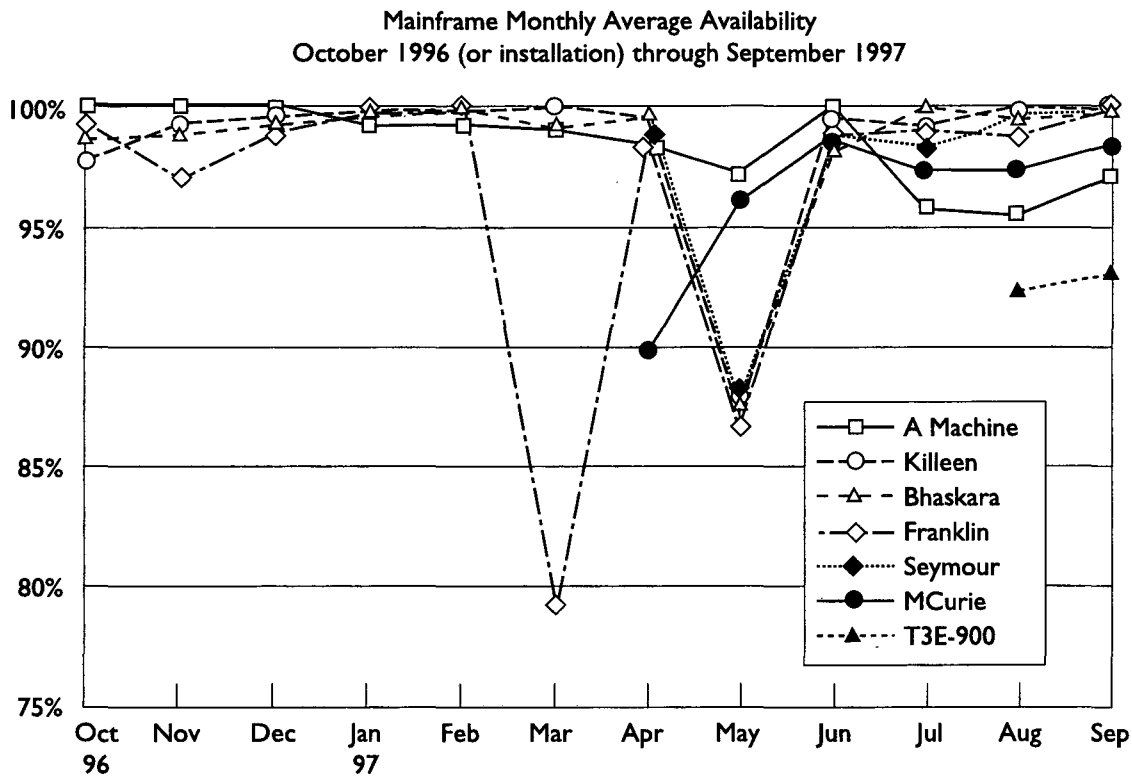
System Metrics Since Move
Measured (Goal)

	% Availability		Mean Time Between Interruptions	Mean Time to Repair
	Scheduled	Overall	(Hours)	(Hours)
Systems				
Vector Systems	**99** (96)	**98** (95)	**239** (96)	**2.4** (4.0)
Storage Systems	**99** (96)	**99** (95)	**92** (96)	**0.2** (4.0)
Parallel Systems	**98** (90)	**95** (85)	**43** (96)	**2.7** (4.0)
Workstations				
Servers (fs/gw)	**100** (96)	**100** (96)	**2856** (316)	**0.1** (8.0)
Clusters	— (96)	— (95)	— (96)	— (8.0)
SAS	**99** (99)	**99** (97)	**571** (490)	**1.1** (4.0)
Composite (VS, Storage, Local Network)	N/A (90)	N/A (90)		
Composite (VS, Storage, FS)	**98** (90)	**98** (90)		

This graph shows the monthly availability of each of the NERSC mainframe computers for the year from September 1996 through September 1997.

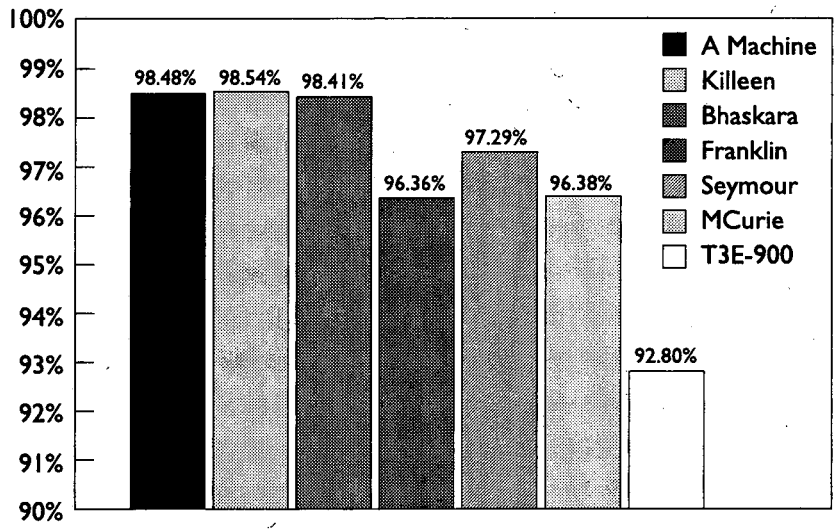
Two points worth noting:

- The J90 "franklin" dropped to about 80 percent availability in March 1997 when a series of six outages covering 145 hours were required to allow a file system reconfiguration.
- All four J90 machines were taken out of service May 28-31 to allow for upgrades to the cluster. The upgraded machines were returned to service two days sooner than anticipated.



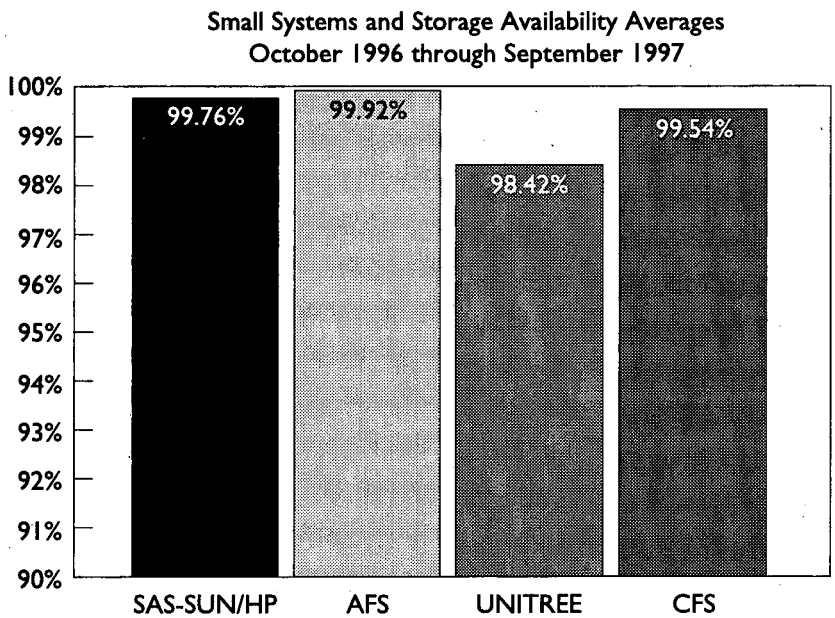
* MCurie was the name assigned to NERSC's Cray T3E-600 until the T3E-900 went on line in September 1997. At that point, the name MCurie was given to the larger machine and the smaller one was removed from service.

This graph shows overall average availability of NERSC mainframe computers for the year from September 1996 through September 1997.



* MCurie was the name assigned to NERSC's Cray T3E-600 until the T3E-900 went on line in September 1997. At that point, the name MCurie was given to the larger machine and the smaller one was removed from service.

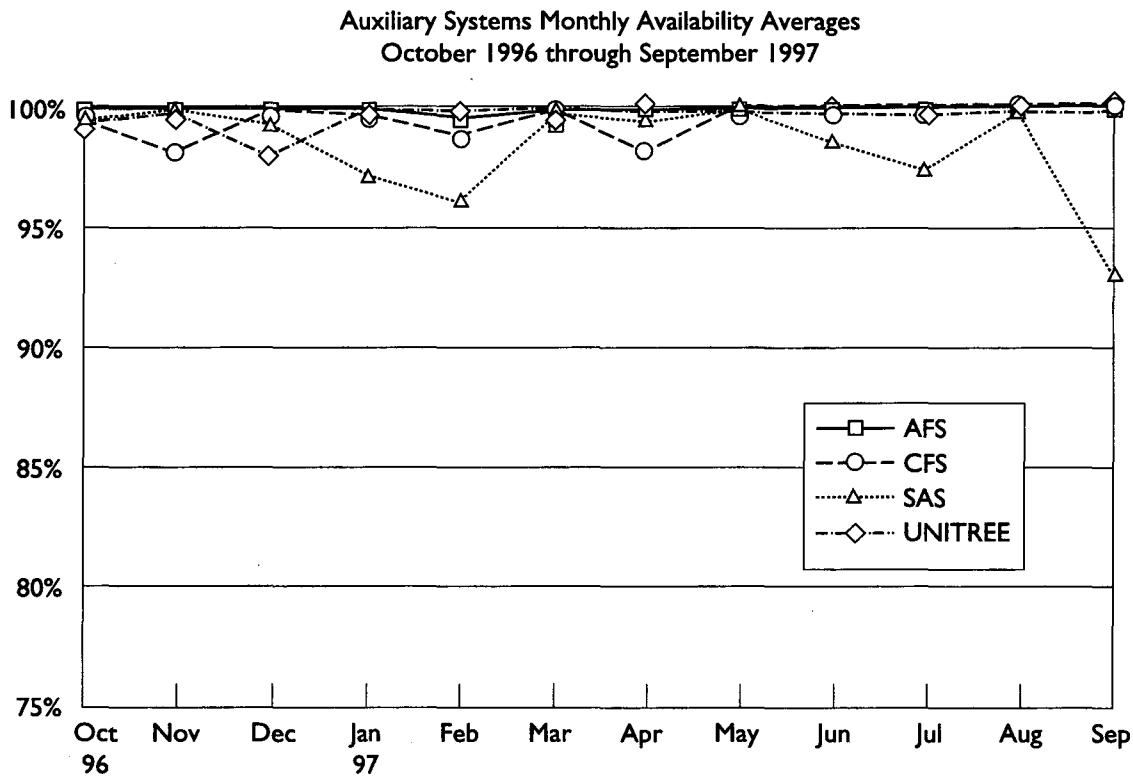
This chart shows yearly availability of small computing systems and storage systems.



The chart on this page illustrates monthly availability of small computing systems and storage systems.

The Unitree system's availability dropped to 96 percent in February 1997 due to problems with a disk array. The problems occurred sporadically over three days before they could be diagnosed and fixed.

In September 1997, Unitree availability was reduced to 93 percent due to a series of small outages and scheduled downtime. Together, these outages totaled 48 hours for the month.



NERSC's service goals are to respond to clients' problems within four working hours and to resolve at least 90 percent of those problems within two working days. The times referred to in this section refer to working days and do not include weekends or after-hours periods.

Here's how we're doing in terms of meeting those goals.

- Spot checks confirm that NERSC meets the goal of responding to problems within four hours.
- Between July 1, 1996, and May 15, 1997, 75.3 percent of all problems were resolved within two days. As the NERSC staff got up to speed, however, we made significant progress in meeting the 90 percent goal: Between March 11 and May 15, 1997, 93.1 percent of all problems were resolved within two days.

Not all problems can be resolved within two days, and in some cases these problems are put "on hold" as longer-term solutions are needed. Reasons for putting a problem on hold include software requests,

ongoing coding projects, "bugs" waiting for a vendor-supplied fix, and user not responding to a request for input within two days. These problems obviously take longer to resolve.

Problems not resolved within 72 working hours are escalated for more in-depth review. This escalation, now done manually, will be automated by January 1998 and this process will help ensure that outstanding problems are addressed. NERSC staff also periodically reviews problems and client requests to ascertain areas needing attention with an eye toward fixing them with minimum disruptions in service.

INNOVATIVE ASSISTANCE

Goal: NERSC aims to provide its clients with new ideas, new techniques and new solutions to their scientific computing issues.

In making the commitment to enhance clients' abilities to perform computational science, NERSC has initiated a range of activities to provide innovative assistance. From individual outreach to general tutorials, these programs demonstrate NERSC's commitment to become the scientific computing center of choice for the ER community. Here are some examples of our innovative service:

On the Lookout

As one of the NERSC consultants who routinely handles inquiries about computing problems, Majdi Baddourah provides timely help to clients with difficult computing problems. But Majdi, a member of NERSC's User Services Group, decided to go a step further. He took the initiative and began monitoring jobs that were performing slowly on NERSC's C-90 and J-90 machines. He then called some of the users and asked if they needed help in getting their source code optimized. Codes were optimized both for I/O and the CPU, then returned to the users for evaluation. This extra effort resulted in poor-performing codes running up to 10 times faster, benefiting all users of the machines.

In another case, NERSC consultant Peter Tang has suggested new algorithms to improve work done by clients. When one Grand Challenge research group asked Peter for help in making a central piece of code run faster, he studied their code and went a step further by suggesting use of a different algorithm. The suggestion resulted in speeding up the work by a fact of four, and Peter is continuing to look at other algorithms to achieve even greater improvements.

In another case, Jonathan Carter of User Services has also provided individual attention to solving various problems. For example, when one group was having trouble moving their FORTRAN 90 code to a new compiler, Jonathan went through 20,000 lines of code piece by piece and found bugs in the compiler. He then created a smaller test case, which led to a fix for the compiler problem. In another case, Jonathan worked with a researcher using a long-running code to create an automatic checkpoint/restart feature so that the data would be automatically saved every 20 minutes, thereby saving hours or days of work in the event of a system shutdown. In a third case, Jonathan set up an autotasking feature for fluid dynamics researchers whose code used 512 megawords of memory, about half the machine's capacity, but whose code used only 20 percent of the processing power (leaving 80 percent of the processors idle. By distributing the loop over more processors, the machine can now be fully utilized.

On a Roll

In fact, creating special outreach efforts to help minimize problems before they get started is becoming a way of doing business at NERSC. With DOE's announcement that NERSC would be a partner in helping multi-disciplinary teams from around the country solve seven of 12 "Grand Challenges" as identified by DOE's Office of Energy Research, Berkeley Lab computer scientists are rolling out a virtual red carpet to participating researchers around the country.

Through our "Red Carpet Program," NERSC staff members are working to build individual working relationships with clients at other national labs and universities tackling such issues as cleaning up nuclear waste, supporting international research in magnetic fusion energy, designing particle

accelerators, understanding the structure of the smallest building blocks of matter, and improving analysis of the growing body of data on human genetic makeup.

From providing new user training to integrating separate physics software packages into a cohesive program to developing new algorithms, members of NERSC's staff will meet and work with grand challenge researchers. NERSC staff is holding site meetings with each team to assess needs on both sides of the equation and to deliver the needed services and support.

The goal is to make NERSC an indispensable partner in solving the grand challenge problems, according to Scientific Computing Group Leader Tammy Welcome. Each group of grand challenge researchers has been assigned a scientific computing expert from NERSC to both learn about the groups' computational requirements and to better inform the researchers about NERSC's capabilities and the level of collaboration the center can provide. In addition to being experts in scientific computing, the assigned NERSC staff members also have extensive knowledge of related scientific fields, such as physics, chemistry and materials science.

Robert Ryne, a physicist at Los Alamos National Lab and the principal investigator in a grand challenge using NERSC to develop computational tools for designing the next generation of particle accelerators, has found the red carpet treatment to be smoothing his research path. As an example of how the program is working, Ryne said NERSC staff member showed him how to solve a programming problem with much less effort than Ryne had thought possible.

On a Learning Curve

As part of our efforts to provide innovative assistance, we have experimented with different approaches. Some have worked well, while others sent us back to the drawing board. For example, to reduce travel time and costs, we tried using videoconferencing as a training tool. On our end, there were problems in learning how to present material via video. On clients' ends, there were difficulties in scheduling facilities, and these increased as we tried to scale up the sessions to reach more sites. As a result we have decided to pursue Web-based technologies. This will allow clients to tap our expertise at their desktop and on their own schedule.

TIMELY AND ACCURATE INFORMATION

Goal: Provide timely and accurate information and notification of system changes to the client community so they can most effectively use the NERSC systems.

The NERSC staff strives to provide our clients with timely and accurate information which may affect those clients' research efforts. Not only do we give adequate notice of changes and outages, we also try whenever possible to provide an explanation of the reasons behind the change and the expected impact on clients. As an example of timeliness:

- Planned system changes were announced at least seven days in advance (except in one instance); all planned system outages announced at least 24 hours in advance.
- All system changes and planned outages were announced in advance on the NERSC "What's New" Web page. Some major changes were also announced by e-mail to PIs.

NERSC has also presented various workshops, training sessions and talks to proactively help clients adapt to using the newest computing technologies. These talks, held primarily at Berkeley Lab, but also in New Jersey, drew audiences ranging from half a dozen to more than 50 people. Here are some examples:

- When NERSC's new Cray T3E was coming on line, NERSC held a series of T3E training classes prior to and during the period the machine was being put through its acceptance testing. During October, November and December 1996, a series of six classes were held on various aspects of the T3E to help prepare clients for using this new machine.
- In January 1997, a two-day training workshop was held at Princeton Plasma Physics Lab in conjunction with the ERSUG meeting. In addition to providing an overview of parallel computing, the workshop included four sessions on optimizing code for the T3E.
- In February 1997, a two-day training workshop was held at Berkeley Lab for about 50 researchers in the Laboratory Directed Research and Development program utilizing NERSC.
- The NERSC staff also gave a talk for researchers currently using the Cray C-90 who were to begin using the J-90 cluster. This presentation, given before the changeover, highlighted the differences between the two machines and helped prepare clients for the transition.

Our goal was to provide an average of one training activity per month. Over the course of the year, we exceeded this by providing nine seminars and six workshops.

Finally, Web-based communications allows the NERSC staff to immediately inform clients of pending changes and provide access to the latest information.

NEW TECHNOLOGIES, EQUIPMENT, SOFTWARE, AND METHODS

Goal: Ensure that future high-performance technologies are available to ER computational scientists.

Since its inception, NERSC has provided its client community with the most up-to-date computing resources available. Since moving from Livermore, two Cray T3E supercomputers, a cluster of four Cray J90se computers and a high-performance storage system have been added to NERSC's equipment roster, giving the center one of the most powerful lineups of computer power in the country. As new machines are introduced to the center, systems experts carefully analyze performance and work with manufacturers to ensure that the equipment meets the high-performance needs of NERSC clients.

- Cray T3E

  To provide state-of-the-art capabilities in scalable parallel computing, NERSC received delivery of the 512-processor CRAY T3E-900 on July 14. The NERSC T3E-900 is the largest I/O system built to date and is a first-of-its-kind configuration. The fully configured machine offers 1.5 terabytes of disc storage, a read/write capability of 800 megabytes per second, and 128 gigabytes of memory. The final acceptance test on the T3E-900 will require the system to have two FDDI interfaces, two HIPPI interfaces, 170 Fiber Channel Disks and 96 SCSI disks.

  On August 20, NERSC announced a milestone in high-performance computing: the successful checkpointing/restart of the T3E-900, which was achieved twice in one week. This is believed to be the first time checkpointing has been accomplished on a massively parallel processing system. Successful checkpointing will allow the NERSC staff to suspend system work for maintenance and upgrades with minimal disruption and downtime for clients. Checkpointing will also allow us to efficiently move jobs between processors or make larger pools of processors available for bigger jobs.

  NERSC has also reached an agreement with Cray Research to buy the T3E-600 on the floor to improve the mid-range computing capability at Berkeley Lab. The system should be up and running by December 1997.

- Storage

  NERSC integrated the new hardware procured in the FY96 major hardware upgrade. We installed uninterruptible power and increased the UniTree disk cache ninefold from 94 GB to 846 GB. We expanded the archival storage bandwidth from 27 MB/s to a potential 252 MB/s and expanded capacity from 33.6 TB to 106 TB (uncompressed). We added a second storage control processor and disks to store UniTree/High Performance Storage System (HPSS) databases. We purchased hardware and software to allow use of the EMC disk in future storage environments and acquired six IBM processors to use as "mover" machines for HPSS.

  After completing a market and benchmarking survey of storage systems and evaluating the options, NERSC began developing an HPSS capability for the user community. Over the next two years, this will involve setting up an HPSS test environment and converting the UniTree and CFS environments to HPSS. HPSS should go into production in February 1998. We continue providing support for the existing CFS environment until it can be merged with the upgraded capabilities under development.

- Cray J90s

  The J90s *franklin* and *seymour* now have queues that can accept jobs up to 1.6 Gigabytes, while *bhaskara* has an additional queue with a limit of 4 Gigabytes, which represents half the physical memory on the machine. These expanded queues allow the J90 cluster to accept very large batch jobs. These batch machines mount the interactive machine *killeen*'s home directories via NFS so that user executables and files are available to each machine.

- SuperHome environment

  NERSC created the SuperHome environment on the Cray J90s to simplify file system access and increase file quotas. SuperHome provides one user home on the interactive J90 and exports that file system to the batch J90s. This environment allowed NERSC to increase the amount of permanent J90 disk space to 5 GB per user and simplified management of the systems. Over Labor Day weekend, SuperHome was also installed on the C90, increasing the size of users' home directories 1,000 fold. NERSC is also working to develop a Common Super Home for all platforms and we expect to deploy this system in the near future.

- NOW/COMPS/Millennium
  - NERSC is evaluating technology produced by the UC Berkeley NOW (Network of Workstations) project.
  - NERSC is developing a communication infrastructure for COMPS (Cluster of Multiprocessor Systems) and evaluating the proposed VIA (Virtual Interface Architecture) standard.
  - The UC Berkeley/Intel Millennium project will establish a number of workstation clusters on the UC campus and a small cluster at NERSC.
  - NERSC purchased and began installing equipment as an early deployment site for Sun Microsystems' next-generation distributed shared memory system. The system is scheduled to be operational by February 1998.

- PDSF

  NERSC has established the Particle Detector Simulation Facility as a full-production system used by several collaborations in the high-energy and nuclear physics community.

- UNICOS 9.0 for C90

  On May 20, 1997, the C90 (*A machine*) was upgraded to the UNICOS 9 operating system, which is also used on the J90 cluster. This upgrade provides a consistent environment across several Cray platforms.

- Requirements-based batch scheduling

  The Network Queuing Environment (NQE), with its intelligent scheduler, is designed to automatically assign the required processing resources to each computing task while balancing the workload among the PEs. We have adopted NQE and are currently working

with the system and clients to improve the efficiencies of the T3Es and J90s. NQE, however, is not yet living up to its full potential.

- Visualization software inventory for NERSC systems

  The Visualization Group evaluated various applications and published their recommendations on the Web. A new remote visualization server — an SGI Onyx 2 — has been installed to provide an environment for maintaining consistent and current visualization software across NERSC systems, as well as supporting our clients' scientific visualization needs across wide area networks.

WISE TECHNOLOGY INTEGRATION

Goal: Insure all new technology and changes improve (or at least do not diminish) service to our clients.

Buying the "latest and greatest" machinery on the market is no guarantee of success (remember the Edsel). To ensure that the equipment we put on line meets our demanding criteria as well as our clients' needs, NERSC computer scientists carefully evaluate and rigorously test all systems under consideration. Our staff evaluates and tests new technologies developed by industry, national laboratories and universities to determine which systems should be considered for use by NERSC, and which should be discarded. Only when they prove to be effective enough to handle their designated workload are the systems put on line. Various NERSC groups then work with clients and stakeholders to put the systems into production, obtain feedback and further improve the technologies. Below, we list our subgoals and the actions we have taken to meet them.

- Thoroughly test systems and software. Examples:
  - The Cray T3E-600 machine initially did not meet performance criteria and was therefore subjected to much longer and more rigorous testing. NERSC and Cray representatives worked closely during this time to identify and resolve problems. The end result was that NERSC clients had to wait a little longer than expected, but were rewarded with a truly scalable MPP machine. The solutions generated by the NERSC-Cray cooperation benefited other centers with T3Es as well.
  - The T3E-900 underwent rigorous acceptance testing, which included several scientific problems designed to fully test—and exploit—the machine's capabilities. Also during this test period, NERSC and Cray staff successfully executed two checkpoint/restart operations on the T3E, the first time such a milestone was achieved anywhere on an MPP. The rigorous testing and the checkpoint/restart signaled the successful transition of the MPP machine into a production environment. To ensure that NERSC clients weren't left in the lurch, the T3E-600 remained available during most of the T3E-900 acceptance period.
  - The User Services Group conducted a major evaluation of existing NERSC third-party applications and library software. After a thorough review of applications for physics, chemistry, mathematics and visualization, underutilized software packages were eliminated to make better use of NERSC resources. A similar review is being carried out for NERSC utilities. The User Services Group also evaluates new software on an ongoing basis.
  - The Visualization Group tests and evaluates new software on an ongoing basis.
  - As a step toward implementation of HPSS (High Performance Storage System), NERSC began using HPSS for system dumps.
- Fully inform the user community about the impact of changes and enhancements of delivered systems. Examples:
  - T3E classes (listed above)
  - The "J90 Cluster Changes" page on the Web informed users about memory upgrades, queue configuration changes, etc.

- The UniTree system was introduced with little disruption.

- Creating workable solutions for clients.
  - The User Services Group has developed many workaround solutions.
  - NERSC has carefully and fully tested the T3E, moving a MPP technology into a full production environment.

- If a detrimental impact is unavoidable, NERSC has the responsibility to ensure that the benefits of the changes will outweigh the detriments. Examples:
  - Since FORTRAN 77 is no longer supported by Cray Research, NERSC is working with clients to HELP convert their codes to FORTRAN 90. The resulting benefits include programming in a modern language that is vendor-supported, can be used in parallel processing and is still compatible with FORTRAN 77.
  - When a conflict was found between NFS3 (Network File System) and FORTRAN 90—a bug that caused programs to crash—the J90 cluster was switched back to NFS2 until the problem was diagnosed and corrected.

PROGRESS MEASUREMENT

Goal: As a national facility serving thousands of researchers, NERSC has a responsibility to our clients to measure and report on how we're doing in terms of providing service, support and facilities.

This report is one of several documentation efforts on how we're doing in terms of meeting the goals we have set for ourselves. We also report—and get feedback—regularly at biannual ERSUG meetings and post statistics on the Web.

To track our progress in meeting these goals, we have established a set of metrics to measure our performance over a wide range of efforts. These measures range from the readily apparent, such as how long it takes to resolve a client's problem and the percentage of time our systems are available to clients, to the less obvious, such as how often our staff members transfer their technological expertise to the high-performance computing community. These metrics and resulting data are presented throughout this report.

Essential to meaningful measurement is taking the perspective of our client community. For example, one possible metric for system availability is the time it takes between a system shutdown occurring and the point at which corrective action begins. It could be argued—and some organizations do—that this is an accurate account of the down time. From a researcher's point of view, however, the meaningful measurement is the time between something going wrong and when it gets fully back on track—and that's the measurement we have adopted.

This set of metrics is the first phase of what is intended to be a comprehensive system for evaluating all aspects of NERSC-related activities. Eventually, we plan to have a method for evaluating the scientific research carried out using NERSC resources.

As the unparalleled leader in unclassified computing in the United States, NERSC is committed to providing our clients and DOE sponsors with the most reliable and productive scientific computing resources. An integral part of this is repeatedly taking stock of our efforts, using this as a benchmark for further improvements, and then continuing the process. We also plan to share these results with our client community on a regular basis.

HIGH-PERFORMANCE COMPUTING CENTER LEADERSHIP

Goal: Improve methods of managing systems within NERSC and be the leader in large-scale computing center management.

As the most powerful unclassified supercomputing center in the nation, NERSC is a leader in implementing new technologies, helping define the supercomputing center of the future and delivering the scientific, engineering and management expertise to keeps us at the forefront of this fast-moving field.

As one of the world's first computing centers to put a Cray T3E into a demanding production environment, we are regularly contacted by others considering such a move. Such organizations as the U.S. Army Corps of Engineers, the National Computational Science Alliance and Georgia Tech in this country and computing centers in France and Korea have tapped our expertise and experience in planning their future facilities.

Some other examples of our leadership include:

- "Reinventing the Supercomputer Center"

  Three managers and six staff gave a tutorial with this title at Supercomputing 96. Also, "Reinventing the Supercomputer Center at NERSC" by Horst Simon was published in *IEEE Computational Science and Engineering*, Vol. 4, No. 3, July/September 1997.

- "The New NERSC Program in FY 1997: An Overview of Current and P roposed Efforts"

  This document, written by Horst D. Simon, C. William McCurdy, William T. Kramer, and Alexander X. Merola outlines how Berkeley Lab plans to establish itself as both a center of excellence in scientific computation and a leading player in defining the future shape of high-performance computing.

- "Best Practices for Supercomputing Centers"

  NERSC was a major example site in this Fall 1996 study by IDC (International Data Corporation).

- Survey of storage systems

  After completing a market and benchmarking survey of storage systems and evaluating the options, NERSC began developing an HPSS capability for the user community as part of our long-range plan to continue providing state-of-the-art storage capabilities.

- First center to successfully checkpoint/restart jobs on an MPP machine

  In August 1997, NERSC achieved a long-sought milestone by successfully checkpointing jobs on the Cray T3E-900, shutting down the system and then restarting it without any loss of data. The procedure was accomplished twice in one week and will further boost the productivity of the 512-processor T3E.

- Development of a wide-area distributed storage system

  The Distributed-Parallel Storage System (DPSS), a scalable, high-performance, data storage system has been developed by the Data Intensive Distributed Computing Group with DARPA funding. A project of the NERSC Division, the DPSS is a collection of wide-area distributed disk servers which operate in parallel to provide logical block level access to large data sets. Operated primarily as a network-based cache, the architecture supports cooperation among independently owned resources to provide fast, large-scale, on-demand storage to support data handling, simulation, and computation in a wide-area inter-networked environment. We are also working to integrate the system into the Particle Detector Simulation Facility (PDSF) and High Performance Storage System (HPSS) as NERSC.

- Providing expertise for new technology development

  NERSC staff members participate in various organizations setting the pace for new technology development. For example, the head of our Systems Group is a member of Silicon Graphics' customer advisory board. Members of our Mass Storage Group are members of the HPSS executive and technical committees, and our Future Technologies Group leader is a member of the Message Passing Interface standards committee.

  The Mass Storage Group is also collaborating with LBNL's Scientific Data Management Research Group (led by Arie Shoshani) to develop and evaluate improvements for large-scale data management.

In short, many organizations look to NERSC to provide essential expertise in the fields of high-performance computing systems.

TECHNOLOGY TRANSFER

Goal: Export knowledge, experience, and technology developed at NERSC, particularly to and within NERSC client sites.

NERSC has recruited a number of experts in a number of fields related to scientific computing. In assembling the staff at Berkeley, NERSC has hired employees from some of the nation's pre-eminent scientific and research facilities. Their expertise is being shared with NERSC clients, as well as other scientific communities. Here are some examples of how we're transferring our technological expertise:

- Released the software "Parallel supernodal method for sparse LU-decomposition," Xiaoye "Sherry" Li.
- Authored papers:
  - "Tracer transport in a stochastic continuum model of fractured media," Y. W. Tsang, C. F. Tsang, F. V. Hale, and B. Dverstorp, *Water Resources Research* 32:10, October 1996.
  - "An asynchronous parallel supernodal algorithm for sparse Gaussian elimination," Xiaoye "Sherry" Li, *SIAM Journal on Matrix Analysis and Applications* (to appear).
  - "Prediction of a New Type of Strain Induced Conduction Band Minimum in Embedded Quantum Dots," A.J. Williamson, A. Zunger and A. Canning, Phys. Rev. B, Rapid Communications.
  - "Fault Tolerant Matrix Operations for Networks of Workstations Using Diskless Checkpointing," Youngbae Kim, Journal of Parallel and Distributed Computing.
- Presented tutorials:
  - "Reinventing the Supercomputer Center," Horst Simon et al., Supercomputing '96.
  - "Material science simulations on Parallel Computers," Andrew Canning, NERSC Technical Seminar Series (videoconference).
  - "Memory locality and parallelism in sparse Gaussian elimination," Xiaoye "Sherry" Li, NERSC Technical Seminar Series (videoconference).
  - "Heterogeneous Network Computing Using Parallel Virtual Machine (PVM)," Youngbae Kim, NERSC Technical Seminar Series.
  - "Diskless Checkpointing and Fault Tolerance," Youngbae Kim, NERSC HPC Technical Seminar Series.
  - "Virtual Reality in Khoros," Visualization Group, Khoros Symposium.
  - "A Practical and Intensive Introduction to MPI," Bill Saphir, SC 96.
  - "Virtual Reality in Khoros," Wes Bethel, Khoros symposium.
- Conference presentations and proceedings:
  - "So optimization breaks your code ...," R. K. Owen, *Proceedings: 38th Semi-Annual Cray Users Group Meeting*, Charlotte, North Carolina, Oct. 14–19, 1996.

- "Fast and accurate evaluation of the radon transform," Peter Tang, *Proceedings of the 1997 International Conference on Imaging Science, Systems, and Technology.*
- "New Results and Implementations for the NAS Parallel Benchmarks II," Bill Saphir, SIAM Parallel Processing for Scientific Computing conference.
- "Fault tolerant matrix operations using Checksum and Reverse Computation," Youngbae Kim, Frontiers '96.
- Conference presentations and proceedings (continued):
- "Fault tolerant matrix operations for networks of workstations using multiple checkpointing," Youngbae Kim, HPC Asia '97.
- "Parallel supernodal method for sparse LU-decomposition," Xiaoye "Sherry" Li, SIAM Sparse Matrix Meeting '96.
- "Climate data assimilation," Chris Ding, Supercomputing '96.
- Memory locality and parallelism in sparse Gaussian elimination, Xiaoye "Sherry" Li, 8th SIAM Conference on Parallel Processing for Scientific Computing.
- "A parallel sparse block matrix solver and climate data assimilations," Chris Ding, 8th SIAM Conference on Parallel Processing for Scientific Computing.
- "Direct and cached parallel quantum chemistry algorithms," Adrian Wong, West Coast Theoretical Chemistry Conference.
- "Quantum Simulations Using Linear Scaling Methods," G. Galli, J. Kim, A. Canning, and R. Haerle, Materials Research Society Fall 1997 Meeting proceedings.
- "Assembling Small Fullerenes: A Molecular Dynamics Study," G. Galli, A. Canning and J. Kim, Materials Research Society Fall 1997 Meeting proceedings.
- "COMBAT: The Cosmic Microwave Background Analysis Tools Project," Julian Borrill, Advanced Information Systems Research Projects Workshop, NASA Goddard.
- "Use of VRML for scientific visualizations," "An Overview of NERSC/LBNL's visualization efforts," "Development of a Threaded Haptic Interface Driver for Scientific Computing," and participation in a PC Graphics panel discussion, Wes Bethel, Stephen Lau, Kevin Campbell, 1997 DOE Computer Graphics Forum.
- NERSC staff presented various demonstrations and displays at the SC 96 conference in Pittsburgh.

• Workshop presentations and working groups:

- Sisira Weeratunga, M3D Developers Workshop, PPPL, January 1997.
- Tammy Welcome, Agency-wide software tools center, Workshop on Software Tools for HPC Systems.
- Chris Ding, "An evaluation of HPF for practical scientific algorithms" (poster), First HPF Users Group Conference.
- Chris Ding, 1997 Petaflops Algorithms Workshop.

- Osni Marques and Kesheng "John" Wu, Workshop on the Use of Iterative Methods for Large-Scale Eigenvalue Problems.
- Kesheng "John" Wu, "Restarted Versions of DQGMRE," International Workshop on Computational Science and Engineering, China.
- Source code and T3E BOFs led by NERSC staff, Silicon Valley Cray User Group (CUG) Meeting 1997.
- Osni Marques, "Large Scale SVD Computations in Linear Geophysical Inverse Problems," II Pan-American Workshop in Computational and Applied Mathematics, Brazil.
- Youngbae Kim, "An Experience with HPF on a Massively Parallel Machine," HPF User Group Meeting, Santa Fe.
- Stephen Lau, "Overview of Visualization Techniques and Software", May 1997 ERSUG meeting.

- Invited talks :
  - Chris Ding, "Climate modeling techniques," NASA Goddard.
  - Xiaoye "Sherry" Li, "Memory locality and parallelism in sparse Gaussian elimination," University of California, Santa Barbara, Computer Science Department.
  - Xiaoye "Sherry" Li, "Sparse matrix packages, Boeing.
  - Wes Bethel, "Implementing Virtual Reality Interfaces for the Geosciences," Virtual Reality in the Geosciences conference.
  - Andrew Canning, "Parallel computing for material science applications," NASA Ames Laboratory.
  - Andrew Canning, "Parallel Material Science Calculations on the T3E," Ames Laboratory, Iowa.
  - Andrew Canning, "Parallel Algorithms for Material Science Applications," Lawrence Livermore National Laboratory.
  - Youngbae Kim, "Diskless Checkpointing and Fault Tolerance, KwangJu Institute of Science and Technology, South Korea.
  - Bill Saphir, "Introduction to MPI," University of California, Berkeley, Computer Science 267 class.
  - Terry Ligocki, "Scientific Visualization," University of California, Berkeley, Computer Science 267 class.

- Sponsored workshop:
  - In conjunction with the U.S. Environmental Protection Agency and the International Standards Organization, NERSC staff conducted a three-day workshop aimed at creating standards for organization metadata to facilitate better sharing of information.

- Project report:
  - Wes Bethel, Janet Jacobsen, Nancy Johnston (all LBNL), Andy Austin, Mark Lederer (BP Exploration), and Todd Little (Landmark Graphics Exploration), "Advanced Flux Visualization and Virtual Reality for Reservoir Engineering." This April 1997 report capped a project funded by the Advanced Computational Technology Initiative to leverage computing facilities at LBNL to produce software for visualizing multi-phase fluid flow data.

- Hosted visitors from numerous other sites and facilities, including Department of Energy, Oak Ridge National Laboratory, National Renewable Energy Laboratory, Idaho National Engineering & Environmental Lab, Ames Laboratory, Princeton Plasma Physics Laboratory, University of California, University or Nebraska, University of Arizona, Science University of Tokyo, University of Victoria, Korea Institute of Science & Technology, Institut National des Telecommunications, IRRMA Lausanne, Stanford Linear Accelerator Center, California Institute of Technology and the National Center for Atmospheric Research.

STAFF EFFECTIVENESS

Goal: Leverage staff expertise and capabilities to increase efficiency and effectiveness.

An important part of the agreement to move NERSC from Livermore to Berkeley was a commitment to do more with less. The "less" side of this equation is easy to measure. NERSC is currently operating on a smaller budget than before and with fewer employees. For example, NERSC had a full-time technical staff of 79 employees in FY 1994. In FY97, we have a staffing level of 62 (though not all positions have been filled). The annual budget has also been reduced, from a high of $40 million to the current level of $29 million. On the other side of the equation, NERSC now addresses expanded expectations and responsibilities. In order to meet the "more" requirement, NERSC staff and resources must be carefully deployed for maximum effectiveness and efficiency. Here are some examples of how we're achieving that goal:

At LLNL, NERSC's hardware consisted of three production vector supercomputers, one production MPP machine and one storage system. The current configuration includes five production vector supercomputers, two production MPP machines, two production storage systems and a third developmental storage system.

Previously, NERSC printed a bimonthly newsletter/magazine called "The Buffer" for clients and staff. This required two full-time employees and part-time contributions from many others. This has been replaced with a monthly Web-based NERSC Research newsletter, produced by the equivalent of a quarter-time position. The electronic format saves printing and design costs, conserves resources and delivers more timely information.

NERSC has also adopted an e-mail notification approach to keep clients informed. By electronically delivering monthly accounting reports, allocation notices and other information, NERSC is saving time, money and materials. This process also requires less staff time and provides faster distribution of information.

For FY98 allocation requests, NERSC refined and integrated with other systems its Web-based ERCAP request system. As with any wide-scale application which much serve the needs of a very varied client community, the ERCAP route hit an occasional bump. However, we expect the lessons learned this year to translate into a much smoother procedure next year. We expect there to be significant time savings as the allocation request procedure becomes fully automated from what was once a time-consuming manual activity.

Training for NERSC clients is also being improved with an eye toward greater effectiveness. In addition to on-site training, we have also implemented a Web-based training program for the Cray T3E and are exploring the use of videoconferencing tools for remote training sessions.

We also hold monthly conferences with ERSUG members, producing more frequent two-way communication with representatives of our key client groups.

Finally, we believe we have "built a better qualified staff" in Berkeley and have achieved a more flexible workforce to quickly support new technologies. For example, in 1994, NERSC's 79-member staff included 42 with degrees, 21 with advanced degrees and 11 Ph.Ds. In 1997, 44 of the 62 staff members have degrees, with 24 having advanced degrees and 21 having Ph.Ds. Although having a degree is not alone a measure of effectiveness, the total picture is one of an intellectual center positioned to help clients in new areas and with new perspectives.

The bottom line is that we believe NERSC continues to offer excellent service to clients and is constantly looking for new ways to meet that goal with approaches that are cheaper, better and faster.

PROTECTED INFRASTRUCTURE

Goal: Provide a secure computing environment for NERSC clients and sponsors.

The security of NERSC's systems are a paramount concern to the Berkeley Lab staff, our DOE sponsors and our national client base. To date, there have been no security incidents involving NERSC.

To ensure the security of the NERSC systems, we have taken various precautions, including:

- Identifying and eliminating 3,000 unused accounts to limit access only to active clients.
- Responding to security issues as needed.
- Installing software to monitor and warn of potential security incidents.

CONCLUSION

According to Bill McCurdy, former head of NERSC and now associate laboratory director for Computing Sciences at Berkeley Lab, NERSC's move from Lawrence Livermore to Lawrence Berkeley National Laboratory has entered the realm of DOE folklore.

The job was done ahead of schedule, within budget and in a manner transparent to our extensive client base. In a room where scientists once handed in reams of 80-column computer cards for batch processing sits one of the world's largest combinations of computing and networking resources. Since its arrival in Berkeley, NERSC has added seven large-scale computers, a high-speed storage system and more than 30 new employees from some of the top computing science organizations in the country. In all, the outlook is excellent for the organization.

But a bright future matters little to a client whose job won't run for an unknown reason, or a researcher whose work is interrupted when a system shuts down. To document the quality of our services, as well as our proactive responses, we have instituted the system of metrics outlined in this report. We believe the data demonstrate that we're doing a good job in meeting clients' needs, and that there is also room for us to do better. We plan to do just that, and will report back to you on our further progress next year.

ACKNOWLEDGMENT

This work was supported by the Office of Energy Research, Office of Computational and Technology Research, Mathematical, Information, and Computational Sciences Division, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

This report was compiled by Jon Bashor. For more information, call 510-486-5849, or write to JBashor@lbl.gov. To learn more about the National Energy Research Scientific Computing Center, visit our web site at: http://www.nersc.gov.

Ernest Orlando Lawrence Berkeley National Laboratory
One Cyclotron Road | Berkeley, California 94720

Prepared for the U.S. Department of Energy under Contract No. DE-AC03-76SF00098