LBNL/PUB-5465, Vol. 2

Implementation Plan

National Energy Research Scientific Computing Center

FY2002–FY2006

Ernest Orlando Lawrence
Berkeley National Laboratory

May 25, 2001

LBNL/PUB-5465 Vol. 2 Rev. 2

NERSC Strategic Implementation Plan

2002–2006

William Kramer, Wes Bethel, James Craw, Brent Draney, William Fortney,
Brent Gorda, William Harris, Nancy Meyer, Esmond Ng, Francesca Verdier,
Howard Walter, Tammy Welcome

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

September 1, 2002

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

NERSC Strategic Implementation Plan 2002–2006

Abstract

NERSC, the National Energy Research Scientific Computing Center, is DOE's premier scientific computing facility for unclassified research. Over the last five years, NERSC — located at the Ernest Orlando Lawrence Berkeley National Laboratory — has built an outstanding reputation for providing both high-end computer systems and comprehensive scientific client services. At the same time, NERSC has successfully managed the transition for its users from a vector-parallel to a massively parallel computing environment. Building on a foundation of past successes, the NERSC center submitted a proposal to DOE for expanding its vision and activities with new directions over the next five years. In November 2001, DOE accepted the proposal and committed to five more years of funding. This Implementation Plan expands the proposal, confirms NERSC's continuing commitment to providing *high-end systems* and *comprehensive scientific support* for its users, and describes how these two core components will be enhanced.

CONTENTS

1 INTRODUCTION

The NERSC Strategic Proposal for FY2002–FY2006, in a separate document, provides NERSC's vision for its activities and new directions over the next five years. This Implementation Plan provides strategy and explains how NERSC will make that vision a reality.

1.1 NERSC's Strong Support of Its User Community, and Emerging Changes in the Practice of Computational Science

NERSC's fundamental mission is to support and advance the scientific research of its clients. The Center is, and has been historically, strongly user oriented. The NERSC Strategic Proposal commits to continuing that support and user orientation.

The Center provides a range of core user services, using a professional staff, via face to face, video, telephone and a variety of Web and email mechanisms. The core services span the range from consulting on common problems in using NERSC's supercomputers to intensive help with debugging, algorithm performance, and special needs of specific projects. They are available to all users, and generally involve short-term interactions with staff members. In addition, the center provides long-term collaborative support by computational specialists in algorithms, computer science, and applied mathematics to selected projects that goes well beyond the core services. This sort of support was developed for the DOE Grand Challenge teams of the late 1990s.

To maintain strong connections with the user community, NERSC established several mechanisms that will be continued. The NERSC User Group (NUG) meets twice a year, and the Executive Committee of the NUG has monthly telephone conferences with the Center staff. We will continue to carry out annual surveys and publish their results together with other data in our annual self-assessment.

NERSC plans to maintain its basic strategy of High End Systems (HES) and Comprehensive Scientific Support (CSS). The NERSC Strategic Proposal explains that there are two ways to augment and improve this approach. Those are: (1) to provide intensive support of multidisciplinary teams of collaborators called "Scientific Challenge Teams," such as those funded by DOE's SciDAC initiative, and (2) to incorporate NERSC's computational and data resources in the emerging DOE Science Grid. The two new directions are not fully described in this plan since they are unfunded at this time. However, since changes are clearly emerging in the national scientific community, this implementation plan shows the NERSC Center is responding, to the extent possible with level resources, to the two new, evolving thrusts in DOE science.

The first and most obvious change in the DOE computational science community, in all its disciplines, is that scientific groups are organizing into large multidisciplinary teams. They are called ***"Scientific Challenge Teams"*** in this plan, a phrase taken from the DOE initiative in Scientific Discovery through Advanced Computing (SciDAC). Some of these teams were formally recognized and funded by the Grand Challenge program of the mid 1990s at DOE, but most were formed earlier, often independent of formal recognition. This change away from the single principal investigator model for computational science, which has dominated computing in the natural sciences for most of the second half of the last century, has been driven by necessity as well as opportunity. The transformation became most apparent after massively parallel computers came to dominate the high end of available computing resources. The gap between the peak performance of machines with hundreds or thousands of processors and the performance attained by simply porting an application to run on them increased dramatically by comparison with the situation that had been seen on the vector supercomputers of the 1980s and early 1990s. Exploiting massively parallel computers as scientific tools frequently requires more effort and a larger range of skills than can be brought to bear by a single PI working with a few graduate and postdoctoral students trained in a single scientific discipline.

The necessity of introducing the ideas, tools, and practices of modern computer science into the practice of computational science is only part of the reason for the trend towards large multidisciplinary teams. The scale of problems being attacked also has provided an opportunity for such teams to gain a distinct advantage over single-PI efforts. DOE's scientific portfolio abounds with examples. In the field of climate prediction, for example, the complexity of the coupled system of atmosphere, oceans, and a host of energy transfer mechanisms, has resulted in the aggregation of an extensive and well organized army of researchers in fields ranging from fluid dynamics and chemistry to applied mathematics who are building the Community Climate System Model. It is perhaps the best-known example of a "community code" that is the intellectual property of an entire community of researchers and forms a shared basis for further work. The DOE Office of Science also funded the development of NWChem, a community-code infrastructure for quantum chemistry and molecular dynamics that is changing the practice of the discipline of theoretical chemistry. Similar efforts are just beginning in materials science with the DOE Computational Materials Science Network. In fusion research, the tradition of large teams of code developers may be longer than in any other area, but we have seen the recent emergence of progress toward total device modeling, first in the form of the Numerical Tokamak Turbulence Project and now with the formation of distributed national teams to exhaustively explore new concepts for fusion devices. In high energy and nuclear physics, the formation of international collaborations for constructing the software for data analysis and detector modeling is a longstanding practice, but recently we have seen the formation of a national consortium to produce a comprehensive terascale accelerator simulation environment for the U.S. particle accelerator community.

The second change occurring in the portfolio of the Office of Science is the ***convergence of computing, experiment, and theory*** in scientific endeavors that make use of all three components simultaneously in a single effort. One example is the Supernova Cosmology Project. The search for type Ia supernovae is yielding experimental evidence which is changing our understanding of the expansion and geometry of the universe. Briefly, to find the supernovae, telescope images are analyzed using supercomputers, and the results are used to quickly task telescopes around the world to observe and measure spectra of candidate objects. Simulations and predictions of the spectra based on the elemental abundances of each candidate are performed simultaneously to refine the identification. The entire effort requires the availability and scheduling of resources across the country and throughout the world. It was for such applications that the concept of *computational Grids* was invented.

There are numerous other examples beginning to emerge of the convergence of computing, experiment, and theory in the science of the DOE. The fact that the Office of Science builds and operates major experimental facilities — including the light sources and neutron sources used by the nation's chemists, materials scientists, and biologists — focuses this trend in the mission of the DOE. Already the possibility of theoretical simulations being used to steer experiments at the Advanced Photon Source and Advanced Light Source is being discussed at the same time that the requirement of remote telepresence of researchers is being developed as a mode for routine use of those facilities. Computational Grids are being built in high energy and nuclear physics, like the GriPhyN project. Individual disciplines are developing experimental efforts in the area of Grids, and the computer science tools for such structures are still largely works in progress.

For the Office of Science to be able to exploit the technology of computational Grids, its largest and most powerful computing facility must become an active node on the DOE Science Grid. That transformation is not accomplished by merely connecting to the Grid; it must be accomplished by the staged development of a new architecture for high-end computing that incorporates, schedules, and manages the computing and data resources at NERSC into a new and evolving national infrastructure.

1.2 A Flagship Facility at LBNL to Support the Programs of the DOE Office of Science

The strategy of maintaining a flagship, high-end facility that serves all its programs is both necessary and wise for the DOE, regardless of what other investments it makes in computing. The National Science Foundation is pursuing the same strategy, now with three such facilities. A centralized facility, properly staffed and managed, provides the best possible mechanism for technology transfer between the computational efforts of different research programs. For example, the expertise developed in supporting materials science benefits chemistry. The expertise in cluster computing gained while supporting the nuclear physics community has become the basis of collaborations with other communities, including the accelerator design community.

Moreover, a concentration of computing resources provides a more flexible mechanism to address changing priorities. DOE’s priorities for its programs sometimes change quickly because it is a mission agency. In the past, climate change emerged as a pressing national problem, and now issues of efficient energy production and technologies for locating and exploiting fossil energy sources are emerging as priorities. A general-purpose facility, with a staff prepared to support the broadest possible array of scientific disciplines, allows DOE to switch priorities and quickly apply its most powerful computing resources to new challenges.

This strategy is also optimal in view of the recent evolution of hardware technology (as the National Science Foundation experience, for example, also suggests). While the specific concerns of each discipline are somewhat different, the computers available to them are increasingly similar. The most powerful and useful massively parallel computers, regardless of vendor, have converged to a single architecture, collections of symmetric multiprocessors connected by fast networks. Of course, a facility serving many disciplines must pay attention to the needs of all the disciplines it will support when choosing the configuration of large computational resources. The extensive efforts outlined in Section 2 on high-end systems are therefore directed towards fielding the most balanced and general-purpose facilities possible.

A set of advantages accrue to NERSC specifically because it is in the Computing Sciences organization at the Lawrence Berkeley National Laboratory (LBNL). DOE funds a portfolio of research in computer science and applied mathematics in the Berkeley Lab Computational Research Division within Computing Sciences that can be strongly coupled to the NERSC Center’s efforts in advanced development. The Mathematical, Information and Computational Sciences (MICS) office at DOE funds work at LBNL in a number of areas including data management, computational Grids, collaborative technologies, imaging, adaptive mesh refinement, level sets, partial differential equations, and linear algebra. Those research programs are involved with the NERSC Center, and their physical and organizational proximity to the Center allows NERSC to take full advantage of their presence. In addition, the Advanced Light Source, a major experimental facility whose experiments are candidates for the Unified Science Environment, is located at LBNL. And finally, the Berkeley Lab is home to a large number of NERSC users from its programs in materials science, chemistry, high energy physics, nuclear physics, and the physical biosciences.

NERSC at LBNL is located immediately adjacent to the campus of the University of California at Berkeley. A number of faculty at U.C. Berkeley have joint appointments in the LBNL Computational Research Division, including Professors Jim Demmel, David Culler, Kathy Yelick, Ming Gu, Jamie Sethian, and Alexandre Chorin. As a result, active collaboration with the campus in computer science and applied mathematics is the norm for the NERSC Center, and its effectiveness is thereby greatly enhanced.

Finally, as the home of Silicon Valley, the San Francisco Bay Area is the center of the computing revolution. It is the location of choice for major computing enterprises. It provides a rich talent pool from which to recruit staff members. It provides the opportunity for collaboration with Stanford University and the University of California at Davis as well as with Lawrence Livermore National Laboratory. It provides the opportunity for close collaboration with the computing efforts at NASA's Center of Excellence in Information Technology at Ames Research Center. Coordination with the Information Power Grid program is made possible by the simultaneous leadership of LBNL's and NASA's efforts by William Johnston.

1.3 The Implementation of NERSC's Strategic Model

Two of the four components in the model proposed in the NERSC Strategic Proposal 2002–2006 are core components for operating the NERSC Center for the next five years, and their implementation is described in this plan. The activities that characterize the Center's effort at present are:

- *High-End Systems* — Balanced introduction of the best new technology for complete computational systems — computing, storage, and networking — coupled with the advanced development activities necessary to incorporate the technologies developed in the other DOE laboratories and elsewhere. NERSC also proposes to increase the size and capability of the computers it fields by increasing the annual budget for hardware acquisition.
- *Comprehensive Scientific Support* — The entire range of support activities, from high-quality operations and user services to direct collaborative scientific support, that NERSC will continue to provide to enable a broad range scientists to effectively use the NERSC systems in their research.

These two continued strategic thrusts should be complemented by two new ones, more limited in scope and financial investment but important to the future of the flagship enterprise and its ability to fully serve the DOE scientific community in a timely manner:

- *Support for Scientific Challenge Teams* — A concentration of the Center's resources on multidisciplinary, multi-institutional teams focused on solving the most challenging problems in computational science. This concentration of resources is needed to realize the promise of terascale computing for scientific discovery in this emerging new model for computational science.
- *Unified Science Environment* — The architectural and systems enhancements required to make NERSC the most powerful computational resource on DOE's Grid. Over five years, a capability will be deployed at NERSC designed to meet the needs of new computational science by facilitating access to computing and data resources as well as to large instruments across the DOE complex using Grid technology.

The two core activities alone provide a flagship level of service to DOE's most demanding scientific activities. This implementation plan discusses the major components of the NERSC Center strategy. This plan concentrates on the core areas of High End Systems and Comprehensive Scientific Support. The other two areas in the proposal — plans for specific support for *Scientific Challenge Teams* and plans for the *Unified Science Environment* — require additional funding and resources to accomplish the full vision of the proposal. However, because of the critical importance of these efforts, the NERSC Center implementation plan incorporates efforts to accomplish some objectives in these two areas that are within the limited resources of the Center.

The remainder of this document describes ongoing and future efforts in the various functional areas. Since this is a response to a five-year strategic proposal, the milestones and efforts described will go beyond the one- to two-year time frame where appropriate. In some areas, describing specific milestones beyond certain time frames becomes more general, as less is known about the future of technological development and progress. In these cases, care is taken to describe a general direction or approach to the issues in the distant future so that the plan is not invalidated if certain developments outside of NERSC's control do not take place. This plan will be revised as needed to meet the changing evolving technological environment and DOE priorities and needs.

1.4 NERSC Guiding Principles

As described in the following sections, the NERSC Center provides early, large-scale production computing and storage capability to focused groups and projects. The systems will be of such a scale as to be unique or nearly unique in many aspects (e.g., computational abilities, storage). NERSC provides Comprehensive Scientific Support for its clients to make it easy and practical for DOE computational scientists to use the advanced systems by

- providing consistent, high-quality service to the entire NERSC client community through the support of the early, production quality, large-scale capability systems
- aggressively incorporating new technology into the production NERSC facility by working with other organizations, vendors, and contractors to develop, test, install, document, and support new hardware and software
- ensuring that the production systems and services are the highest quality, as well as stable, secure, and replaceable within the constraints of budget and technology
- participating in other work to understand and address the unique issues of using large-scale systems.

High End Systems, Comprehensive Scientific Support, and related high performance computing (HPC) development and integration efforts create a balanced enterprise to enable NERSC to provide exceptional systems and support of its clients. It is the heart of the strategy that sets NERSC apart from other centers and greatly enhances the impact of the physical technology.

1.4.1 NERSC Quality of Service

In order to be successful, NERSC must provide much more than just raw hardware cycles and bytes of storage. The DOE computational science community requires and demands high quality services and support to help them be the most effective they can be.

NERSC has specific processes for establishing its yearly goals for the Center. It is an iterative process that involves the entire staff, NERSC clients, and DOE. It is both top-down and bottom-up. A schematic for the process is shown in Figure 1-1. Essentially, once a year, NERSC Center staff, working within their groups, develop a set of goals for the next one to two years. Goals are both quality of service and project milestones, measurable not just in the amount of activity or meeting a deadline, but also in the quality of the accomplishment. As the NERSC Center groups develop their goals, so does the NERSC Center management team. Once each group develops its goals, the groups share their goals with each other and with management, which in turn shares its goals. As can be expected, it takes discussion and iteration to make sure the lower-level goals are consistent with higher. For example, one group may have a goal that relies on goals of another group. It is critical to ensure that the goals of the latter are coordinated.

At times, it is necessary to resolve inconsistencies between the management goals and the group goals. If management has a goal but few or none of the groups have related goals, then the disconnect must be addressed. Likewise, it is possible that one or more groups have goals that do not relate to a management goal. In both cases, the reasons for this must be explored, and either the management goals and/or group goals are added or dropped.

Once all the groups have a consistent set of goals, they are presented to the NERSC User Group, Berkeley Lab management, and DOE program officers for validation. All goals are in some way measurable. The goals may be adjusted at this point as well. Once validated, NERSC will work to meet the goals and milestones, using metrics to assess success. Once a year, NERSC does a self-assessment. The result is reported to DOE, Berkeley Lab management, and NERSC clients. Figure 1-1 summarizes this process.

The following section presents the NERSC goals for FY2002/3. Many of these high-level goals evolve over time, and often performance level increases, so these are representative of the goals NERSC will use during this period. NERSC will continue to follow this process every year, to develop goals and define expectations with NERSC clients.

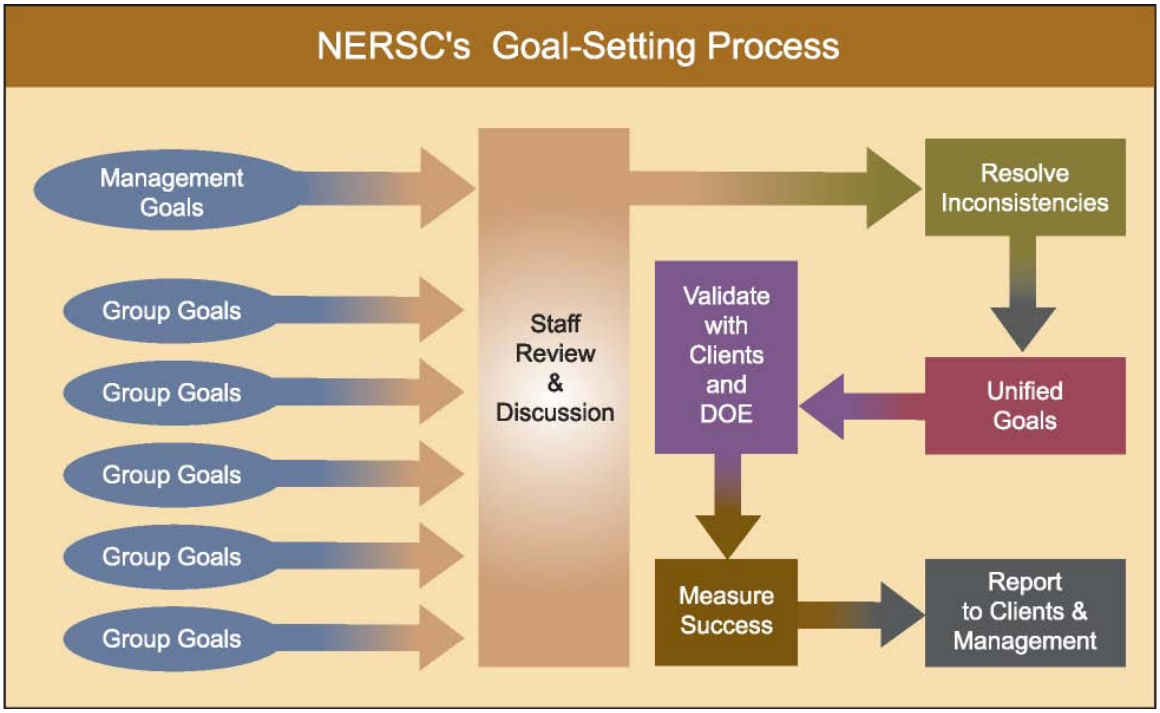


Figure 1-1. NERSC’s goal-setting process.

1.4.2 Systems and Service Goals from the Client’s Viewpoint

1. Provide 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 production service (Table 1-1).

Table 1-1
System Metrics For FY01
FY01 Goal (Measured in 1999)

Systems	Scheduled Availability %	Overall Availability %	Mean Time Between Interruptions (Hours)	Mean Time to Repair (Hours)
Computational Systems	98	97	240	4
Storage Systems	98	97	120	4
Network Systems (routers, servers, etc.)	99	99	340	4
Servers (fs/gw)	99	97	340	8
Clusters Computational and other)	97	96	340	12

Systems metrics definitions:

- Scheduled availability is the percentage of time a system is available for users, except for any scheduled downtime for maintenance and upgrades.
- Overall availability is the percentage of time a system is available for users or special purposes. In NERSC’s 24 × 7 environment, 100 percent availability would be 168 hours per week.
- A service interruption is any event or failure (hardware, software, human, environment) that disrupts full service to the client base. Full service is defined system by system.
- 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 being restored to service, it is treated as one continuous outage.

2. Provide Effective, High Quality Client Support

The end measure of a computer center is how much productive scientific work users accomplish. Centers 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.

A multi-tier service architecture, discussed below, ensures response to client problems within four working hours and timely resolution:

- Resolve at least 80 percent of problems within two working days.
- Escalate unresolved problems for management review within five working days.
- Install accounts within one working day.
- Provide timely and accurate electronic information. For full production systems, this means all system outages announced at least 24 hours in advance and all planned system changes announced at least seven days in advance.
- Make continual improvements and enhancements. Improve system usability by testing, working with vendors, and monitoring and tuning performance and system parameters.

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

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

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, the NERSC Center has the responsibility to ensure that system changes have a maximum benefit and minimal detrimental impact on the clients’ ability to accomplish scientific progress. This responsibility includes:

- thoroughly testing systems and software
- informing the user community of the impact of changes and enhancements of delivered systems
- creating workable solutions for clients when they exist.

If detrimental impact is unavoidable, NERSC has the responsibility to ensure that the benefits of the changes significantly outweigh the detriments.

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

The NERSC Center 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-changing environment. NERSC will help design strategies and integrate and develop technologies to enable our clients to improve their use of our systems and to more effectively accomplish their science.

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

NERSC is uniquely placed to establish methods and procedures that enable research products and knowledge to smoothly flow into production. Responsibilities include:

- designing research and development projects to facilitate transition into full-scale service
- establishing methods to allow NERSC client requirements to be taken into consideration in developing research plans and activities
- establishing evaluation steps to ensure products are developing properly
- establishing well-understood test and acceptance criteria that include real support costs for potential products
- evaluating the impact of research products and improving their serviceability
- establishing peer relationships between service providers, researchers, and developers.

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

As DOE’s largest unclassified scientific computing facility, the NERSC Center continually provides leadership and helps shape the field of high performance computing. As HPC technology evolves at an increasing rate, it is crucial that the NERSC Center remains at the forefront of getting the most out of these systems.

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

For NERSC to be a leader in large-scale computing, it must export experience, knowledge, and technology. Transfer must be made to other client sites, supercomputer sites, and industry. Methods of technology transfer include:

- presentations of papers at conferences
- publication of articles in scientific/technical journals
- tutorials for users and the HPC community
- code releases for the user community
- modifications being adopted by vendors and other sites
- technical interchanges
- site visits with other centers
- providing expert opinions and analysis to the DOE
- leadership roles in appropriate conferences, committees, and publications
- serving on review and advisory committees for other organizations.

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

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

Thus, the NERSC Center will continue to evolve, change, and grow. NERSC will continue to facilitate a paradigm change in computing within DOE’s Office of Science and the HPC community. NERSC will be able to quickly take advantage of appropriate new opportunities while maintaining its core focus.

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

Every staff member of the NERSC Center has a stake in the success of NERSC, and they are encouraged to contribute their ideas for helping all of us succeed. Staff members interested in additional training to increase their effectiveness are encouraged to talk to their group lead or department head. To help facilitate the professional exchange of ideas and information, NERSC has adopted a series of guidelines and recommendations which are posted on the staff Web pages.

1.4.3 System Improvements

Goals 3 and 4 create a healthy tension between too much and not enough change. NERSC does not change for change's sake. Rather NERSC only changes its systems or services when the benefit outweighs the impact. NERSC staff explore and evolve new functionality for systems. They are experts at testing and integrating new system hardware and software with little or no service disruption.

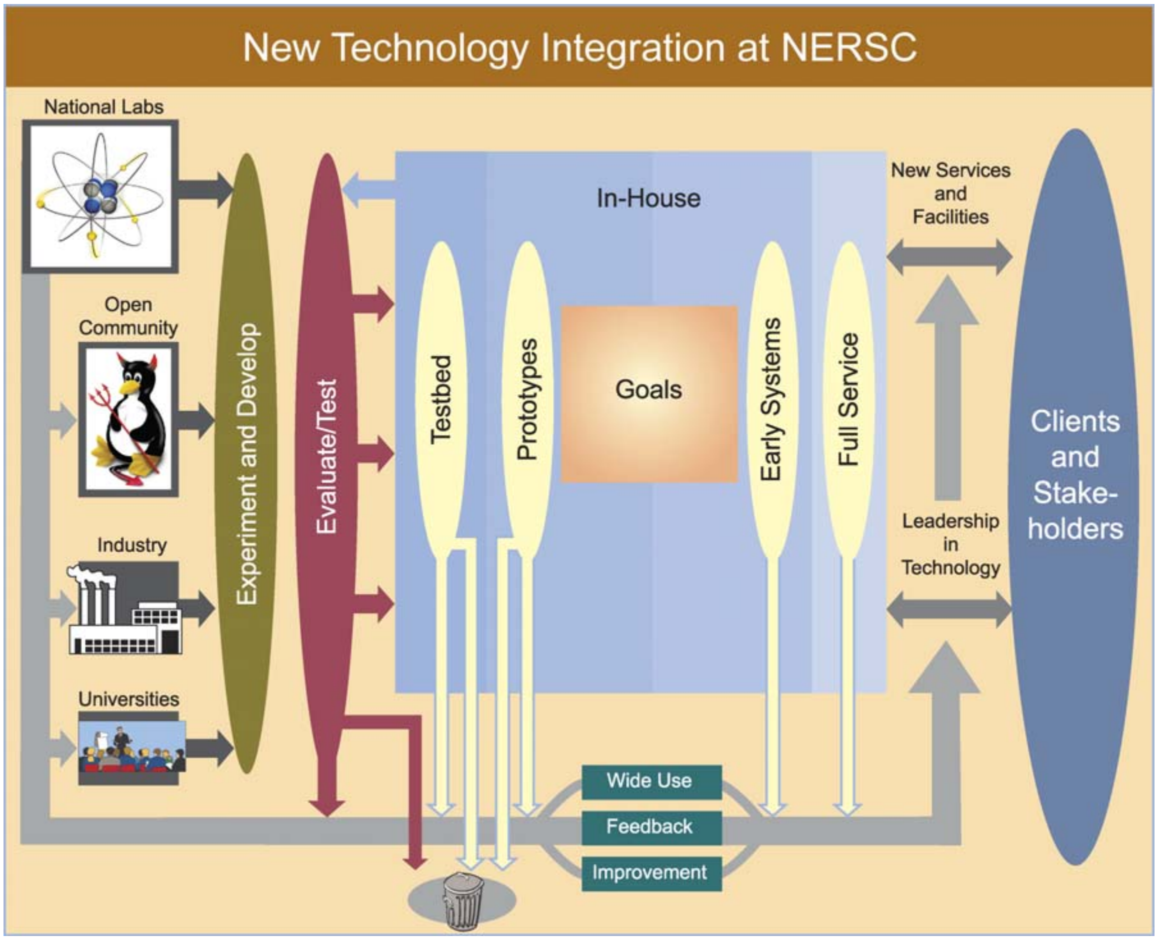


Figure 1-2. New technology integration at NERSC.

Figure 1-2 shows the process the NERSC Center uses to implement new technology, both software and hardware. Technology comes from vendors, the open source community, the academic community, national laboratories, NERSC staff, and other sources. It is the job of the NERSC Center to wisely deploy the best and most appropriate technology in a cost-effective manner. Technology, whether hardware, software or a combination, enters the process and goes through different phases based on its source, maturity and potential function.

The phases that technology goes through are experimentation, research, and development; evaluation, observation and external testing; prototyping; testbed; early use; general or special use; and final full service. At each phase, the technology is evaluated for

- readiness to progress to the next phase
- potential impact on NERSC clients (both immediate and long-term)
- overlap with existing functions
- costs (both initial and ongoing)
- benefits and risks.

Reviews take place for technology to progress to the next phase. Throughout the phases of the process, NERSC staff provide feedback analysis of the technology to the supplier. NERSC may assist in developing new requirements for vendors and other groups, and when appropriate develop key technology important to the success of NERSC clients. NERSC staff interface with vendors and with other sites and visitors in this process.

NERSC’s systems will be leading edge and so may be missing vital functions in system software when first deployed. NERSC addresses this by working with vendors, the SciDAC Integrated Software Infrastructure Centers (ISICs), and the broader HPC community to provide the missing functions. Particular areas of need and expertise are in job management, networking, algorithm development, large global filesystems, development of hierarchical storage management systems, data management, and benchmarking. NERSC will selectively explore and develop some of this technology for both its own needs and the needs of the DOE HPC community.

2. HIGH-END SYSTEMS

Providing the most effective and most powerful high-end systems possible. This is the foundation upon which NERSC builds all other services in order to enable computational science for the DOE/SC community. High-end systems at NERSC mean more than highly parallel computing platforms — they also include a very large-scale archival storage system, auxiliary and developmental platforms, and networking and infrastructure technology. Our successful high-end system strategy includes advanced development work, evaluating new technologies, developing methodologies for benchmarking and performance evaluation, and acquisition of new systems.

NERSC plans to introduce the NERSC-4 system in 2003 and the NERSC-5 system in 2006, each with at least a factor of 3 increase in capability over the previous-generation base system, bringing NERSC-5 to approximately 27+ teraflop/s peak performance. Figure 2-1 shows the proposed peak computing power of NERSC. But computing power is only one measure of capability. NERSC will continue to increase the capacity of its storage system, reaching at least 15 petabytes (PB) of capacity in 2006. A major effort will be made in developing and deploying a Global Unified Parallel File System (GUPFS). The development of GUPFS takes existing work as its point of departure, including work in industry, universities, and other DOE laboratories. This capability will not only increase scientific productivity by simplifying file access, but will also be one of the foundations for the Unified Science Environment (USE), which is described in Section 3.

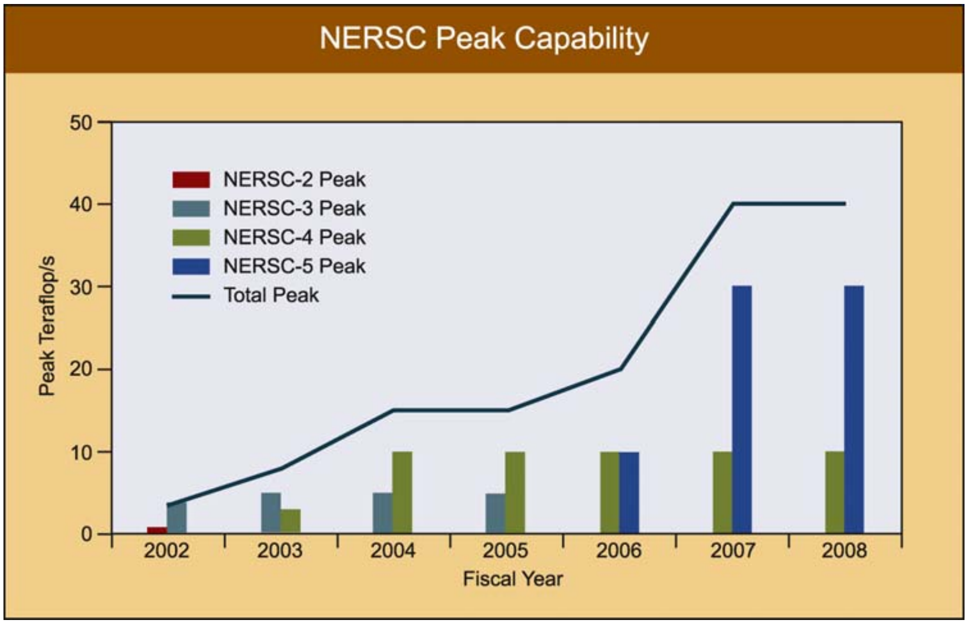


Figure 2-1. Computational capability growth of the NERSC Facility.

The following subsections describe the details of NERSC's strategy in high-end systems. First is background material summarizing the technology trends and user requirements that influence NERSC system architecture. The current NERSC architecture is described, followed by a description of the evolution of NERSC high-end systems over the next five years.

2.1 Technology Trends and User Requirements

Both technology trends and user requirements influence NERSC current and future system architecture.

Technology opportunities and constraints provide parameters within which NERSC must develop its high-end systems strategy. Because the future of NERSC is intimately connected with advances in a wide range of information technologies, it is imperative that NERSC closely monitor these advances and carefully assess their potential impact, both on the systems we acquire and the services we provide. A snapshot of the advances, including semiconductor technology, networking, open software development model, aggregation and centralization of computing resources, commoditization of computing technology, and commercial service models, are discussed in an LBNL Technical Report.

NERSC high-end systems strategy is further influenced by user requirements. It is essential that NERSC understand the requirements, both near-term and long-term, of the computational scientists using its systems. These requirements come from DOE strategic and tactical programmatic goals, from leading-edge research teams, and, most importantly, from the NERSC User Group Greenbook (http://hpcf.nersc.gov/about/NUG/DOE_Greenbook.pdf). DOE special projects, such as SciDAC and Big Splash, are also carefully considered when planning future purchases. They define, in large measure, the strategic applications areas for DOE. Therefore NERSC pays particular attention to the applications and methods used by the Scientific Challenge Teams.

The latest Greenbook argues that the DOE SC community must increase computational resources capable of supporting very large-scale applications, since this remains the primary requirement for the scientific projects. Most projects anticipate substantial online storage requirements, with high bandwidth to the computational platforms. They also require higher network bandwidth to their sites. Some projects have very large data archive requirements, while others need large amounts of memory for their computational applications. Finally, all clients and projects want access to continued and expanded scientific support.

NERSC must support a diverse workload. NERSC’s highest priority is to support *capability computing*, which we define here as the need to use more than one-fourth of an entire computing resource over an extended time period. NERSC also supports *large-scale computing*, which is defined as the use of more than one-eighth of the entire resource over an extended time period. Finally, NERSC supports a very small amount of *related capacity computing*. NERSC’s systems and service architectures are currently designed to support a capability workload of about 3–8 very large strategic projects, 10–15 large-scale projects,[1] 50 modest size projects, and 50 startup projects.

2.2 NERSC Today

A complete description of NERSC’s systems and activities is available in the NERSC Annual Report and on the NERSC web site (www.nersc.gov). But is it appropriate to briefly summarize the system here in order to put the proposed changes in context. Figure 2-2 is a diagram of the current system architecture.

[1] Class A projects are very large projects on the scale of SciDAC projects. Class B projects are ordinary time requests. Startup projects are focused on parallel porting and development projects.

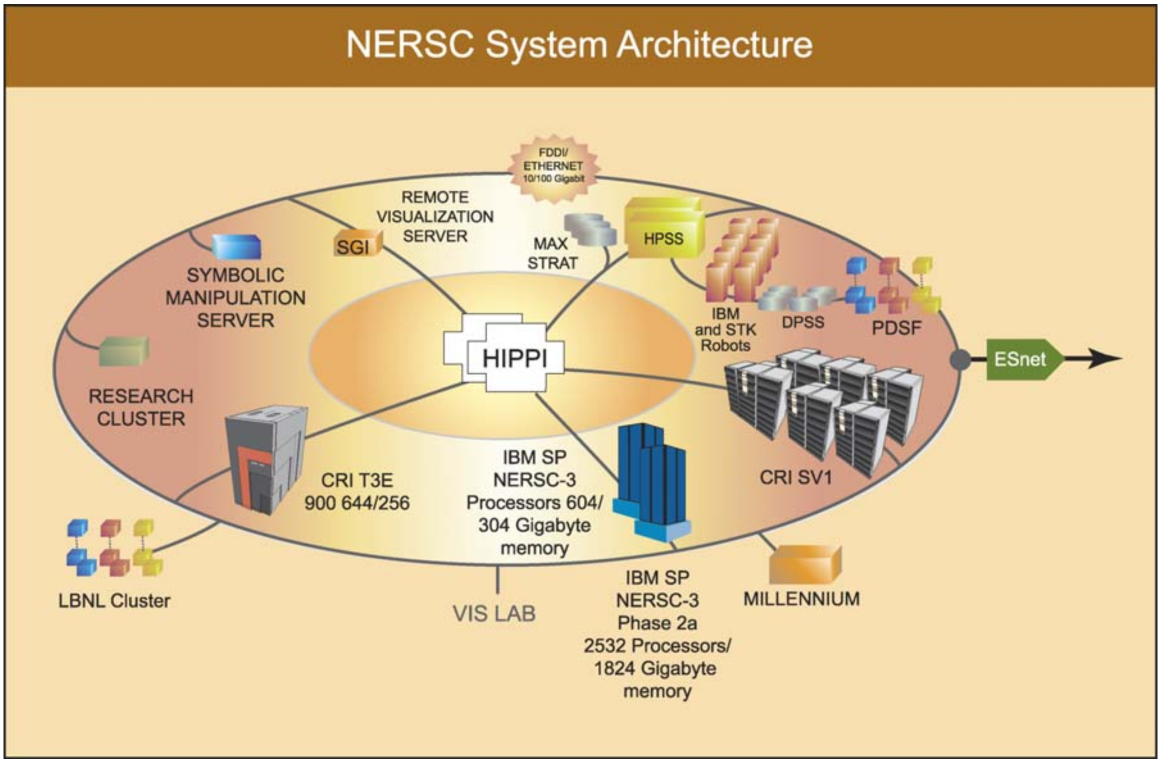

NERSC System Architecture
FDDI/ ETHERNET 10/100 Gigabit
REMOTE VISUALIZATION SERVER
SGI
MAX STRAT
HPSS
SYMBOLIC MANIPULATION SERVER
IBM and STK Robots
DPSS
PDSF
HIPPI
ESnet
RESEARCH CLUSTER
CRI T3E 900 644/256
IBM SP NERSC-3 Processors 604/ 304 Gigabyte memory
CRI SV1
LBNL Cluster
VIS LAB
IBM SP NERSC-3 Phase 2a 2532 Processors/ 1824 Gigabyte memory
MILLENNIUM

Figure 2-2. A schematic of the NERSC Systems as of July 2001.

NERSC consists of two generations of computational systems, currently called NERSC-2 and NERSC-3. NERSC-2 is a Cray T3E installed in 1996 and upgraded in 1997. NERSC-3 is an IBM SP system. Phase 2 of the IBM system was accepted in June 2001 and is currently in production.

In addition to the large parallel systems, which generate 97% of the computational resources in FY2001, NERSC also provides a cluster of 68 Cray SV1 CPUs in three systems. These systems are used by codes that have not yet been ported to distributed memory implementations, as well as commercial packages not available in parallel form.

NERSC has been a development site for the High Performance Storage System (HPSS) and now operates one of the largest HPSS systems, with approximately 340 terabytes (TB) of actual data. Since 1996, NERSC has operated the Particle Detector Simulation Facility (PDSF), a production Linux cluster with almost 400 CPUs and over 35 TB of online disk storage. PDSF supports high energy and nuclear physics experimental programs. NERSC also operates a range of servers and networks to support its user community. Table 2-1 summarizes NERSC’s system capabilities.

Table 2-1
A Summary of Major NERSC Systems in February 2002

System	Description
IBM SP RS/6000	NERSC-3/Phase 2a, Seaborg, a 3,328-processor system using 16 CPU SMP nodes, with the “Colony” double/single switch. Peak performance is ~5.0 Tflop/s. 16–64 GB memory per computational node, 20 TB of usable globally accessible parallel disk, and 11 TB of local disk space for system usage.
Cray T3E	NERSC-2, Mcurie, a 696-processor MPP system with a peak speed of 575 Gflop/s, 256 MB of memory per processor, and 1.5 TB of disk storage. A peak CPU performance of 900 Mflop/s per processor.
Cray Vector Systems	NERSC-2, three Cray SV1 machines. A total of 68 vector processors in the cluster, 4 gigawords of memory, and a peak performance of 83 Gflop/s. Killeen is used for interactive computing; Bhaskara and Franklin are batch-only machines.
HPSS: High Performance Storage System	HPSS is a modern, flexible, performance-oriented mass storage system used at NERSC for archival storage. It was designed and developed by a consortium of government and commercial entities. The NERSC archive has a peak capacity of 2.5 PB, buffer cache of 15 TB, and a theoretical maximum speed of 6.4 Gigabits/sec.
PDSF: Particle Data Simulation Facility	The PDSF is a networked distributed computing environment — a cluster of workstations — used by six large-scale high energy and nuclear physics investigations for detector simulation, data analysis, and software development. The PDSF includes 390 processors in compute nodes with 15 TB of local scratch disk and 39 disk vaults with file servers for 20 TB of shared disk.

2.3 Future Plans

Several factors will drive fundamental changes to the NERSC system architecture:

- increased computational resources
- the move to more data-oriented applications
- more complete integration of the architecture
- new technology that allows improved online storage solutions.

There are three major areas of system design and implementation at NERSC: the computational systems, the storage system, and the network. The balance of the entire Center is determined by the requirements that evolve from the increased computational capability, plus independent requirements for other resources. Storage system improvements must be designed to support not just current work, but future workloads as well. Figure 2-3 shows the evolution of the NERSC system architecture between 2001 (left) and 2006 (right), with the introduction of the Global Unified Parallel File System (Section 2.3.2 below) and the Unified Science Environment (Section 3.5.5) integrating the discrete computational and storage systems. The following paragraphs provide detailed descriptions of NERSC’s strategy for its computational systems, storage systems, and network.

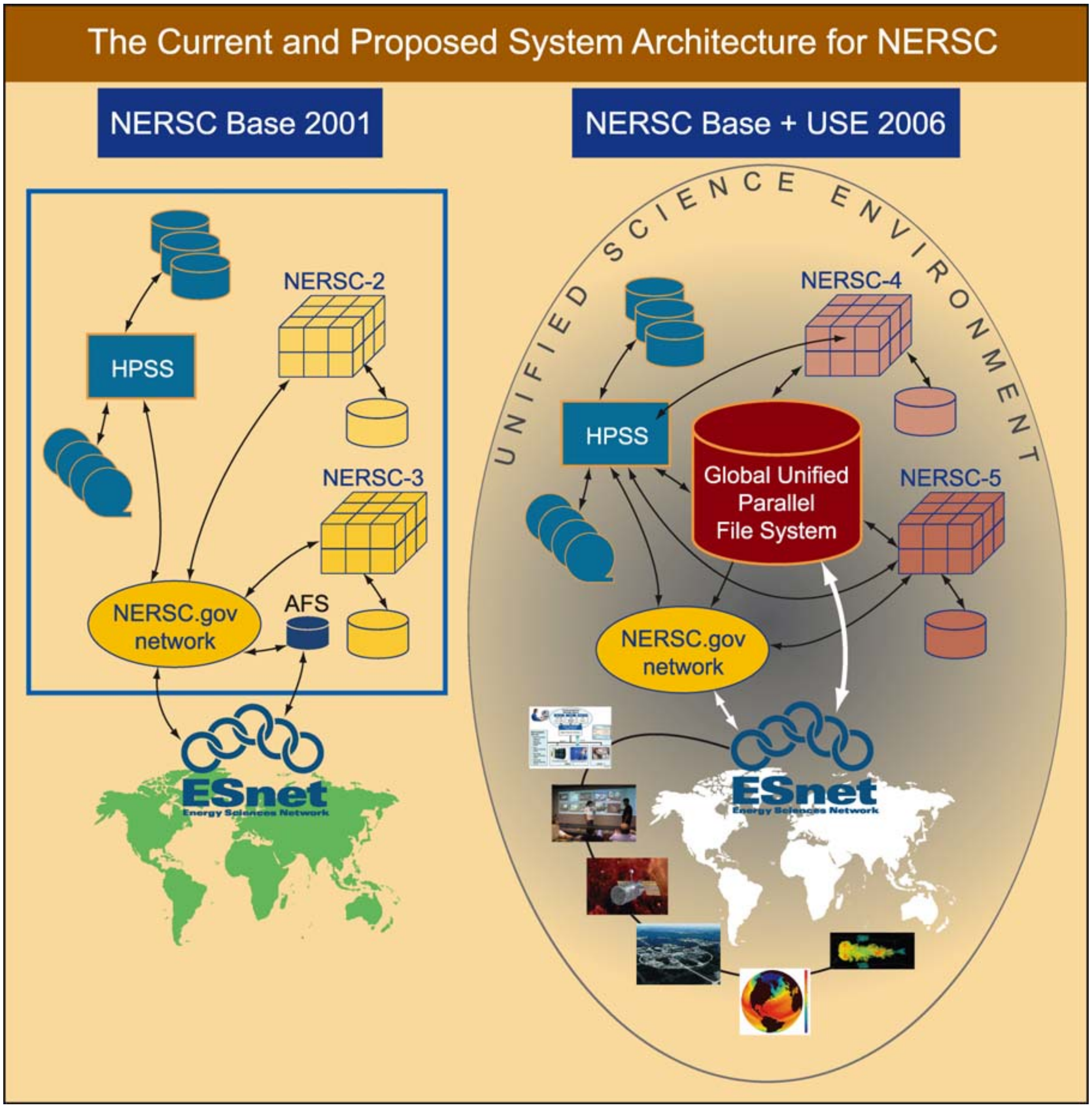

The Current and Proposed System Architecture for NERSC
NERSC Base 2001
NERSC Base + USE 2006
HPSS
NERSC-2
NERSC-3
AFS
NERSC.gov network
ESnet
Energy Sciences Network
UNIFIED SCIENCE ENVIRONMENT
NERSC-4
NERSC-5
Global Unified Parallel File System

Figure 2-3. Evolution of the NERSC system architecture between 2001 (left) and 2006 (right).

2.3.1 Computational Systems Strategy

NERSC acquires a new capability-focused computational system every three years. A 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 has 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.

The goals of DOE computational projects, and the requirements that derive from that work, mean that NERSC has to remain one of the top ten most powerful computational facilities (as shown in Figure 2-4). Indeed, it is reasonable to expect that NERSC will have the most powerful unclassified system at times — based on system delivery to NERSC and other sites. To do this, NERSC-4 and NERSC-5, which are the two generations of new computational systems in this proposal's timeframe, have to have a three- to four-fold increase in peak capability over the previous generation.

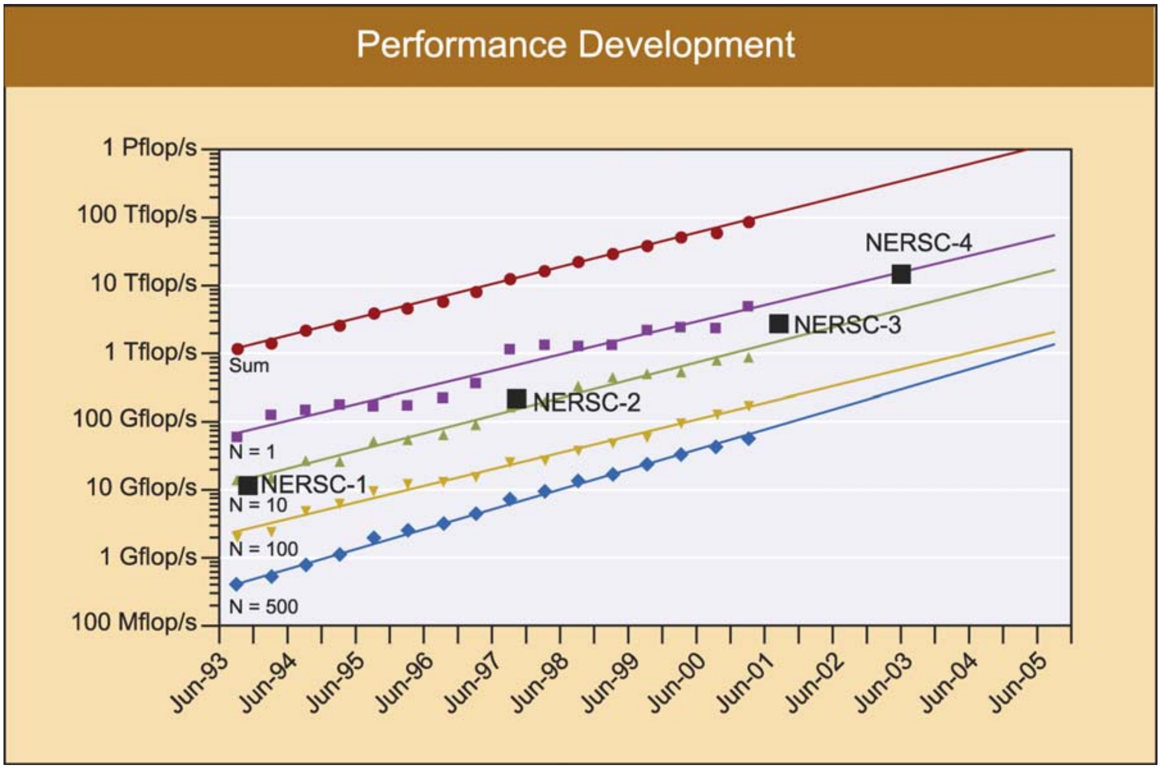


Figure 2-4. The DOE computational requirements and NERSC's leadership in high performance computing mean it will have one or more of the 10 most powerful systems.

Achieving this level of capability is possible within the budget in the associated strategic proposal, although the price/performance goals are very aggressive, depending as much on market forces as on technology. These systems follow Moore's law in general, and hence it is feasible to increase price performance by a factor of 3 to 4 every three years. Table 2-2 shows the approximate cost per teraflop/s of the most recent NERSC acquisitions.

The total annual investment in the latest supercomputer system alone will be approximately one-third of the total NERSC annual funding. As in the past, lease-to-own payments will be spread over three years, and it is possible that technology availability will dictate a phased introduction over one year to 18 months.

Table 2-2
Projected Future High-End Computer Price Points

System	Date of Initial Delivery (Includes Phase 1 of multiple phased system delivery)	Approximate Peak Performance of Final System (Tflop/s)	Approximate Price (Millions of Dollars)	Cost per Teraflop/s (Millions of Dollars)
NERSC-2 T3E	August 1996	0.58	$26.00	$44.87
NERSC-3 IBM SP	August 1999	3.80	$27.00	$7.11
NERSC-3 Augmentation (Power3+ nodes added to the NERSC-3 Phase 2 system)	January 2001	0.14	$1.00	$6.94
LBNL Midrange Intel Cluster	February 2001	0.14	$.075	$5.40
NERSC-4	February-May 2003			Estimate $2.50–3.00
NERSC-5	February-May 2006			Estimate $1.00

Best Value Source Selection

NERSC uses the *Best Value* process for procuring its major systems. Best Value source selection was developed at LLNL and refined at LBNL for the purchase of very large systems. Rather than setting mandatory requirements and using a quantitative rating scheme, the Best Value method requests minimum requirements and performance features. 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 because it provides them more flexibility as well in crafting the best solution.

There are 22-steps in the Best Value process, allowing considerable flexibility for NERSC and also providing an opportunity for significant innovation by suppliers. The steps are:

1. Accumulate and evaluate possible benchmark applications and software.
2. Using the candidate codes, create a benchmark suite and test run it on several different systems for portability and performance.
3. Draft the set of rules for benchmark codes.
4. Set basic goals and options for procurement and create a draft RFP document.
5. Conduct market surveys (vendor briefings, intelligence gathering, etc.). This is done after the first items so we can look for the right information and also tell the vendors what to expect. It is often the case that we have to “market” to the vendors on why they should be bidding, since it costs them a lot.
6. Evaluate alternatives and options for RFP and tests. This is where a technology schedule (when what is available) and estimated prices (price/performance) are developed.
7. Refine RFP and benchmark rules for final release.
8. Go through reviews of RFP as needed.
9. Release RFP.

10. Answer questions from vendors.
11. Evaluate proposals from vendors.
12. Determine best value; present results and get concurrence from necessary parties.
13. Prepare to negotiate the contract.
14. Negotiate the contact based on the proposal and identified issues.
15. Put contract package together.
16. Get concurrence and approval of contract.
17. Vendor builds the system.
18. Factory test of the system.
19. Vendor delivers system.
20. Acceptance testing and resolving issues found in testing. (First lease payment is two months after acceptance.)
21. Preparation for production.
22. Production.

Steps 17 through 22 may be repeated one or more times if the system is delivered in phases, as happened with both NERSC-2 and NERSC-3.

The principal task of the acquisition team is to decide the best alternative among the available choices. NERSC uses only measured performance to evaluate systems, not peak performance (Figure 2-5). Vendors are asked to run a set of benchmarks and tests on existing systems that are related to the systems being proposed. Then the vendor projects performance (in a way that is both justifiable and measurable). Those performance projections are then used as contract deliverables. A balance must be struck between the desire to measure every aspect and the costs and uncertainties to both NERSC and vendors in doing so. Hence NERSC uses between 35 and 45 discrete measures in this process. Three key measures are described below: the Sustained System Performance metric, NERSC applications that are part of the workload running on NERSC Center systems, and the Effective System Performance test.

The Sustained System Performance (SSP) metric is based on benchmark performance integrated over three years. The SSP uses well-known, easily understood applications to set a single performance indicator of the amount of scientific computation that a system can deliver. The measurements are related to the expected workload and the benchmarks, but they also depend on system functionality. For NERSC-3 the SSP (Version 1) value consisted of the average performance of the six floating point NAS parallel benchmarks. The NERSC-4 SSP (Version 2) value is calculated using five application codes selected from those running on NERSC-3. For NERSC-5, the NERSC Application Performance Suite might be used as a basis for this value. NERSC uses this performance level to establish an integrated metric, teraflop/s-years, for the first three years of the contract. This allows NERSC to evaluate the impact of different performance levels and delivery timetables. The goal is to maximize the integrated performance-time value for the NERSC clients. Figure 2-5 shows the NERSC-3 SSP curve as of June 2001. Since major functionality for the interconnect is implemented by the software about six months after the CPU hardware is delivered, just looking at the peak teraflop/s rate would be misleading in terms of what the scientist can accomplish on the system.

The SSP-2 of the NERSC-3 final system is ~11% of peak performance. This means it is an accurate number and a true measure of how well systems can support the NERSC workload.

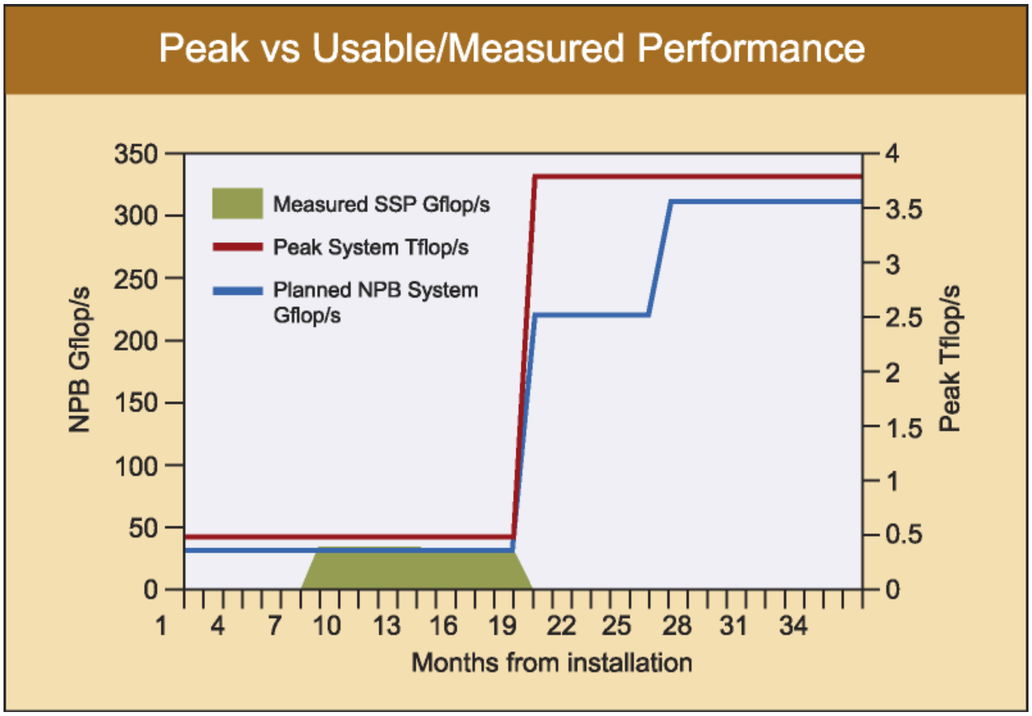


Figure 2-5. Peak vs. usable/measured performance.

NERSC selects real applications as part of the system evaluation as well. These applications come from the client community codes and represent the future workload that is strategic to the DOE science community. These applications will most likely be drawn from the SciDAC teams, since they implicitly indicate the strategic investments DOE is making. Additional applications may be selected from projects in other discipline areas, although practical considerations limit the number of full applications feasible to consider between five and eight. Added to the applications will be specific functionality tests, reliability tests, and benchmarks for storage and network I/O. NERSC also considers specialized tests that measure internal communication bandwidth in the memory subsystems of the processors, or across the communication fabric. The main point is that NERSC does not just use simple kernels that provide only limited insight into how well a system will do for NERSC applications.

The productive work that can be extracted from a computational system is dependent not only on computational performance but also on the software infrastructure. In particular, resource management functionality (e.g., scheduling, job launch, and checkpoint/restart) has become an increasingly important issue, given the difficulty of managing large-scale parallel computers. The Effective System Performance (ESP) test[2] was created by the NERSC Center specifically to provide a standardized metric of resource management that is independent of computational performance and representative of production usage. The objective is to provide a tangible measurement for vendors and administrators that will focus attention on aspects of system software that are important to NERSC and other large supercomputing facilities. Prior to this work, little attention has been directed to quantifying the effects of resource management or predicting the realizable utilization of a system.

[2] http://www.nersc.gov/aboutnersc/esp.html

The structure of the ESP test resembles a throughput test where the objective is to process a predetermined workload in the minimum time, T. The workload consists of several hundred jobs where the relative sizes, p_i, and run times, t_i, have been modeled on the NERSC usage profile. The ESP metric is given by $E = (\Sigma_i p_i t_i)/(P(T + S))$, where P is the total number of processors and S the observed time to shutdown and reboot. This ratio tends towards unity for increasingly optimal resource management and is independent of computational performance. Two special jobs, equal in size to the total system, are submitted at distinct points during the test. One of these jobs must be run before any other pending job is launched. The other must complete before 90% of the total wall clock of the test has elapsed. This stipulation rewards systems that can expedite large jobs and penalizes systems that are not responsive to dynamic changes in the workload, both of which are of particular concern in production environments. Figure 2-6 shows the logical flow of the test.

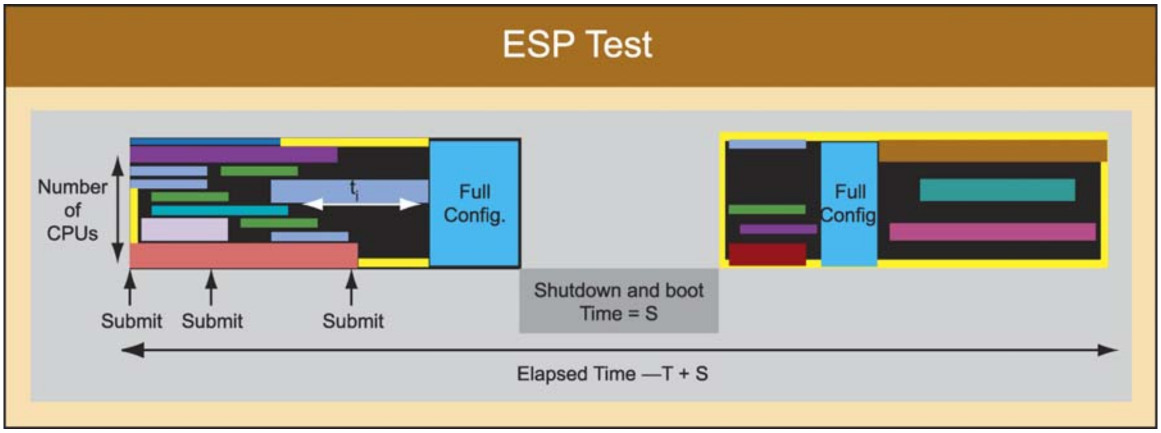


Figure 2-6. The Effective System Performance Test

The ESP test has already proven useful in comparing disparate systems, assessing the efficacy of preemption and scheduling strategies, and tracking improvements in system updates. In preparation for NERSC-4, this test has been recently updated so that various system sizes can be compared, and it is now scalable from tens to thousands of processors. The immediate goal is to obtain results on the widest variety of systems that are available. In particular, ESP will be used as a stress test for Linux cluster batch systems in order to spur further development and improvements. Further insight into very large system management will be provided from ESP runs on Advanced Simulation and Computing Initiative (ASCI) platforms. Over the longer term, improvements in ESP itself will be incorporated. Specifically, a new scalable application will replace the current applications to ease the configuration and execution.

The Expected Result of the Best Value Process

As mentioned above, we expect that NERSC-4 and NERSC-5 will very likely be commercial integrated SMP cluster systems. The type of processors within the SMPs may have a wide range of implementations including RISC, CISC, and vector. Special architectures will be considered, but it is not likely that these will be ready for high-quality production usage in the proposal’s time frame. For example, projects such as Blue Gene and Blue Light will be available (if at all) only after 2005, since their first full-scale systems prototypes are slated for completion after 2004. Entirely commodity cluster systems will also be considered, but based on technology assessments, it is less likely that these systems will be able to support the diverse and communication-intense applications at NERSC in this timeframe. Cluster hardware will at best have a modest performance-per-dollar advantage, but cluster software in particular is significantly less mature than vendor-supplied software. This situation is expected to persist for the next three to five years. If NERSC were to procure a large cluster system, we would have to allocate significantly more of our personnel resources on system integration and support work than proposed.

2.3.2 Storage System Strategy

Archival Storage

Over the period of time covered by this proposal (2001 to 2006), NERSC plans to greatly augment both the aggregate capacity and the transfer rate to and from the mass storage system. The aggregate capacity will increase more then tenfold, from 2 PB today to over 15 PB in 2006. Transfer bandwidth will increase more than tenfold, from approximately 1.5 TB per day today to over 20 TB per day in 2006. If network bandwidth improves as planned, the NERSC storage system can handle flows of 70 TB/day.

Functionality in archive storage will be driven by the deployment of the Unified Science Environment (USE), as well as by integration with computational systems. NERSC will continue collaborating in HPSS development in order to improve archive technology. In particular, NERSC will help develop schemes to replicate data over long distances, to import and export data efficiently, and to deploy Grid functions.

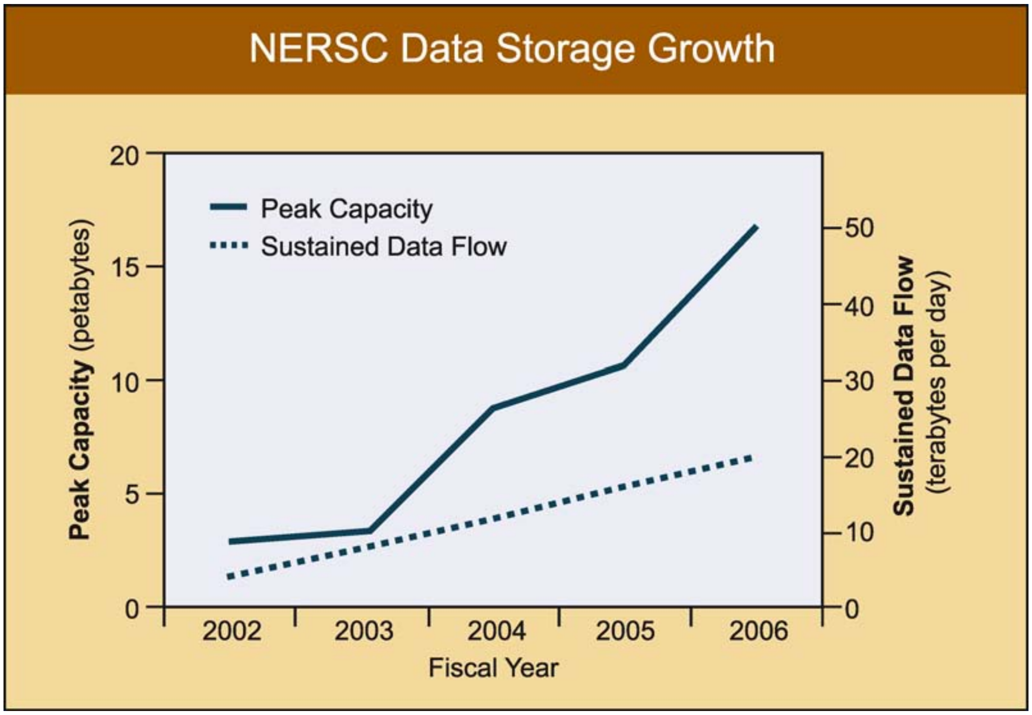


Figure 2-7. Predicted NERSC data storage growth, 2001–2006.

Figure 2-7 shows NERSC’s predicted growth in both capacity and transfer rate for the period covered by this proposal. The increase in capacity is driven by the introduction of new, denser media. NERSC currently has 20 and 60 GB tape cartridges. Over the lifetime of the proposal, media will increase in density to 1 TB per cartridge. NERSC will deploy 200 GB cartridges in FY2003.

NERSC will continue to incrementally improve the storage system. Each year, new tape technology will be added to the system that increases both the capacity and the bandwidth. NERSC will design and deploy an archive storage system to handle data flows of 6.1 TB/day in FY2002 to more than 25 TB/day in 2006. To achieve this, both the network to and from the storage system and the bandwidth within the storage system must be able to maintain these rates. NERSC will use Jumbo Gigabit Ethernet and Fibre Channel networks in the near term. Both are capable of close to 100 MB/s. NERSC will increase the archive cache disk from 10 TB to over 100 TB in 2006.

In order to use new, higher-density media, NERSC will deploy a different tape drive technology. At any one time, the media that exist in the storage system will be a combination of densities. The mix of media will be related to the relative number of tape drives capable of handling the density. NERSC will have to continually migrate data from older media to newer, denser media. The introduction of new media and drives also enables high-bandwidth drives, therefore increasing the bandwidth. However, while the media density will increase from 20 to 1,000 (50 times), bandwidth will only increase from 10 MB/s to 70 MB/s, a factor of 7. Thus more parallelism is needed for a balanced archive that provides sufficient performance as well as capacity.

NERSC will continue to keep all data within STK robots and not return to a shelf operation for tapes, because a shelf operation is inefficient, prone to error, labor intensive, and slow. Currently there are eight STK robots in the facility, each capable of holding 5,500 3480-style cartridges. That is expected to be sufficient with current technology trends to meet our capacity and bandwidth goals.

The archival capacity will grow from 1.3 PB this year to over 15 PB in five years. This will be done by adding and upgrading the tape units to handle higher-capacity tapes. By 2006, NERSC expects to have 30 drives capable of handling 1 TB tape cartridges. Figure 2-8 shows an example of the distribution of tape drive units by density. This takes into account expected drive and media costs, and the upgrade paths of different technologies.

NERSC will use storage area network (SAN) technologies for storage to bring fiber tape drives into the storage environment. NERSC will look for opportunities to work with other groups (e.g., Lawrence Livermore National Laboratory, Indiana University, and IBM) to develop and deploy a high performance interface between HPSS and IBM’s GPFS filesystem on SP and Linux systems.

This will provide a transparent hierarchy of storage for NERSC-3, with GPFS as the lowest level and HPSS the higher levels. NERSC will work on integrating the HPSS archive to NERSC’s Global Unified Parallel File System when it becomes available.

NERSC will track commercial storage systems, as they may provide interesting Web/SAN/database solutions. Currently, except for HPSS, these systems are not scalable or large enough to meet our needs, but the Web and Internet service provider marketplace is driving innovation and performance.

NERSC will continue to be an active development member of the HPSS collaboration. NERSC staff serve on both the Executive and Technical Committees, as well as on the development team.

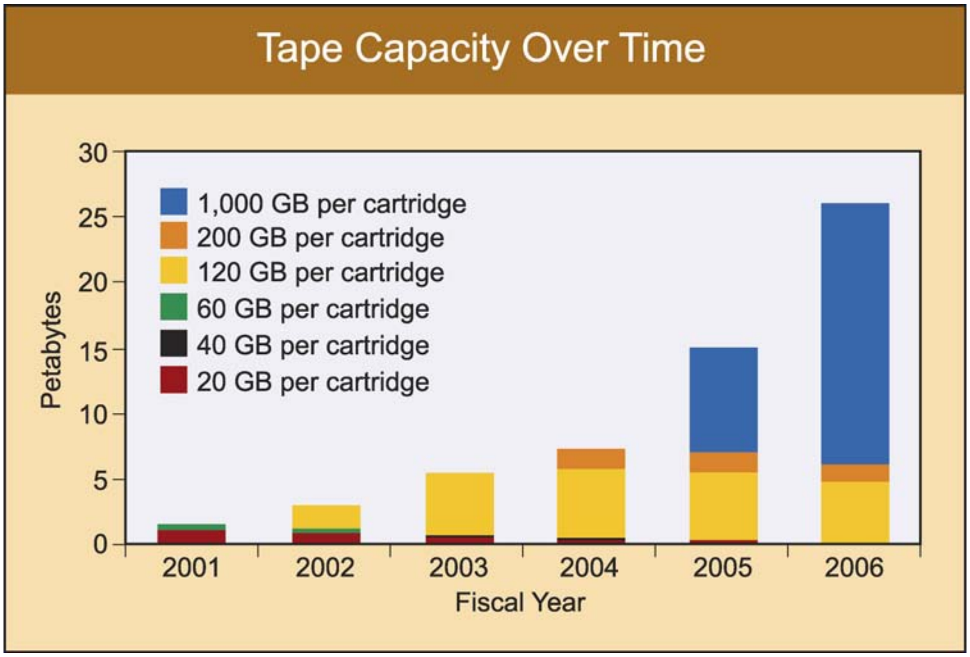


Figure 2-8. Contributions of different density media over time.

The Global Unified Parallel File System

The Global Unified Parallel File System (GUPFS) project aims to provide a scalable, high-performance, high-bandwidth, shared-disk filesystem for use by NERSC high performance production systems. GUPFS will be integrated with HPSS to provide hierarchical storage management (HSM), archival, and backup capabilities. The project will also explore the feasibility of Grid/USE distribution of the filesystem, but implementation of this capability is not planned at this time within the limits of Center resources. Hopefully, additional funds will become available to accomplish this as well.

NERSC plans to use a shared-disk filesystem, either commercial, open source, or from the ASCI PathForward Scalable Global Secure File System (APF SGSFS) program. NERSC does not plan to independently design and develop such a filesystem. However, the GUPFS project will likely include development of extensions to the available filesystems to incorporate or accelerate the incorporation of missing functionality needed for NERSC clients, such as tight integration of the filesystem with HPSS and support for a large number and diverse range of file types and sizes.

We will begin the project with a multi-path approach consisting of both an independent technology evaluation and joint/collaborative activities. Shared-disk filesystem, SAN fabric, and storage technologies will be evaluated on testbed systems. At the same time, joint/collaborative activities with other DOE Office of Science labs, the HENP community, and the APF SGSFS program will be explored and initiated. Based on the information and expertise obtained during these technology evaluations and collaborative activities, assessments of technological trends, and the viability, projected design, and implementation schedule for available shared-disk filesystems, a decision will be made as to the most promising solution no later than early FY2005. This solution will be deployed in a phased manner, and will be brought into production on the NERSC systems by the end of FY2006.

While there are a number of alternative solutions for GUPFS, the ASCI PathForward SGSFS program's goals and requirements are currently well aligned with those of the NERSC GUPFS project. This makes the filesystem being developed under this program a potential candidate for the GUPFS filesystem. However, since there are a number of other viable alternatives that warrant evaluation, and because the SGSFS is far from being implemented, the GUPFS project will simultaneously pursue an independent evaluation of other available shared-disk filesystems during the first three years of the project to identify alternative options. If at any point during this period SGSFS or another project appears to be converging on a successful and desirable solution, it may be reasonable to focus GUPFS resources more directly on that implementation. This could result in direct collaboration with the developers, or development of areas important to NERSC but not fully implemented under that project. Upon selection of a shared-disk filesystem technology, associated storage technology, and supporting middleware software, this integrated solution will be implemented in the NERSC production environment in a phased approach. This approach will begin with testbed implementations, followed by early use implementations on selected computational systems and storage servers and the building of a consolidated storage infrastructure, leading to an implementation on all production systems by the end of FY2006. Under this plan, it is expected that as part of the NERSC-5 procurement, a large part of the storage that would traditionally be acquired as local storage for the system could be acquired separately as shared-disk storage to be added to NERSC's consolidated storage and accessed by GUPFS.

In FY02, NERSC plans the following:

- Expand technology evaluation testbed.
- Identify joint/collaborative activities.
- Evaluate different shared-disk filesystems and SAN technologies (e.g., InfiniBand).
- Establish preliminary evaluation criteria and develop shared-disk filesystem specific benchmarks accordingly.
- Conduct preliminary wide-area network (WAN) storage distribution tests.

FY03:

- Begin HPSS integration.
- Continue evaluation of shared-disk filesystem and SAN technologies (e.g., APF SGSFS, other filesystems, iSCSI).
- Implement second generation of SAN fabrics and technology evaluation testbed.
- Complete identification of evaluation criteria.
- Start scalability tests on auxiliary systems.

FY04:

- Begin HPSS HSM integration.
- Conduct APF SGSFS beta testing.
- Select GUPFS solution shortlist candidates.
- Begin final evaluation and selection.

FY05:

- Complete final evaluation and select GUPFS solution.
- Generate deployment plans for consolidated storage and GUPFS.
- Begin phased consolidated storage buildup.

- Begin phased GUPFS deployment.

FY06:

- Continue phased storage buildup.
- Continue phased GUPFS deployment.
- GUPFS production on NERSC production system.

Post FY06:

- GUPFS is a standard architectural component of NERSC.

2.3.3 Networking and Data Communications Strategy

NERSC must expand its networking and data communication capacity as applications become more bandwidth intensive. The USE paradigm will change the common practice from bulk data transfers with some TCP/IP interactive traffic to many more bulk data transfers combined with inter-process communications. Since higher production bandwidth assumes bandwidth increasing every year, and since clients indicate the need, NERSC will move to what is effectively a new OC bandwidth level (increasing bandwidth by a factor of 4) at the NERSC LAN/WAN production gateway every three years.

As high-performance computing continues to become more network-centric (the Grid, HPSS, cluster interconnects, etc.), the network will become the “glue” that holds everything together. NERSC will continue to develop expertise in network engineering; as this is the only way to be able to deliver the full capability of NERSC systems to users.

Local Area Network

The NERSC network is designed to be the most cost-effective solution that will meet NERSC clients’ state-of-the-art requirements. Within five years, NERSC users will require more than 20 PB per year of permanent archive storage. With an average of one write and four reads for each byte, 80 to 100 PB will traverse the LAN each year, requiring at least 20 GB/s of average real network bandwidth. NERSC plans to have an aggregate peak network capacity of 200 GB/s of network bandwidth dedicated to storage. Most likely, the storage network will consist of 10 Gigabit Ethernet nodes using jumbo frames (packets sizes greater than 1,500 bytes). However, Fibre Channel or its successor or InfiniBand may be viable technologies and will be explored.

The introduction of GUPFS will substantially increase the ability to perform research but will also substantially add to the local network requirements. Initially such filesystems will require their own independent network that mirrors the capacity of the mass storage network. It is more likely this network will be based on Fibre Channel or InfiniBand technology. Once the full GUPFS functionality is deployed, and it is integrated with HPSS, NERSC will investigate ways to merge both the mass storage and the GUPFS networks so that, within five years, they become a seamless environment.

Wide Area Network

Data availability across the country and around the world has become much more important in the last five years, and we see this trend accelerating for the next five years. Many existing DOE projects have increasing data requirements, and new DOE projects will only accelerate them. NERSC expects that a significant fraction of storage traffic will traverse the wide area network so that sustained WAN rates of at least OC-48 (2.5 gigabits/s) will be required by FY2003 and OC-192 (10 gigabits/s) within three years.

Currently NERSC network statistics show that we need the capability to burst WAN traffic at ~8 times sustained rates in order to accommodate peak hours and minimize network contention. This requires the effective bandwidth of multiple OC-192s, which even five years from now could be costly. Since a higher peak rate improves reliability and efficiency while lowering the required sustainable rate, it is possible a more cost-effective solution would consist of a dedicated OC-192 with on-demand bursting of OC-768 (40 gigabits/s). Another option would be to participate with networks that provide on-demand lambda switching.

End-to-end network tuning is increasingly critical to providing high performance network connectivity to NERSC clients. NERSC will take advantage of the latest enhancements in networking systems and protocols, both in NERSC systems and at remote sites, to enable NERSC clients to access the system and move data and to continue to provide consulting services to optimize end-to-end performance.

3 COMPREHENSIVE SCIENTIFIC SUPPORT

Comprehensive Scientific Support recognizes the need to incorporate the standard service architecture/ client support model with high performance computer development and integration efforts. This model illustrates major functional areas and interrelationships necessary for the facility's successful support of DOE's science mission. The model serves as the introduction to detailed discussions of each area in this chapter.

3.1 The NERSC Service Model

There are several key concepts that illustrate the functional relationships for NERSC. These are identified simply as the client base, the service model, the technology development and integration model, and the goal setting process, which ties it all together. The combination of these relationships provides consistent, high quality service to the entire NERSC client community through the support of early, production-quality, large-scale capability systems.

The NERSC client base represents the varying and different roles clients/users play, from users of resources to participants in developing new technologies. NERSC recognizes its clients play different roles at different points in time. For example, a user of NERSC systems may at times be a part of a team effort to incorporate a new technology or service at NERSC, or at minimum be a part of identifying a needed service or technology. In almost all cases, clients are consumers of NERSC resources, both systems and human. Clients in this role are seeking supportive technologies and compute and storage resources to support the scientific progress for various programs and projects. In this way they behave and act as customers in most standard service models. Clients have integral roles in the next two concepts.

NERSC employs a decentralized, open but organized methodology for providing system time, resolving service problems, and interacting with its clients (Figure 3-1). From right to left of the model, clients interact with frontline service interfaces for usage time, to answer questions, and to resolve problems. This interaction takes place along a porous border on the right. Depending on the situation, a client's needs may depend on multiple or single functional areas to resolve problems and receive service. When a client's need requires more then the frontline service interface, the client can "reach back" (going right to left) to other functional areas in the model for support. Clients may flow freely in and out of this model as they request and receive service/support.

The development of high-performance computer systems and technological integration plays a central role in advancing and balancing core and peripheral systems at NERSC in support of its clients. From the top and sides of the model, technology requirements are developed, articulated, and prioritized down through the three stages of technology planning. By incorporating the requirements from its core systems area (networking, systems, and storage), services and other supportive technology areas, as well as clients, NERSC ensures that improvements and advances in technology remain responsive and balanced in a complex operating environment.

Finally, at the center of this model is NERSC's goal setting process, which ties all the functions together. This process is described in detail in Section 1.4.1 and is illustrated in Figure 1-1. Essentially this a process by which NERSC management and staff, using client input, prioritize system improvements and advances for further development and integration with the production environment. Other service goals needed for the smooth, reliable operation of high performance compute resources are also prioritized in this process.

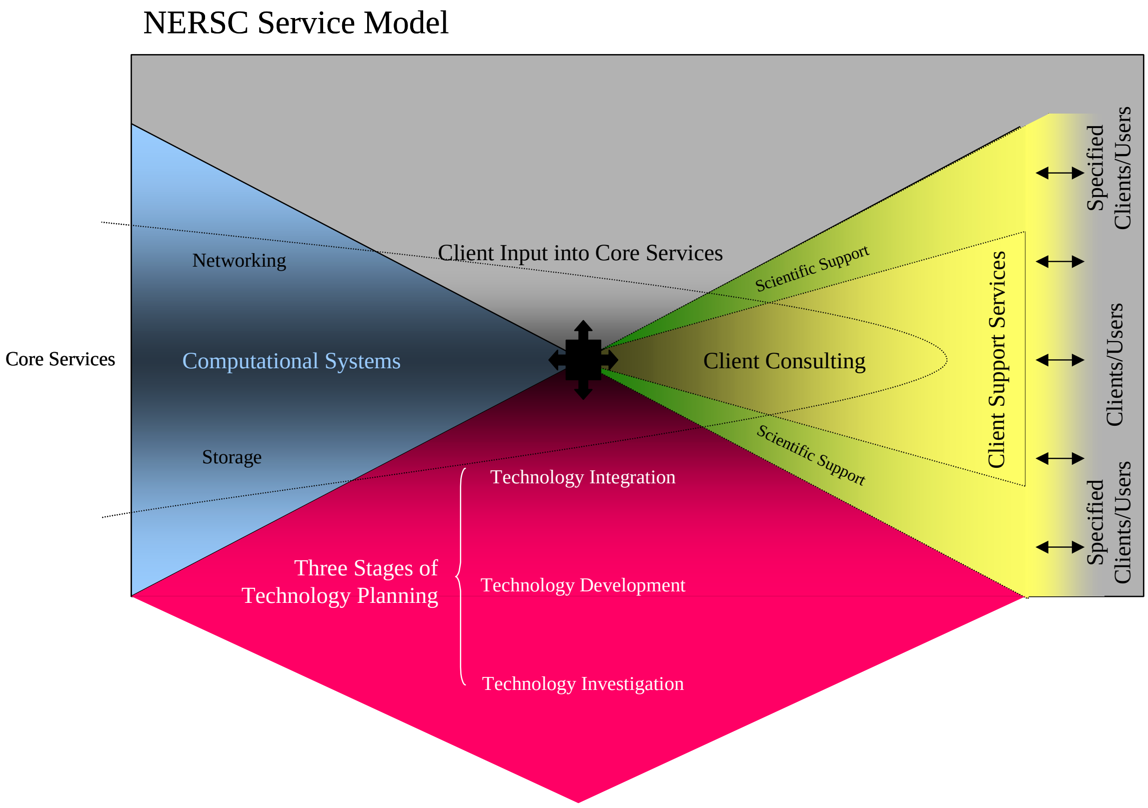


Figure 3-1. A conceptual model of NERSC services

3.2 Core Client Support

3.2.1 System Monitoring and 24 by 7 Support

System monitoring and operations staff support all systems on a 24-hours, 7 days a week, 365 days a year schedule. The tasks involve system monitoring, initial system troubleshooting, system backup, and management of the near-line and offline storage media, in or out of robots. Equally important, NERSC supports productivity-improving tools and techniques designed to automate or improve the operational tasks that must be acquired, created, or imported. This activity also recognizes that vendors often do not have complete and usable operational documentation, in part due to the leading-edge work of NERSC and the early delivery of new technology. The creation and update of all the relevant documentation concerning the operation of systems is important as a reference, a training tool, a method of configuration documentation, and a way to distribute much of the NERSC experience to other sites.

Operations analysts are always available to NERSC clients, 24 hours a day, 7 days a week. They can resolve basic requests (such as changing a password) immediately. Other requests are logged for the consultants or account support staff to answer during business hours.

NERSC systems and storage staff provide basic system administration and remedial maintenance 24 hours a day. Each system has a point of contact assigned who responds to system issues and problems. Also, a system manager is responsible for the overall operation and support of each system, as well as being an expert on the particular hardware and software. Vendor personnel, sometimes on site, are also available to make sure the systems operate well and to provide high reliability.

3.2.2 Consulting Support

Consulting support focuses on directly helping the DOE scientific community become more productive in their computational and data management work. It provides direct client assistance, managing and resolving client problem reports. It is important that the client community is able to interact and ask for assistance in the way most comfortable and effective for them — not just in the most efficient way for NERSC. Thus, NERSC supports telephone, email and Web interactions with very timely acknowledgment and response resolution. Once a client reports a problem, the consultants manage it until it is resolved. They will not send the client to another group or have clients manage their own problems. NERSC uses problem-tracking systems across the organization to document and manage problems reported by NERSC clients. NERSC consultants document the solutions to problems, and these documented solutions can turn into reference material for other users.

Consulting staff solve and manage client problem reports and requests for assistance, particularly with regard to programming, and debugging and optimizing codes. NERSC uses live phone coverage from 8 a.m. until 5 p.m. Monday through Friday (local time). There are a number of areas in which the consultants provide assistance:

- *Debugging:* Probably the most labor intensive part of NERSC user services. Sophisticated debugging and code analysis can be intimidating. Clients often need to be walked through the process. Consultants who use these sophisticated tools over and over again know how to apply them to particular debugging tasks. NERSC consultants are directly available to work one-on-one with clients in debugging.
- *Determining what the problem really is:* Clients don't always properly identify their problem. E.g., the user thinks the problem is with the routine generating the error message, whereas the problem is really a memory leak elsewhere in the code. Since NERSC consultants are full-time professional staff, they have accumulated multiple years of experience. This experience helps tremendously in quickly identifying the cause of a problem.
- *Optimizing I/O:* I/O is usually very specific to each machine. Clients unfamiliar with I/O might not even consider the system-specific options available to them. NERSC has documented features and differences of I/O libraries in online documentation.
- *Optimizing communication patterns:* The user might request faster communications methods, whereas what they really need to do is reduce the number of communications through algorithm improvements.
- *CPU optimizations:* These are specific to the compiler (e.g., user wants to use cache bypass directive for complex data).
- *Parallelizing:* Ten to fifteen years ago, computer centers were teaching vectorization techniques that were later incorporated into compilers. Message passing parallelization is still a manual process. Consultants assist with the assessment that needs to be identified. Clients need to be trained and updated on the latest methods. Furthermore, NERSC advocates improvements to vendors' compilers and libraries to make them more efficient.

- *Conversions from one platform to another:* NERSC consultants consolidate information on the differences between platforms, and decide whether porting tools need to be written (e.g., data format converters). This information is then posted on the Web.
- *Assistance in choosing library software:* Consultants are well versed in helping with the selection of software, especially software that is not standardized, e.g., FFT routines, random number generators, sparse matrix solvers.

Since consulting means working closely with NERSC clients, it is important that the staff assess and understand client needs and requests, and act as advocates for them to the rest of NERSC and to the vendor community. The consulting staff is intimately involved in testing and evaluating the changes brought on by new functions. They assist in developing and maintaining benchmark suites that are used in system testing and testing of new application software, such as compilers, libraries, and tools for advanced programming.

3.2.3 Account Support

Account support staff create and manage user and group accounts. They input and maintain the data in NERSC's Information Management (NIM) system: contact information for people, project and repository information, usernames, filegroups, etc. They act as the NERSC postmaster, creating and maintaining email aliases and lists, and resolving issues with bounced email.

NERSC Information Management (NIM)

During the past year and more (FY2000–FY2001), NERSC has completely redesigned, reimplemented, and modernized its administrative information for account requests, accounts, and allocations management. The current system is called the NERSC Information Management (NIM) system. NIM is designed to provide an easy-to-use, effective system for account management, accounting and usage information, and other informational tasks. It is designed as a client-server system, with a central database for all data about projects, users, machines, and usage. User accounts are managed by NIM, with client fingers automatically installing accounts on systems once authorized. This not only improves efficiency and reduces manual effort, but increases accuracy. The system maintains knowledge of what the projects are, which organization has granted time to the project, and which users are associated with these projects.

NIM accumulates usage and charges by user and project. It is capable of enforcing limits on use when necessary and is used to provide usage reports: Principal investigators and users can do ad hoc queries of NIM for usage and allocations. Program managers can do summary queries. Additionally, NIM produces periodic reports of usage.

NIM greatly facilitates integrating a new system — or cluster of systems — into the NERSC system and service architecture, which is critical since NERSC is always changing. A new system is added to the NIM data structures, and client code is ported to do both account management and usage reporting. The client is designed to use standard vendor-supplied accounting information, such as UNIX process or LoadLeveler accounting, to collect the machine-specific information. Then the system-specific clients process that information and provide it to NIM in appropriate form, via encrypted transfer.

Account management works in the reverse manner. NIM provides the necessary information to create an account on a system via encrypted transfer, and then the client on the system invokes a script or program that is system specific to do the actual installation.

NERSC staff have created a set of relational databases using an architecture that permits management of SQL and PHP Web interface code. This architecture facilitates better code maintenance, controlled evolution, systematic testing, and automated quality control. This new environment makes it much less painful to upgrade to new versions of the underlying database management system (Oracle), Web server (Apache), and Web-database application language (PHP).

For NIM's users (including scientists, DOE program managers, and NERSC staff), these changes have resulted in a new, unified Web interface for user account management, usage information, and the Energy Research Computing Allocations Process (ERCAP) for NERSC resources.

NIM Phase 1 was deployed in the beginning of FY2001 after a one-year development effort. In FY2001 NERSC completed the integration of ERCAP into NIM: the PI allocation request form, staff interfaces for managing the request forms, the reviewer evaluation and award allocation forms, and initial distribution of resources for FY2002 allocations. NIM was also enhanced to provide more general usage reports, in particular daily usage (or usage by time period) and usage by machine (or groups of machines).

In FY02 NERSC will add the first installment of "User Self-Service" facilities that will enable members of the NERSC community to change most personal information (address, telephone, email, organizational affiliation, etc.) with post hoc staff review. Other plans for NIM in FY2002 are:

- Incorporate administration of user accounts and data on HPSS and PDSF.
- Complete work on facilities to delete users who no longer are affiliated with NERSC and migrate their files to HPSS via NIM.
- Streamline and automate management of UNIX file groups.
- Automate the allocation management policies (removing resources from projects that have made insufficient use of them at certain points throughout the year).
- Provide a better interface for specification of business rules, with direct connections to the code that implements them, be it SQL, PHP, or scripts.
- Migrate the user account management functions from Oracle forms to a PHP Web interface with appropriate security and access control in order to have a unified interface platform (PHP).
- Provide more flexible, comprehensive, user-controllable reporting tools.
- Enhance NIM security: use better authentication procedures on both the database and Web server sides.

Phase 2 of NIM development will be completed at the end of FY2002, and the basic NIM system will be in a maintenance, new-system integration, and small-improvement mode.

In summary, NERSC actively manages authorized accounts and provides a system-independent infrastructure for PIs and DOE to authorize and audit the usage on the system to assure it is appropriate and planned. NIM is the center of that system. NIM provides an excellent interface and system independence for this function.

3.3 Computational Scientific Support

The core of NERSC service architecture is a team of professional, highly expert consulting staff. These large-scale computational specialists are able to assist the diverse NERSC user community in using the NERSC systems in the most effective manner.

3.3.1 Training and Documentation

Training

People learn and assimilate information in different ways. The limits of having to read a document or attend a class in a certain location will be overcome as a more multisensory approach to information delivery is provided. Training will be accomplished with integrated material: video and audio broadcast of the lecturer, presentation slides made available electronically, and possibly backup written information.

NERSC provides advanced training and client instruction in the use of the latest technology. NERSC staff develop skills in new areas and share them with clients by creating, updating, and presenting all the relevant external and internal training information related to using NERSC systems. These activities include teleconferences, multiple days of intensive classes, lectures, seminars, and symposia presented in collaboration with other groups. Remote training increases the difficulties by separating the learner and instructor. Immediate feedback is not always available to the remote learner or to the teacher. Because NERSC serves a remote clientele, we must constantly work to minimize the effects of distance.

NERSC uses traditional videoconferencing and teleconferencing to provide the timeliest information to the clients, rather than using infrequent face-to-face meetings. But until recently, each method has had shortcomings. In the coming years, in part because of the development and deployment of the Unified Science Environment and collaborative tools, it will be possible to cost-effectively deliver training and information largely independent of where the provider and consumer are. NERSC will capture its training content — classes, seminars, and lectures — in a form that allows digital distribution. Real-time video broadcast (not studio production quality, but nonetheless competent and effective) of training and other events will make it far easier for clients to have the most current information on large-scale systems, programming, and algorithm development. Once captured, the content can be played later (although without interactive discussion and questions), so a library of training information will be available.

NERSC trains highly proficient computational specialists in specialized tasks that use the most advanced computer systems to solve the most difficult technical problems. Such people are challenging students: they demand exactly what is needed for the solution to their individual problem or class of problems, and nothing less. NERSC’s past approach to distance training has been to schedule lectures on single, well-defined topics using audio, video, and online methods. Broadcast lecture materials are available on the NERSC Web server. From Web statistics and from lecture-related questions to NERSC consultants, it is clear that archived lectures are used by the NERSC clients. For example, in March 2001, 25% of the Web “hits” were to the training area that includes these archives.

Ordinary conference telephone calls suffice for audio coverage, providing short lectures to remote attendees. One approach has used full video conferencing, via ESnet infrastructure. This involves coordinators at each participating site scheduling rooms containing specialized equipment for the lecture time slots. NERSC users have not always been able to book such a specialized facility (and often they are not at a site with such a facility).

We have also used the Real Media format for recording lectures. They can be displayed with the Real Player plugin from Real Networks. NERSC’s video archive is at http://hpcf.nersc.gov/training/classes/video/.

Another critical part of base services is how NERSC will increase and improve its delivery of information and services via the Web. NERSC plans to improve its Webcasting methodology. Webcasting involves sending medium-quality video and high-quality audio streams to commonly available software applications used with or within Web browsers. Several of these are available, from Microsoft, Real Networks, Apple, and others. There is active competition among the various formats, with no clear standard as of mid-2001. They are all cheap (or free) to the end users, and the only real expense involved is in the encoding and streaming servers that would be at NERSC. Web browsers, of course, are ubiquitous. Each of the Webcasting technologies allow material to be encoded and transmitted in real time, to be encoded and archived for later on-demand transmission, and to be digitized from video tapes.

Frame rates in outgoing streams can be adjusted for differing receiver bandwidths. For instance, clients with only modem access could be provided with a stream with lower frame rates and less frame data, sacrificing video for audio quality, while users with full network support could be provided a higher bandwidth stream that potentially could deliver quality equivalent to broadcast television. An encoding and streaming server has been set up by the LBNL telecommunications department for Real Networks Webcasting; we will take advantage of this service. This protocol carries no stream constraints and will be fairly cheap to implement. Finally, based on our experience with the above, we may also elect to set up a server for Microsoft’s Windows Media. In the future, other protocols and clients will be evaluated and selected for use.

The advantages of Webcasting are several. First, the capital outlay for equipment is lower than for full video conferencing. Second, its usage is more flexible, requiring less in the way of scheduling and being more available at the user’s convenience. Third, it is more available than specialized video conference setups. Finally, presentation archiving allows for asynchronicity between presenter and attendees.

The most important area of Webcasting to focus on is not the broadcasting *per se* but the usefulness of the content generated for the training archive. The lectures need to be indexed so that it is easy for the user to find the exact material they wish to review. Clients should be able to choose the display format that best meets their learning style. Examples of three such choices are audio, video, and presentation material; audio and presentation material; and presentation material only.

Several options are available for providing the remote learner feedback on the material she/he is learning. First, the user can call a consultant during business hours. Second, self-testing units can be provided. Third, the training lecture archives can be linked with a NERSC knowledge database that the user can search for more information. By expanding its use of the Web, NERSC will be able to provide more information and assistance than ever before.

In FY2002 and FY2003, we will concentrate on finding and deploying software that allows us to produce high quality audio/video lectures for our video archive. The top candidate today is Virage’s MediaSync (http://www.virage.com/products/mediasync.html), which provides an integrated, end-to-end solution for rapidly assembling, synchronizing, and publishing streaming video with PowerPoint presentations. The steps required to capture and encode video, synchronize it to PowerPoint, annotate it, and then publish the results to our Web site can be orchestrated in one automated process.

In FY2002, NERSC is planning to conduct training classes using Access Grid nodes that will incorporate the infrastructure necessary for both full video conferencing and Webcasting, as well as multimedia presentation and local attendance. This will be a traditional lecture room/classroom augmented with equipment for presenting and broadcasting lectures in real time, as well as for digitally capturing them. Access Grid Nodes can be used for both NERSC user training and for more general purposes. For instance, lectures by NERSC clients and staff on state-of-the-art science can be presented. They can be aimed at audiences ranging from university classes to college and secondary school classes, and to the lay public. This will greatly enhance our educational outreach and our visibility to the public.

In FY2004 and FY2005, NERSC expects to work on more capable back-channel methods for use in real-time events. Video telephony will become more practical as available bandwidth increases. This technology will be used to allow remote users to ask questions and receive immediate feedback. We will look for platform independence, simplicity, appropriateness, and reliability. This technology will allow us to conduct highly specialized training sessions customized for small groups of users.

In summary, NERSC will continue to provide high-quality state-of-the-art training information, but will invest in improving its tools and infrastructure to enable delivery of the training content in new ways. Much of this work will be in collaboration with the Unified Science Environment development and deployment going on at the same time.

Online Information and Web Technology

The creation and update of all the relevant documentation for using NERSC systems and environments are extremely important since they serve as reference and training tools. While NERSC leverages to the greatest extent the information provided by vendors as well as that available from other sources, a large amount of work needs to be done to provide the value-added information necessary for the clients to make the best use of NERSC systems. NERSC documentation can be found at http://hpcf.nersc.gov/.

Over the next five years, NERSC plans to implement new Web technologies in the areas of (1) server reliability and security, (2) Web access and authentication, (3) Web tools, and (4) multimedia. These plans are described in more detail below. In addition to these new areas, we will devote considerable effort to keeping the technical content of the NERSC Facility Web site (http://hpcf.nersc.gov/) up to date and writing new technical documentation for our users. User surveys (http://hpcf.nersc.gov/about/survey/) have shown that maintaining technical documentation is very important to the NERSC client base.

The implementation plan for Web services includes the following areas.

Web Partitions

By the end of FY2002 we plan to implement a partitioned Web architecture for improved reliability, security, and integrity. This is prompted by several recognized needs:

- The need to support information retrieval at all times (24 × 7).
- The need to continually improve the Web site through software changes, content changes, and system configuration adjustments places increasing pressures on the site for reliability and service.
- A partitioned architecture can aid our cyber security posture.

To meet these needs, NERSC is designing and implementing a partitioned architecture for our Web systems. The partitions include public and private areas. Each area will be improved and expanded over the next five years. Expansion will include increase in performance, expansion of content, and functional improvements.

The public partition systems are open to the entire Internet: the initial configuration is two systems in the public partition. These systems will act as a primary Web server and a backup in a hot fail-over. At this time a load-balancing system is not required either to provide sustainable service or to enhance system integrity. This assessment may change over a five-year time frame. We are certainly actively following load balancing technologies and will implement such a scheme when it is cost effective.

The public partition systems will not be used for Web authoring. The filesystems will be updated from the private partition systems (below) as information becomes ready for user consumption. The filesystem on the backup system will be automatically kept in synch with those on the primary system for its role as a hot backup.

On the public system, there will be authentication mechanisms that allow restrictions to the information on the systems as appropriate (e.g., vendor manuals and documentation). For example, some information will be available only to authorized NERSC users. Other information may be available only to NERSC PIs. NERSC expects that most of the information will be generally available.

The initial configuration of the private partition systems will essentially duplicate the public partition configuration for the private partition but will be restricted to the portion of the NERSC network open only to NERSC staff. There will be a primary and a backup system. The primary system in the private partition will be used primarily for Web authoring. This primary Private system will have all the relevant NERSC staff as users for Web content development (as our main Web server does now).

NERSC is in the deployment phases for components of this partitioned architecture. We anticipate that as we roll it into production during the next year, we will be able to provide more reliable, security hardened service while at the same time providing a bit more consistency/integrity for our users by moving only content onto the production system when it is fully configured and checked out.

Sitewide Authentication

When NERSC is able to put a facility-wide authentication mechanism into place for all its clients, it will incorporate Web authentication as well as system authentication. This capability will facilitate a number of Web-based applications (such as a MyNERSC Web portal) that are either awkward or impossible now, including services such as:

- Monitoring the status of jobs, submitting jobs, retrieving output, etc. Such facilities may initially be limited to file transfers for submission or status monitoring. Ultimately, applications can be developed that make use of authenticated Web interfaces for active monitoring and control, even to the extent of dynamically adding piped processing facilities such as those for visualization of partial results, adjustments to add processing for hot spots or to modify processing parameters for more accurate final results, etc.
- With more general Web user authentication available, some of the technology that NERSC has developed for NIM services (e.g., HTML-embedded actions through PHP active Web content and database services) can be deployed more widely for general services (e.g., more interactive user surveys, user-specific Web communication/consulting). We can glimpse some such services in the Right Now Web consulting service that NERSC has deployed; see http://hpcf.nersc.gov/help/helpdesk.html. Such services are currently limited to some extent by their lack of effective (safe, shared, trusted) authentication.
- Web-based authentication for NERSC staff can also help enable a new generation of Web facilitated Web authoring. Such Web facilitated authoring can provide more effective shared authoring, quicker turnaround for Web content, and other indirect service improvements.
- With user authentication and database integration, it is possible to consider developing data flow between user applications and databases that can be managed through Web interfaces.

- User Web authentication can enable mechanisms for sharing data and other capabilities between users through a Web interface.

Web Tools

We monitor a number of Web tools on a regular basis for improvement opportunities. Among these are:

- Web spidering, indexing, and searching/structuring tools to enable users to find information more effectively.
- Link checking, spell checking, and other site consistency checking to improve site content quality.
- Log analysis to help NERSC staff tune the site to better respond to users needs as seen in their patterns of access.
- Web authoring/content sharing with facilities for more functional use of style sheets, meta tag information (e.g., for better search identification), active content (improving user interactivity), etc.

NERSC will be able to employ many of these new technologies to enhance the Web information access/control experience for NERSC clients during the next three to five years.

Multimedia

At present we have relatively little multimedia (audio/video) information on the NERSC Web site, except in the form of recorded training classes or seminars. We plan to develop more interactive multimedia Web applications (including online Web conferences or training sessions, Web facilitated multimedia interactions between consultants and users, etc.).

3.3.2 Visualization

The mission of NERSC visualization is to apply scientific visualization principles and practices to domain-specific projects as part of a multidisciplinary team, and to anticipate, define, and develop new visualization technologies that are appropriate for contemporary and future applications. In order to fulfill this mission, the NERSC staff pursue activities in a number of related areas.

First, NERSC maintains and supports a diverse collection of visualization software on NERSC production systems. The software ranges from simple plotting packages through high performance 3D applications. Some software is commercial and some is freeware. The maintenance activities are focused upon keeping this collection of software up to date and operational in the production environment. The support activities include consulting projects and providing accurate and up-to-date online documentation. Some consulting projects are short, consisting of only simple questions from users and succinct answers. Others are more substantial, involving a Visualization consultant generating complex visualizations. In some cases, these more involved consulting engagements evolve into long-term collaborations that seek to find solutions to problems that are not easily solved with existing software or approaches.

Second, NERSC performs outreach activities which are intended to increase awareness in the user community of the NERSC visualization services and capabilities.

Third, NERSC engages in investigations that are intended to produce solutions to today’s hard visualization problems, as well as to evaluate approaches and solutions to tomorrow’s problems. These activities often involve collaborations with staff in other NERSC groups or other institutions. A recent example of such projects is Visapult, which is targeted at remote and distributed visualization of data that is too large to fit in any one machine’s secondary storage, and also couples visualization to a live-running simulation. Visapult also shows the need for substantial network bandwidth, and has been used to win the SC Bandwidth Challenge two years in a row (2000, 2001).

Priorities and Plans

In the years ahead, NERSC envisions significant growth in visualization capabilities in the NERSC Center. Growth in visualization capabilities mirrors growth in fundamental technology areas that impact visualization: visualization represents the application of diverse resources that are brought to bear on a single task.

One growth area is leveraging the continuing improvement of the quality, speed, and reliability of desktop machines. Historically, remote visualization from NERSC has focused on the use of remote X11 display technology. This approach benefits from a high degree of portability — any user who has an X-based workstation (X-terminal) can use any of the X-based visualization applications installed on NERSC production systems. For many types of applications, this arrangement is adequate. For graphics and visualization applications, inadequate network bandwidth and high latency rates contribute to make remote interactive visualization all but impossible. As desktop machines increase in capability, a new approach becomes possible that was not feasible a few years ago. NERSC intends to begin to provide a service known as *application license serving*. This service will allow users to download graphics and visualization applications from a NERSC distribution server. The user will then install the application on their local workstation, and as a result, interactivity of visualization applications will not suffer from network-related problems. Commercial applications that use FlexLM licenses will connect to a central license server at NERSC, where a pool of floating licenses will be made available to users. This approach requires additional consideration of license management and the potential for contention.

Another growth area will be streamlining and simplifying access to visualization capabilities at NERSC, with particular emphasis upon techniques for remote visualization. Remote and distributed applications, such as Visapult, require the user to manage the launch and execution of multiple, distributed components on several machines. This mode of use is at best inefficient, and at worst, unfriendly and not conducive for user adoption of new technology. In order to address challenges posed by an increasing array of resources, we are focusing on the deployment of a Grid portal that simplifies and streamlines access to visualization resources at NERSC. Research prototypes can be transitioned into production use within the portal framework. As more NERSC resources become Grid enabled, the number and diversity of remote visualization resources will likewise increase, and be presented to users via a portal interface. In most cases, the user will have access to a significant amount of visualization technology using only a Web browser on a laptop or workstation. In the future, we expect to be able to provide services such as one-button MPEG creation by using the portal infrastructure to coordinate execution of several NERSC resources. The portal acts as a broker on the user’s behalf to find and control resources.

The NERSC Center maintains a dedicated production visualization server with powerful graphics hardware. This server will be more effectively utilized by several research prototype visualization applications currently under development that will perform hardware-accelerated volume rendering with the resulting images delivered to the remote user’s web browser. NERSC will increase the scope of applications that take advantage of this unique resource and will keep the resource up to date by regularly upgrading the system infrastructure. Like other NERSC production systems, the visualization server is monitored by the NERSC Operations staff, and is supported 24/7.

Outreach Efforts

To better address the visualization needs of the NERSC user community, NERSC will need to better understand their requirements. NERSC will approach this task in a number of ways: attending NUG teleconferences and meetings to disseminate information regarding visualization, delivering presentations, describing and soliciting responses regarding usage of visualization capabilities.

NERSC will establish a process to engage projects for visualization services. The process will be iterative and will evolve over time. An example of this process is described below.

> NERSC staff will work with projects that have visualization needs to understand their data requirements, including the source (experimental vs. theoretical), potential size, and nature of the data (images, volumes, vector data, etc). It is critical to know what types of analyses are performed, and for what purpose, as well as what tools the projects are currently using (commercial, domain specific, or custom codes they have developed themselves). Large datasets pose special challenges for analysis and visualization, and these projects may require substantial investigation in order to yield a solution.
>
> Next, NERSC staff will strive to gain a better understanding of what type of visual presentation of data will provide the greatest benefit to the investigator. This representation could be anything from a line or bar chart to complex multidimensional volumetric time series or a combination of multiple interactive representations. We will want to know what they are currently using to perform their visualization activities, such as domain-specific tools (e.g., molecular modeling, materials science programs) or tools combining other analytical functions (e.g., statistical or symbolic capabilities).
>
> The location of user data can have a substantial impact on the choice of tool or resource for visualization. Do the project members need to move the data to specialized equipment to analyze it? For example, does the data need to be moved to specialized visualization servers with higher visualization capabilities than general purpose computational resources provide, or do they wish to perform visualization at the same location where they store or process their data? Is there adequate network bandwidth between the data source and the computational resource for performing visualization?
>
> The evaluation of visualization services by groups that have already used these resources is fundamental to assess whether the directions set in our work plan are the right ones for the user community. Do investigators feel that they have sufficient support to adequately use the visualization resources, or are they limited by a lack of adequate documentation, support, training, personnel, equipment, time, or some other restrictions?
>
> What software have they already used and on what platforms? Is the software easy to access? Is the documentation regarding its use both comprehensive and easy to understand? This would include gnuplot, xv and commercial software (AVS Express, AVS5, IDL, PV-Wave). We realize that no one package begins to meet all the requirements of a user or even a particular project, and will try to suggest what additional options may make the investigators’ visualization more productive.
>
> Based upon user feedback, are there common recurring problems utilizing this software? What approaches can be used to solve these problems (e.g., examples provided on our Web site or other ideas suggested by the user community)?

In response to the information gathered from working with these projects as well as information already gathered, NERSC will announce new capabilities, such as software upgrades or software availability on new hardware platforms. NERSC will improve the usability, availability, and scope of online documentation. This documentation will include links to vendors' documentation, other users of the software, the location of downloads, examples of usage, and helpful tips.

NERSC is currently increasing the availability of software resources through wider deployment of these software resources on more platforms. This will include deployment of visualization software on more NERSC platforms than just the visualization-specific servers, particularly platforms where it would be more expedient to visualize the data in situ. In addition, we expect to deploy visualization resources on users' equipment utilizing remote license management.

To better understand how much improvement has occurred, we will use software asset management software to monitor software usage metrics. This will be particularly important as we deploy software more widely through multiple novel approaches, including portals and distributed license serving.

3.3.3 Application Software Support

Consulting staff provide software support for a complex set of applications, libraries, tools, and environments that exist on some or all the NERSC systems, such as Gaussian, NAG, MPI, Totalview, and performance analysis tools. A variety of software packages are supported, including visualization and client interface software, some of them commercial and some from the open scientific community. The list of available software is documented at http://hpcf.nersc.gov/software/.

Software Management

The overall goal of NERSC's software management process is to ensure that users have easy access to up-to-date software to maximize their productivity and research capacity while at the same time balancing support costs. To accommodate the function of NERSC as a production scientific computing center as well as a center to promote new cutting-edge technologies in computational sciences, software installed at NERSC is viewed as either *production* software or *experimental* software. Production software is further characterized as *recommended* (fully supported), *acceptable* (continued support may end), or *refrain from using* (scheduled for deletion).

Production software packages are generally mature products, most usually supported by a commercial vendor, and are most suited to be used in a production environment. Some, however, may be mature public domain packages with a widespread appeal. NERSC's policy is to provide at least the following level of support for *recommended* or *acceptable* software:

- Online and/or Web documentation.
- Consulting support on access, invocation, and general functionality.
- Source and/or vendor maintenance agreement.
- The ability to build or reinstall by NERSC staff across system upgrades.
- Testing to assure overall functionality.

Software upgrades for fully supported software are made within three months of announced availability and are installed within one month of delivery to NERSC.

Adding New Software

The consulting staff keep abreast of developments in software available for high performance computers. For example, staff participation in the Ptools consortium resulted in deploying Zerofault (a memory debugging tool) and PAPI (Performance API) on the IBM SP. Consultants monitor developments in our vendor-provided software and select new offerings to bring to NERSC (past examples are Parallel Gaussian and Parallel NAG, and more recently the Guide and Assure tools from Kuck & Associates). New software goes through a pilot stage (staff testing followed by user testing) before being installed as production software. Users may submit requests for new software, which are reviewed by the staff. The decision to support a given package depends on its expected use, its significance to support the computing needs of the user base, its quality, and its purchase and ongoing maintenance costs.

During the next five years, NERSC expects to expand its software offerings in the following areas:

- Performance analysis and optimization tools
- Selected parallel application packages
- Tools developed by the DOE Integrated Software Infrastructure Centers (ISICs)
- Visualization software.

3.3.4 Core Cyber Security Support

NERSC's computer security strategy was discussed in Section 1. In order to maximize our clients' ability to conduct science and mitigate the effects of computer security incidents, NERSC provides the following services:

1. **Vulnerability Detection.** NERSC performs host level scanning on a periodic basis to detect and remove vulnerabilities. The rate at which NERSC scans its hosts will depend on the severity of the vulnerabilities. In some instances, host services may be blocked until a resolution or alternative is developed and deployed. Scans are performed with minimal to no impact on our clients.

2. **Intrusion Detection.** NERSC will continue to pursue a course of analyzing network traffic to detect and block intrusion attempts. Sensors are deployed throughout the NERSC network to analyze traffic in real time or to store network traffic information for analysis at a later date or to assist forensic analysis of computer security incidents.

3. **New Technology/Vulnerabilities.** The field of computer security is constantly evolving and rapidly changing. NERSC keeps abreast of these changes by monitoring known discussion groups for system vulnerabilities, by maintaining close contact with system vendors, and by analyzing usage patterns for new attacks or exploits. NERSC also investigates new technologies and techniques that may provide a more secure environment. New services are deployed after their impacts on the NERSC client base and our operations are taken into consideration.

4. **Client Awareness.** A key component in maintaining a secure environment is having a staff and client base aware of secure computing practices. NERSC works with our clients, migrating them away from insecure protocols, and developing documentation about secure computing practice. Additionally, NERSC staff are available 24/7 for computer security related incidents.

5. **Incident Response.** It is impossible to eliminate all computer security related incidents at a site such as NERSC. The goal of incident response is to evaluate an incident, determine if any unauthorized access or usage occurred, and mitigate those instances in which a violation of NERSC policy

occurred. In incidents that involve a NERSC client, NERSC works closely with the client to determine the pathway for the incident and to prevent future occurrences. Additionally, NERSC provides information to the client so that they can inform their local site regarding the incident.

3.3.5 Direct Scientific Collaborations

NERSC scientific staff members have extensive experience working in direct research collaboration with NERSC clients on computational science projects, with the goal of promoting and enhancing the use of the facility. With expertise in applied mathematics, computer science, and various scientific disciplines, the scientific staff can effectively collaborate with NERSC clients and the broader scientific community. They are involved in timely development of state-of-the-art methodologies and strategies for computational sciences that are suitable for massively parallel computation, thus opening ways to do new science that is otherwise impossible. They also perform research in the design and implementation of highly efficient computational kernel algorithms for current and future NERSC architectures and applications. The work of NERSC's scientific staff in the last five years has resulted in a number of significant accomplishments.

One of the accomplishments is the modeling of metallic magnet atoms, which involved a collaboration with Oak Ridge National Laboratory, the Pittsburgh Supercomputer Center, the University of Bristol (UK), and others. This collaboration received the 1998 Gordon Bell Prize for the best achievement in high performance computing.

Another example, in climate modeling, was a collaboration with the Geophysical Fluid Dynamics Laboratory (GFDL). The goal in this project was to develop a massively parallel version of GFDL's Modular Ocean Model (MOM) code, which is used by researchers worldwide for climate and ocean modeling.

A third example is in the area of computational kernel algorithms. A sparse linear equations solver developed by the NERSC scientific staff played an important role in solving a 50-year-old problem of electron-impact ionization with three charged particles. The result was featured in a cover story in the journal *Science* (December 24, 1999).

In the next five years, NERSC scientific staff will continue their tradition of engaging in high-quality direct scientific collaborations. They will expand their scope and seek new research collaborations whenever possible. Potential new areas for collaboration include fusion calculations, accelerator physics simulations, lattice quantum chromodynamics calculations, computational astrophysics and cosmology, and computational biology. There are challenging computational problems in all these areas. The objective is to help NERSC clients perform their calculations and simulations effectively and efficiently on the terascale computing platforms. When appropriate, the scientific staff will collaborate with NERSC clients to develop necessary tools and technology needed to solve the computational problems mentioned above.

Under the Scientific Discovery through Advanced Computing (SciDAC) Program, NERSC staff will be actively involved in selected Scientific Challenge Teams and their corresponding Integrated Software Infrastructure Centers (ISICs). NERSC will interact closely with those ISICs that deal specifically with numerical tools and performance issues. Scientific staff will provide high-level support in porting, evaluating, and deploying tools developed by ISICs on the terascale computing platforms at NERSC.

They will help in understanding and resolving issues with the goal of enhancing the performance of such tools. Support for the Scientific Challenge Teams will be discussed in detail in the next section.

3.4 Scientific Challenge Team and Collaboration Support

In upcoming years NERSC will place an increased focus on supporting Scientific Challenge Teams. This section describes how NERSC will create a management focus on the high end, in what NERSC considers the community with the greatest potential for most significant progress in future years.

3.4.1 The NERSC Spectrum of Usage

NERSC distinguishes:

- Scientific Challenge Teams
- High-end capability users
- New users transitioning from midrange computing.

The relative positioning of these groups and NERSC's role in supporting the different types of usage is sketched in Figure 3-2.

Scientific Challenge Teams

The arrival of large, highly parallel supercomputers in the early 1990s fundamentally changed the mode of operation for successful computational scientists. In order to take full advantage of the new capabilities of these parallel platforms, scientists organized themselves into national teams. Called *Grand Challenge Teams,* they were a precursor to the *Scientific Challenge Teams* that NERSC anticipates as its leading clients in the next decade. These multidisciplinary and multi-institutional teams engage in research, development, and deployment of scientific codes, mathematical models, and computational methods to maximize the capabilities of terascale computers. NERSC's support model for the Scientific Challenge Teams is discussed in more detail in Section 3.4.3.

High-End Capability Users

These users are characterized by single principal investigator teams consisting of a leading researcher and his or her group of collaborators, postdocs, and students. These groups are usually at a single university or laboratory, working on a research level code, which is not shared outside the collaboration, or using well established third-party applications software. In NERSC's experience there is a continuum of usage, which can be represented by a pyramid: there are a few projects with very large computational or storage requirements at the top, and a much larger number of projects, which have smaller, but still very high-end requirements.

NERSC distinguishes between three classes of users.

- Strategic Projects: The extremely large and most critical projects for DOE science. About 15% of the system usage will be devoted to SciDAC and other major areas.
- Class A: Very large projects of the scale of Grand Challenges. About 50% of all the NERSC resources are expected to be dedicated to this class.
- Class B: Ordinary time requests. About 30% of all NERSC resources are expected to be dedicated to this class.
- Class S: Startup allocations. No more than 5% of all NERSC resources are expected to be dedicated to this class.

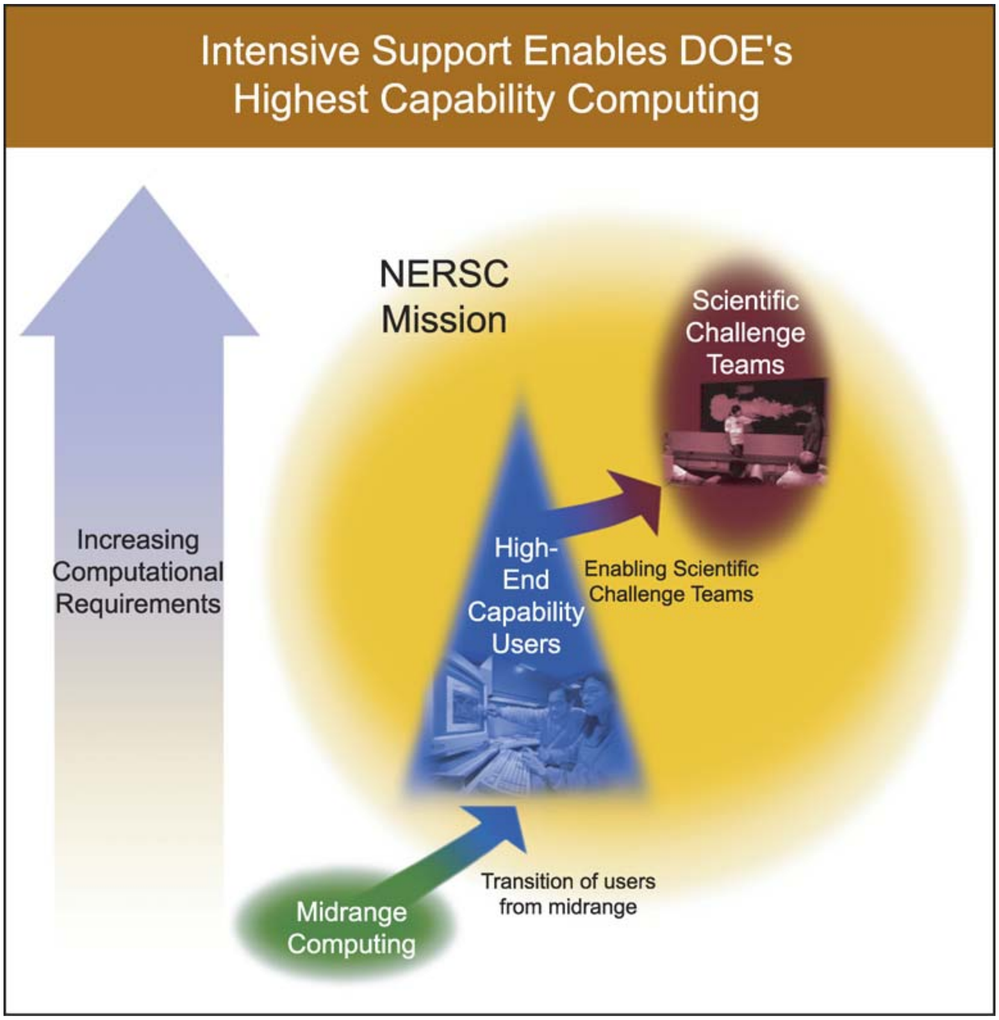

Intensive Support Enables DOE's Highest Capability Computing
NERSC Mission
Scientific Challenge Teams
Increasing Computational Requirements
High-End Capability Users
Enabling Scientific Challenge Teams
Midrange Computing
Transition of users from midrange

Figure 3-2. NERSC facilitates the transition to high-end capability computing, and enables Scientific Challenge Teams through intensive support.

Table 3-1
Distribution of Resources by Class

Strategic	15% of resources	3-4 projects
Class A	35% of resources	About 15 projects
Class B	30% of resources	about 50 projects
Class S	5% of resources	about 50 projects

Note that Classes A, B, and S refer to overall resources, i.e., a combination of computational and storage usage. In general terms there are several high-end capability user requirements, which set the NERSC high-end capability users apart from midrange capacity users:

- computational requirement of a sustained performance 500–1000 times of what is available on a typical desktop
- requirement for permanent archival storage of simulation or experimental data, in particular shared data, which is a community resource

- requirement to access special software (scalable community codes and strategically selected third-party application codes) which is not easily available on desktop or midrange platforms.

NERSC will continue to support the high-end capability users through its comprehensive scientific support described in detail in Sections 3.2 and 3.3.

Transitioning New Users from Midrange Computing

NERSC considers midrange computing and supporting capacity users to be outside of its mission. Their needs should be met by institutional or departmental servers, and possibly by the emerging topical centers. However, NERSC sees it as an important part of its mission for the DOE Office of Science community to promote computational science of scale and facilitate the transition of users from the midrange to the high end. All of the services described in sections 3.2 and 3.3 are available for supporting new users transitioning from midrange to high performance computing.

3.4.2 A Model for the Formation of National Teams: The Impact of SciDAC

In March 2000, DOE launched a new initiative called “Scientific Discovery through Advanced Computing” (SciDAC). SciDAC defines and explicitly calls for the establishment of Scientific Challenge Teams. These teams are characterized by large collaborations, the development of community codes, and the involvement of computer scientists and applied mathematicians with the discipline scientists. In addition to high-end computing, teams will also have to deal increasingly with issues in data management, data analysis, and data visualization. The expected close coupling to scientific experiments supported by the USE environment will be an essential requirement for success for some of the teams. Scientific Challenge Teams represent the only approach that will succeed in solving many of the critical scientific problems in SC’s research programs. These teams are the culmination of the process of users moving to ever-higher computing capability, and NERSC’s new structure enables that entire process.

3.4.3 NERSC’s Strategy for Supporting the Scientific Challenge Teams

NERSC’s strategy for the next five years is to build a focused support infrastructure for the Scientific Challenge Teams consisting of three components:

- integrated support and collaboration from the NERSC staff
- deployment of software developed by the ISICs
- deployment of Grid and collaboration technologies (USE).

By leveraging its comprehensive scientific infrastructure, NERSC will integrate all three components into a comprehensive support structure for the Scientific Challenge Teams. NERSC’s plan for each of these components is described in the following paragraphs.

Integrated Support and Collaboration from the NERSC Staff

During the second round of the Grand Challenge Program, from 1997 to 2000, NERSC served as the computing facility for eight Grand Challenge projects. NERSC provided focused support by developing the “Red Carpet” plan for the Grand Challenge teams. This plan revolved around building individual relationships with the users, as well as providing NERSC staff who serve as a focal points to expedite any problems or concerns. In order to extend the same level of support and collaboration for the Scientific Challenge Teams, NERSC will continue to use this Red Carpet plan, and add some new elements, described below. In the Red Carpet plan, each science area is provided with two focal points named from the NERSC staff, one from the User Services Group and the other from the Scientific Computing Group. The two focal points handle all requests by the Scientific Challenge Teams from the corresponding science area, and facilitate special requests, such as access to special queues and early access to new systems.

NERSC consultants play an important role in making Science Challenge Teams successful. They work with the science team members to accommodate special requests, such as adjusting system limits as needed (e.g., long 2,048-processor jobs), raising priorities for jobs at critical times, or making software modifications to allow for larger calculations. They can provide access to services dedicated to the needs of these teams, such as CVS repositories, database services, Web hosting, and collaboration tools such as email lists and email archives. NERSC consultants can provide training customized to meet the team’s needs, and they deliver specialized talks at conferences for specific user areas. Major efforts in the past have included porting software that otherwise would not have been available, e.g., CERNLIB on the T3E, and parallelizing public domain software and optimizing it for the center’s platforms (e.g., NetCDF).

The Scientific Challenge Teams also have direct access to the staff in the Visualization and Scientific Computing groups. The focal point from the Scientific Computing Group is responsible for identifying needs in special technical areas, such as visualization tools and algorithmic development, and communicating these needs to other staff within NERSC. This allows the science teams to have direct access to the expertise and resources available at NERSC. When appropriate, the focal points will leverage the connection between NERSC and several of the SciDAC Integrated Software Infrastructure Centers to ensure that the science teams have access to the software developed by these centers.

Deployment of Software Developed by Integrated Software Infrastructure Centers

SciDAC has established a set of new centers in computer science and applied mathematics, called Integrated Software Infrastructure Centers (ISICs). Their goal is supporting research, development, and deployment of software in order to accelerate the development of and protect long-term investments in scientific codes, to achieve maximum efficiency on terascale computers, and to enable a broad range of scientists to use simulation in their research.

Since the success of these ISICs will be measured by their impact on applications and Scientific Challenge Teams, NERSC will establish a close collaboration with these centers. The focus of the ISICs is the development and deployment of software in direct support of the Scientific Challenge Teams. NERSC will facilitate ISIC software deployment through a number of activities; however, the ultimate responsibility for deployment rests with the ISIC staff. In particular, NERSC expects that ISIC staff with the task of deployment will install and support the ISIC tools on the NERSC platforms and will provide the advanced support in the use of the tools for the Scientific Challenge Teams using the NERSC platforms. NERSC will facilitate management with these teams.

ISIC support can be seen as a direct extension of the Advanced Computational Software (ACTS) Collection support model, which provides documentation, assistance, and training, as well as second- and third-tier support. Since some of the NERSC staff are also involved in several ISICs, NERSC will not only have firsthand information on some of the tools that are to be developed under the SciDAC Program, but NERSC will also be directly engaged in the design and implementation of new numerical algorithms. In addition, these staff will be included in the evaluation and deployment of these tools on the NERSC platforms. Since they are also involved with making effective use of compilers, libraries, and tools for parallel computing, and performance evaluation and tuning on high-performance hardware, they are in a very good position to determine the extent to which these tools are applicable to the broader scientific community.

Deployment of Grid and Collaboration Technologies

The tools developed and deployed by the Unified Science Environment (USE) activity (see Section 3.5.5) will support Scientific Challenge Teams. Uniform access to computing and data resources across NERSC and Science Grid sites, as envisioned in the USE, will facilitate the collaboration among geographically distributed teams. Grid middleware services to support problem-solving environments and workflow frameworks are critical for large, integrated teams. The USE will deploy tools for integrating human collaboration with computing tools and tools for remote access (e.g., visualization and data). These themes are described in more detail in Section 3.5.5.

3.5 HPC Development and Integration

As a center, NERSC must remain vigilant in its technology investments in order to provide the best possible service. These investments represent a diverse and balanced technology portfolio that ranges from cooperative activities with vendors through bringing new technology to the center.

In the next several years, NERSC faces challenges in several key areas. One is the strengthening trend of vendors to respond to the needs of the consumer sector, and is exemplified by acute bandwidth bottlenecks. CPUs increase in speed at a much greater rate than memory or secondary storage bandwidth. To respond, the Center must investigate solutions that help to achieve balanced systems by alleviating these bottlenecks, and must promote the needs of computational scientists to the computing industry. Another challenge is the proliferation of resources across the wide area, and the need to weave these heterogeneous and distributed resources into a common fabric that is accessible to the scientific computing user. The fundamental technology to create this fabric is an emerging technology area, as is the presentation of the technology to users in a way that increases user productivity. Concurrent with growth and deployment of new technology, the Center faces an ongoing challenge of providing an environment that has strong defenses against the increasing number and sophistication of attacks from cyberspace. At the same time, these defenses should not create an impediment to the usability of the Center.

3.5.1 What Is HPC Development and Integration

In order to accomplish its goals and to provide the systems and services expected of the DOE’s flagship Center, NERSC must work to understand and overcome issues for systems and architectures of the future. Some of the most critical areas to ensure that future systems are able to support the computational needs of DOE science are listed below.

- Internal interconnect bandwidth and latency
- Network protocols that work in both the local and wide area networks and make TCP/IP effective at 10 Gbps and beyond
- Memory/CPU bandwidth and latency imbalances
- The expanding facilities needs of large-scale, commodity-based systems, including space, cooling, and electrical power
- The increasing complexity of large-scale systems

- The challenges of getting high performance from large-scale systems consisting of commodity hardware and software
- Limitations that inhibit getting data in and out of very high end, scalable systems
- I/O to CPU bandwidth balance and I/O functionality
- New network protocols
- High-bandwidth network interfaces
- Application performance analysis tools
- Self-configuring and self-healing systems and applications

The NERSC Program cannot afford to address all these areas within its current resources and expertise. Likewise it cannot afford to deal with many of them in depth. The NERSC plan is to address prioritized areas with focused activities, while positioning the Center to understand and leverage new technology coming from other efforts in the DOE and other organizations. The sections below discuss the NERSC Center efforts in these areas. Section 3.5.2, Future Technology Investigation, discusses internal interconnect bandwidth and latency; limitations that inhibit getting data in and out of very high end, scalable systems; memory/CPU bandwidth imbalances; and the expanding facilities needs of systems. Section 3.5.3, Numeric Algorithm Development, discusses the increasing complexity of systems and the challenges of making large-scale systems consisting of commodity hardware and software perform at high rates. I/O to CPU bandwidth imbalance and I/O functionality imbalance are addressed with the Global Universal Parallel File System and other efforts in Section 3.5.4. Network protocols that work in both the local and wide area networks and make TCP/IP effective at 10 Gbps and beyond are discussed in Section 3.5.5.

Unfortunately, at this time NERSC does not have the resources to conduct significant development efforts in the areas of new network protocols, high bandwidth network interfaces, application performance analysis tools, self-configuring and self-healing systems and applications, and other areas. The NERSC Center will continue to track progress of other organizations, including ASCI, National Science Foundation, and SciDAC projects. When results become available in the general marketplace, we will implement them as appropriate. NERSC will also periodically reevaluate these areas and may at some point in the future shift resources in areas not currently covered.

3.5.2 Future Technology Investigation

NERSC’s investment in high-end computational systems is critical to its existence. The Center is able to maintain its world-class capabilities through careful planning and by selection of effective technologies and suppliers. Support for this capability is provided through ongoing investigation into future technologies in computer architecture and interconnect networks. Activities such as participating in vendor-led non-disclosure meetings facilitate exchange of information, helping both parties ensure the best possible solutions for the scientific community.

Much of the focus in this area is on the long term. In the timeframe of the NERSC-5 and NERSC-6 systems, computational architectures face several challenges. Current architectures consume too much power and space and therefore require too much cooling. In terms of performance, the memory bottleneck threatens to choke off future gains in processing speeds. Finally, interconnection technologies appear not to scale to the numbers needed. While processors continue to follow Moore’s Law, the rest of the system is evolving at slower rates.

Various groups are working on new architectures that could provide solutions to some of these concerns. Much of the activity centers on highly capable processors that use less power and take up less space; most of these are aimed at the consumer market. Commercial success and commoditization of these technologies has the potential to produce discontinuous innovation, which may enable performance gains above current extrapolations.

To make use of these new architectures, or even to scale current CPU architectures, interconnection technology must make advances. Unfortunately, less work appears to be going on in this area because vendors are focused on the consumer market rather than scientific computing. This suggests an area of risk in the long term, and the Center is initiating activities intended to monitor and encourage progress.

3.5.3 Numeric Algorithm Development

NERSC staff will continue to engage in the development and deployment of high performance numerical algorithms that will benefit a wide range of scientific applications using the system in the NERSC Center. Making high performance numerical algorithms available to NERSC clients enables application scientists to focus on the development of their models and their application codes, instead of re-implementing many of the underlying numerical algorithms. NERSC will continue to focus on algorithms in the area of numerical linear algebra, which is at the heart of many simulation codes. For example, the solution of sparse systems of linear equations appears in accelerator simulations, supernovae simulations, and fusion energy calculations, to name just a few. Eigensolvers are needed in electromagnetic simulations, fusion energy calculations, and quantum chemistry calculations. The numerical linear algebra algorithms under development include the direct solution of sparse linear systems, iterative methods, preconditioning techniques, and eigenvalue calculations. Staff associated with the NERSC Center also participate in the Terascale Optimal PDE Simulations (TOPS) project, which is one of the SciDAC Integrated Software Infrastructure Centers. Thus, NERSC will be able to leverage much of the work on linear solvers and eigensolvers from the TOPS ISIC.

By interacting with NERSC clients, NERSC staff will also determine other numerical algorithms that may be common to several scientific applications. NERSC will either interact with appropriate algorithm developers to make such algorithms available to the NERSC clients, or will develop such algorithms if none exist already. Mesh generation and mesh refinements are two good examples, but these two areas are covered by the Algorithmic and Software Framework for Applied PDEs Project and the Terascale Simulation Tools and Technologies Project, which are two ISICs funded by SciDAC. NERSC staff have experience working with algorithm and tool developers to deploy tools and libraries on the NERSC systems. NERSC is already supporting the Advanced Computational Software (ACTS) Collection, which is a set of DOE-developed tools that make it easier to write parallel scientific programs. NERSC is responsible for the evaluation and deployment of ACTS tools. Hence, NERSC will be able to leverage the ACTS Collection effort to help deploy tools from developers on the NERSC systems.

3.5.4 Parallel I/O

The need for parallel I/O comes with the massively parallel computing systems that dominate HPC. Older HPC systems such as vector machines and SMPs of moderate scale maintained a single system image, giving every task on the system equal access to system resources such as memory, filesystems, I/O, and networking. Message passing libraries such as MPI make it easier to deal with the lack of a shared memory image, but a shared view of the filesystem is critically important for effective use and management of a system.

There are many mature solutions currently available, such as NFS, which provides a shared view of a filesystem across a large number of machines, but they have severe performance limitations. The key feature that differentiates parallel I/O from systems like NFS is that an *ideal* parallel filesystem should have performance that scales linearly with the size of the MPP system. So, if a scalable parallel I/O system provides a single-node job *X* Mbytes/sec of I/O performance, then an ideal parallel I/O system would approach *N***X* Mbytes/sec file writing performance to an application running on *N* nodes of the MPP. Such scalable I/O performance is particularly important for HPC centers because of the huge amount of data used by simulation codes.

Even more importantly, a parallel I/O system should provide simultaneous shared access to a single file from all processors of the parallel job. A brute-force way of doing I/O within a parallel application is to write one data output file per processor. However, this can generate thousands of files on a system the size of NERSC-3 (a management nightmare), and the file contents and configuration become job-size dependent. So, a job that executes on 128 processors produces 128 files. Before running the job with a different number of processors, the user must redistribute the contents of these files. A parallel I/O system should provide a means to write data efficiently into a single file in a manner that is independent of the number of processors employed in the simulation. This would greatly simplify file management and data analysis issues.

Categories of Parallel I/O

Parallel I/O solutions have been introduced in the past five years. They can be grouped roughly into three different categories based on the level of the software infrastructure they inhabit (these categories are not exclusive). The categories are:

1. **Application Level:** Here, a library/API has been produced to orchestrate I/O across the nodes of an MPP for each individual application. Users of these APIs must program specifically for these APIs in order to take advantage of them. Many Grid-related efforts are taking this approach in order to manage I/O over the wide area network as an issue distinct from HPC-center-level I/O issues. Examples include MPI-IO/ROMIO, PANDA, PHDF5, HDF4-EOS, and Globus-IO.

2. **Filesystem Level:** Much like NFS, the operating system is able to mount a specialized filesystem, so rather than managing the parallel I/O requirements for each individual application, the system is managing the parallel I/O requests for all applications running across the entire system. In many cases, the application-level APIs are still employed to provide hints and scheduling information to the filesystem in order to improve performance (e.g., MPI-IO implementations that sit on top of IBM’s GPFS and the T3E’s GigaRing I/O systems). Examples include IBM GFS/OpenGFS, GPFS, Intermezzo, PPFS-2, Berkeley TARDIS, and various SANs.

3. **Block/Device-Level:** Each node of a parallel system essentially sees a big multiport disk drive. The internal architecture of this disk drive may well be a fabric of interconnected I/O devices, but from the standpoint of the operating system, it primarily sees a block I/O device. So file integrity and other such semantics are maintained at a block level rather than filesystem level (although some significant interaction and design at the filesystem level is required to create a viable solution). In many of these cases, custom hardware is involved in the implementation of the architecture. Examples include GPFS and various SANs.

Filesystems and device-level filesystems are typically investigated for non-production supercomputing systems. Their tuning, testing, and installation can cause extreme disruptions to production system operations. Therefore, in order to participate in the evaluation of these technologies, HPC centers must either focus on application-level solutions or set aside a fraction of resources dedicated to testing lower-level solutions on an experimental, non-production basis.

The NERSC Center should participate in evaluation and hardening of these technologies because:

1. The research projects that tend to produce these libraries do not have the capital resources to invest in the large-scale machine architectures required for NERSC. In order to bring new and experimental systems to production quality, it requires some set-aside of relevant resources to perform the experiment. Having a centralized proving ground for these technologies is critical to getting them into real-world use.

2. The NERSC Center has the relevant users with real applications to test experimental parallel I/O technologies. The scientific operations that produce parallel I/O systems by necessity must limit themselves to one or two driving applications around which the architecture is constructed. This can lead to unrealistic expectations for general-purpose performance of these technologies.

Challenges of Current Architectures

There are several areas where current and even near-term implementations are limited in what they offer. These areas include:

Static Tuning Parameters. The mere existence of a parallel I/O API and/or parallel filesystems does not guarantee adequate performance. Considerable tuning is required for transfer sizes, data layouts, and synchronization operations. It is not feasible to automate these functions. There are many research efforts, for example PANDA, that allow the I/O subsystems to orchestrate and direct the I/O so as to maximize their bandwidth utilization. Still other efforts use prediction engines to reconfigure the I/O subsystem to match job characteristics. The ARIMA system under PPFS-2 watches file access patterns in running jobs and automatically adjusts the tunable parameters of the parallel filesystem to maximize performance as the job is running.

NERSC and other HPC centers can do a lot to address this problem. To first order, an HPC center can characterize the performance of its various parallel I/O solutions so as to help users accelerate their performance tuning efforts. HPC centers also need to provide access to libraries that automate performance tuning of I/O and to encourage vendors (such as IBM) to incorporate such technologies into their production systems.

Vendor/OS Specific. One of the great successes of NFS is the ability to cross-mount I/O devices on machines from totally different vendors and different UNIX implementations. Solutions so far for parallel I/O are generally locked into a particular vendor’s solution and have extremely limited interoperability. There are some SAN vendors who offer shared I/O solutions that work across platforms, but these are limited in scale as well as openness.

Lack of Market. The market does not have incentive to support the special needs of HPC applications. So solutions tend to be oriented towards sharing many distinct smaller files and optimizing random access read performance (e.g., SANs for sharing documents in a workgroup or for sharing large databases). HPC simulations tend to produce massive data outputs in a short period of time.

Levels of Access

Parallel I/O systems target three levels of access:

1. **Single Machine/Homogenous (e.g., GFS):** Vendors of more integrated MPPs such as the Cray T3E and IBM SP offer shared/parallel filesystems as part of the delivered machine. However, in the realm of Linux clusters, these solutions are still very experimental and require much more work to harden them. Furthermore, the parallel systems on the vendor systems could be advanced greatly if they provided more adaptive/predictive performance tuning as well as higher-level APIs for incorporating efficient parallel I/O into applications.
2. **Multi-Machine (e.g., Computing-Center SAN):** While many SAN solutions on the market offer scalable performance by attaching additional storage servers to the fabric, there are not many that work across the sorts of system architectures that are critical to the NERSC Center (i.e., Cray, HPSS, and IBM SP systems).
3. **Global/Multisite (e.g., Data Grids):** Data Grids must offer scalable performance, but it is not clear that they require the level of integration of a parallel I/O system. It is important that they interoperate efficiently with any parallel I/O system and that any SAN or parallel I/O solution provide scalable/efficient data migration and replication across multiple HPC sites. We cannot concentrate exclusively on optimizing performance within the HPC center to the exclusion of wide-area, data-intensive computing issues.

Currently, NERSC has distinct systems (SP2, HPSS, various Linux clusters, the T3E, and SGI visualization server) that each individually offer a working parallel I/O system. Scp or HSI is used to explicitly move data between these machines, creating a lot of redundant data storage. For instance, data on the scratch disk of the SP2 also resides on HPSS for archival storage, and an additional copy might be on the visualization server for data analysis.

From a standpoint of efficiency and convenience, it would be much better if all of these systems could share a single I/O resource such that we have 100 TB of storage that is universally visible to all of our supercomputers and tertiary storage systems within NERSC (a storage-area network). More importantly, such a SAN must provide full parallel I/O scalability such that writing to this shared filesystem from all nodes of the SP2 will not incur a large performance penalty in comparison to writing to the SP2’s internal GFS parallel filesystem. It is generally agreed that the architecture of such a system will be a fabric of interconnected I/O devices, but it is not yet clear when, how, or even if such a hardware/software solution that works across the entire Center will be available if NERSC simply waits for commercial market pressures to generate it.

3.5.5 Enabling a Unified Science Environment: The Integration of High-Capability Computing and Science

The DOE science community is comprised of interdependent components which currently lack the necessary infrastructure to provide an efficiently integrated science environment. Enabling such an environment would allow researchers greater access to an ever growing collection of these geographically dispersed components, such as supercomputing and large-scale data storage services and facilities, experimental instrumentation, and even access to the researchers themselves.

With the advent of Grid technologies, there is a potential infrastructure for integrating all these components into a global framework of human and technological resources. The opportunity exists to provide a standard large-scale science environment, a *Unified Science Environment*, that ties together all these components of scientific research.

The Unified Science Environment (USE) is the integration of computation, storage, theory, and experimentation into a tightly knit environment adapted to the processes of modern science. The core of the USE will be constructed using NERSC's unique supercomputing and large-scale data storage facilities and integrated into the DOE Science Grid with Grid middleware (Figure 3-3).

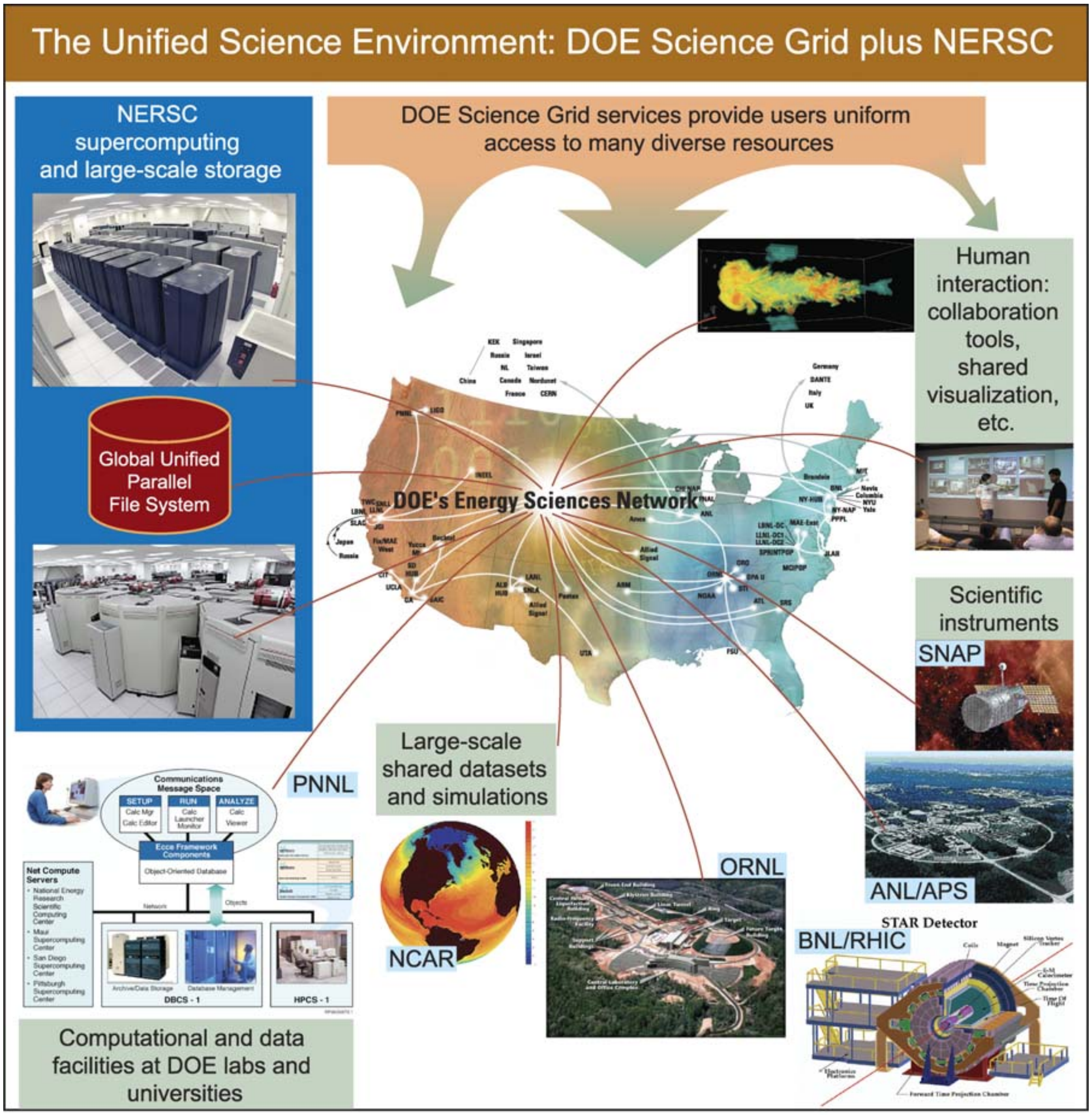

The Unified Science Environment: DOE Science Grid plus NERSC
NERSC supercomputing and large-scale storage
Global Unified Parallel File System
DOE Science Grid services provide users uniform access to many diverse resources
Human interaction: collaboration tools, shared visualization, etc.
Scientific instruments
SNAP
DOE's Energy Sciences Network
KEK Singapore
Russia Israel
NL Taiwan
China Canada Nordunet
France CERN
Germany
DANTE
Italy
UK
Japan
Russia
PNNL LIGO
INEEL
LBNL LLNL SLAC JGI
Yucca Mt
Bechtel
SD HUB
CIT UCLA CA SAIC
ALB HUB LANL SNLA Allied Signal
Pantex
UTA
CHI-NAP FNAL ANL
Ames
Allied Signal
ARM
ORNL ORO OPA U STI NOAA
ATL SRS FSU
MIT Brandeis BNL NY-HUB NY-NAP PPPL
Nevis Columbia NYU Yale
LBNL-DC MAE-East LLNL-DC1 LLNL-DC2 SPRINTPOP MCIPOP JLAB
Large-scale shared datasets and simulations
NCAR
ORNL
ANL/APS
BNL/RHIC
STAR Detector
Coils
Magnet
Silicon Vertex Tracker
E-M Calorimeter
Time Projection Chamber
Time Of Flight
Electronics Platforms
Forward Time Projection Chamber
PNNL
Communications Message Space
SETUP Calc Mgr Calc Editor
RUN Calc Launcher Monitor
ANALYZE Calc Viewer
Ecce Framework Components
Object-Oriented Database
Network
Objects
Net Compute Servers
• National Energy Research Scientific Computing Center
• Maui Supercomputing Center
• San Diego Supercomputing Center
• Pittsburgh Supercomputing Center
Archive/Data Storage
Database Management
DBCS - 1
HPCS - 1
Computational and data facilities at DOE labs and universities

Figure 3-3. In the future, science will depend on the interaction and interoperation of simulation (computing), data (large-scale archives), scientific instruments, and collaborators at many different institutions. Uniform access, large-scale distributed system construction tools, security, and coupling NERSC to DOE Office of Science's other facilities, will produce a Unified Science Environment.

USE will bring together the resources required to create and sustain distributed application environments. It will provide persistent infrastructure, high-end facilities such as NERSC, the DOE Science Grid (which includes DOE and non-DOE institutions and facilities), and the Energy Sciences Network (ESnet). The overall goal of USE is to make routine the solving of very large-scale, compute-intensive, data-intensive, and/or collaboration-intensive science and engineering problems in widely distributed, collaborative environments.

Overall Requirements

USE must support these applications and processes in a way that has the following attributes:

A. Integrated and coupled simulation, experiment computational analysis, and experiment control.

B. Seamless and efficient access to large and highly distributed data sets.

C. Automated and distributed management of continuous and complex computing and data handling processes.

D. Creation, use, and management of collaborative tools and environments.

E. Interactive access to supercomputing resources.

F. Infrastructure stability and widespread adoption of a single USE approach.

These attributes give rise to three categories of functional requirements that are summarized below:

Collaboration and Workflow:

- Scientific workflow management systems that are easily adaptable to existing and disparate data processing models and are sufficiently flexible to allow ad hoc reconfiguration in response to research needs (from A).
- Collaborative tools and frameworks that allow close interaction between physically and, to some extent, programmatically separate communities cooperating on common, computationally intensive research topics (from A).
- Software tools for coupling supercomputers and the Grid environment with scientific instruments (from A).
- Collaborative frameworks that foster and support day-to-day operations of distributed, multi-institutional collaborations. As operational activities for science experiments take on a 24 × 7 character, these tools must support task description and delegation within a continuous, around the globe work day (from C and D).

Data Grid:

- Tools that support predictive data set staging in collaboration with compute schedulers (from B).
- Active data directories and catalogs that will support large-scale data archives distributed over multiple remote storage and computing facilities (from B).
- Access to optimized, high speed data transfer mechanisms, such as parallelized, Grid-enabled FTP or data transfer tools that stripe storage accesses across multiple sites (from B).

Compute Grid:

- A software environment that will support remote execution of analysis and simulation codes at a remote data storage location (i.e., “transport the program to the data”) when optimal (from B and C).
- Deployment of secure, distributed mechanisms that allow scientific collaborations to monitor, control, and alter large-scale computing processes that are essentially in continuous operation — over 10 or more years (from C).
- Computing and storage system scheduling tools need to allow both prioritized and preemptive execution of designated tasks, where this is supported by the underlying resources (from E).

USE Implementation Plan

NERSC has prioritized the identification and deployment of USE services and capabilities so that they can provide early impact coupled with continually increased capability over the next five years.

In analyzing the Grid requirements, there is dependency ordering that becomes apparent. The collaboration and workflow requirements depend on the underlying computation and data Grids. The computational Grid relies on the data Grid to provide the datasets used in computation. As a consequence, our roadmap reflects this dependency.

The overall development roadmap is:

FY2002

- Data Grid pre-production activities
- Track computational grid, collaboration, and workflow development

FY2003

- Focus on data Grid production rollout
- Pre-production compute Grid
- Track collaboration and workflow development

FY2004

- Focus on compute Grid production rollout
- Pre-production collaboration and workflow

FY2005

- Focus on collaboration and workflow production rollout

FY2006

- All major USE components on NERSC production systems

FY02 Grid Development and Testing

FY02 will be devoted to the development and testing of USE components and Grid services on NERSC non-production systems. By the end of FY02, NERSC will have completed testing of the major USE components and started partial production on the HPSS. Activities will include:

- Identify and cultivate the NERSC human and computing resources to be incorporated into the Grid.
- Identify the systems administrators for all systems that will be involved (computing and storage).
- Identify systems that will provide production Grid services.

- Establish a NERSC Grid Production Working Group (“ProGrid WG”) that involves the system administrators, cybersecurity, networking, and User Services staff.
- Understand Grid applications development.
- Build Globus 2.0 on the following NERSC test systems:
  - Probe — the HPSS test system
  - Dev2 — the NERSC 3 IBM AIX development system
  - Alvarez — a Linux cluster
  - PDSF — a Linux cluster.
- Use PKI authentication and Globus or other temporary certificates for this test environment.
- Evaluate for production-level quality and security.
- In conjunction with ESnet, build and test the security infrastructure, which will include a registration authority (RA) and, if appropriate, a certificate authority (CA), and identify staffing needed to maintain it.
  - NERSC will generate PKI certificates for staff.
  - The RA will not be integrated with NIM (at this time).
  - Integrate RA into DOE Science Grid PKI infrastructure.
  - Establish the conventions for the Globus mapfile.
  - Map user Grid identities to system UIDs — this is the basic authorization mechanism for each individual platform.
  - Establish the connection between user accounts on individual platforms and requests for Globus access on those systems. (Initially this will consist of a non-intrusive mechanism such as email to the responsible sys admin to modify the mapfile.)
- Build and test basic Globus services such as GIS/MDS.
- Build and test basic data mover services.
  - Single stream (non-parallel) gsi-ftp clients and servers.
  - Port to production HPSS and PDSF.
  - Build and test Grid HPSS servers and clients.
- Investigate intrusion detection system and firewall issues.
  - Globus can be configured to use a restricted range of ports, but it still needs a number of open ports that is proportional to the number of expected simultaneous, running Globus jobs, and these must be negotiated with system managers.
  - GIS/MDS also needs some open ports.
  - Investigate how the BRO intrusion detection system used at NERSC must be updated to understand Grid services.
- Track developments in Grid job tracking and monitoring, as well as collaboration and workflow.
- Identify early users and have the Globus application specialists assist them in getting applications running on the Grid.

FY03 Initial Production Capabilities

Basic Grid services will be available on some set of NERSC production resources which will interoperate with the DOE Science Grid. Client-side services will be available for UNIX workstations to submit jobs to NERSC and other DOE Science Grid resources for testing. Activities include:

- Build, test and place into production data Grid servers.
  - GridFTP service on Alvarez, PDSF, Seaborg
  - Web-based portals to Grid
  - MyProxy and related Grid authentication support
  - NIM integration
- Apply Grid security infrastructure for single sign-on
- Test and operate pre-production compute Grid services
  - GateKeeper
  - Condor-G
  - Individual cluster batch scheduler integration
    - LoadLeveller on SP
    - LSF or PBS for Linux
    - Integration with NERSC IV scheduler
- Track developments in collaboration and workflow technologies
- Put up one of the various Web portals for Grid resource monitoring.
- Maintain and update existing production data Grid services

FY04 Expanding Production Capabilities

In FY04 we will see a maturation of the production Grid support to include support for the computational Grid, making NERSC cluster resources available for global job scheduling. Initial testing for collaboration and workflow services built on top of data Grid and compute Grid services begins.

- Roll out production computational Grid services
- Finish NIM integration
- Test and operate pre-production collaboration and workflow services
- Maintain and update existing data Grid services

FY05 Complete USE Production Rollout

FY05 results in the rollout of the final major components of USE, collaboration and workflow services built on top of the data Grid and compute Grid foundations already in place.

- Roll out collaboration and workflow services
- Maintain and update existing data Grid and compute Grid services

FY06 Unified Science Environment Complete

In FY06, all major components of the Unified Science Environment will be in production, supporting global scientific collaborations with NERSC’s unique resources.

3.5.6 HPSS Development

NERSC has been an HPSS development site since 1996. NERSC contributes to core server development for new technology devices. This allows NERSC to incorporate new technology into the archive as it becomes available. As a development site, NERSC contributes to the technical direction of HPSS as a member of the technical committee and also contributes to long-range HPSS planning as a member of the HPSS executive committee. HPSS is deployed at sites that have very aggressive requirements for high performance data transfer and large bulk storage, and we benefit as a result.

Over the past three years we have developed physical volume repository (PVR) servers for new tape technology drives as they became available: The Eagle tape drive technology from StorageTek and linear tape-open (LTO) tape technology from IBM increase possible HPSS archival capacities from 1 GB/ cartridge to 100 GB/cartridge and also add support for IBM’s new LTO library. Over the past three years, HPSS development has incorporated numerous performance and functionality improvements such as support for parallel transfers, ability to stream and stripe data over multiple devices, support for very large devices, and the ability to link multiple storage systems together over the wide-area network.

In the next few years, HPSS will incorporate support for SAN, better and faster database structure and query support, support for new IEEE parallel transfer protocols, and support for high performance networks and devices as they become available.

NERSC will be working to develop transparent access to shared high performance filesystems, support for applications and development tools that enable data sharing, techniques that use the storage system to replicate data, and database recovery utilities and tools.

3.5.7 High Performance Cyber Security Development and Improvement

Continued development, testing, and deployment of new and existing computer security technologies is essential in maintaining an open yet secure computing environment at NERSC. NERSC will have to adapt its computer security tools and techniques to reflect the changing computer security landscape and the increasing capabilities of NERSC clients and the NERSC center. This may include the deployment of new host-level scanning tools, new intrusion detection systems, further filtering of network traffic, and deployment of more secure applications.

As Grid-related applications are made available for production use, NERSC will migrate toward an environment where certificates are the primary means for user authentication. NERSC will need to adapt its existing authentication and authorization mechanisms to fit within this framework. This will include modifications to the NIM system and to our high performance platforms and resources to allow certificate-based authorization. NERSC will have to work closely with the DOE Science Grid community to ensure cross-site interoperability. The deployment of Grid-related services requires an evaluation of their impact across all of NERSC’s computer security levels. This may result in modifications of the various tools in use at NERSC and also deployment of new tools.

Specific security improvements during this period will include:

1. **2.4 Gbit/s reactive monitoring system for intrusion detection (BRO).** The current computer security model for NERSC is proactive monitoring of traffic and reacting to potential threats and intrusions. This has been accomplished by using LBNL’s BRO reactive monitoring system, which monitors network traffic inbound and outbound from NERSC and coordinates access control changes

with network routers. Increases in network bandwidth dictate a subsequent increase in the capability of BRO's monitoring. This will require both hardware and software changes to keep up with OC-48 network speeds. Network speeds greater than OC-48 will require research and development efforts outside the NERSC program.

2. **Sitewide Public Key Infrastructure (PKI) deployment.** The DOE is moving toward using PKI for user authentication and service authorization. Although it is still unclear what direction DOE will eventually take with PKI, it is clear that DOE sees PKI as the future method for authentication and authorization within DOE facilities. A NERSC PKI will give our users the capability to securely authenticate and access NERSC services. The challenge in developing a NERSC PKI will be to ensure compatibility with the various DOE PKI efforts.
3. **NERSC production systems capable of supporting PKI.** A challenge in deploying a sitewide PKI is to ensure that all of NERSC production systems are capable of handling PKI. Although DOE has funded several PKI efforts, NERSC will still need to ensure that all production-system platforms support PKI. This may require porting and integrating nonproduction level software and working closely with PKI developers within LBNL and the rest of the PKI community.

3.5.8 Portals

The term *portal* is used to refer to a Web-based collection of resources. In the most generic sense, portal resources present a collection of information and services that are useful to members of a specific community. Etrade.com and Amazon.com are familiar examples of portals. Both provide a vast array of services and information to a large number of people. These services and resources are not intrinsically part of the Web browser, but are provided by the portal developers. Portals have grown in popularity in the past few years because they are exceptionally useful ways to empower users with access to information and other resources using a simple interface. They also characterized by centralized administration and wide applicability to diverse domains.

In the generic implementation, users interact with a portal using only a Web browser. On the Web server side, matters can quickly become much more complex, and herein lies the real power of portals. The server side of portals consists of a Web server that has been augmented with additional services to enable it to provide specialized services to users. Some portals, such as Etrade and Amazon, focus on providing a high-volume transaction-processing infrastructure. Because of the high degree of flexibility in configuring Web servers, it is possible to configure portal server software to provide essentially any kind of service. It is the responsibility of portal developers to "glue together" the back-end services in such a way that they may be accessed using a Web browser.

The efforts related to the development and deployment of Grids, which are described elsewhere in this document, have several key benefits to users. These benefits include single sign-on authentication, and resource location and access. From a high-level view, these services are intended to allow users of resources on Grids to more quickly and efficiently access those resources, regardless of their geographic location or site-specific access policy. Several efforts are already under way that provide access to Grid-based resources using a Web portal model. These efforts focus on creating *Grid portals.*

Grid Portals

A Grid portal is one that uses special tools on the Web server side to implement Grid features, such as single sign-on authentication, and resource discovery and management. The Grid portals may be narrowly focused on a particular application or domain, or may be completely general. Grid portals will provide capabilities that:

- allow you to monitor job status from any location with a browser
- enable you to submit and delete jobs
- permit you to browse services that are available at different sites
- provide a convenient interface for examining the metadata in these services (fileset catalogs, host information, queue status, etc.)
- provide a simple interface to the collaboration tools being envisioned for the Grid.

Portals are typically implemented as a base framework that handles authentication, authorization, miscellaneous state information, and presentation. The code that talks to back-end services is usually implemented using some set of components that plug into the framework; as new services are integrated, the necessary components are added. Integration of the Grid tools that transform a Web server into a Grid portal is conceptually a straightforward activity, but fraught with pitfalls. At the time of this writing, early portal development activities at NERSC have required a determined and cooperative effort involving multiple groups in order to bring online a Grid portal with minimal capabilities. We expect that this process will become more streamlined and easy to perform in the years ahead: the direction of current NERSC portal development is to create a framework that allows rapid and efficient construction of domain-specific portals. The eventual goal is to create a portal framework that would allow a member of User Services to configure a portal on the fly for new collaborations, taking into account their workflows and other usage-related factors.

Visualization Portals

NERSC is focusing on the use of Grid and portal technology for a number of related activities. One is to provide the means to transition research prototype visualization software into the production environment. Many new visualization software prototypes are composed of several software components, each of which runs on a different platform.

In order to improve the quality of information available to NERSC users, a rich online environment contains a substantial body of documentation germane to the visualization products and services. The products consist of a diverse array of visualization software that is installed on NERSC production machines. Services consist of close collaborations that involve focused application of visualization know-how, use of software, and input from users to create cutting-edge visualizations. This activity is representative of traditional website construction, but accurately reflects that portals and generic websites do not differ significantly from a user’s perspective.

Looking further into the future, we envision use of portal technology enabled with Grid services to provide a new class of tools for production visualization. One of the primary difficulties we hear from users is the difficulty in using multiple complex software tools to perform a relatively simple task, such as creating an MPEG movie from a dataset. A portal can simplify this process for a user. In an ideal case, the user would be able to select a dataset, a visualization process or procedure, click a single “go” button, and then have an MPEG created on their behalf by software components controlled by the portal back end.

While NERSC has a number of visualization projects planned that make use of portals, a number of other possibilities may prove useful to the NERSC user community as well as for the NERSC Center itself. These projects provide the users with increased and simplified access to center resources, while portals provide a way for the center to operate more efficiently because they encourage centralized control and maintenance.

Building and Maintaining Portals

The technologies that go into making a portal are finite in number, yet may be combined and populated with content to yield an extremely large number of permutations. For example, one group may desire a portal that focuses on documentation. This can be done with only a generic Web server and some disk space. Another group may wish to deploy a portal that uses Grid service that provide the means for a user to select data sources and computational resources to perform a specific task, such as visualization. A portal of this type requires a Web server enabled with Java server pages along with appropriate hooks into Grid services.

We can identify elements that are common to both types of portals, and we can borrow from concepts used by commercial Internet service providers to provide context for discussion. In the commercial ISP world, new Web hosting users are offered a menu of services for their new Web-hosted accounts. The ISP makes appropriate configurations to a virtualized Web server that provide these features. From the customer’s perspective, there is no need for concern about administration of the Web server, nor need for concern about configuring and operations management of a physical server machine.

In our portal example, a NERSC should be able to quickly spin up new portal frameworks that can then be populated with content relevant to NERSC services and clients. The customer, in this case, might be NERSC staff or NERSC users who have need for a portal. Some portal technologies, such as Web servers, are today very stable. Other technologies, such as Grid services, are still immature. The challenge for the NERSC portal providers will be to hide the complexity of integrating diverse back-end technologies from the portal customers. The goal should be to create a portal kit that can be quickly deployed in a fast and reproducible fashion. NERSC will need to track changes in emerging and changing technologies, such as Grid services, so that the most advanced capabilities are available to those who require them.

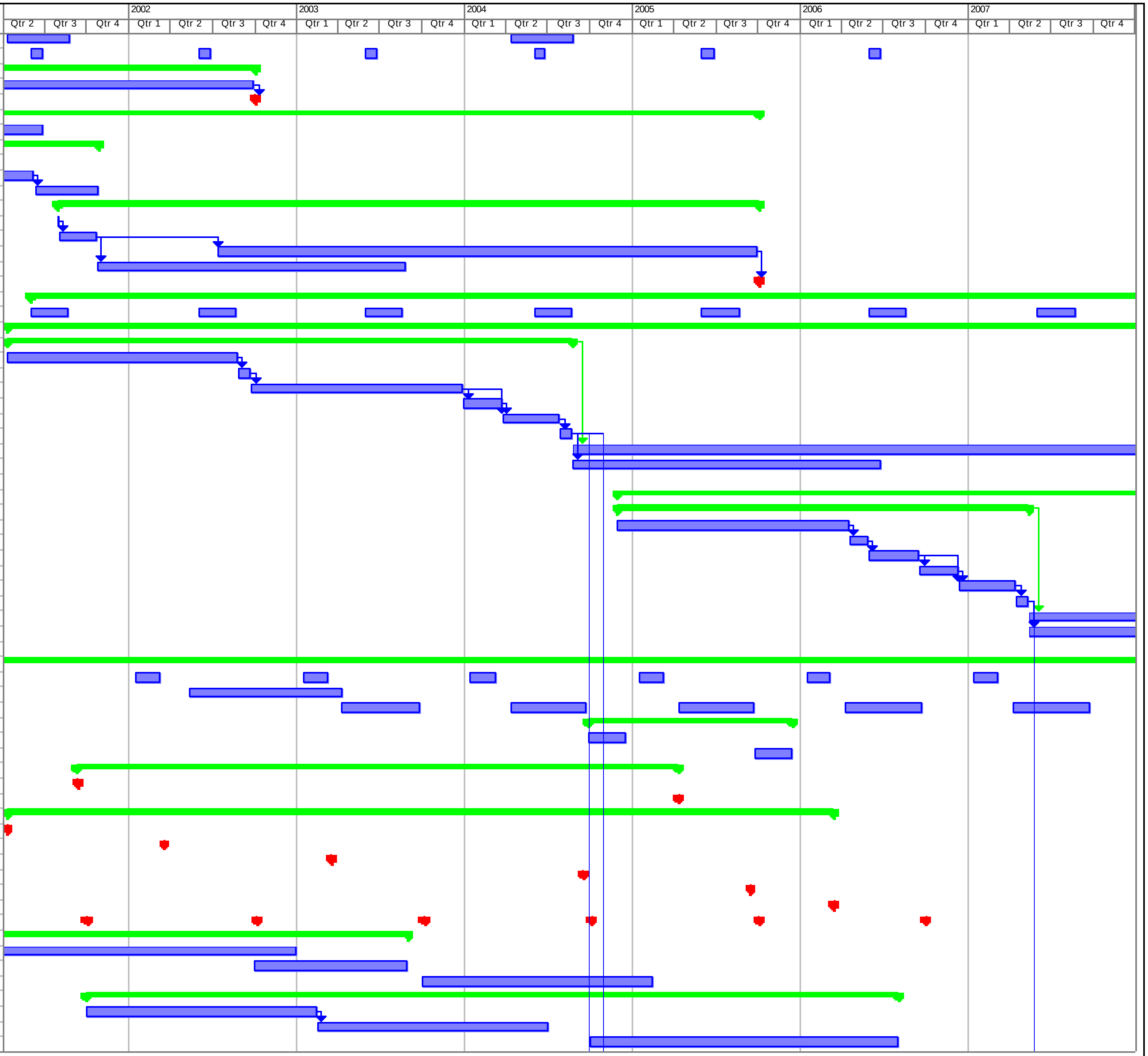
4 SUMMARY

This Implementation Plan outlines the strategy and approaches NERSC will use to achieve the goals proposed in the NERSC Strategic Proposal approved by DOE in November 2001. The plan depends on appropriate levels of funding and support as outlined in the Strategic Proposal.

The result of this plan when it is fully accomplished should dramatically enhance the NERSC Center and enable the DOE Office of Science researchers to be highly effective in pursuing their scientific goals.

5 IMPLEMENTATION SCHEDULE

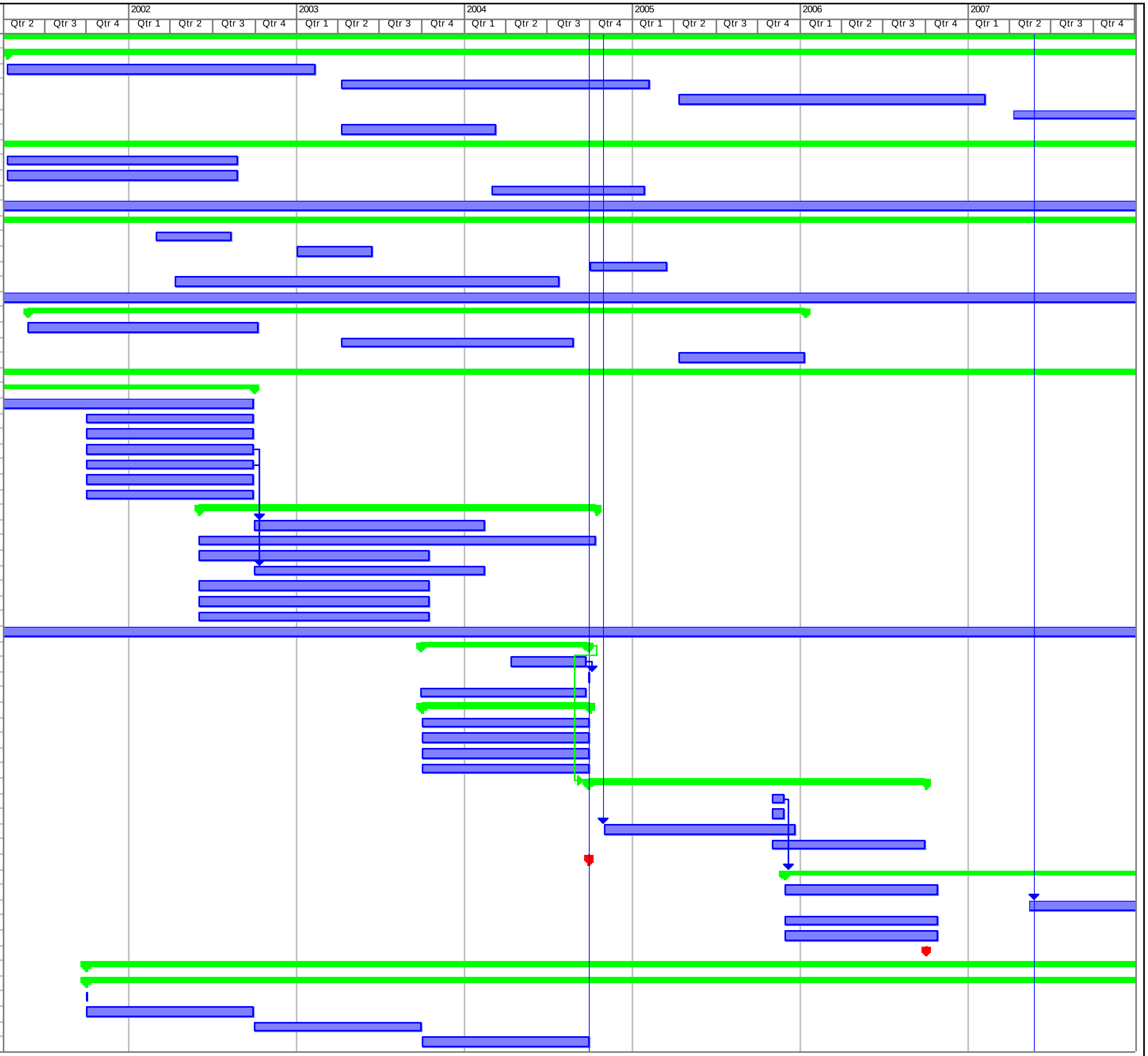
ID	Task Name
1	**NUG Green Book**
4	**NERSC Annual Goals**
11	**NERSC-2**
12	Production Usage - SSP-1 30 Gflop/s
13	Decommission
14	**NERSC-3**
15	Phase 1 Production Usage - SSP-1 33 Gflop/s
16	**Phase 2a**
17	System Build, Factory Test, Ship
18	Acceptance Test
19	Production Service - SSP - 239 Gflop/s
20	**Phase 2b**
21	Software Delivery
22	Acceptance Test
23	Production Service - SSP-1 - 270 Gflop/s
24	Improve Effectiveness
25	Decommission
26	**PDSF**
27	**Yearly Upgrade - double system**
36	**NERSC-4**
37	**System Procurement**
38	15 Step Procurement Process
39	Award
40	Build, Factory Test, Ship
41	Site Preparation
42	Installation and Acceptance Testing
43	Prepare for production
44	Production Services - SSP-2 - 750 Gflop/s
45	Improve Effectiveness
46	Decommissioning
47	**NERSC-5**
48	**System Procurement**
49	15 Step Procurement Process
50	Award
51	Build, Factory Test, Ship
52	Site Preparation
53	Installation and Acceptance Testing
54	Prepare for production
55	Production Services - SSP-3 2,000 Gflop/s
56	Improve Effectiveness
57	Decommissioning
58	**Mass Storage Upgrades**
59	**HPSS Upgrades**
68	Web Portal for HPSS files
69	**Media device reader (e.g. new tape drives)**
76	**High Density Cartridges**
77	40 GB/cart-20 mb/sec-FC
78	120 GB/cart-40 mb/s-FC
79	**Cache Disk**
80	20 TB on-line
81	100 TB on-line
82	**Maximum Capacity Storage**
83	1.3 Petabyte
84	3 Petabytes
85	6 Petabytes
86	9 Petabytes
87	10 Petabytes
88	15 Petabytes
89	**HPSS Releases**
96	**Wide Area Movers**
97	Develop and Test in Probe
98	Deploy on Production HPSS
99	Explore Fibre Channel third Party transfer
100	**Hierarchical Resource Managers**
101	Implement Prototype on Probe
102	Integrate HRMs into production HPSS
103	Evaluate and Deploy SDM-ITIC technology as appropriate



Project: NERSC- Implementation Plan F
Date: Thu 8/15/02

Task	🟦	Milestone	◆	Rolled Up Task	🟦	Rolled Up Progress	▬	External Tasks	▭	Group By Summary	▬
Progress	▬	Summary	🟩	Rolled Up Milestone	◇	Split	┈	Project Summary	▬		

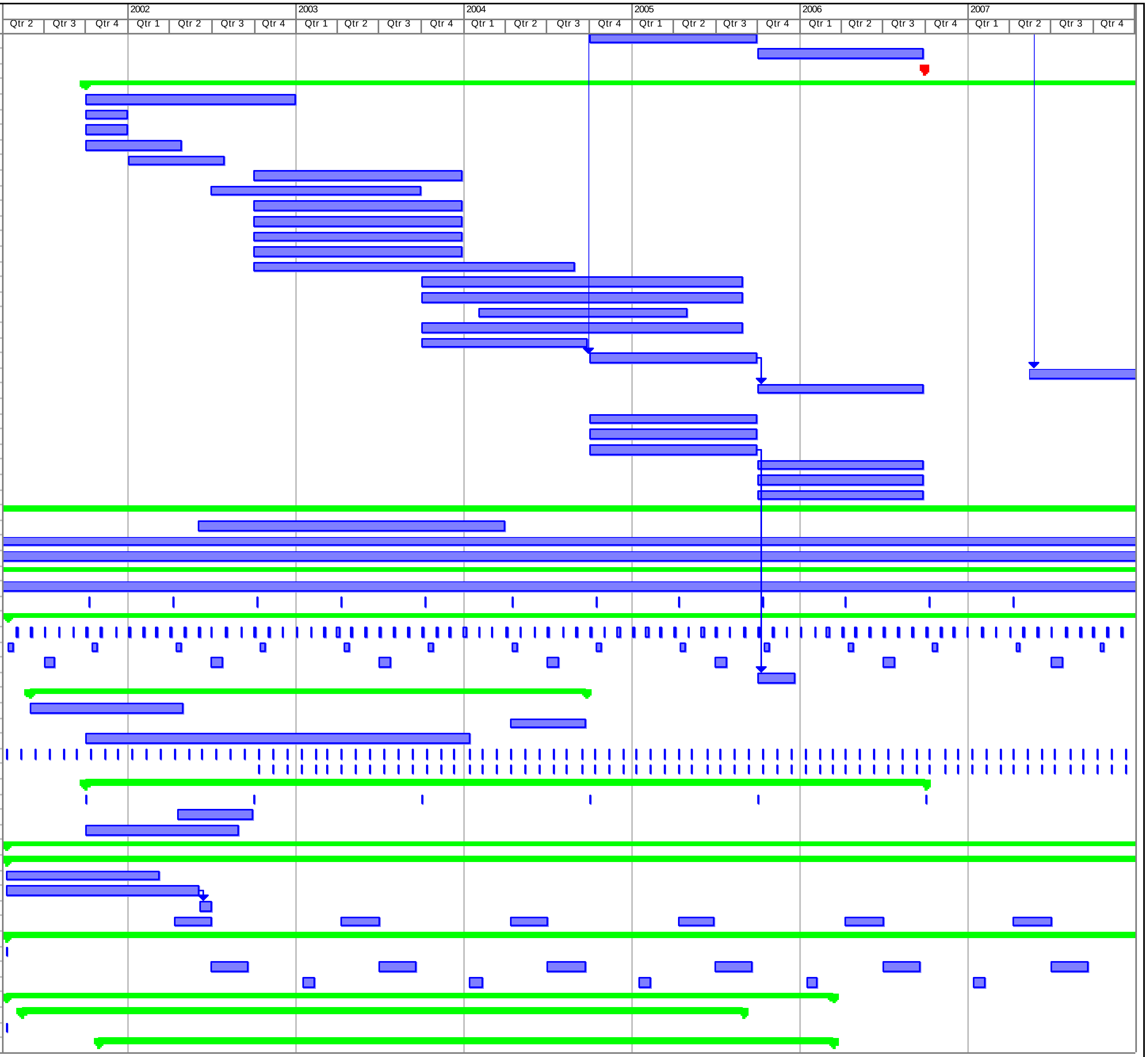
ID	Task Name
104	**Communications**
105	**Production Wide Area Network**
106	Upgrade to OC 12
107	Upgrade to OC 48
108	Upgrade to OC 192
109	Upgrade to OC 768
110	Quality of Service
111	**Local Area Network**
112	Replace HiPPI with GigE - Jumbo
113	Replace FDDI with Gig E
114	10 Gigabit Ethernet
115	Network Experiments and Projects
116	**Security**
117	Upgrade and Implement BRO on OC 48
118	Upgrade and Implement BRO on OC 192
119	Upgrade and Implement BRO on OC 768
120	BRO Automatic monitoring and analysis
121	Security Monitoring and Improvements
122	**Testbed System Evaluations**
123	Alverez Cluster
124	Next Testbed
125	Next Testbed
126	**Global Unified Parallel File Systems**
127	**Technology Evaluation**
128	Early Assessment
129	Fibre Channel Testbed
130	Arrange Collaborations - ASCI SPGFS, HENP,…
131	Test Shared disk file systems
132	Linux HPSS mover with gfs file system
133	Sharing files on Fibre Channel SAN
134	Evaluate iSCSI and Inifiniband
135	**Testbed Implementation**
136	Install LINUX based HPSS
137	HSM Interface to HPSS via DMAPI
138	Wide Area testing
139	Acquire advanced SAN and storage devices
140	System integrate with Linux Cluster (e.g. PDSF)
141	Explore file system bridging
142	Evaluate and document shared disk file system
143	AFS Services
144	**Detailed Design and Assessment**
145	Detailed Assessment if preferred GUPFS file system
146	Detailed Assessment report
147	USE Integration
148	**Phase 1 Implementation**
149	Select and implement preferred prototype GUPFS system
150	HSM Integration with shared disk file system
151	Beta Test SPGFS on test bed system
152	USE integrations with shared disk file systems
153	**Phase 2 Implementation**
154	Final selection for implementation
155	Building disk storage infrastructure
156	Integrate with NERSC 4
157	Integrate with auxiliary systems
158	Production
159	**Enter Production**
160	Complete integration with all production platforms
161	NERSC 5 Integration
162	Transition to New GUPFS compatible archive system
163	GUPFS access via Grid
164	Production
165	**Unified Scientific Environment**
166	**Application / User Milestones**
167	project begin
168	Initial User Capabilities: Grid job and data management
169	Pre-production USE: NERSC in Grid and integration w/ user resources
170	Production Phase I: NERSC supports large-scale science collaborations



Project: NERSC Implementation Plan F
Date: Thu 8/15/02

Task	🟦	Milestone	◆	Rolled Up Task	🟦	Rolled Up Progress	▬	External Tasks	▭	Group By Summary	▬
Progress	▬	Summary	🟩	Rolled Up Milestone	◇	Split	┈	Project Summary	▬		

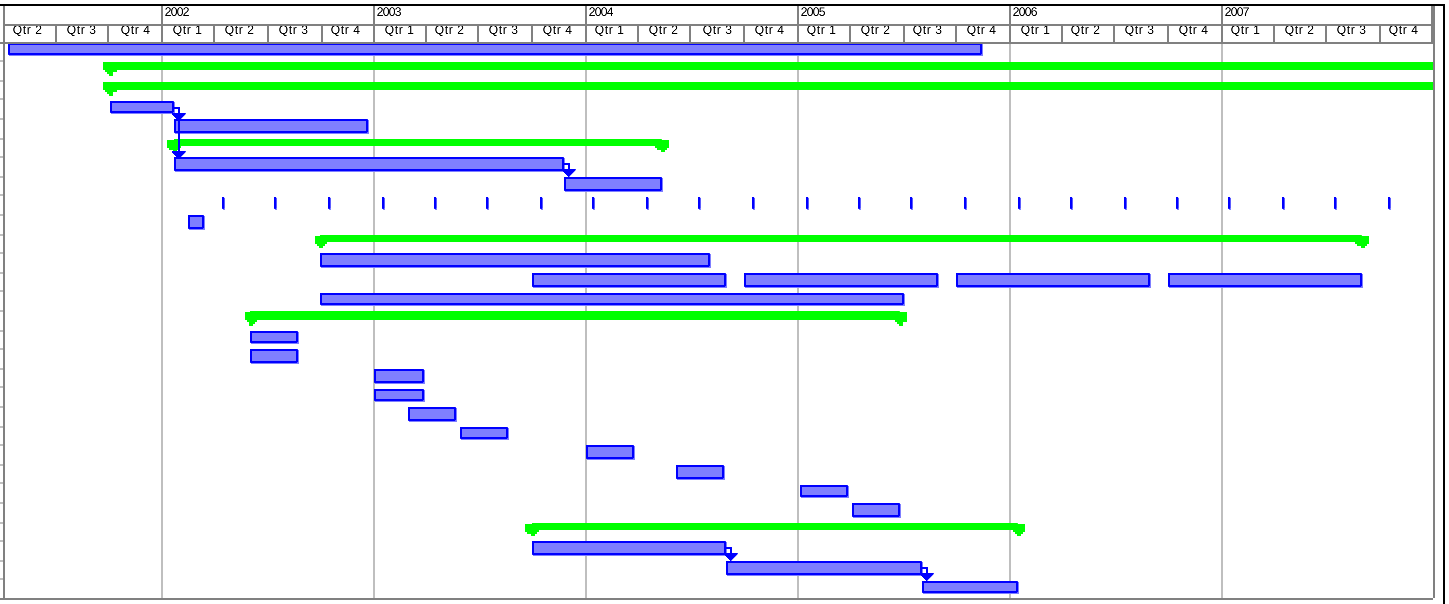
ID	Task Name
171	Production, Phase II: NERSC support for complex science workflow
172	Production, Phase III: Integration with instrument based experiments
173	USE project end
174	**Deployment and Development**
175	Grid Information Services
176	Certification Authority
177	Deploy Globus
178	Security Infrastructure
179	Integrate with Science Grid
180	Grid Tertiary Storage
181	Globus on all NERSC systems
182	User Services and Documentation
183	NERSC User/Partner deploy package
184	Web Portal Access to USE
185	Condor-G Job Manager
186	Distributed Authoring and Versioning
187	Collaborative Grid Services
188	Virtual Organization Support
189	Global Data Catalogue Support
190	Global Event Services
191	Basic Workflow Management
192	Globus Service on NERSC 4
193	Globus Service on NERSC 5
194	Advance CPU Reservation on NERSC 4
195	Advance CPU Reservation on NERSC 5
196	Generalized Directory and Data Services
197	High-Speed Remote Visualization
198	Full Workflow Management Services
199	Quality of Service for HPSS
200	High Performance Global File system in Grid
201	Automated Data Replica Management
202	**Comprehensive Scientific Support**
203	Problem Tracking - New methods
204	24 by 7 Help Desk
205	24 by 7 monitoring and operations
206	**Consulting Services**
207	Ongoing
208	**Software Reviews**
221	**Web**
222	**Upgrade Content**
346	**Review Content**
368	**Upgrade Server**
380	Introduce Collaborative Consulting
381	**VAC-BF - Video and audio capture and broadcast facility**
382	VAC-BF Initial Implementation
383	VAC-BF Upgrade
384	Desktop Video Tools
385	**Monthly Training**
509	**Monthly Video Seminar**
615	**NIM**
616	**Allocation Turnover**
623	Phase 2
624	Deploy User Self Service Facilities
625	**Benchmarking and Performance Analysis**
626	**NERSC Application Performance Test suite**
627	Work Load Analysis
628	Develop Benchmark
629	Release benchmarks
630	**Reports/Updates**
641	**Effective System Performance Test**
642	Initial Development
643	**New version**
653	**Release Report**
663	**Development Activities**
664	**Cluster Interconnect Networks**
676	BLD - Release 2
677	**Checkpoint Restart for LINUX**



Project: NERSC- Implementation Plan F
Date: Thu 8/15/02

Task	▬	Milestone	◆	Rolled Up Task	▬	Rolled Up Progress	▬	External Tasks	▬
Progress	▬	Summary	▬	Rolled Up Milestone	◇	Split	┈	Project Summary	▬
								Group By Summary	▬

ID	Task Name
684	On-going Investigations
685	**Scientific Challenge Team Support**
686	**Software Engineering Infrastructure**
687	Establish code repositories
688	Methods for community sharing
689	**S/W Engineering Management Tools**
690	Evaluate, Acquire, Integrate New Tools
691	Training
692	**S/W Engineering Training Seminars**
737	Establish ETC Points of Contacts
738	**Extend Scientific Challenge Team Support**
739	SCT Support for Priority Teams
740	**Support one more SCT team**
745	Integrated support and collaboration
746	**Algorithms**
747	Efficient and scalable incomplete factorizations
748	Direct Sparse Matrix Solver - to shift invert eigensolvers
749	Benchmark Eigensolvers for SciDAC applications
750	Benchmark ILU preconditioners
751	Evaluate ordering strategies for ILU
752	Multiblock ILU codes
753	Multilevel ILU preconditioners in eigensolvers
754	Benchmark Multilevel ILU preconditioners in eigensolvers
755	Accuracy Control for J-D algorithms
756	Implement block variants of eignsolvers
757	**ISIC Tools**
758	Evaluate and Select
759	Test, diagnose and improve for NERSC systems
760	Deploy to NERSC clients



Project: NERSC- Implementation Plan F
Date: Thu 8/15/02

Task	🟦	Milestone	◆	Rolled Up Task	🟦	Rolled Up Progress	▬	External Tasks	▭	Group By Summary	▬
Progress	▬	Summary	🟩	Rolled Up Milestone	◇	Split	┄	Project Summary	▬		

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.

Ernest Orlando Lawrence Berkeley National Laboratory is an equal opportunity employer.