LBNL-44337

ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

Building the Teraflops/Petabytes Production Supercomputing Center

Horst D. Simon, William T.C. Kramer,
and Robert F. Lucas

**National Energy Research
Scientific Computing Center**

September 1999

Presented at the
EuroPar '99 Conference,
Toulouse, France,
August 31–September 3, 1999,
and to be published in
the Proceedings

DISCLAIMER

LBNL-44337

Building the Teraflops/Petabytes Production Supercomputing Center

Horst D. Simon, William T.C. Kramer, and Robert F. Lucas

National Energy Research Scientific Computing Division
Ernest Orlando Lawrence Berkeley National Laboratory
University of California
Berkeley, California 94720

September 1999

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

Building the Teraflops/Petabytes Production Supercomputing Center

Horst D. Simon, William T. C. Kramer, and Robert F. Lucas

National Energy Research Scientific Computing Center (NERSC), Mail Stop 50B-4230,
Lawrence Berkeley National Laboratory, Berkeley, CA 94720
{hdsimon, wtkramer, rflucas}@lbl.gov

Abstract. In just one decade, the 1990s, supercomputer centers have undergone two fundamental transitions which require rethinking their operation and their role in high performance computing. The first transition in the early to mid-1990s resulted from a technology change in high performance computing architecture. Highly parallel distributed memory machines built from commodity parts increased the operational complexity of the supercomputer center, and required the introduction of intellectual services as equally important components of the center. The second transition is happening in the late 1990s as centers are introducing loosely coupled clusters of SMPs as their premier high performance computing platforms, while dealing with an ever-increasing volume of data. In addition, increasing network bandwidth enables new modes of use of a supercomputer center, in particular, computational grid applications. In this paper we describe what steps NERSC is taking to address these issues and stay at the leading edge of supercomputing centers.

1 Introduction

In just one decade, the 1990s, supercomputer centers have undergone two fundamental transitions which require a rethinking of the basic tenets of their operation and their role in the high performance computing (HPC) world. The first transition in the early to mid 1990s was a result of a technology change in high performance computing architecture. The introduction of highly parallel distributed memory machines built from commodity parts increased the operational complexity of the supercomputer center, and required the introduction of intellectual services as equally important components of the center.

We have only recently completed this transition and developed the tools necessary to bring the revolution of the mid-1990s to a successful conclusion. Now three new developments are appearing which will again force us to step up to new challenges in supercomputer center management: (1) yet another change in the architecture of supercomputing platforms, (2) the increasing importance of managing large volumes of scientific data, and (3) the deployment of a new generation of high-speed wide-area

networks. After reviewing the two transitions in Section 2, in the remainder of this paper we will discuss these three technology developments and their likely impact on high performance computing. We will describe how the U.S. Department of Energy's National Energy Research Scientific Computing Center (NERSC) [1] is preparing itself to meet these challenges.

The first issue we are facing is another technology change in high performance computing systems. The next generation of supercomputers will be built from commodity components, in all likelihood from shared memory multiprocessor (SMP) systems. These cluster-of-SMP systems not only add an additional level of complexity for user applications development, but also lack most of the robust systems software that a high-quality, production-oriented center relies on. Furthermore, these systems also require an order of magnitude increase in available floor space and power consumption. NERSC is planning to install such a system during 1999, and we will report in Section 3 on our initial experiences.

The second issue we are facing is increasing rates of data generation both from computer simulations and from experiments. For example, the high-energy physics community is bringing new experiments on line that will generate data at rates exceeding 250 terabytes per year in 1999 and 1 petabyte per year in 2005. The bioinformatics community points out that the number of base pairs output by the various genome projects is growing faster than Moore's law. These projected data rates force us not only to reevaluate tertiary storage and bandwidth requirements, but also to develop new tools in scientific data management. NERSC has developed a data-intensive computing strategy to deal with these issues in a comprehensive way. We will describe this strategy and some of its technology elements in Section 4.

Lastly, we are facing incredible increases in the available bandwidth for wide-area networks. This enables new modes of using both compute and data resources at the centers. Computational steering and remote visualization and exploration of data will become commonplace. While many of these technologies are already at the prototype or advanced development stage, substantial work needs to be invested to provide these tools on a routine basis. These developments are often summarized under the term "grids." We will discuss NERSC's potential role in the data grid and how it may benefit our user community in Section 5.

2 The Two Technology Transitions of the 1990s

NERSC recently announced the successful completion of the NERSC-3 procurement, resulting in the acquisition of an IBM SP-3 with more than 3 Tflop/s peak performance. The machine will arrive at NERSC in two phases.

Phase I installation, scheduled to begin in June 1999, will consist of an RS/6000 SP with 304 of the two-CPU POWER3 SMP nodes that were recently announced by IBM. This system will be the first implementation of the POWER3 microprocessor, with two

processors per node. The 64-bit POWER3 can perform up to two billion operations per second and is more than twice as powerful as its predecessor. In all, Phase I will have 512 processors for computing, 256 gigabytes of memory, and 10 terabytes of disk storage for scientific computing. The other 48 nodes will be used for interactive, network, parallel file system, and other services. The system will have a peak performance of 410 Gflop/s, or 410 billion calculations per second.

Phase II, slated for installation no later than December 2000 (but likely much sooner), will consist of 152 16-CPU POWER3+ SMP nodes, utilizing an enhanced POWER3 microprocessor. The entire system will have 2,048 processors dedicated to large-scale scientific computing. The system will have a peak performance capability of more than 3 Tflop/s.

While this configuration may appear as a logical continuation of the trend toward highly parallel systems, established during the mid 1990s at many U.S. supercomputer installations, there are significant differences between NERSC-3, the new IBM SP system, and NERSC-2, a 640-processor SGI-Cray T3E/900. These differences are summarized in Table 1.

Table 1. Comparison of typical high-end platforms in the 1990s

	NERSC-1 Cray C90	NERSC-2 Cray T3E	NERSC-3 IBM SP-3
Year of Installation	1991	1996	2000
Number of Processors	16	640	2048
Processor Technology	Custom ECL	Commodity CMOS	Commodity CMOS
Peak System Performance	16 Gflop/s	580 Gflop/s	3000 Gflop/s
Measured System Performance	3.75 Gflop/s	29.6 Gflop/s	365 Gflop/s
Architecture	Shared memory, parallel vector	Distributed memory (shared address space)	128 nodes with 16-processor SMPs
System	Fully integrated custom system	Fully integrated custom system with commodity CPU and memory	Loosely integrated system with commodity system components
System Software	Vendor supplied, ready on delivery	Vendor supplied, completed after nearly three years development	Vendor supplied, contract for delivery in about three years
Footprint	588 ft	360 ft	1440 ft
Power Consumption	500 kW	288 kW	<1 MW

The three major NERSC systems of the 1990s are representative of the changes in high-end technology which were experienced in this decade. Between each successive generation, a major technology shift occurred that had an immediate impact on the production supercomputer centers. The first transition, which happened from about 1994 to 1996, is commonly characterized as the transition to massively parallel computing based on commodity microprocessors. This transition was widely anticipated as the "attack of the killer micros" and has been well documented, e.g., in various analyses of the TOP500 list [2]. The two important consequences of this transition for the supercomputer center were a reinvention of the center as a balance of production capabilities and intellectual services, together with an increase in the effort to develop system tools and software for highly parallel machines. The reinvention of NERSC has been discussed previously [3], and we will review some of the system software efforts at NERSC in Section 3.

The second transition is the result of an attempt to exploit not just commodity processors and memory, but whole commodity systems as the building blocks of the supercomputer. These cluster-of-SMP supercomputers have higher peak performance, greater memory capacity, and better cost performance than their predecessors. They also have higher inter-node communication latency, a larger footprint as they are air-cooled, and greater power requirements. They must be integrated with external disk caches and tape archives to manipulate the increasing volume of scientific data. They must also be closely coupled to wide-area networks to provide seamless access to a nationwide user base as well as other large computational assets. What is not yet fully appreciated is that this second technology transition (represented by NERSC-3) will be as fundamental as the previous one, and potentially even more dramatic (Table 2).

Table 2. Impact of the two technology transitions of the 1990s

	1994–1996 transition	**1998–2000 transition**
Economic Driver	Price performance of commodity processors and memory	16–64 CPU "sweet spot" for SMP technology in the commercial market place
Advantages of Transition	Higher performance and better price performance	Higher performance
Challenges of Transition	1. Applications transition to distributed memory, message passing model (MPI) 2. More complex system software (scheduling, checkpoint restarting)	1. Applications transition to hierarchical, distributed memory model (threads + MPI) 2. New development efforts for even more complex systems software 3. Increased cost of facilities

The changes are threefold:

- applications need to be written that can tolerate an increase in communication latency and parallelism as well as a distributed, hierarchical memory model;
- system software will have to be developed anew for increasingly complex, more difficult to manage, one-of-a-kind systems; and
- center management will be forced to take creative new approaches to solve the space and power requirements for the new systems.

These changes in the high-end platform technology will be accompanied by the previously mentioned increases in data storage requirements and the integration of the supercomputer center into a computational grid utilizing high bandwidth, wide-area networks.

3 Challenges of Clusters of SMPs as a Teraflop Production Platform

This section will discuss three challenges: the increase in parallelism and a distributed hierarchical memory model, the need for new system software for high-end platforms, and increased space and power requirements.

3.1 Increase in Parallelism and a Distributed Hierarchical Memory Model

In order to facilitate the transition to the new programming paradigm of massively parallel computing, NERSC in 1996 changed its service model to include both excellent facilities and excellent intellectual services. In this change, several new groups were added to NERSC, so that the computer center occupies the tension field between computational science and computer science (Fig. 1). For both areas, a strategy was developed for bringing the latest research results to bear on the effectiveness of the center.

One new group for which this change in strategy was particularly important was the Scientific Computing Group [4]. In the early 1990s, focus was on an applications group with experts in parallel computing representing the different applications areas, whereas the Scientific Computing Group consists of experts in computational techniques, which are relevant to a wider set of applications. This model is shown in Fig. 2.

The Scientific Computing Group at NERSC has been operating under this principle since 1996, and staff have been involved as research partners in many of the Grand Challenge applications at NERSC. One example of this successful work is the Gordon Bell Prize-winning collaboration between NERSC and Oak Ridge National Laboratory. In this collaboration, NERSC staff developed a new implementation of fast Fourier transforms (FFTs) on spherical data for distributed memory machines. The group was recognized for their simulation of metallic magnet atoms (Fig. 3), which was run on

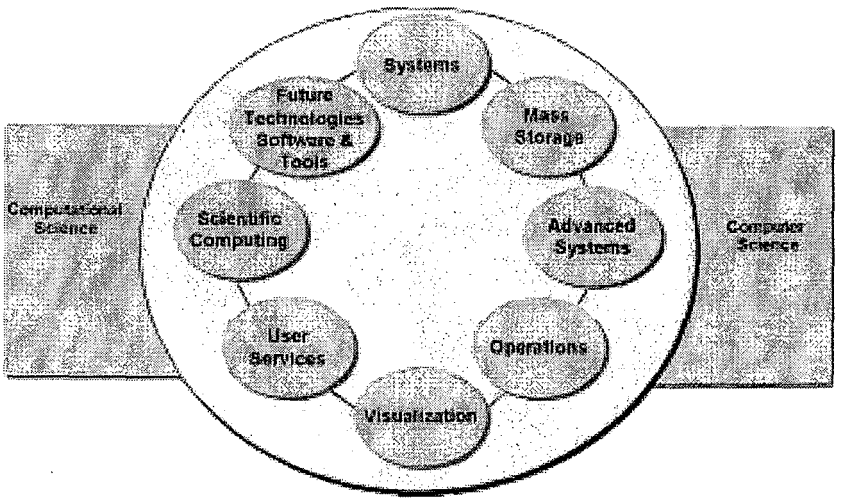

Fig. 1. The intellectual home of NERSC

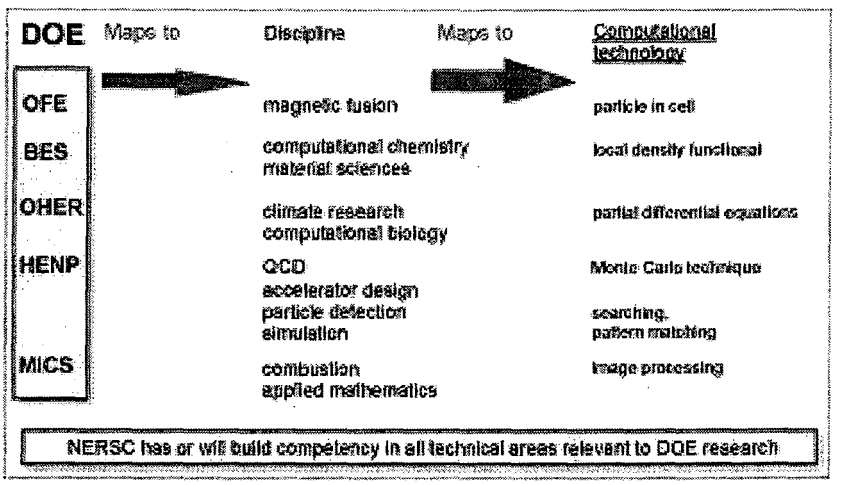

Fig. 2. Building computational competency at NERSC

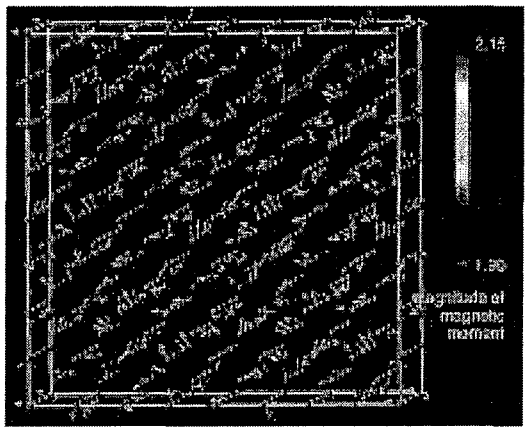

Fig. 3. Modeling of metallic magnetism – 1998 Gordon Bell Prize winner

increasingly powerful Cray T3E supercomputers. They started with NERSC's 512-processor machine, and once the code was optimized, moved to ever larger T3Es at other centers. They won the prize with a sustained performance of 657 Gflop/s using a 1024-processor T3E-1200. The group later topped that performance by achieving 1.02 Tflop/s on a 1480-processor T3E-1200. This was the first complete scientific application to have exceeded a sustained 1.0 Tflop/s performance [5].

This and many other success stories indicate that the Scientific Computing Group at NERSC is on the right track. For the next couple of years, the responsibilities and organization of the group will stay the same. The technology focus will be on the continued development of generic algorithmic techniques, which will support large-scale applications on platforms with larger numbers of processors and with a hierarchical memory structure. The group is well positioned to exploit the next generation architecture and produce algorithmic techniques that will benefit the NERSC client community at large.

3.2 System Software for High-End Platforms

The second challenge brought about by the next generation of high performance computing platforms is the decreasing maturity and lack of production quality of the system software. In the past NERSC has made pioneering contributions to bring first vector and later massively parallel machines into a full production environment. Based on work carried out in the Computational Systems [6] and Advanced Systems [7] groups, NERSC was the first site to demonstrate checkpoint/restart on a highly parallel system in

the fall of 1997 [8]. Continued work in collaboration with SGI/Cray led to the first installation of the complete Psched software system in the spring of 1999. Psched combines a number of software components that allow NERSC to manage the T3E very effectively. In early April 1999, NERSC demonstrated a utilization of more than 93% of the T3E for a sustained period of time (Fig. 4) [9]. Again this was a first for highly parallel machines. This is even more remarkable considering that the NERSC operating environment includes a wide range of jobs, ranging from interactive and debugging jobs to 512-processor Grand Challenge runs of up to 12 hours. The combination of these development efforts has made a cumulative impact on the utilization of the T3E, which has been consistently increasing over the last three years.

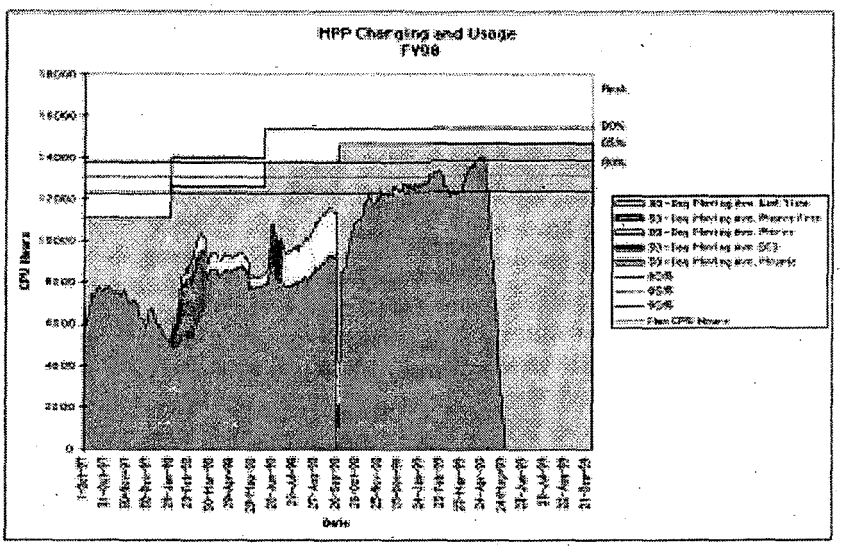

Fig. 4. Utilization of the Cray T3E at NERSC

After these successes with the T3E, the question needs to be raised, can NERSC continue with a similar strategy on the new IBM platform? A first evaluation of the IBM system software reveals that it is far less mature than the Cray T3E. Many features which have been developed with great effort on the T3E over the last couple of years do not exist yet, and may become available only halfway through the expected lifetime of the IBM SP3. For example, checkpoint/restarting, which NERSC believes to be critical for the success for a production parallel machine, will only arrive in late 2001. This is symptomatic for the high performance computing industry in the U.S., which has focused

most development efforts on 16 to 64 processor platforms. These constitute a "sweet spot" in the market and can be sold with larger margins to industrial and commercial users. Consequently software development efforts that are of interest only to a handful of high-end users, such as the government labs, have taken a backseat.

Looking a few years into the future, Paul Messina [10] has characterized the situation for high-end software by the "above the line–below the line" argument. He believes that vendors will only deliver system software up to the line of profitability, that is up to the "single box" SMP. Any software beyond the single system SMP, e.g., schedulers, file systems, communication protocols, accounting tools, etc. will have to be developed by the HPC community. The Department of Energy's Accelerated Strategic Computing Initiative (ASCI) will take the lead here, since they will have the first and largest of these configurations.

While NERSC is certainly willing to participate in any joint development effort, we believe that the "line" may be drawn too low. By lowering our expectations too much, we risk letting our vendor deliver hardware without the system software necessary to integrate it into a true production computing system. For the current IBM SP-3 system, NERSC and IBM have jointly developed an approach for developing high-quality production system software. Here we only want to mention one of our many joint efforts, which we believe has the potential to set a new standard for measuring utilization in the HPC community.

Currently there is no standard test to measure improvements in system utilization. While we have in general terms argued that developments such as checkpoint/restarting and scheduling algorithms will increase utilization, there is no quantitative assessment of the achieved improvements available today. As a matter of fact, most performance measurements today focus simply on improving speed of applications, i.e., on how much scientific work can be done for a given quantum of CPU time. In terms of a production system there is a second dimension, which we call effectiveness, i.e., how many quanta of CPU time can be made available to scientific programs. In order to maximize utilization, we must improve performance *and* effectiveness (Fig. 5).

The tool NERSC is planning to use to measure the impact of system software improvements on utilization is the SUPER benchmark [11]. This test uses a mix of NERSC test codes that run in a random order, testing standard system scheduling. There are also full configuration codes, I/O tests, and typical system administration activities. The setup of the SUPER benchmark is explained in Fig. 6. We expect at least 5% improvement per year on the IBM system.

NERSC continues on an aggressive acquisition strategy, which will lead to the installation of NERSC-4 in 2002 and NERSC-5 in 2005. One recent development that we are tracking closely is the rapid development of cluster computing with commodity processors and open software based on Linux. The Future Technologies Group at NERSC [12] has entered a CRADA (cooperative research and development agreement) with Intel for the development of M-VIA. M-VIA is an implementation of the Virtual Interface Architecture (VIA) for Linux [13]. VIA is a new standard being promoted by Intel,

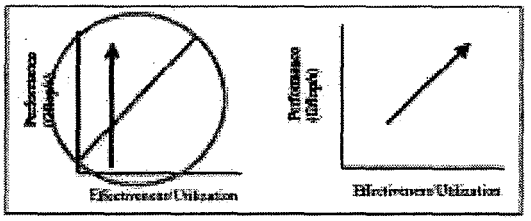

Fig. 5. The goal of high quality system software: improving both performance and utilization

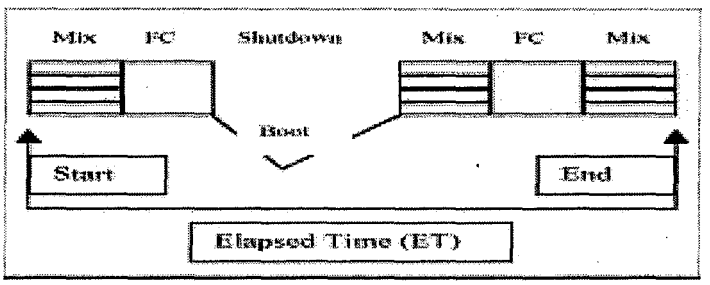

Fig. 6. Schematic organization of the SUPER benchmark

Compaq, and Microsoft that enables high performance communication on clusters. M-VIA is being developed as part of the NERSC PC Cluster Project [14]. The goal of this project is to enable NERSC and NERSC clients to build PC clusters for scientific computing. Small clusters are useful for parallel code development, special-purpose applications, and small- to medium-sized problems. Large clusters show promise of replacing MPPs such as the NERSC T3E for certain applications. While we are not yet ready to claim that NERSC-4 or NERSC-5 will be a cluster-based platform, the rapid changes of the last decade teach us to anticipate yet another technology transition. Work done in the Future Technologies group will help us to position ourselves for change in the three to five years.

An example of NERSC's M-VIA work, shown in Fig. 7, is a benchmark comparison of the NERSC PC cluster versus the T3E. For small numbers of processors (up to 32), the cluster is clearly competitive; but for larger numbers of processors, the well-designed interconnection network of the T3E shows its strength. Our data show that PC processors are even today not far from MPP processors in performance, and PC clusters with weak interconnects (fast Ethernet) are already a good alternative for some of our applications.

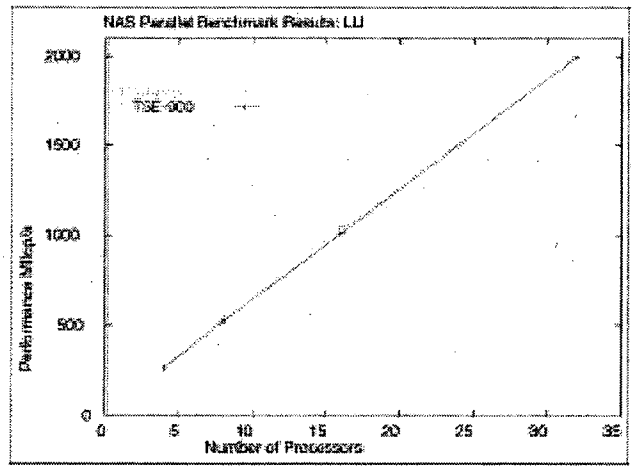

Fig. 7. PC cluster performance with M-VIA versus Cray T3E

3.3 Space and Power Requirements

The transition to commodity systems in the next generation of HPC platforms comes at a significant cost, which is usually hidden and is rarely mentioned when discussing the merits of different architectural approaches. In the transition from NERSC-2 to NERSC-3, we face a threefold increase in the machine's footprint and a twofold increase in its power requirements. These translate into substantial increases in ongoing facility costs for leasing space and purchasing electric power. NERSC is approaching this challenge by expanding to a new building. In an urban environment with a well-developed commercial real-estate market, this is a costly but manageable task. Our colleagues in the ASCI program have been forced to resort to new buildings at a scale previously unheard of for supercomputer centers. New buildings at LANL and LLNL are planned with a price tag in the $100M range. While NERSC will be able to meet the challenge at only a fraction of this cost, these high costs raise an interesting perspective on the transition to systems composed of commodity SMPs. Would not the HPC community in the U.S. be better off paying a higher price to our computer vendors in exchange for more tightly integrated systems that can be maintained within our existing facilities? As things stand today, we are spending tens of millions of dollars for non-technology-related expenses such as new buildings.

4 The Petabyte Data Challenge

While the challenges on the computing side are already quite formidable, supercomputer centers must also cope with an ever-increasing amount of data. In the past it has been correct to say that the amount of data generated *by computer simulations* was usually limited by the available computational technology. Thus the increase in archival storage was comparable to the increase in computational capability. In 1999 this view is no longer correct. What has changed is the fact that we will have to deal increasingly with *experimental data* which are generated from new technologies. One such example is the recent success in automating gene-sequencing technology. The rate at which data from the various genome projects will become available for further analysis, and the corresponding time it takes to search the genome database, are increasing faster than Moore's Law (Fig. 8).

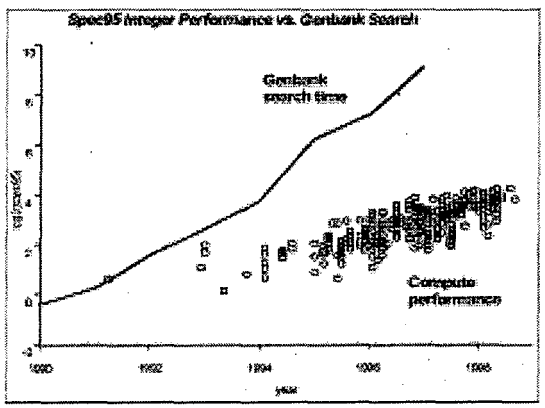

Fig. 8. Genome database search time is increasing at a faster rate than Moore's Law [15]

NERSC expects the increasing data it will be handling to come from several sources:

- genome data, e.g., the Human Genome Project at the DOE Joint Genome Institute;
- climate data, e.g., from the Program for Climate Model Diagnosis and Intercomparison (PCMDI) at Lawrence Livermore National Laboratory;
- high-energy physics data from new experiments such as the STAR experiment at Brookhaven's Relativistic Heavy Ion Collider or the ATLAS experiment at CERN's Large Hadron Collider.

Estimates of the high energy physics data produced by the Large Hadron Collider at CERN are on the order of a petabyte per year in 2005.

There are two efforts at NERSC to respond to this challenge. The File Storage Group [16] will continue to provide the storage media and baseline technology for large amounts of data. This group has increased the tertiary storage capacity at NERSC at an exponential rate, and so far has done an outstanding job of keeping our available storage capacity ahead of the demand (Fig. 9).

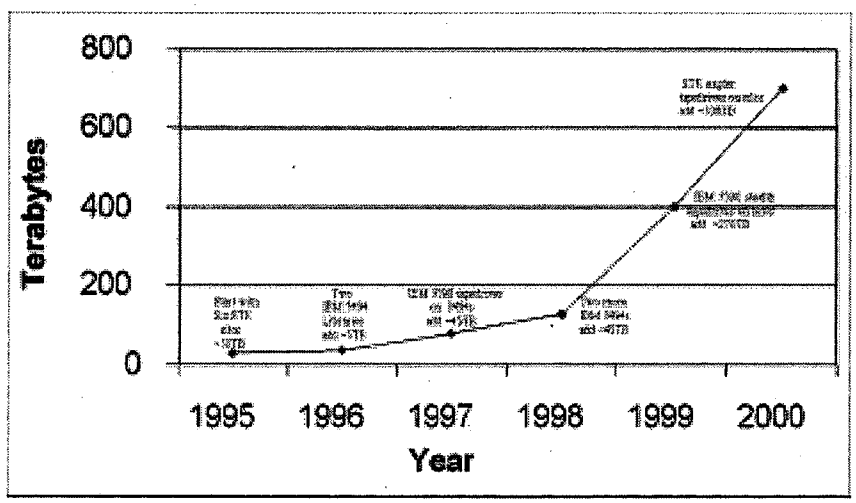

Fig. 9. NERSC storage capacity (raw)

At the same time, while raw capacity was increasing at an exponential rate, NERSC transitioned its storage management system completely to HPSS. As a developer site, NERSC is able to influence the HPSS consortium to provide tools to meet the requirements of our data intensive applications. Given the flood of future data, this will be a significant advantage for NERSC clients.

The second thrust in meeting the petabyte data challenge at NERSC is to provide tools for scientists to manage their data more effectively. There are two groups that work in this area. The Center for Bioinformatics and Computational Genomics (CBCG) [17] provides tools for the analysis of biological sequences, protein structure and function prediction, and large-scale genome annotation, as well as tools for access to biological information (database integration, data mining). The Scientific Data Management Group [18] is involved in various projects including tertiary storage management for high energy physics applications, data management tools, and efficient access to mass storage data.

Here we can highlight just one of the many projects of the Scientific Data Management Group (SDM). Typically, climate simulations and assimilated observational climate data

are large multidimensional datasets in space (represented as meshes) and time. Accurate models require that these meshes are as dense as possible. When scientists return to analyze these datasets, they often need to access only one of the fifty or more parameters associated with each node in the mesh. If the entire dataset must be accessed from tape in order to extract the fields of interest, the time required can be many minutes or hours. This slows down the effectiveness of data analysis to the point that much of the data is never analyzed. Increasing access time of subsets from hours to minutes is the key to effective analysis (Fig. 10).

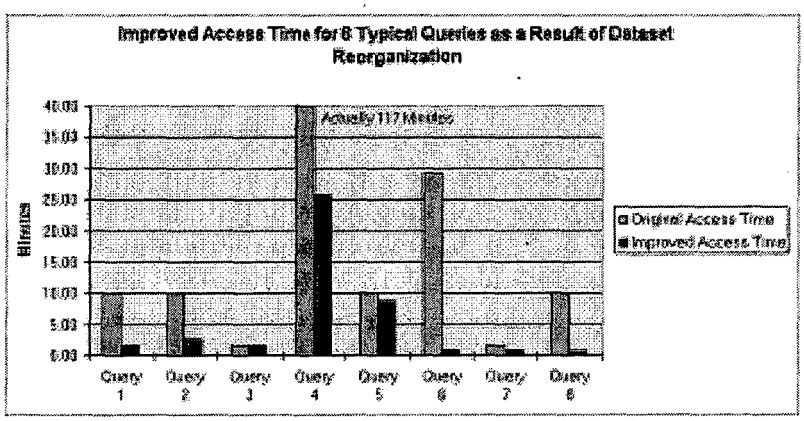

Fig. 10. Improved access time for climate data from mass storage

The NERSC SDM group has developed optimization algorithms for reorganizing the original datasets to match their intended usage. Further, the SDM group has designed enhancements to current storage server protocols to permit control over physical placement of data on storage devices. At the analysis level, this involves the application of clustering algorithms and organization of data into bins.

The work of the SDM group is unique among supercomputing centers, and we are not aware of a comparable research effort elsewhere. Projects in the SDM group and in CBCG will result in efficient new tools that NERSC clients can use to deal with their petabyte datasets.

5 Preparing for the Data Grid

In the last two years, the vision of a computational grid has gained broad acceptance. The grid is envisioned as a unified collection of geographically dispersed supercomputers,

storage devices, scientific instruments, workstations, and advanced user interfaces (Fig. 11). The recent book *The Grid* [19] is an excellent summary of the current status of efforts to build such a grid.

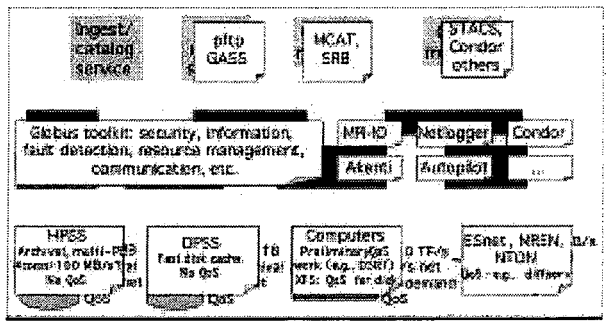

Fig. 11. Components of the data grid [20]

The most significant aspect of the grid for a supercomputer center like NERSC is the new concept of a data grid, enabling transparent access to data by scientists widely distributed across the United States. The petabyte datasets discussed in the previous section are community resources, which will be shared by researchers who are geographically distributed yet participating in collaborative projects. We do not expect these data to reside exclusively at one site, nor do we expect access to be restricted to a local set of users. Therefore, NERSC is investigating research issues related to large datasets distributed over a wide-area network. An example is the Distributed Parallel Storage System (DPSS) [21]. DPSS is a distributed disk cache which provides high-performance data handling as well as an architecture for building high-performance storage systems from low-cost commodity hardware components. This technology has been quite successful in providing an economical, high-performance, widely distributed, and highly scalable architecture for caching large amounts of data that can potentially be used by many different users.

One recent project that builds on DPSS, and which can be considered a prototype instantiation of data-grid technology, is the China Clipper Project (Fig. 12) [22]. In this project, high energy physics data which are generated at the Stanford Linear Accelerator Center (SLAC) are shared among storage systems at SLAC, NERSC, and Argonne National Laboratory. One of the early successes was a sustained data transfer rate of 58 Mbyte/sec from SLAC to the data archive in Berkeley [23].

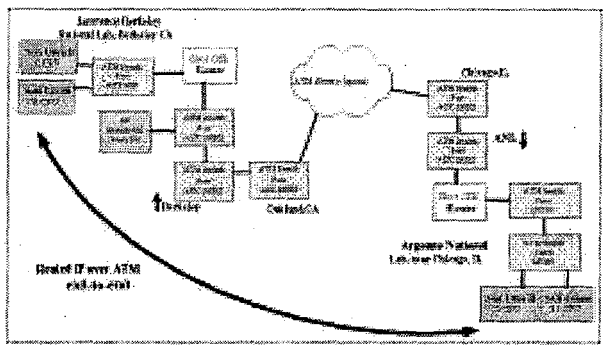

Fig. 12. The China Clipper testbed

6 The High Performance Organization

One of the driving factors for the continuous change in high performance computing is Moore's Law. Moore's Law postulates exponential growth in technology—a performance doubling of microprocessors every 18 months. Normally Moore's Law is plotted on a semi log scale and appears to us as a straight line, as in Fig. 13.

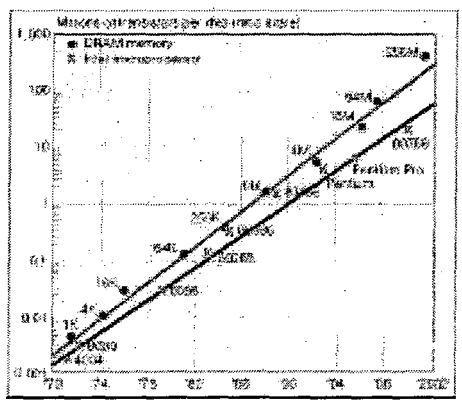

Fig. 13. Moore's Law – the customary straight-line view

However, human experience cannot deal well with a logarithmic scale, or intuitively grasp well the true effects of exponential growth. In the real world, we are sitting at the "bend" of an exponential curve. From our perspective, Moore's Law appears more like a straight wall, climbing steeply up into the infinite (Fig. 14). The significance of the "wall" is that in a few years, technology will be again completely different, and we have no clue what the future will be. For anyone not accustomed to change, Moore's Wall will be impenetrable and bewildering.

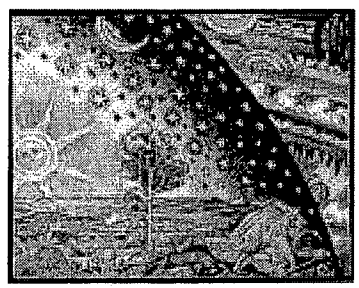

Fig. 14. Moore's Wall – the true exponential point of view (16th century version)

Thus the last and grandest challenge for a high performance computing center is to be a high performance organization. By this we mean an organization where staff can thrive and perform well under the stress of constant technology change and an unpredictable future.

Acknowledgements. We would like to thank all NERSC group leaders and NERSC staff, who helped in building an excellent organization. In particular, we acknowledge contributions by David Bailey, Andrew Canning, Jim Craw, Brent Draney, Keith Fitzgerald, Bill Saphir, Arie Shoshani, Brian Tierney, Mike Welcome, Tammy Welcome, and Manfred Zorn (all at NERSC) as well as contributions by Ian Foster (ANL) and Bill Johnston (LBNL and NASA). Their ideas, research results, and data were used in writing this paper. This work was supported by the Director, Office of Advanced Scientific Computing Research, Division of Mathematical, Information, and Computational Sciences of the U.S. Department of Energy under contract number DE-AC03-76SF00098.

References

1. http://www.nersc.gov
2. Horst D. Simon: High Performance Computing in the U.S. in 1994. Supercomputer **11** (1995) 21–31

3. Horst D. Simon: Site Report: Reinventing the Supercomputer Center at NERSC. IEEE Computational Science and Engineering Vol. 4 No. 3 (1997)
4. Esmond Ng et al.: http://www.nersc.gov/research/SCG/
5. B. Ujfalussy, X. Wang, X. Zhang, D. M. C. Nicholson, W. A. Shelton, G. M. Stocks, A. Canning, Y. Wang, and B. L. Gyorffy: High Performance First Principles Method for Complex Magnetic Properties. IEEE Proceedings of SC98, Orlando, Florida (1998)
6. James Craw et al.: http://www.nersc.gov/aboutnersc/sys.html
7. Tammy Welcome et al.: http://www.nersc.gov/research/annrep98/29advanced.html
8. Jon Bashor: NERSC First to Reach Goal of Seamless Shutdown, Restart of Supercomputer. http://www.lbl.gov/Science-Articles/Archive/Cray-checkpointing.html (1997)
9. Jon Bashor: NERSC Achieves Breakthrough 93% Utilization on Cray T3E. http://www.nersc.gov/news/t3e-utilization4-19-99.html (1999)
10. Paul Messina: The 30Tflops Procurement at Los Alamos. Presentation (January 1999)
11. David H. Bailey et al.: System Utilization Performance Effectiveness Rating (SUPER) (in preparation, 1999)
12. Bill Saphir et al.: http://www.nersc.gov/research/FTG/
13. Bill Saphir et al.: http://www.nersc.gov/research/FTG/via/
14. Bill Saphir et al.: http://www.nersc.gov/research/FTG/pcp/
15. David J. States: http://www.ibc.wustl.edu/~states/mooreslaw.html
16. Keith Fitzgerald et al.: http://www.nersc.gov/groups/FSG/
17. Sylvia Spengler and Manfred Zorn: http://www.nersc.gov/cbcg/
18. Arie Shoshani et al.: http://gizmo.lbl.gov/DM.html
19. Ian Foster and Carl Kesselman (eds.): The Grid: Blueprint for a New Computing Infrastructure. Morgan-Kauffman (1998)
20. Ian Foster: Large-Scale Data Grids. DOE Conference on High Speed Computing, Salishan Lodge, OR (1999)
21. Brian Tierney et al.: http://www-itg.lbl.gov/DPSS/
22. William Johnston et al.: http://www-itg.lbl.gov/Clipper/
23. Jon Bashor: New Technology Demonstrates Priority Service for Internet Traffic Between Lawrence Berkeley and Argonne National Laboratories. http://www.lbl.gov/CS/Archive/headlines4-14-98.html (1998)

ERNEST ORLANDO LAWRENCE BERKELEY NATIONAL LABORATORY
ONE CYCLOTRON ROAD | BERKELEY, CALIFORNIA 94720

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