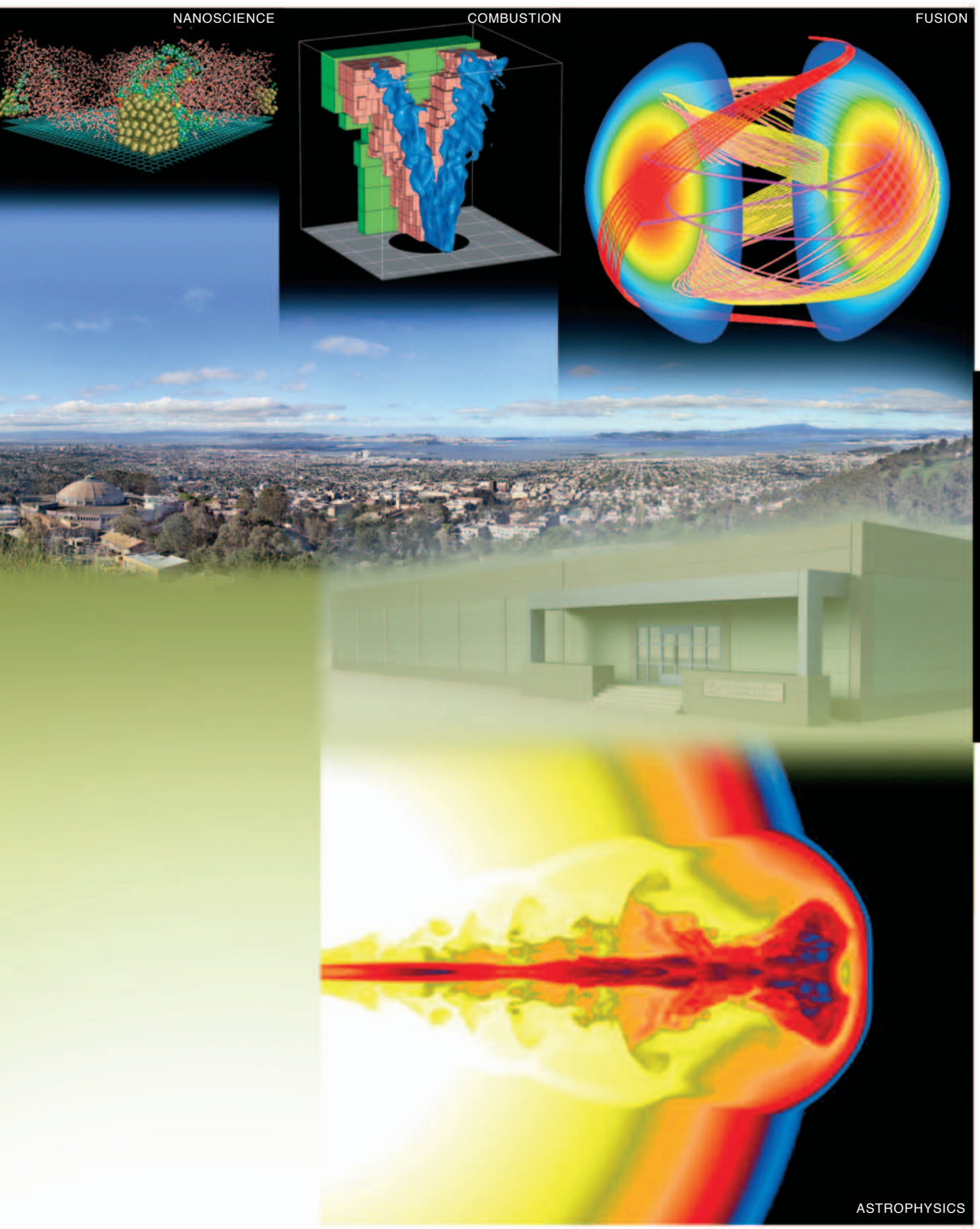
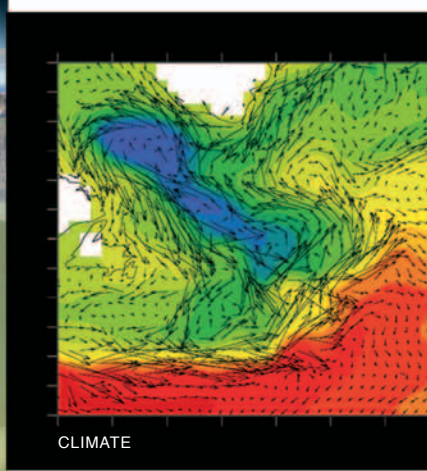
National Facility for Advanced Computational Science: A Sustainable Path to Scientific Discovery

A Proposal to the DOE Office of Science
from Lawrence Berkeley National Laboratory





Ernest Orlando Lawrence Berkeley National Laboratory
NERSC National Energy Research Scientific Computing Center
Office of Science U.S. Department of Energy

LBNL/PUB-5500

National Facility for Advanced Computational Science: A Sustainable Path to Scientific Discovery

A Proposal to the DOE Office of Science from Lawrence Berkeley National Laboratory

April 2, 2004

Horst Simon, William Kramer, William Saphir, John Shalf, David Bailey, Leonid Oliker, Michael Banda, C. William McCurdy, John Hules, Andrew Canning, Marc Day, Philip Colella, David Serafini, Michael Wehner, Peter Nugent

This work was supported by the Director, Office of Science, Office of Advanced Scientific Computing Research of the U.S. Department of Energy under Contract No. DE-AC 03-76SF00098.

Table of Contents

Project Summary

Lawrence Berkeley National Laboratory (Berkeley Lab) proposes to create a National Facility for Advanced Computational Science (NFACS) and to establish a new partnership between the American computer industry and a national consortium of laboratories, universities, and computing facilities. NFACS will provide leadership-class scientific computing capability to scientists and engineers nationwide, independent of their institutional affiliation or source of funding. This partnership will bring into existence a new class of computational capability in the United States that is optimal for science and will create a sustainable path towards petaflops performance.

In order to effect a fundamental change in computer architecture development, we have established a national Leadership Computing Consortium (LCC), composed of a number of federal agencies and U.S. supercomputing facilities, including the National Science Foundation's SDSC, NCSA, and TeraGrid consortium, as well as Lawrence Livermore National Laboratory. The LCC, in collaboration with NFACS, will change the way that high performance computers are designed and deployed through a new type of development partnership with the computer vendor based on the concept of *science-driven computer architecture*. The role of the LCC is twofold: first, it will provide input from science to drive computing technology forward; second, it will incorporate best practices from our partners into NFACS. We will develop a robust nationwide support infrastructure that ensures effective use of the facility.

We propose a partnership with IBM, the leading U.S. vendor of high performance computing systems for science, to design a science-driven computer architecture with a sustained performance of 50 Tflop/s on a broad spectrum of applications of national importance. This will be achieved through the phased development and installation of two Leadership Class Systems (LCS-1 and LCS-2). These will be designed with assistance from computational scientists, with prototype pre-commercial configuration and testing under way at IBM. This architectural approach achieves the highest sustained performance across a broad range of key scientific applications for the lowest cost. It provides the best national investment for scientific productivity, demonstrates continued U.S. leadership in computational science, and forges a path to petaflops computing.

Applications scientists have been frustrated by a trend of stagnating application performance, despite dramatic increases in claimed peak performance of high-performance computing systems. Our strategy reverses that trend by engaging those scientists well before an architecture is available for commercialization. The partnership with IBM is based on a collaborative approach to designing computer architecture that will enable heretofore unrealized achievements in computer-capability-limited fields, including nanoscience, combustion modeling, fusion, climate modeling, and astrophysics. The unprecedented level of computational capability that will be made available to researchers in these fields will result in scientific breakthroughs on issues of national importance.

We will draw on the demonstrated track record of the NERSC Center at Berkeley Lab in acquiring and fielding high-end systems that meet user requirements and lead the country in unclassified scientific computing. Leveraging NERSC will significantly reduce operational costs of NFACS, while ensuring a timely installation and operation of the new facility.

The establishment of NFACS will result in a scientific computing capability that durably returns the performance advantage to American science. A new class of computer designs will not only revolutionize the power of supercomputing for science, but it also will affect scientific computing at all scales. This proposal represents our vision for achieving outstanding computational science by providing for the continued development of science-driven computer architecture.

PI: Horst D. Simon, Associate Laboratory Director for Computing Sciences
Mail Stop 50B-4230, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
(510) 486-7377, fax (510) 486-4300, hdsimon@lbl.gov

1. INTRODUCTION: A SUSTAINABLE PATH TO LEADERSHIP COMPUTING

This proposal presents a plan that will maintain and strengthen U.S. leadership in high performance computing, initiate a new wave of scientific discovery, and enable the solution of problems of national importance. Lawrence Berkeley National Laboratory (Berkeley Lab) proposes to create a National Facility for Advanced Computational Science (NFACS) and to establish a partnership between the American computer industry and a national consortium of leading laboratories, universities, and computing facilities. In this proposal we are guided by the following analysis:

1. This U.S. Department of Energy (DOE) investment must lead to widely deployable new technology for high-end scientific computing. If it leads merely to a series of experiments or the purchase of a single machine, it will not have a lasting impact.
2. The technology we need will not spontaneously appear on the market. By taking a passive approach that relies on existing vendor offerings, the high performance computing community has ceded leadership to other players whose requirements are increasingly incompatible with the needs of high-end computing.
3. Several national panels have concluded that the rules of engagement between the scientific community and the American computer industry must be revised. Scientific applications must directly influence machine design in a repeating cycle: (a) scientific applications input to designers, (b) computer design with increased performance, (c) deployment and delivery of new systems to the scientific community, (d) repeat.
4. Successfully changing the rules of engagement requires a partnership with the American computer company with the most resources, the best track record of research and development, and proven success in delivering in high performance computing. To justify the necessary commitment from the company, we will form a national consortium of laboratories, computing facilities, universities, and researchers equally committed to changing the future of the computing capability available to the scientific community.
5. Berkeley Lab and our partners have evaluated a representative array of scientific applications to establish precisely their algorithmic characteristics. From those algorithms we have derived a clear understanding of the limitations of current high-end systems of all designs, from clusters to vector computers.
6. Over the past two years, the Blue Planet partnership led by Berkeley Lab has worked closely with IBM to design a machine that better meets the needs of scientific applications. The goals and methodology of this partnership were validated by the successful design and implementation of the $100M ASCI Purple system at Lawrence Livermore National Laboratory (LLNL), based on the Blue Planet design.
7. We propose to continue and expand the Blue Planet process, bringing in a national consortium of partners to guide the process. IBM has committed to participate, delivering a Leadership Class Computing System based on an extension of the Blue Planet work, with possible evolution into a hybrid with the IBM Blue Gene designs currently under research and development.
8. Berkeley Lab, through its effective management of NERSC, has earned the reputation for delivering the best high-end computing to the national scientific community. Leveraging these resources and experience will produce the greatest return on DOE's investment and provide the greatest opportunity for a successful scientific program.

Berkeley Lab has already started on the path to establishing this new facility, deploying the optimal architecture, and building the national coalition necessary for the successful realization of this vision.

2. SCIENTIFIC APPLICATIONS AND UNDERLYING ALGORITHMS DRIVE ARCHITECTURAL DESIGN

The central goal of this proposal is to deliver new scientific results on computations of a scale that greatly exceeds what is possible on current systems, with sustained performance rates of 50 Tflop/s in 2007 on applications of scientific and national importance. To that end, we have identified the following application classes as being ripe for breakthrough science using very high-end computing, and relevant to some of the most important national objectives: nanoscience, combustion modeling, fusion energy simulations, climate modeling, and astrophysics.

Appendix D describes in detail Berkeley Lab's scientific collaborators for this proposal, their scientific applications, computational requirements, and the achievements that can be expected when the applications are run on this Leadership Class System. Table 1 summarizes the goals, computational methods, and example applications of each science area.

Table 1
Science Breakthroughs Enabled by Leadership Computing Capability

Science Areas	Goals	Computational Methods	Examples of Breakthrough Applications
Nanoscience	Simulate the synthesis and predict the properties of multi-component nanosystems	Quantum molecular dynamics Quantum Monte Carlo Iterative eigensolvers Dense linear algebra Parallel 3D FFTs	Simulate nanostructures with hundreds to thousands of atoms, as well as transport and optical properties and other parameters
Combustion Modeling	Predict combustion processes to provide efficient, clean and sustainable energy	Explicit finite difference Implicit finite difference Zero-dimensional physics Adaptive mesh refinement Lagrangian particle methods	Simulate laboratory-scale flames with high-fidelity representations of governing physical processes
Fusion Energy	Understand high-energy density plasmas and develop an integrated simulation of a fusion reactor	Multi-physics, multi-scale Particle methods Regular & irregular access Nonlinear solvers Adaptive mesh refinement	Simulate the ITER reactor
Climate Modeling	Accurately detect and attribute climate change, predict future climate, and engineer mitigation strategies	Finite difference methods FFTs Regular & irregular access Simulation ensembles	Perform a full ocean/atmosphere climate model with 0.125 degree spacing, with an ensemble of 8–10 runs
Astrophysics	Determine through simulation and analysis of observational data the origin, evolution, and fate of the universe; the nature of matter and energy; galaxy and stellar evolution	Multi-physics, multi-scale Dense linear algebra Parallel 3D FFTs Spherical transforms Particle methods Adaptive mesh refinement	Simulate the explosion of a supernova with a full 3D model

The most effective approach to designing a computer architecture that can meet these scientific needs is to analyze the underlying algorithms of these applications, and then, working in partnership with vendors, design a system targeted to these algorithms.

From this list of important scientific applications and underlying algorithms, several themes can be derived that drive the choice of a large-scale scientific computer system: (1) multi-physics, multi-scale calculations; (2) limited concurrency, requiring strong single-CPU performance; (3) reliance on key library routines such as ScaLAPACK and FFTs; (4) the use of particle methods, with couplings to grid-based methods that lead to large-scale interaction of two regular, but unaligned, data structures; (5) widespread usage of finite difference computations, requiring good performance on fairly regular accesses in multiple dimensions and high main memory bandwidth; (6) an increasing usage of sparse,

unstructured, and adaptive mesh (AMR) methods, which entail some irregular control sequences that do not perform well on vector systems; and (7) ubiquitous data parallelism providing the opportunity for fine-grained operation concurrency; (8) irregular control flow inhibiting fine-grained symmetric operation concurrency. Table 2 presents a qualitative summary of this information:

Table 2
Algorithm Requirements

Science Areas	Multi-physics & multi-scale	Dense linear algebra	FFTs	Particle methods	AMR	Data parallelism	Irregular control flow
Nanoscience	X	X	X	X		X	X
Combustion	X			X	X	X	X
Fusion	X	X		X	X	X	X
Climate	X		X		X	X	X
Astrophysics	X	X	X	X	X	X	X

The characteristics summarized here point to the need for a flexible system — one that can perform well both on random memory access calculations as well as regular memory access problems and that combines strong single-node performance (to minimize the required concurrency in the application) and a powerful system-scale network.

Of the two principal classes of high performance systems in widespread usage — superscalar systems and vector systems — each has a different set of advantages and disadvantages for these applications. Superscalar, cache-memory-based systems tend to do well on problems with spatial and temporal data regularity. These systems also do relatively well on irregularly structured algorithms and codes with heavy usage of conditional branching in inner loops. However, many cache-based systems feature low or over-subscribed main memory bandwidth, since they are not primarily designed for scientific computation. Thus, codes with low computational intensity typically do not perform well on these architectures.

Vector systems exploit regularities in the computational structure to expedite uniform operations on dependence-free data. Many scientific codes are characterized by predictable fine-grained data-parallelism and thus allow vectorization. However, vector systems tend to do poorly on codes with irregularly structured computations. These codes are characterized by irregular control flow, intensive scalar operations, and significant conditional branching — operations that inhibit vectorization. Performance on vector architectures degrades significantly even when a small fraction of the work is non-vectorizable, as described by Amdahl's Law. This is particularly true for newly emerging multi-method, multi-physics codes that can only leverage vectorization for a subset of the numerical components.

These considerations suggest that an architecture that combines the best features of high-end superscalar and vector systems would be best suited for the workload that we project for future high-end computing of national importance. To that end, we will describe in the following sections a system that is being developed by IBM, in collaboration with NFACS and the Leadership Computing Consortium, that targets this broad range of scientific computing.

3. A SCIENCE-DRIVEN SYSTEM ARCHITECTURE

Applications scientists have been frustrated by a trend of stagnating application performance despite dramatic increases in claimed peak performance of high-performance computing (HPC) systems. This trend has been widely attributed to the use of commodity components whose architectural designs are

unbalanced and inefficient for large-scale scientific computations. It was assumed that the ever-increasing gap between theoretical peak and sustained performance was unavoidable. However, recent results from the Earth Simulator (ES) in Japan clearly demonstrate that a close collaboration with a vendor to develop a science-driven architectural solution can produce a system that achieves a significant fraction of peak performance for critical scientific applications. The key to the ES success was the long-term collaborative development strategy between the scientists of JAMSTEC (Japan Marine Science and Technology Center) and NEC Corporation.

Realizing that effective large-scale system performance cannot be achieved without a sustained focus on application-specific architectural development, Berkeley Lab and IBM have led a collaboration since 2002 that involves extensive interactions between domain scientists, mathematicians, computer experts, as well as leading members of IBM's R&D and product development teams. The goal of this effort is to change IBM's architectural roadmap to improve system balance and to add key architectural features that address the requirements of demanding leadership-class applications — ultimately leading to a sustained Pflop/s system for scientific discovery. The first product of this multi-year effort has been a redesigned Power5-based HPC system known as Blue Planet [1] and a set of architectural extensions referred to as ViVA (Virtual Vector Architecture). This collaboration has already had a dramatic impact on the architectural design of the ASCI Purple system [2], and has resulted directly in the strong Leadership Class Systems (LCS-1 and LCS-2) offering presented in this proposal.

Blue Planet design is incorporated into the new generation of IBM Power microprocessors that are the building blocks of the LCS-1 and LCS-2 configurations. These processors break the memory bandwidth bottleneck, reversing the recent trend towards architectures poorly balanced for scientific computations. The Blue Planet design improved the original power roadmap in several key respects: dramatically improved memory bandwidth; 70% reduction in memory latency; eight-fold improvement in interconnect bandwidth per processor; and ViVA Virtual Processor extensions, which allow all eight processors within a node to be effectively utilized as a single virtual processor.

The approach described in this proposal — a two-stage deployment of a Leadership Class System, LCS-1 and LCS-2 — is a continuation of the work that started in the Blue Planet initiative. We will expand upon this successful collaborative effort, starting with the baseline configurations discussed below and in Appendix A. The purpose of this collaborative approach is not just to produce the most effective scientific computing platform in the LCS-2 timeframe, but also to begin moving on a longer-term roadmap towards successful Pflop/s computing.

3.1 Building on the Blue Planet Collaboration: Addressing the Memory Bandwidth Bottleneck

We propose to continue and expand the Blue Planet process to develop further improvements to the LCS-2 system and beyond. This continued collaboration will lead to a set of enhancements known as ViVA-2. Leadership Computing Consortium (LCC) partners, including the science application collaborators, will participate in the system design and refinement process. We will hold quarterly meetings to review progress, create ideas, and refine the design decisions. These meetings will integrate application scientists, system designers, HPC performance experts, and computer scientists. This community approach of directly engaging vendors in the collaborative process of designing leadership HPC systems was laid out by the High End Computing Revitalization Task Force (HECRTF) [3] and the DOE SCaLeS Workshop [4], and demonstrated successfully by the Earth Simulator, initial Blue Planet effort, and Red Storm effort [5].

There is an opportunity to incorporate the ViVA-2 scientific enhancement technology into future Power processor design. During FY04 and FY05, IBM and the LCC partners will evaluate various enhancements to the LCS-2 processor, node, and interconnect design, including assisted processing capabilities and their impact on the associated components (e.g., compilers, libraries, tools, etc.). The

LCC will advise IBM on how to incorporate the resulting technology into LCS-2 and subsequent systems, to maximize its impact on scientific discovery. Thus, the LCS-2 system described in this document should be considered a minimum base from which improvements will evolve.

IBM's willingness to work with Berkeley Lab (see letter by IBM in Appendix D) to develop modifications to its hardware that further enhance performance of scientific applications clearly demonstrates their commitment to scientific computing and the importance of the IBM's partnership with the computational science community. IBM is the only company that both demonstrates a clear commitment to make such deep changes to their design and offers the immense resources required to meet those commitments.

ViVA Design Targets

ViVA and ViVA-2 are specialized enhancements to the Power architecture designed to significantly improve sustained performance on a wide range of scientific applications. ViVA is a compiler-supported programming model that combines processors to form more powerful virtual processors by making use of fast barrier synchronization technology available in Power5 and Power6 processors. ViVA will be available on both the LCS-1 and LCS-2 systems.

ViVA-2 is envisioned as a set of extensions to the Power6 architecture that will accelerate scientific applications by supporting deeper pipelining of memory requests in order to hide memory latencies. These extensions will improve the efficiency of memory accesses on both vectorizable and non-vectorizable codes. ViVA-2 is superior to strictly vector designs because it offers the flexibility of achieving high performance on non-vectorizable algorithms using state-of-the-art superscalar technology, while efficiently processing data-parallel code segments that are amenable to vectorization. These enhancements address a variety of scalar memory performance degradations often attributed to irregularities in the data-access patterns. Examples include ineffective hardware prefetching, load/store instruction issue-rate limitations, and wasted bandwidth due to partially used cache lines.

3.2 Leadership Computing Systems (LCS)

Our goal is to build an architecture balanced for leadership-class science requirements as described above in Section 2, which presents the computational science applications that will be of critical importance to U.S. scientific leadership in 2007 and are able to take advantage of an ultra-scale computing system.

The key science requirements for leadership class computing can be distilled into three main system features: processor performance, interconnect performance, and software. Processors should have excellent sustained single-node performance across the spectrum of applications. The interconnect should provide high per-link performance (both latency and bandwidth) as well as high bisection bandwidth. Effective system utilization requires proven system software scalability and optimized numerical libraries.

The goal of NFACS is to enable new science discoveries. Implicit in this is a requirement for real working systems. Our plans take into account both credibility and risk in vendor roadmaps. As part of its management of NERSC, Berkeley Lab recently released a request for information (RFI) to the entire high performance computing and storage industry. The RFI went to over 40 high-end computing vendors.

From an analysis of the responses, we concluded that there are only two U.S. vendors with a credible roadmap to provide leadership-class computing capability in the 2007 time frame: IBM (Power6) and Cray (vector systems). Other vendors have competitive offerings at smaller scales, but not for the largest system scales. Additionally, software for cluster architectures is not sufficiently robust at this time to effectively manage a leadership-scale system.

After analyzing the latest system architecture and pricing information from IBM and Cray, we concluded that the IBM solution will ultimately be the best way to meet needs for a leadership class

system in 2007. This architectural approach achieves the highest sustained performance across a broad range of key scientific applications for the lowest cost. The IBM solution builds upon a stable foundation of technology development, whereas the technical discontinuities in the Cray roadmap create high risk to both schedule and performance.

We have had access to an early Power5 system to run benchmarks and validate our assumptions about the ability of this processor to sustain a relatively high percentage of peak performance. Our tests confirmed that the Power5 will sustain high performance. Details of the technical rationale and benchmark results are discussed in Appendix C.

We propose a two-phased approach — LCS-1 and LCS-2 — to achieve a Leadership Class System in 2007 with a sustained performance of 50 Tflop/s. LCS-1 will be installed in June 2005. LCS-2 will be installed in November 2007. For the complete schedule, see Appendix A.

LCS-1

LCS-1 is a Power5 system with eight single-core CPUs per node and [proprietary information deleted] nodes ([proprietary information deleted] CPUs). The CPUs run at [proprietary information deleted] GHz ([proprietary information deleted] Gigaflop/s peak). The system will feature more than 31 Tflop/s peak and 6 Tflop/s average sustained performance. The system will have 16 terabytes (TB) of main memory and 500 TB of disk (Appendix B). The interconnect will be IBM's high performance Federation switch. It is expected that average application performance will be 20% of peak, with several key applications well above that range. Key innovations in the Power5 architecture that allow it to obtain a much higher percentage of peak performance than its predecessors, such as the Power4, include:

- **High-memory bandwidth per processor,** including a memory architecture that achieves [proprietary information deleted] bytes/flop, comparable to vector architectures.
- **"Single core" node design.** IBM's original roadmap called for two processor cores on a single chip to share the same memory system. Going to a single core design effectively doubles the memory bandwidth per processor.
- **Small node design.** With eight-processor nodes, it is possible to put the processors closer to memory, reducing memory latency. Furthermore, by reducing the number of processors per node, effective network bandwidth per processor exceeds IBM's original 32- or 64-way SMP roadmap.
- **ViVA Virtual Processing** that allows the eight processors in a node to be treated as a single processor with peak performance of 64 Gigaflop/s. Codes that benefit from Cray X1 multistreaming, for example, will directly benefit from ViVA capabilities. See Appendix A for more details.

The LCS-1 network will be based on IBM's "Federation" interconnect. Two "planes" of this network will provide 8 GB/s of bidirectional bandwidth per node, or 1 GB/s per processor. Federation topology is a modified fat-tree that provides full-bisection bandwidth. Unlike systems that employ mesh and torus networks, the fat-tree network allows any processor to communicate with any other processor in the system free of bandwidth contention. This offers the most flexibility and the highest performance of any comparable system, resulting in a gross bisection bandwidth of 4 TB/s for LCS-1.

The global file system on the LCS systems will be IBM's General Parallel File System (GPFS), a mature parallel file system that provides excellent performance and functionality. GPFS is the only parallel file system that has been demonstrated to support a diverse scientific parallel workload at the scale of the LCS systems.

A robust software environment is a critical component of a leadership architecture. System software on the LCS systems will be an improved version of IBM's current SP system software, which powers one-third of the Top 500 computers in the world and is the result of thousands of person years of effort.

IBM's SP software has proven its robustness and reliability at NERSC by consistently enabling utilization of 90 to 95% of available computational resources.

LCS systems will have optimized mathematical and scientific libraries, including ESSL, MASS, and FFTW. Many codes poised to run at this scale depend on the availability of such libraries to extract maximum performance from the architecture.

LCS-2

LCS-2 is a Power6 system with eight single-core CPUs per node, running at [proprietary information deleted] GHz ([proprietary information deleted] Gigaflop/s). It has [proprietary information deleted] nodes, for a total of [proprietary information deleted] processors, with a total of 42 TB of main memory and 1.6 petabytes (PB) of shared global disk. The system will feature 210 Tflop/s peak and 50 Tflop/s average sustained performance. The Power6-based LCS-2 system will have an impressive memory performance of [proprietary information deleted] bytes/flop ([proprietary information deleted] GB/s per processor), allowing increased sustained performance across a broad spectrum of leading scientific applications.

The LCS-2 network will be based on [proprietary information deleted]. The aggregate bandwidth of the full bisection network achieves 31.5 TB/s, allowing for efficient execution of large-scale applications with global communication requirements.

LCS-2 will have the same basic file system and software as LCS-1, with improvements.

LCS-2 Refinements and Beyond

The ViVA-2 extensions being studied for LCS-2 are intended to benefit scientific codes that are characterized by the kind of predictable data parallelism that is typically associated with vector processing. Since the superscalar core performs all computations on operands fetched by ViVA-2, its advantages are available even for non-vectorizable algorithms. The LCC will investigate design tradeoffs in collaboration with IBM and define the final ViVA-2 architecture.

Additionally, IBM is developing [proprietary information deleted].

Based on the expertise gained from LCS-2 system design, and the extensive application knowledge represented by the application partners, we will leverage the collaborative effort to assess the most effective and timely system options for a sustained Pflop/s system. IBM currently has the most diverse HPC research portfolio of any company in the world, including: BlueGene/L, DARPA HPCS PERCS, "cell" microprocessor technology, and Osmosis optical interconnect. The current roadmap, which is from IBM's RFI response, is described in Appendix B and depicted in Figure 1. The LCC will be involved early in this process in order to drive IBM and the community to an effective Pflop/s design for state-of-the-art scientific applications.

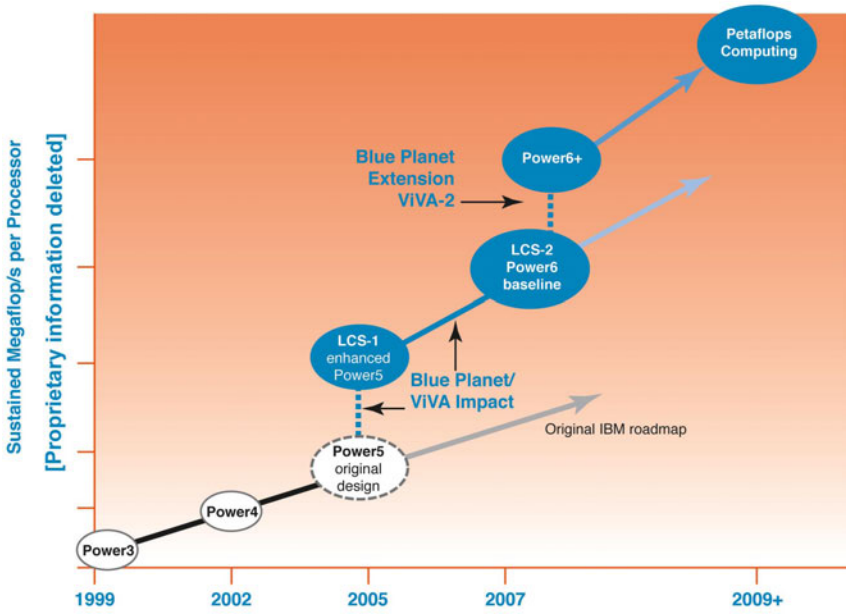


Figure 1. Science-driven architecture advancements.

4. THE BERKELEY LAB ADVANTAGE

Berkeley Laboratory enjoys an open and unrestricted intellectual environment that is easily accessible to scientists and visitors worldwide. Berkeley Lab is located in the heart of the San Francisco Bay Area, which is home to a large number of universities, laboratories, major facilities, and a vibrant scientific and research community. The Bay Area offers the critical mass of resources and intellectual leadership necessary to assemble an institution of international prominence like NFACS.

4.1 Berkeley Lab's Experience Managing National User Facilities

Berkeley Lab has been a leader in science and engineering research for more than 70 years. Located adjacent to the Berkeley campus of the University of California, Berkeley Lab is a DOE National Laboratory managed by the University of California. Many of its scientific staff hold joint faculty appointments with the University.

Berkeley Lab conducts only unclassified research across a wide range of scientific disciplines with key efforts in fundamental studies of the universe; quantitative biology; nanoscience; new energy systems and environmental solutions; and the use of integrated computing as a tool for discovery. In addition to its 17 scientific divisions, Berkeley Lab hosts four DOE national user facilities. The focus of Berkeley Lab in managing these large, leading-edge scientific facilities is to provide the best resource, service, and support to the general scientific community, both university and national laboratory based. Berkeley Lab will draw on this extensive experience to manage NFACS.

The user facilities at Berkeley Lab support thousands of users throughout the country. In addition to the Advanced Light Source (a third-generation synchrotron light source), the National Center for Electron Microscopy, and the Molecular Foundry (under construction but with an initial research/outreach program), Berkeley Lab manages two networking and computational facilities directly applicable to the management of the Leadership Class Computational Facility.

The Energy Sciences Network, or ESnet, is a high-speed network serving thousands of DOE scientists and collaborators worldwide. A pioneer in providing high-bandwidth, reliable connections, ESnet enables researchers at national laboratories, universities, and other institutions, both nationally and

internationally, to communicate with each other using the collaborative capabilities needed to address some of the world's most important scientific challenges. Managed and operated by Berkeley Lab, ESnet provides direct connections to all major DOE sites with high performance speeds, multiple high-speed cross-connection with Internet2/Abilene and the major European and Japanese research and education network, as well as fast interconnections to more than 100 other networks to provide global connectivity.

Berkeley Lab also manages the National Energy Research Scientific Computing (NERSC) Center, a world leader in accelerating scientific discovery through computation. NERSC provides high-performance computing tools and expertise that enable computational science of scale, in which large, interdisciplinary teams of scientists attack fundamental problems in science and engineering that require massive calculations and have broad scientific and economic impacts. NERSC is the foremost resource for large-scale computation within DOE's Office of Science. NERSC has built an impressive record of technological leadership, unparalleled user support, and scientific achievement. Since its early days, institutions across the country and around the world have tapped NERSC's expertise and followed its model as they work to establish their own scientific computing centers.

4.2 NERSC: A Proven Track Record Delivering High-End Computing to the National Community

NERSC is the premier open and unclassified computing facility for the Office of Science. Operating the NERSC Center has enabled Berkeley Lab to acquire unsurpassed expertise in operating large computational and storage systems, integrating them into high-speed networks, and providing comprehensive scientific support that enables researchers to make the most productive use of these resources. NERSC supports more than 2,000 users nationally and internationally. Over 50% of the users are from universities. NERSC's success is measured by the scientific productivity of its users.[1] Its staff and management are adept at balancing and satisfying the diverse needs of researchers within the constraints imposed by programmatic missions, goals, and requirements. NERSC is known worldwide for the quality of its computing services. Berkeley Lab intends to leverage this expertise and infrastructure to the maximum extent possible.

Normally, staffing a facility like NFACS from scratch requires 45 to 60 people. However, because of heavy leveraging of NERSC infrastructure and expertise, NFACS will use just 13 staff for direct support of LCS — only 20% of what is needed for a separate facility. The details of the leveraging are in Appendix F.

NERSC's user services and scientific support are highly regarded. NERSC provides a 24 by 7 help desk where it is possible for a user to talk to NERSC staff. They also produce leading-edge training and state-of-the-art documentation for NERSC systems. The NERSC support staff provides specialized services ranging from supporting unique software needs, to special processing and scheduling, to long-term collaborative interaction in order to build and optimize science codes. The support staff are highly trained (most are Ph.D.s), not just in computational methods but also in scientific disciplines.

NERSC has a sophisticated account management and allocation management system that has automated many routine tasks. It gives instantaneous access to usage data, not just for users and PIs but for DOE program managers.

NERSC proactively engages with the user community through the NERSC User Group, which meets monthly by phone and semi-annually face to face. NERSC also measures a number of quality metrics ranging from typical reliability/availability/serviceability measures to job throughput, system efficiency,

[1] For 96 pages of citations of publications resulting from computations at NERSC in 2003, see http://www.nersc.gov/research/ERCAPpubs03.htm.

and responsiveness solving problems reported by users. The most telling evidence is the annual NERSC User Survey that assesses user satisfaction.[2] One user characterized the quality of NERSC as follows:

> *"NERSC simply is the best-run centralized computer center on the planet. I have interacted with many central computer centers and none are as responsive, have people with the technical knowledge available to answer questions, and have the system/software as well configured as does NERSC." —2003 NERSC User Survey Respondent*

4.3 Intellectual Resources

Berkeley Lab: A Proven Track Record in Computational Science and Computer Science

NFACS will directly benefit from the pool of talent available in the two research departments of the Computational Research Division at Berkeley Lab. There are also increased opportunities for technology transfer from other DOE/OASCR-funded projects elsewhere at Berkeley Lab, especially the Scientific Discovery through Advanced Computing (SciDAC) programs led by Berkeley Lab PIs.

- **The High Performance Computing Research Department (HPCRD)** addresses long-term research and development questions in HPC. With more than 125 staff and expertise in computer science, computational science, and applied mathematics, HPCRD can provide additional resources and talent for the advanced development needs of NFACS and for focused high-end support of the application areas.
- **The Distributed Systems Department (DSD)** focuses on issues in distributed computing, Grid technologies, networking research, collaborative tools, and security. With more than 25 staff, DSD develops and prototypes technologies and testbeds to facilitate solving scientific problems that require complex and large-scale computing and data handling environments involving geographically and organizationally dispersed components. DSD can provide additional resources and talent for enabling the distributed infrastructure for NFACS applications areas.
- **SciDAC Centers:** Berkeley Lab is the leader of four SciDAC centers and eighteen SciDAC projects. NFACS will leverage the activities of these projects, in particular the APDEC and PERC Integrated Software Infrastructure Centers.

The University of California reinvests the laboratory management fee in projects that foster collaborations between the laboratories and campuses. NFACS has access to these funds in support of the LCC collaborations. We anticipate at least one FTE will be supported by those funds.

University of California, Berkeley

Berkeley Lab's location, only a short walk or shuttle bus ride away from the campus of the University of California at Berkeley (UC Berkeley), facilitates numerous formal and informal collaborations. Currently, there are seven joint appointments of faculty from the EECS and Math Departments at UC Berkeley with Berkeley Lab Computing Sciences: David Culler, James Demmel, Susan Graham, Ming Gu, Arie Segev, Jonathan Shewchuck, and Katherine Yelick. Jim Demmel is also the Chief Scientist for the Center for Information Technology Research in the Interest of Society (CITRIS), a four campus, 200+ faculty research institute centered at Berkeley. The combination of NFACS facilities and Berkeley Lab and campus computing efforts creates a vibrant community for cross-institution and cross-discipline efforts in research in algorithms, architectures, and applications, and in training of future computational scientists (for details see Appendix D).

[2] For the latest survey results, see http://hpcf.nersc.gov/about/survey/2003/.

Katherine Yelick leads the Berkeley Unified Parallel C (UPC) team, a collaborative effort centered at Berkeley Lab, which is working to produce more efficient and productive programming models. Yelick is also working with Berkeley Lab scientists on the evaluation of advanced architectures for scientific computing, including processor-in-memory, streams, VLIW, and vectors. This architecture evaluation team worked closely with IBM in the early stages of ViVA design to understand the benefits and limitations of vectors, and what type of memory system was needed to support the more challenging DOE applications. The close research interactions between the UC Berkeley campus and Berkeley Lab have had tremendous impact on DOE science and technology development thus far. Locating NFACS in close proximity to these intellectual resources will ensure that these benefits are conferred to the national user community.

4.4 Infrastructure and Capabilities

Networking

Berkeley Lab and NERSC are located near the primary switching point for national networks in Northern California at Sunnyvale — home to both the Qwest and Level3 networking hubs. The Qwest hub is the transit point for the backbones of major production networks such as DOE's ESnet, NSF's Abilene, NASA's NREN, and the NSF TeraGrid, while the Level3 hub carries experimental dark-fiber networks such as the National Lambda Rail, the DOE Ultranet, and the CENIC/Pacific Light Rail. The proximity allows NFACS easy and cost-effective access to each of these networks. In order to promote interaction with and outreach to scientists in industry, academia, and other federal programs, Berkeley Lab will work closely with ESnet to create network peering arrangements that will maximize the effective remote access to NFACS users regardless of their institutional affiliation and facility location.

In order to ensure the highest performance network access to NFACS, Berkeley Lab will immediately upgrade its connection with Sunnyvale to OC-192 so as to match the existing backbone bandwidth of the ESnet and Abilene production networks. In order to provide more effective access to the NSF user community, ESnet and Abilene are implementing high-speed peering between their networks at each of these co-located hubs at Sunnyvale, Chicago, New York, and Atlanta to create a common network backplane that provides very high-speed connectivity between the labs and universities, comparable to what either backbone alone can provide among their primary sites. In addition to its support of production network infrastructure, the NFACS system will tie in to major experimental and dark fiber networks, such as the TeraGrid, DOE Ultranet, and National Lambda Rail, in order to add its capabilities to a vibrant research community that combines sensors, archival data, and supercomputers to accomplish large multidisciplinary scientific projects. These upgrades coincide with Berkeley Lab's migration to a 10 Gigabit internal network infrastructure, which is already under way. Both the upgraded internal network and wide area network infrastructure will be immediately available to the first generation NFACS system and will continue to be expanded to match the scale of successive systems and continuously match the performance improvements of the production network backbones.

In the first year of NFACS operation, ESnet will deploy an MPLS-based QoS service that operates initially between ESnet border routers which will be expanded within 2 years to allow dynamic provisioning of circuits across both Abilene and ESnet as envisioned by the Internet2 Hybrid Optical/ Packet Infrastructure (HOPI) working group. These "bandwidth corridors" will support NFACS global file system (WAN GPFS) and storage peering arrangements between other laboratories and Leadership Computing Consortium Partners such as the National Science Foundation's Partnership for Advanced Computational Infrastructure (NSF-PACI) supercomputing centers in order to support our vision of a nationwide supercomputing infrastructure.

Systems Management

The NFACS management team will draw on the expertise of NERSC, and then customize the support model for the LCS systems to be more tightly integrated with the selected science projects. NERSC already manages three distinct systems with different user communities and requirements.

The LCS systems will be operated in a dynamic manner in close collaboration with the users. NFACS expects to support approximately 20 projects. NFACS will provide a custom and flexible environment that supports the unique requirements of each project — not just for system management but for all support functions. For example, libraries and middleware will be selected and installed in close collaboration with the projects. Users will be able to request to use large amounts of the resource interactively for debugging and computational steering. At times, it will be possible for a user to be given the entire system in a dedicated manner. Because there is a smaller, more manageable set of projects, NFACS staff will be able to coordinate the system scheduling to meet computational science project goals in a custom manner.

NFACS will involve users in the discussion of system management changes — in particular queuing, priorities, and disk-space management — through the monthly conference calls. The Leadership Computing Applications Team (LCAT) points of contact will be the advocates for the computational science applications areas and will communicate their requirements.

Data Storage and Archives

NERSC's High Performance Storage System (HPSS) has enough capacity to serve both NERSC and NFACS clients. NERSC currently stores approximately 1,050 TB of data (30 million files) and handles between 3 and 6 TB of I/O per day. The current maximum capacity of NERSC's archive is 8.8 PB at current tape densities; the buffer (disk) cache is 35 TB; and the maximum transfer rate is 2.8 GB/s. NFACS will require large amounts of archival storage and will invest in new tape technology. For LCS-1, 500 GB tape drives and cartridges will be added to the NERSC HPSS, giving a total maximum capacity of 4.5 PB just for LCS-1. For LCS-2, 1 TB cartridges will be deployed, adding 5 PB a year (10 PB total for the time period of the proposal) to the potential NFACS storage capability, for a total of 15 PB of storage.

The NFACS HPSS will be federated with archival storage systems (both HPSS and Unitree) across all sites involved in the LCC. Users of the LCS systems will have equal access to archival data across NFACS, NERSC, and NSF-PACI facilities through the LCC storage federation. Close coordination of certificate management between DOE Science Grid, TeraGrid, and NSF-PACI sites will enable single-sign-on access across facilities and seamless transfer of data between archival storage systems. Also, the "bandwidth corridors" described above will support dedicated high-speed data transfers between the sites for efficient mirroring and staging of massive datasets between their respective storage systems.

In addition to archival storage systems, NFACS will be part of a wide-area shared file system that will link together all LCC partner sites including the NSF-PACI supercomputing centers, NERSC, and Louisiana State University (LSU). The file system will be based initially on WAN GPFS, which is being developed through a partnership between IBM Research and the San Diego Supercomputer Center (SDSC), and will be usable across both Linux and IBM-SP supercomputing infrastructure at LCC partner sites. In demonstrations conducted by SDSC this past year, GPFS sustained well over 900 MB/s over a wide-area 10 Gigabit link. The shared file system will enable more flexible migration between the systems for users who have shared accounts and will help the LCC consortium form a well-integrated computing environment that better serves a national scientific user community

Grids

As the home of ESnet and NERSC, the lead site for the DOE Science Grid, and one of the original six development sites for the HPSS, Berkeley Lab has already made significant progress in integrating high-

end computing, storage, and data management into the Grid environment. We will do the same for NFACS, thereby facilitating large-scale science for DOE and the nation. NERSC has established ties with all major Grid efforts in DOE and NSF and is closely collaborating with the DOE Science Grid and all its partners. The NFACS center staff will leverage the NERSC Center staff's broad experience with Grid software and services. We will work in close coordination with the LCC members to establish the peering of Certificate Authorities and trust relationships necessary to support coordinated access to Grid services. An interface to the NERSC Information Management (NIM) system will make it easy for NFACS users to get Grid authentication certificates. Coordinated management of Grid certificates will support single-sign-on access to Grid services across all LCC partner sites including the NSF-PACI centers, NERSC, and the TeraGrid Consortium.

Visualization and Data Analysis

High-end visualization and data analysis tools will be essential to turn raw simulation data into scientific discoveries. NFACS will work closely with its LCC partners to apply technologies developed across the coalition and make them available to the user community. In particular, we will work closely with LLNL to share, test, debug, and deploy the latest ASCI tools for visualization of massive datasets, including commercial technologies that offer new levels of graphics performance, the LLNL/Stanford-developed distributed parallel rendering software (Chromium), and proven parallel, scalable end-user applications (like VISIT and Blockbuster movie player), and the Terascale Browser. The Berkeley Lab/NERSC visualization group will also provide LCS users and LCC members with access to the VisPortal, which automates complex workflows like the distributed generation of MPEG movies or scheduling of file transfers, mediates access to limited hardware resources like off-screen graphics pipes, and controls the launching of complex multicomponent distributed visualization applications like Berkeley Lab's Visapult — an application used for remote and distributed, high performance interactive volume rendering of massive remotely located datasets. All of these tools will be tightly coupled with the high-speed networks, coordinated Grid services, storage federation, and WAN GPFS capabilities deployed across the LCC sites. This powerful set of tools and services will enable LCS users across the nation to rapidly understand the enormous amount of data they generate at NFACS. Without tools of this caliber and computer scientists available to support these tools, the huge data generation engines that NFACS will be deploying would be less useful.

Security

As an unclassified facility, Berkeley Lab makes its facilities available for use by investigators from institutions throughout the nation and the world. To sustain its scientific mission, Berkeley Lab protects its resources and assets, both intellectual and material. Only necessary technical staff have access to computer rooms and computer facilities. The general staff and the public do not have physical access to these computer resources. All Laboratory assets are tracked and protected by Laboratory security services. NFACS users will access the system remotely, subject to all Berkeley Lab cyber security policies, controls, and restrictions.

Berkeley Lab and NERSC have an outstanding security record and are recognized as leaders in cyber security within the DOE and beyond. This expertise will make NFACS both secure and easily accessible. In order to maximize our ability to conduct science and mitigate the effects of computer security incidents, the Laboratory provides non-invasive advanced monitoring and automatic reactive tools using components that are embedded in the network as well as in every computational and storage system. Berkeley Lab's active security infrastructure is able to detect cyber attacks, detect vulnerable or compromised hosts, and initiate a large-scale coordinated response to cyber-security incidents without resorting to methods that impede legitimate system access. For example, firewalls are creating significant roadblocks to pervasive deployment of Grids. Berkeley Lab's active intrusion detection system offers a compelling alternative to standard firewalls as a means to defend against cyber attacks. DOE is funding efforts to extend this system to sites other than Berkeley Lab. The Laboratory will continue to use and

improve these advanced monitoring tools to provide NFACS with the best level of security with minimal impact on performance and function.

4.5 Building and Physical Infrastructure

Berkeley Lab's Oakland Scientific Facility (OSF) includes a 20,000-square-foot computer floor. Currently, there are 5,000 square feet of computer floor available for an additional system. The LCS-1, described in this proposal, will require 2,400 net square feet and will readily fit in the existing OSF.

The follow-on system, LCS-2, will be housed in a new dedicated computer building in the center of the Berkeley Lab campus on a cleared site adjacent to the Bevatron, whose external beam hall was recently demolished. The 20,000-gross-square-foot building will contain a computer room and utility support space. This cleared site also provides for the ability to expand into a second adjacent 20,000 square feet of computer floor, yielding a 40,000-square-foot computer complex. Site plans and conceptual building renditions are shown in Appendix E.

Recently, the DOE has encouraged third-party financing approaches to facilities construction, and these approaches will enable Berkeley Lab to provide the requisite NFACS building to accept delivery of LCS-2. Because Berkeley Lab is located on University of California-owned land, this process is actually less complicated for Berkeley Lab than for those national laboratories situated on federal land, which must be transferred via a quitclaim deed to a development entity. The University can simply enter into a long-term ground lease with a developer at a nominal cost. When the building is complete, DOE can approve a UC lease of the facility a year at a time over the life of the building. Berkeley Lab and the University are currently developing a research office building on the main Lab campus, targeted for completion in 2006, through a third-party development. The experience and knowledge gained from this procurement give us every confidence that the NFACS building can be completed on time.

The contingency plan for housing LCS-1 and LCS-2 is expansion of the OSF to gain another 20,000 square feet. The OSF was designed for such a contingency, which can be exercised in time for LCS-2.

Berkeley Lab, therefore, has existing and committed space for NFACS.

5 BUILDING A NATIONAL LEADERSHIP COMPUTING CONSORTIUM

As potentially the largest open computing resource in the nation, NFACS will provide leadership-class computing capability to scientists and engineers nationwide, independent of their institutional affiliation or source of funding. Berkeley Lab has therefore established a national Leadership Computing Consortium (LCC) comprising some of the leading high-end computing centers of the nation, including the NSF-PACI supercomputing centers.

LCC has two functions:

- Technology development: LCC will be the main vehicle for implementing the science-driven architecture development. LCC will engage the major vendor partner, IBM, in an ongoing dialogue of science-driven architecture development.
- National facility operations: LCC will be the vehicle to establish close connections and strategic collaborations with computer science programs and facilities funded by the DOE Office of Science (SC), the National Nuclear Security Administration (NNSA), NSF, and NASA, as well as universities.

Berkeley Lab and the NSF-PACI sites, along with large university computing facilities such as the Louisiana State University Center for Computation and Technology, will forge a close-knit relationship, ensuring that the NSF user community has unencumbered access to NFACS. Recognizing that the typical

workload on a supercomputer follows a power-law-like curve of job sizes in order to satisfy users' development, data analysis, and post-processing needs, LCC members will establish a national computing fabric that will lower barriers to user migration and resource sharing between computing facilities. In particular, LCC sites will define systems that support coordinated access to accounts, federate archival storage devices across sites, establish a federated parallel file system (WAN-GPFS) that spans the U.S., and tie all of these services together with high performance network services to move data between all of these components. NFACS will provide funds to establish these working relationships. The goal is seamless migration across the U.S. computational infrastructure. LCC sites will also collaborate to jointly develop system documentation, mutual training, and support mechanisms, to conduct detailed performance analysis of applications, and to contribute to the direction of future systems development, drawing on their years of combined experience supporting a national user community. This collaboration will greatly reduce duplication of effort and free up resources to ensure that the U.S. supercomputing infrastructure will provide the highest quality platform for advanced scientific applications.

Initial LCC members are listed below. This list is by no means final, and new members will be invited to join as opportunities for collaborations arise in the future. In particular, we will encourage all DOE-SC laboratories to join the LCC. Similarly, the LCC is open to other vendor partners in addition to IBM. The envisioned national computing fabric will lay the foundation for continued U.S. scientific preeminence.

5.1 Charter Members of the LCC

The charter members of the LCC are listed below. Their initial contributions to and collaborations with NFACS are discussed in Appendix D.

- IBM (Vendor Partner)
- Lawrence Livermore National Laboratory (LLNL)
- NSF-PACI: National Center for Supercomputing Applications (NCSA) and San Diego Supercomputing Center (SDSC)
- Argonne National Laboratory (ANL)
- National Center for Atmospheric Research (NCAR)
- NSF TeraGrid
- Louisiana State University Center for Computing and Technology (LSU-CCT)
- NASA Goddard Space Flight Center (GSFC)
- Pacific Northwest National Laboratory (PNNL)

Membership in the LCC is open to the computational science community, and charter members obtain no privileges not available to others who may wish to join later.

5.2 Leadership Computing Applications Teams (LCATs)

NFACS has identified five computational science applications areas that will require a leadership-class computing capability to make major computational advances: nanoscience, combustion, fusion, climate, and astrophysics. In each of these applications, project teams have been assembled who will collaborate with NFACS to accomplish their computational goals. These projects include a total of 21 research groups from 17 different universities and research laboratories. (See Appendix D for a complete list.) There is no promise of any privileged access to these applications scientists, and a competitive, peer-reviewed access and use process will be used (see section 6.4).

NFACS will interface with the applications areas in several ways. One NFACS staff computational scientist will be assigned to each of the applications areas as point of contact (POC, see Appendix H). The POC will develop a deep understanding of the algorithmic techniques and computational requirements of the applications areas and will communicate these to NFACS and to the vendor partner in the quarterly

NFACS progress meetings. This input from the science community is an important element in the process of driving future technology developments.

5.3 Communications and Outreach

NFACS, as potentially the largest open computing resource in the nation, has developed collaborations with computational scientists in universities, research labs, and industry. In order to maximize the dissemination of information, and to promote and support computational science and computer technology for high-end computing, NFACS has far-reaching plans for collaborations, outreach, and dialogue with stakeholders. In the areas of technology development, NFACS will engage its major vendor partner, IBM, in an ongoing dialogue of science-driven architecture development. The LCC will build on close connections and strategic collaborations with computer science programs and facilities funded by DOE-SC, DOE-NNSA, NSF, and NASA, as well as universities.

The following events will facilitate this outreach:

1. *Monthly meeting/conference call with LCATs*. Leadership applications teams will hold a monthly conference call with NFACS staff to discuss operational issues, progress towards system and software deliverables, applications porting and performance issues, etc.
2. *Quarterly progress meetings*. NFACS staff, LCC members, the vendor partner, and LCAT representatives will meet quarterly to report on progress with their tasks. The quarterly progress meeting will also serve as the main communication mechanism for the implementation of the science-driven architecture development.
3. *Annual "all-hands" meeting.* NFACS will organize an annual event that will be open to all stakeholders and the community at large. It will include scientific presentations from LCAT members, updates from the vendor partner, and computer science and technology presentations from the LCC members and NFACS staff.
4. *Workshops and planning meetings.* As new and important topics arise, NFACS will hold workshops and planning meetings for interested stakeholders.

5.4 NFACS Education and Workforce Development

NFACS will develop a leadership computing community through integrated educational and training components that build skilled computational scientists, with a focus on graduate and undergraduate students. Outreach to underrepresented students will be integral to building the educational pipeline. The education program will include:

- Seminars on leadership computing capabilities targeted to specific research topics
- Short courses on specific computational topics
- Consulting services, including course assistance to ensure up-to-date user information
- Web site resources with comprehensive technical, information, and course content
- Internships available to qualified applicants for summer and semester appointments
- Graduate and undergraduate research on all phases of leadership computing
- Faculty sabbaticals to update computing courses and curriculum

A key benefit of the NFACS Consortium is the combined educational resources of members, such as the internationally recognized education program in computing science and applied and computational mathematics at UC Berkeley. The NFACS ties to the NSF-PACI supercomputing centers (SDSC and NCSA) and the NSF TeraGrid consortium bring national university educational resources to bear on training and future workforce development. NFACS will provide training and internships for the DOE Workforce Development for Teachers and Scientists program.

6. MANAGEMENT PLAN

NFACS will be managed as a national scientific resource with full and complementary support to the programs of the Office of Science and mission of the U.S. Department of Energy. Facility management will focus on sound annual planning, cost-effective line management, comprehensive review, and a highly consultative management advisory framework. The management system will be coupled to the Office of Science program evaluation process for program oversight and Laboratory management. The management efforts will reinforce the NFACS mission of demonstrating continued U.S. leadership in computational science through performance at the largest scale of computational problems. NFACS will be planned and implemented though a Project Management Framework to assure the completion of facilities components on schedule, scope, and budget in a manner that is consistent with all affected stakeholders.

6.1 The NFACS Management and Organization

NFACS will be led by Horst Simon, Associate Laboratory Director for Computing Sciences and High Performance Computing Facilities (HPCF) Division Director. As NFACS Director, he will be accountable for all aspects of the NFACS program. The Director will recommend strategic programmatic directions and the development of programmatic ties to other laboratories, universities, and industrial partners, as appropriate. The Director will be responsible for scientific productivity and maintaining the leadership role of the facility.

Division directors at Berkeley Lab are direct appointees of the Laboratory Director and are members of the Director's planning team, participating in the Laboratory's oversight and review activities. NFACS will have direct access to the Laboratory Director and high visibility at the Laboratory. The organizational arrangement for NFACS is similar to that of the Advanced Light Source, the National Center for Electron Microscopy, the Molecular Foundry, ESnet, and NERSC, the other major national user facilities at Berkeley Lab.

6.2 Key Technical Personnel

The NFACS Director will be supported by a management team composed of a General Manager, the LCS Team Leader, the LCS Lead System Analyst, the LCS Lead Performance Analyst, and the LCS Lead Scientific Support Analyst. The Leads report to the General Manager and will work with other LCS staff to carry out their responsibilities.

The General Manager, William Kramer, reporting to the NFACS Director, is accountable for the NFACS facility, with management responsibility for planning, budgets, enhancements, personnel, vendor and user relations, physical resources, and program and operational integration.

The LCS Team Leader, William Saphir, reporting to the General Manager, is responsible for the development, management, and operations of computing, storage, and networking resources, as well as the support needed by the user community.

The LCS Lead System Analyst, Nicholas Cardo, working with the LCS Team Leader, is responsible for the deployment, management, and operations of computing resources.

The LCS Lead Performance Analyst, David Skinner, working with the LCS Team Leader, ensures that NFACS application and systems performance improvements are determined and implemented.

The LCS Lead Scientific Support Analyst, John Shalf, working with the LCS Team Leader, ensures that NFACS scientific support meets the needs of the user community. He coordinates the Leadership Computing Application Team points of contact.

6.3 National Oversight and Policy

The NFACS management will meet with leadership of the Office of Advanced Scientific Computing Research (OASCR) and with the Mathematics, Information and Computing Sciences (MICS) program management staff. The NFACS management will work with MICS leadership to develop program plans and budgets and to facilitate periodic OASCR and other national reviews of the NFACS program.

The Berkeley Laboratory Director will conduct an annual review of NFACS, which will include an assessment of NFACS long range planning, staffing, quality of programs and operations execution. The review will be conducted by the NFACS Policy Board, which will be appointed by the Laboratory Director in consultation with OASCR. The Board will consist of leading representatives of the high performance computing community, and will provide advice on strategic issues and policy directions to both the Laboratory Director and the NFACS Director. The current members of the NERSC Policy Board (Appendix G) will be invited to serve on the NFACS Policy Board.

6.4 Allocation Review Process

In 2003, DOE initiated a new program entitled Innovative and Novel Computational Impact on Theory and Experiment (INCITE) at NERSC. INCITE awarded 4.9 million supercomputer processor hours and corresponding data storage space at NERSC to three computationally intensive large-scale research projects, with no requirement of current DOE sponsorship. A peer review process for all proposals was established, which involved a Web-based proposal submission system, a review panel of about 50 scientists, and a well-defined process for evaluating both the scientific goals and the computational methods and techniques of the proposal. The NFACS allocations process will be modeled after the successful INCITE program, leveraging both infrastructure and the reviewer pool. The existing Energy Research Computational Application Program (ERCAP) and NERSC Information Management (NIM) systems will be used to implement the mechanics of the allocation process. ERCAP and NIM will need only minor modifications to provide a unique “look and feel” for the NFACS systems and users.

7. BUDGET

[Proprietary and confidential information deleted.]

APPENDIX A
LCS-1 and LCS-2 System Architecture

This appendix provides a detailed description of the system architecture for Leadership Class Systems LCS-1 and LCS-2 as well as an overview of the proposed Virtual Vector Architecture (ViVA) enhancements.

A.1 IBM Proposed System Design for LCS-1 and LCS-2

On March 31, 2004, IBM submitted the document reproduced on the following pages to support the Berkeley Laboratory National Facility for Advanced Computational Science (NFACS) proposal to the Office of Science to supply a leadership-class computing capability system. Over the past five years, IBM has provided Berkeley Lab and the National Energy Research Scientific Computing Center (NERSC) with superior large-scale computing resources to support the nation's scientific discovery efforts. IBM will continue to be a strong technology partner for the leadership-class computing systems that drive scientific research and discovery. By working collaboratively, IBM and Berkeley Lab will be able to raise the threshold of new capabilities to support the research scientific community. We believe the technologies discussed in this response will assure the optimal balance of highly productive systems with minimal technical and commercial risk.

[Proprietary information deleted.]

A.2 ViVA Technical Overview

[Proprietary information deleted.]

A.3 Milestones and Deliverables

The schedule below assumes that an award for the leadership computing system will be made on April 15, 2004.

LCC Quarterly Progress Meetings

These meetings combine two different sub-meetings:

NFACS status meeting
BluePlanet architecture meeting

Every fourth meeting (in July) will be an annual all-hands meeting

07/2004 Kickoff meeting with LCC
10/2004 Quarterly meeting
01/2005 Quarterly meeting
… will be scheduled every three months through
10/2008 Quarterly meeting

LCC Milestone
12/2005 ViVA-2 Design complete

LCAT

01/2005 Monthly LCAT meetings (plus semiannual larger ones).

Systems

06/2005	LCS-1 Installation (1/4 system)
07/2005	LCS-1 First user access
09/2005	LCS-1 Installation (remainder of system)
10/2005	LCS-1 First user access
11/2007	LCS-2 Installation.
01/2008	LCS-2 First user access

Infrastructure

12/2004	Distributed Terascale Facility network connection available for testing
01/2005	WAN GPFS Demonstration
01/2006	Performance enhancements to WAN GPFS
02/2005	10 Gb/s Ethernet infrastructure in place
06/2005	HPSS upgrade to 500GB tape infrastructure
07/2006	Computer floor space available for LCS-2
03/2007	HPSS upgrade to 1 TB tape infrastructure

Contracts

10/2004	Detailed statement of work with IBM for LCS-1
02/2007	Detailed statement of work with IBM for LCS-2

APPENDIX B
IBM Response to Berkeley Lab's Request for Information on High-End Computing Systems, February 24, 2004

[Proprietary information deleted.]

APPENDIX C
Performance Analysis of LCS-1 and LCS-2 Systems

Summary

This Appendix presents a quantitative analysis of LCS-1/LCS-2 performance and a comparison to possible alternatives in the Cray vector line, using the best available data.

The overall conclusion is that per-processor performance of the LCS-1/LCS-2 systems is comparable to per-processor performance in the Cray alternatives at significantly lower cost. This performance result arises from two factors:

- IBM has reengaged with the high performance computing (HPC) market, and the results are paying off in processors that perform much better than their predecessors.
- The performance of Cray processors, while good, is not as good as is commonly thought.

[Proprietary information deleted.]

APPENDIX D
Leadership Computing Consortium (LCC) and Leadership Computing Applications Teams (LCAT)

This appendix provides detailed descriptions of the collaborative efforts to be conducted by the Leadership Computing Consortium (LCC) and Leadership Computing Applications Teams (LCAT). Section D.3 of this appendix reproduces letters of support from collaborators.

D.1 Leadership Computing Consortium (LCC)

Charter members of the LCC and their agreed-upon contributions and collaborations with National Facility for Advanced Computational Science (NFACS) are listed below.

IBM (Vendor Partner)

As the vendor of the Leadership Class System, IBM is committed to providing an innovative, extremely effective, high performance computing system of unprecedented performance and capability, and will be a major partner of NFACS. IBM will continue its innovation to reduce latency and to improve the efficiency of science codes. A wide range of IBM technologies will be brought to bear to assure that NFACS is an outstanding success. The scale, schedule, and requirements for LCS mean that the standard roadmap of technology is not sufficient. IBM has already adopted the concepts of science-driven architecture design in redesigning the Power5/Power6 node to focus on the balance of flop/s and memory bandwidth. This new eight-CPU single-core node is already on the product roadmap and is the basis of the LCS-1 system. IBM will continue the science-driven design approach in collaboration with Berkeley Lab as the details of the LCS-2 system are defined and implemented.

The concept of virtual vectorization — accelerators that will improve the efficiency of codes while still leveraging the cost-effectiveness and balance of IBM's high-volume CPU cores — is something IBM will pursue. IBM will work with NFACS staff and the application-area teams to complete the design of this effort and create effective methods to exploit the new functionality. [Proprietary information deleted.]

In order to monitor progress towards these goals, obtain applications input, and communicate accomplishments and new technology, IBM and LCC will hold quarterly progress meeting. Attendance at these meetings will be open to NFACS staff, LCC members, and computational scientists from the applications areas.

Lawrence Livermore National Laboratory (LLNL)

LLNL has an existing collaboration with IBM to field the ASCI Purple and Blue Gene/L systems. At their introduction, these will be the most powerful systems in the world. LLNL is looking back over a ten-year history of fielding some of the most powerful and innovative systems, often the first of their type. Berkeley Lab and LLNL have collaborated in many ways in the past, most recently in the design of the 8-way node. LLNL will bring the following elements into the LCC:

- Collaboration in standing up and operating the next generation of IBM platforms. With ASCI Purple, of similar design to LCS-1, being installed first, NFACS will be able to learn from the LLNL experience. NFACS and LLNL will exchange staff: staff from Berkeley will work side by side with LLNL staff when ASCI Purple comes on line, and vice versa. In the long term, after LCS-1 is on-line, LLNL and NFACS agree to share operational information, e.g., trouble-tickets etc.
- Share LLNL's planning documents for storage-area network (SAN) architecture, including I/O Blueprints. We will continue to work together with the high Performance Storage System (HPSS)

consortium, using our collective leverage with IBM and our combined HPSS development staffs to assure that the appropriate solutions are rapidly written into the HPSS releases.

- Share, test, debug and deploy together the latest ASCI tools in visualization, including utilization of commercial technologies to achieve new levels of graphics performance, the distributed parallel rendering software stack (Chromium), parallel, scalable end-user applications (like VISIT and Blockbuster movie player) and the blueprint for future Purple 100 TF-related visualization deployment.
- As members of the BlueGene/L (BGL) Consortium, Berkeley Lab will work together with LLNL to evaluate the appropriateness of the BlueGene/L and the BG-family (BG/P follow-on architecture) as a leadership-class investment later in this decade by the Office of Science. BlueGene/L represents a $100M R&D investment by IBM in a machine for science, and employs three separate networks to enhance efficiency and an extremely low power system on a chip design. The results of this shared evaluation effort will likely drive changes in both the BG/P and LCS-2 designs and will have significant importance in defining the road to petaflops. IBM will make available to Berkeley Lab our SLURM and LCRM fair-share scheduling and node-packing software, should Berkeley Lab choose to employ this solution rather than native software.
- Staff from LLNL who are active in the ASCI program will be part of the quarterly progress meetings that LCC will have with IBM.

PACI: National Center for Supercomputing Applications (NCSA) and San Diego Supercomputing Center (SDSC)

Berkeley Lab and the NSF Partnerships for Advanced Computational Infrastructure (PACI) sites will forge a close-knit relationship, ensuring that the NSF user community has unencumbered access to NFACS. Berkeley Lab and the NSF centers will share the effort of supporting and training this diverse nationwide scientific community. Berkeley Lab and PACI will collaborate on development of system documentation, training, and user support, drawing on their years of combined experience supporting a national user community. This collaboration will greatly reduce duplication of effort and free up resources to ensure that the U.S. supercomputing infrastructure will provide the highest quality platform for advanced scientific applications.

In addition, our collaboration will involve the following arrangements to form a virtual machine room that supports seamless migration between PACI and DOE systems for users who have allocations on both the NFACS and NSF centers:

- Accounts on archival storage systems (HPSS and Unitree) to provide equal access to archival data across NFACS and PACI facilities.
- Peering of Grid certificate authorities and coordination of certificate management between DOE Science Grid, TeraGrid, and PACI sites, in order to enable single-sign-on access across facilities and seamless transfer of data between archival storage systems.
- A fast path for data migration between centers using "bandwidth corridors." The bandwidth corridors allow scheduled dedicated access to high bandwidth channels between sites for efficient mirroring and staging of massive datasets between mass storage systems using fixed data-rate protocols.
- Federated Global Parallel File System (GPFS) file systems across systems at NCSA, SDSC, NERSC, and the NFACS system. The software support for wide-area GPFS support for both Linux clusters and SP systems is already in progress (see below). Such a shared file system will enable more flexible migration between the systems for users who have shared accounts.

These arrangements will form a U.S. supercomputing infrastructure that spans the continent.

It is of critical importance that the NFACS system is focused on the largest possible jobs in order to fulfill its role as the preeminent leadership-class system for the U.S. computing infrastructure. In practice, the typical workload on a supercomputer follows a power-law-like curve of job sizes in order to satisfy users' development, data analysis, and post-processing needs. Our PACI partners will investigate Load Leveler configurations that will support routing of smaller jobs to appropriately sized systems across the LCC sites. The tightly integrated operating environment and storage federation offered by LCC ensures that NSF users who have access to NFACS can focus their NFACS allocation on the jobs that scale to the system's full capability, thereby ensuring that NFACS will be focused on the largest capability computing applications that exploit its full potential.

Wide-Area GPFS

The San Diego Supercomputer Center (SDSC) is recognized as a world leader in data-oriented computing with deep expertise at all levels of the "data stack" — from storage and file systems to data-oriented services, visualization, and application portals. The coupling of the NFACS system with SDSC's leading storage expertise and facilities provide an immense opportunity to develop a national facility that will advance both scientific discovery and technology leadership for the U.S.

There is a particular opportunity to work with SDSC on wide-area-network (WAN) global file systems. At the SC02 meeting in Baltimore, SDSC used Nishan FCIP protocol converters to link Sun's QFS file system across an extended storage area network between San Diego and the show floor. Transfer rates of over 700 MB/s were demonstrated for the fastest transfers of that type so far. Between then and the recent SC03 meeting in Phoenix, SDSC worked with IBM to make a pre-release version of their GPFS software WAN-compatible and linked a fast GPFS file system at SDSC with a Linux cluster in the SDSC booth. These transfers went directly across a 10 Gb/s link without the necessity for protocol converters and sustained a remarkable 900 MB/s, winning that portion of the SC Bandwidth Challenge.

Present releases of GPFS require a unified user identification (UID) space between the participants; it is clear that for widespread adoption in the Grid community, a more flexible form of authentication is required, potentially via Grid Security Infrastructure (GSI). SDSC is leading an initiative with the IBM GPFS development group on WAN GPFS authentication, with the purpose of integrating Grid-type authentication into GPFS. SDSC's leadership in this area, and their partnership with TeraGrid, the European DEISA activities, and others, will provide an innovative environment in which to store, manage, and manipulate the data from the NFACS system and other facilities.

Argonne National Laboratory (ANL)

There is a long history of collaboration between Berkeley Lab and Argonne involving data Grids, visualization, PC clusters, collaboratories, and applied mathematics. In the case of data Grids, ANL has been a leader since the I-WAY demonstration at SC95. In recent years, Berkeley Lab and ANL have been partners in a number of DOE projects, including the Particle Physics Data Grid (PPDG), the Earth Systems Grid (ESG), the DOE Science Grid, and the Programming Methods Center. For the NFACS facility, we intend to deepen these collaborations in the areas of computer architecture evaluation; and modeling, simulation, and real-time data analysis relating to large-scale problems in life sciences and nanoscience. Argonne is interested in developing new interactive techniques appropriate for leadership-class computing systems that go beyond today's batch-computing modalities.

In particular, Berkeley Lab has been invited as a founding partner in the Argonne-led Blue Gene consortium. As members of the Blue Gene Consortium, Berkeley Lab will work together with ANL and LLNL to evaluate the appropriateness of the BlueGene/L and the BG-family (BG/P follow-on architecture) as a potential leadership-class investment later in this decade by SC. The results of this shared evaluation effort will likely drive changes in both the BG/P and LCS-2 designs and will have significant importance in defining the road to petaflop/s.

National Center for Atmospheric Research (NCAR)

The National Center for Atmospheric Research (NCAR) is operated by the University Corporation for Atmospheric Research under the sponsorship of the National Science Foundation (NSF) and other agencies. The Climate and Global Dynamics (CGD) Division has principal responsibility for the development of the Community Climate System Model (CCSM). Principal funding for CCSM is provided by the NSF, with supplementary funding from the DOE's Climate Change Prediction Program. Collaborations between NCAR and DOE laboratory scientists are broad and include both formal and informal arrangements. Simulations performed at NFACS would involve multiple institutions and be considered a community resource. Climate-model output data is currently distributed from the NERSC HPSS system. In the near future, the current dataset will be distributed using ESG technology. NFACS-produced climate-model data will be distributed in a similar manner.

Louisiana State University Center for Computing and Technology (CCT) and Louisiana Optical Network Initiative (LONI)

As part of this partnership, the facilities of the Louisiana State University Center for Computing and Technology (CCT), including a 1,024-processor Intel Xeon Linux cluster, will be integrated with the LCC partner sites, including the NFACS and PACI computing centers, to create a computing infrastructure that supports transparent access across the LCC sites. CCT will work closely with LCC partners to support coordinated account creation and Grid certificate-management services. The Louisiana Board of Regents has recently approved funding to become a founding member of the National Lambda Rail (NLR), which will place an NLR access point in Baton Rouge and enable CCT to participate directly in a high-capacity, national global file space. As a further step, the state has announced its intention, endorsed by Governor Blanco, to create the regional Louisiana Optical Network Initiative (LONI), which will connect eight major research sites across the state, and will support the requirements of the storage federation, wide-area file systems, and other network-intensive services required to unify resources and researchers across all these sites, extending dramatically its capacity and user base.

NSF TeraGrid

In view of the benefits to the scientific and engineering community, the NSF Extensible TeraGrid Facility (ETF) Project has extended an invitation to join the TeraGrid.

The TeraGrid project consists of nine sites, with computing centers, storage archives, and advanced scientific instruments tied together with a high performance network backbone using multiple OC-192 links and an integrated software environment. The ETF supports large scale multidisciplinary scientific investigations that require the ability to rapidly compare data collected from all of these resources, such as comparing computational weather-modeling data to a massive nationwide sensor network and archives of similar past weather events. A number of applications arise from this fusion of data resources that have the potential to require petaflops-scale computing resources. Consequently, NFACS can connect to the TeraGrid backbone in order to better support these emerging applications and expand resources available to the TeraGrid community.

To better integrate with the TeraGrid community, this NFACS facility will participate in the definition and deployment of and the Common TeraGrid Software Stack (CTSS), which helps to enable interoperation among all of the facilities connected to the TeraGrid, as well as the accounting, accounts management, and other operational aspects of the TeraGrid.

The net result of the inclusion of the NFACS facility in the TeraGrid as a resource available to the U.S. science and engineering communities will help facilitate the rapid advance of scientific simulation and analysis.

Pacific Northwest National Laboratory (PNNL)

Pacific Northwest National Laboratory (PNNL) plays a key role in the development of software for computational chemistry and biology. PNNL will work with NFACS to port, optimize, and maintain these codes on the LCS systems.

Chemistry plays a critical role in many DOE missions. For instance, understanding chemical transformations is the key to predicting processes associated with catalysis, nanoscale phenomena such as electron/hole generation in oxide films, combustion thermodynamics and kinetics, electron transfer processes in biological systems, and heavy-element chemistry. Computational modeling and simulation hold the promise to provide insight that is not available through experimental techniques; and thus, it has become an invaluable mechanism for providing chemical insight as well as quantitative information to the chemical sciences. However, to achieve the promise of computational chemistry for large, complex molecular systems will require ultrascale computational resources as well as revolutionary physics and chemistry algorithms, novel numerical techniques, and more efficacious computer implementations of the algorithms. For example, a factor-of-1000 increase in the performance of simulations is needed to have major impact in catalyst design [D1]. Toward this end, PNNL will port, optimize, and maintain the computational chemistry code NWChem [D2] for the architectures proposed for NFACS. NWChem is a massively parallel software suite that provides many methods to compute the properties of molecular and periodic systems using quantum mechanical descriptions of the electronic wavefunction or density. In addition, NWChem has the capability to perform classical molecular dynamics and free energy simulations. These approaches may be combined to perform mixed quantum-mechanics and molecular-mechanics simulations.

The modeling and simulation of complex biological systems is the next phase in the computational-biology component of systems biology as part of a number of federally funded major research programs, such as DOE's Genomes to Life program. Microbial systems are the primary targets of the programs of interest to DOE because of the potential role that microbes play in addressing core DOE missions in energy production, bioremediation, and carbon sequestration. There are a number of key areas, such as the assembly and dynamics simulation of molecular machines, for which ultrascale simulation codes are available that will need to ported and optimized to the computer architectures proposed for NFACS. Our focus will be on the porting and the performance analysis of several codes for simulating molecular interaction systems, including the computational chemistry suite NWChem for long-term classical molecular dynamics simulations, and NWGrid and NWPhys for mesh-based reconstruction of cell environments. The goal of current computational-biology software development is to develop and maintain these molecular and cell-modeling capabilities to allow the computational analysis of energetics, dynamics, and electrostatic properties of biomolecular systems of a much larger spatial and temporal scale than has been heretofore possible.

D.2 Leadership Computing Applications Teams

NFACS has identified five computational-science applications areas that will require a leadership-class computing capability to make major computational advances: nanoscience, combustion, fusion, climate, and astrophysics. In each of these application areas, project teams have been assembled who will collaborate with NFACS to accomplish their computational goals. This section lists the collaborators in each team and describes their research, their goals, and their computational methods and needs.

Nanoscience

Collaborators: R. Car, Princeton University; D. Ceperley and R. Martin, University of Illinois Urbana-Champaign; A. Zunger, NREL; S. Louie, UC Berkeley and Berkeley Lab; G. Galli, LLNL.

The fabrication and integration of nanoscale systems promises to revolutionize science and technology, from targeted drug delivery in medicine to ultrafast single-electron devices for computer

technology. Ray Orbach, the Director of the DOE Office of Science, has identified "nanoscale science" as one of the seven highest priorities for the Office of Science. Computational simulations are an indispensable part in nanoscience research, in part because the complexities of nanosystems often make traditional analytical tools inapplicable, and the small size scales make direct experimental measurements very difficult.

Many-body methods in computational nanoscience are based either on a Monte Carlo approach (quantum Monte Carlo) or eigenfuntion-type calculations, which involves the diagonalization of large matrices using dense or iterative solvers. In the single-particle method, the wavefunctions are usually expanded in plane waves (Fourier components), and calculations typically involve parallel 3-D fast Fourier transforms (FFTs) and dense linear algebra in the form of iterative eigensolvers. A third class of computation is classical molecular dynamics codes, which are used to study the synthesis of nanostructures and large nanostructures beyond the scope of quantum calculations.

To accurately study nanostructures with 1,000 or more atoms, as well as their transport and optical properties, with the accuracy of a quantum mechanical approach, a sustained performance of 30–50 Tflop/s is required. Looking to the future, the simulation of nanoelectronic devices will require at least 10 times this performance.

We have assembled an eminent team of leading researchers in materials science and nanoscience computation: R. Car, quantum mechanical molecular dynamics for 200-atom organic systems; D. Ceperley and R. Martin, quantum Monte Carlo simulations for finite temperature liquids; A. Zunger, million-atom simulations for nanodevices and multiple excitations; S. Louie, many-body calculations of optical excited states for nanosystems; and G. Galli, quantum Monte Carlo simulation for optical properties of thousand-atom nanosystems.

First Principles Molecular Dynamics Study of Biomolecules and Organic Nanostructures, *R. Car, J. L. Li, N. Wingreen, E. Steifel, and S. Zilberman, Princeton University; S. Louie and J. Neaton, UC Berkeley.*

An important class of phenomena occurring at the nanoscale are governed by a fine interplay between electronic structure and molecular dynamics. These phenomena include chemical reactions in biological and aqueous environments and the processes that lead to binding selectivity and molecular recognition in organic molecular structures. Given the complexity of these systems and the associated size and time-scale problems, most computational modeling in this area has been limited so far either to molecular dynamics simulations based on classical empirical force fields, or to electronic structure calculations on static configurations of small model systems.

Theoretical progress has made available a number of computational tools to overcome the limitations of the above approaches. These tools include:

(i) first-principles molecular dynamics techniques in which the potential energy surface for atomic dynamic is generated on the fly from the instantaneous ground state of the electrons,

(ii) novel energy functionals to describe accurately the electronic ground state,

(iii) techniques to solve the DFT ground state problem with optimal scaling with size,

(iv) hybrid quantum mechanical/molecular mechanical (QM/MM) methods to describe the effect of the environment on a quantum mechanical active region, and

(v) powerful path sampling and biased molecular dynamics methods to deal with activated processes and rare events.

Limitations in the available computational power have hampered, however, widespread applications of these methods to model biomolecules and organic nanostructures. This situation will change dramatically if a new computional platform capable of 50 teraflop/s performace would be made available

to the scientific community. This machine would make possible first-principles molecular dynamics simulations of systems containing several hundred atoms for time scales of several tens of picoseconds. These simulations would allow us to model small organic molecules (e.g., DNA pairs and aminoacids) in water solution, or the enzymatic active site of a protein coupled to an MM environment. The new computational platform would also make possible studies of reaction pathways in systems of the above complexity, e.g., by means of approaches that use a chain of system replicae distributed between reactant and product states.

In this proposal we will apply the above techniques to study organic nanostructures in biological and aqueous environments. In particular, we plan to study excited-state properties and dynamics of nanostructures in solution. This project will be based on a collaboration between a Princeton group (R. Car, J. L. Li) and a Molecular Foundry – UC Berkeley group (S. Louie, J. Neaton). This project will combine the first-principles molecular dynamics expertise of the Car group in Princeton with the expertise of the Louie group in dealing with excited-state properties. This will allow us to elucidate the effect of the solvent on the optical properties of selected organic molecules. In addition, we plan to investigate the microscopic origin of the hydrophobic interactions between DNA-base pairs by first-principles molecular dynamics methods. This study will involve a collaboration between the Car group (Chemistry) and the group led by Ned Wingreen (Molecular Biology) at Princeton University. Finally, we will apply first-principles molecular dynamics with path-sampling and biased-dynamical approaches to study atomic reaction pathways in the enzymatic processes hydrogenase and nitrogenase (R. Car, E. Steifel, S. Zilberman, and a graduate student, Chemistry Department, Princeton University)

Coupled Electron-Ion Quantum Monte Carlo Simulation of Liquids and Solids, *D. Ceperley and R. M. Martin, Physics Department, University of Illinois Urbana-Champaign.*

Simulation of many-particle systems of atoms and molecules plays a central role in materials science. In 1985, Car and Parrinello (CP) introduced their method, which replaced an assumed functional form for intermolecular potentials with a local density approximation-density functional theory (LDA-DFT) calculation done "on the fly" [D3]. They did a molecular dynamics simulation of the ionic motion of liquid silicon by directly computing the forces due to the electrons at every molecular dynamics (MD) step. It has been a very useful method, particularly since one does not need to have an accurate parameterized potential; the original paper has been cited thousands of times. However, the LDA approximation is not always sufficiently accurate for the many-body effects. On the other hand, quantum Monte Carlo (QMC) methods have been proved to be very accurate to deal with the many-body Schrödinger equation [D4]. It is now feasible to carry out *ab initio* QMC calculations of materials such as crystalline carbon and silicon [D5], large molecules of carbon [D6], and silicon clusters [D7]. Predictions of new phases of matter, including hydrogen at its metal insulator transition [D8] and nonmolecular forms of nitrogen [D9], have also proved to be successful. However, one of the major challenges for the QMC method is to use it for molecular dynamics or Monte Carlo simulations for the atoms. Here we propose an algorithm to do just that, using the much increased computer power proposed here.

In the spirit of the CP method, Coupled Electronic-Ionic Monte Carlo (CEIMC) simulates systems at finite temperature but using QMC instead of LDA for electronic correlation. There are numerous possible applications of the methods to problems in physics and chemistry. That this is possible is the result of both improved algorithms and teraflop/s-scale computational resources. An important aspect of CEIMC runs is that they are naturally parallel. Many independent QMC runs can be performed for a fixed length of time. Roughly 10^3 QMC steps/processor suffices to reduce the noise level enough for one ionic movement step. This typically takes several processor seconds for a large system. The results from different processors are globally communicated and averaged to decide whether the ionic move will be accepted. The only inefficiency of this parallelization strategy is in the warm-up time for each QMC simulation. At high temperatures, this limits the number of processors that can be efficiently used. But given the more powerful computer proposed here, we will be able to tackle low-temperature and many-electron problems where the warm-up time is no longer a problem.

Given enough computer power, a dynamical version of CEIMC, i.e., CEIMD, becomes possible. To generate the correct distribution from trajectories with noisy forces, one simply has to add a viscosity term to remove the energy added by the fluctuating forces, as is done in Brownian dynamics. It is clear that the noise level on the forces must be small, since dynamical trajectories are sensitive to noise. Such an approach is difficult to do with current computers, but will become feasible using the computer proposed here. CEIMD can be used to study the stability of structures, to search the equilibrium geometries, and to find state-to-state transition probabilities. Here we give a few examples of the applications of this method important in material science/nanoscience.

Simulation of hydrogen. The main application to date of CEIMC is high-pressure hydrogen. Hydrogen is one of the simplest elements, but it displays remarkable variety in its properties and phases. It has several solid phases at low temperature, and the crystal structure of phase III is not fully known yet [D10] even though there is much experimental activity using diamond anvil cells [D11]. At high temperature and pressure, the system becomes metallic, but the exact nature of the transition is not known, nor is the melting transition from liquid to solid for pressures above 1 MBar. Theoretical CEIMC simulation provides a way to solve these mysteries. The advances in methodology, such as new boundary conditions and trial functions and increased computer performance, allow, for the first time, completely realistic calculations throughout the bulk hydrogen phase diagram. A key question here is to decide if there is a first-order liquid-liquid transition in the dense fluid (the plasma phase transition). The CEIMC energy resolution, which is less than 100 K (0.01 eV) makes this possible. One particular focus of the effort will be to understand the connection between the metal-insulator and atomic-molecular transitions and the role of zero-point motion and thermal effects on those transitions. The proposed facility will be invaluable in scaling the system size from 54 atoms today to systems with several hundred atoms and for much longer simulation times.

Simulation of liquid water. A follow-on project will be the *ab initio* CEIMC simulation of a many-body system of water. Water is of central importance in chemistry, biology, and nanoscience. However, current simulation methods, both semi-empirical and those based on CP-MD, fail in important details [D12]. CEIMC will remove several of the main approximations in current calculations: errors of the density functional and of the zero-point motion of the ions. Almost all DFT-based simulation methods perform classical molecular dynamics of the ions. CEIMC based on MC will have different ways to move through phase space and test the ergodicity of the path integral molecular dynamics (PIMD) approach and the accuracy of the approximations used, and will be a much-needed benchmark of these other MD approaches. The proposed new facility should allow computations with systems of several hundred water molecules. We will also collaborate with Dr. G. Galli (LLNL) in these comparisons. There are many interesting scientific questions concerning the phase diagram of water at higher temperatures and pressures.

Multiple excitations and systems by design for multicomponent nanosystems, *Alex Zunger (NREL), Alberto Franceschetti (ORNL), Gabriel Bester (NREL).*

Experimental sophistication in nanostructure growth has now reached a new high in terms of its ability to grow dots, wires, and wells of high quality, especially for their crystallinities, purities, size distributions, and surface passivations. The next experimental frontier is to use these individual building blocks to make 3D architectures of nanostructure systems. This includes imaginative 3D assemblies of dots, wires, and wells [D13] into systems exhibiting interactions between the constituents, leading to new functionalities. The motivation to assemble basic nanostructure units lies in the desire to demonstrate devices in the form of "absorbers," "emitters," "memory units," and "spin units," and to observe novel collective phenomena. Recent progress made in both the synthesis and characterization of dots and dot arrays has already led to new device ideas based on such systems, including the observation of stimulated emission in InAs/GaAs dots [D14], colloidal CdSe dots [D15], and even Si [D16]. New devices include LED (light-emitting diodes?) [D17], electronics [D18], and quantum computing [D19]. In the area of photovoltaics, new proposals exist regarding quantum-dot solar cells [D20]. The challenge faced by

theorists is therefore not only the understanding of isolated nano objects, but also 3D architectures of nanosystems. Another exciting area in nanoscience is multiple excitations. Much like in atomic physics, the interactions of multiple excited carries (electron and holes) in a nanostructure give very complicated spectroscopies. Due to the ability to control the overall shape and size of these nanostructures ("artificial atoms"), the phenomena here are much more rich and the potential applications are enormous (partly because they are part of the solid system, not gaseous atoms). Also because of the size differences, the multiple-excitation physics in a nanosystem is very different from the atomic physics; they cover different physical interaction regions.

In the past ten years, we have developed a systematic approach to calculate the electronic structures of thousand- to million-atom systems. In this approach, empirical pseudopotential calculations [D21] are followed by a configuration interaction treatment [D22] for the multiple-excitation many-body effects. This approach has been used to study electronic structures of thousand-atom quantum dots and wires using thousand-processor NERSC computers. However, the current computer capability limits us to calculate the electronic structures of single nanostructures (quantum dots, wires), not assemblies of nanostructures and devices. It also limited us to use a small number of configurations (e.g., up to six-electron and six-hole single state levels) in the configuration interaction study. As the number of excited carriers increases, this becomes a serious bottleneck. Another current trend is to calculate the electronic properties by design. That is, for a prescribed electronic property to be used as a sensor or as a specialized material, one wants to calculate the atomic structure, or nanostructure shapes, or architectures, which give rise to that electronic property. To reach this goal, through optimization algorithms, thousands of high-fidelity electronic structure calculations are needed for given nanostructures or assembles of nanostructures. Thus, massive parallel computations are needed.

A computer 40 times more powerful than the current NERSC computer will enable us to do the following calculations: (1) To calculate the electronic structure of nanostructure assembles, like quantum dot arrays, devices consisting of quantum dots, wires, and substrates. It will also allow us to calculate the electron transports for these devices. (2) To calculate systems where the many-body effects are critical to the observed physical properties. This is essential because in a quantum dot, the electron and hole are physically confined in a small volume; this increases their interactions. For single photon applications, in order to help the device design, the accurate peaks of the many-body optical spectrum must be known. Besides, in this area, accurate quantitative results are often needed for qualitative understanding, and the understanding of such highly accurate and delicate spectroscopy is important. (3) To provide systems by design for nanosystems containing millions of atoms and multiple components. A high-fidelity prediction capability like that can fundamentally change the role of theoretical calculations in system design and the search for new materials.

Our current calculations have proved that we can scale to thousands of processors. What we need is more powerful processors on each node, and faster communication. The new computer architecture proposed here suits our needs ideally.

Large-scale many-body simulations of electron transport, optical excitation and excited state nanomechanics, *Steven G. Louie (Berkeley Lab), James R. Chelikowsky (University of Minnesota), Sohrab Ismail-Beigi (Yale).*

Recent advances in nanoscience have opened up new frontiers for scientific research and given tremendous promise for applications. However, the studies of the structures, properties, and functionalities of nanostructures pose major challenges for both experimental and theoretical investigations because of their size. Understanding and controlling these systems, which are in between the better known molecular and condensed matter limits, require synergetic collaboration between theory and experiment. In the past few years, computation and modeling have proven to be extremely fruitful in explaining and predicting the properties of systems such as nanotubes, atomic wires, nanoparticles, and molecular junctions. However, these studies have been limited to simple systems because of huge demand

in computational resources. Typically, even with the best available methods, the calculations of the various properties of interest (such as the bonding, mechanical, electronic, optical, and transport properties) scale like N to the 3rd or 4th power, with N being the number of atoms in the relevant part of the nanostructure.

With the availability of the Leadership Class System, we will be able to address a number of major issues in nanoscience through computation, which have heretofore not been possible because of limited resources. We propose to carry out *ab initio* calculations for realistic prototype systems on the following topics that are central to nanoscience and the future of nanotechnology.

Electron transport through nanostructures. One of the major areas in nanoscience is molecular electronics. This field is driven by the ultimate goal of fabricating electronic devices with novel properties at the nanoscale, replacing conventional microelectronics. Such future molecular electronic devices are likely to be composed of components such as atomic wires, nanotube junctions, or molecular junctions (i.e., a single molecule attached to two leads). Calculating the electrical transport through these nanoscale objects is a major challenge because it is dominated by quantum effects; and many-body interaction effects, such as those due to electron-electron and electron-phonon interactions, can play an important role. An accurate determination of the quasiparticle excitation energies in such systems is a crucial first step in understanding their transport and tunneling properties. Further, to simulate real devices, one needs go to beyond linear response and calculate the current-voltage characteristics at finite bias. Finally, one requires a self-consistent theory for situations with finite current and open boundary conditions, calculation of forces in the nonequilibrium state, coupling of the current to the vibrational modes and other excitations, and determination of the behavior of noises. Recently, the Louie group has developed a first-principles scattering-state approach to address these issues [D23]. This approach will be used to study electron transport through atomic wires, nanotubes, and various molecular systems of experimental interest.

Photoemission and optical properties of nanostructures. The photoemission and optical properties of nanostructures are significantly altered compared to equivalent bulk systems owing to quantum confinement and enhanced Coulomb interaction effects. For example, the optical response (e.g., the color of the luminescence light) of a semiconductor nanocrystal is a sensitive function of its size, which can be tuned for different applications, and that of a carbon nanotube is also sensitively dependent on its diameter and chirality. These changes form the basis for many existing and potential optoelectronic applications of nanostructures. With the new computational capability, we plan to investigate the photoemission and optical properties of several key nanoscale systems (including nanotubes, nanocrystals, and quantum wires on surface steps) by treating the many-electron interactions at the state-of-the-art level using the GW-Bethe-Salpeter equation approach pioneered by the Louie group and his collaborators [D24].

Excitations and nanomechanical properties. Another exciting development is the use of induced structural changes in nanostructures for mechanical applications, for example, employing them as nanomotors. Owing to the nanoscale dimensions of these structures, such applications involve, in general, conformation transformation of the system after electronic excitation by either photon or charge injection. Understanding these phenomena requires knowing the forces on the atoms in the excited state, going beyond standard ground-state theories. Recent theoretical advances have made possible first-principles computation of such excited-state forces [D25]. We plan to apply these methods to address previously computationally inaccessible questions regarding photo-induced conformation changes, luminescence spectra, molecular dynamics in the excited state, and the microscopic structure of photo-induced defects in nanostructures.

Development and validation of predictive computational design of nanostructured materials, *Giulia Galli, Jeffrey C. Grossman, Eric Schwegler, Andrew Williamson, and F.Gygi, LLNL.*

At present, *ab initio* simulations provide key contributions to the understanding of a rising sea of measurements at the nanoscale: they provide access to numerous physical properties (e.g. electronic, thermal and vibrational properties) at the same time, with a controllable level of accuracy; they complement experimental investigations that are sometimes controversial and cannot be explained on the basis of simple models; and they allow to investigate properties that are not yet accessible to experiment. A notable example is represented by microscopic models of the structure of surfaces at the nanoscale, which cannot yet be characterized experimentally due to the lack of appropriate real space imaging techniques. Quantum simulations can not only provide structural models but also deep insight about the chemistry occurring at nanoscale surfaces and interfaces. .

At present, state-of-the-art first principles molecular dynamics (FPMD) can treat systems with a few hundred atoms (200–500, depending on the number of electrons and the accuracy required to describe the electronic wave-functions) and simulation times of 10–100 ps (depending on the size of the systems involved). State-of-the-art Quantum Monte Carlo (QMC) codes using newly developed linear scaling algorithms can now enable the calculation of the energetics and optical gaps of sp-bonded systems with up to 100–300 atoms.

We estimate that in the next few years, algorithmic developments (e.g., linear scaling methods), along with the anticipated surge in computational power, will enable FPMD simulations of systems comprising 3000–4000 atoms for several picoseconds, as well as of systems comprising 200–300 atoms in the nanosecond range. In addition, QMC calculations of systems containing several thousand atoms, using newly developed linear scaling algorithms, will be made possible with unprecedented levels of accuracy. This will allow one to simulate, e.g., organic/inorganic interfaces found in nanoscale-devices for bio-detection, transport properties of single-molecule electronic devices and semiconductor nanowires, the properties of magnetic systems at the nanoscale and in general of advanced materials.

With the coming of age of first principles theory of matter, as well as the development of powerful algorithms for quantum simulations, the application of FPMD and QMC techniques will extend far beyond the traditional fields of condensed matter physics and physical chemistry into biochemistry and biology. In the next decade we expect quantum simulations to effectively enter the realm of biology and to tackle problems such as microscopic modeling of DNA repair mechanisms and drug/DNA interactions. In particular, nearly exact QMC results will represent valuable theoretical benchmarks that may help overcome some of the current limitations of experimental biology.

Advances in quantum simulations require progress in different areas: (i) theoretical and algorithmic developments are needed, to improve the accuracy and efficiency of both FPMD and QMC techniques; (ii) code optimizations to adapt to new and changing platform architectures (e.g., Blue Gene/L) are required and often imply tackling complex applied mathematics and computer science issues; (iii) as the scope and predictive power of quantum simulations become broader, knowledge on how to best use these techniques in a way fully complementary to experiment and theory needs to be developed and novel approaches to analyze data obtained from quantum simulations (including visualization tools) need to be established.

Combustion Modeling

Collaborators: H. Pitsch, Stanford University; A. Trouve, University of Maryland; J. Chen, H. Najm, J. Olefein, Sandia National Laboratories; A. Ghoniem, Massachusetts Institute of Technology; J. Bell, Berkeley Lab.

Combustion provides more than 85% of U.S. energy. Meeting U.S. energy demands, as well as environmental concerns, require that power-generation, industrial processes, and transportation systems operate at higher efficiency with lower emissions. Large-scale computational simulations are essential for studying the complex interaction of fluid mechanics and chemical processes associated with turbulent

combustion. These computations typically span a huge range in time and length scales (as much as a factor of 10^9), requiring large and adaptively-managed sets of gridpoints.

Combustion codes typically combine a subset of the following computational techniques: (1) explicit finite difference, finite volume, and finite element methods for systems of nonlinear partial differential equations (PDEs); (2) implicit finite difference, finite volume and finite element methods for elliptic and parabolic PDEs (typically utilizing iterative sparse linear solvers); (3) zero-dimensional physics (often heterogeneous in work per point), including evaluation of thermodynamic and transport data, as well as integration of stiff systems of ordinary differential equations (ODEs); (4) adaptive mesh refinement; and (5) Lagrangian particle methods embedded in finite-difference, finite volume, and finite element (FE) algorithms.

The computational resources proposed for NFACS will enable combustion researchers to simulate flames with high-fidelity representations of the governing physical processes. Simulations of this type will provide the insights needed to understand turbulent flame dynamics, turbulence chemistry interaction, and pollutant formation. This type of information is key to designing cleaner, more efficient systems.

For example, NFACs will make it possible to model laboratory-scale turbulent natural gas combustion with comprehensive chemical kinetics and diffusion transport.

Direct Numerical Simulation

An area that is well-primed to exploit increased computing resources is the exploration of fundamental turbulence/chemistry interactions in laboratory-scale flames through the use of DNS. Experimentally validated direct numerical simulation (DNS) results will be critical in the development of physically based subgrid models used in the larger simulation tools (such as large-eddy simulation [LES]) that are ultimately required for large-scale component design and system optimization. DNS is equally critical as an aid to understanding fluid-chemistry interactions at the very basic level. However, reacting flow DNS applications are currently limited to volumes of the order of a cubic centimeter or less, and are difficult to compare directly with relevant experimental data for a number of reasons. A key element of turbulent flow is the "cascade" of energy from the large scales associated with interactions with the device boundary into the small scales at the Kolmogorov length associated with viscous energy transfer and flame reaction chemistry. Unless the entire experiment is of the scale of the simulation domain, the calculation will fail to represent key modes in systems with significant turbulence. As a result, the turbulence intensity that could be represented is likely to be far lower than the levels of practical interest. Moreover, such small experiments are extremely difficult and expensive to diagnose.

The range of current DNS applications related to physical experiments is quite limited. Autoignition studies form one important DNS application involving the fine-scale interaction of reaction chemistry and fluid turbulence in the high-pressure environment of an internal combustion engine. The full autoignition problem is presently intractable for theoretical and laboratory investigations, due to the fine spatial and temporal scales involved and the requirement for very detailed chemical mechanisms to describe the ignition process. However, recent two-dimensional simulations have been used to model regions too small to reliably collect experimental data. The work is yielding important new insights into the mechanisms of hydrogen-air interactions, and the proposed NFACS hardware will enable these simulations to be expanded to accommodate the full range of spatial scales, from those that control the autoignition process to those that are experimentally observable in laboratory engine experiments, opening the way for experimental validation of the results. Additionally, the NFACS system will enable the exploration of hydrocarbon fuels such as n-heptane, and the study of ignition problems in the full time-dependent, three dimensional, high-pressure domains where results can directly impact the understanding of internal combustion engine performance.

Low Mach Number Simulations

Low Mach number simulations offer a unique promise as well with this increased computing capability. When combined with adaptive mesh refinement, low Mach number methods, which analytically filter away the fast time scales associated with acoustic wave propagation, have been demonstrated to be effective for simulating low speed flames, typical of most practical combustion systems, while incorporating detailed models for chemical reaction and molecular transport without explicit subgrid models. With access to the NFACS facilities, low Mach number methods will used to simulate laboratory-scale flames using comprehensive chemistry and transport allowing us to probe the detailed dynamical and chemical properties of these types of flames over the full range of length scales observed and measured in laboratory flames. These simulations will be able to quantify how turbulence alters the chemical pathways in the flame and how chemistry affects the turbulent flame speed. At the most aggressive computing performance levels proposed, advances in capability-class hardware will enable the prediction of the formation and emission of atmospheric pollutants from such flames, and understanding how the presence of larger hydrocarbons affects the flame chemistry. Multiphase and high-pressure aspects can also be incorporated to allow analysis of evaporation and mixing properties of spray fuels in a closed turbulent environment.

Lagrangian-Eulerian Algorithms

Lagrangian-Eulerian algorithms also have potential for dramatically extending the range of combustion phenomena that can be modeled without explicit subgrid models. Lagrangian-Eulerian approaches take advantage of the high resolution and low diffusive errors of Lagrangian methods, and the adaptive refinement and robustness of Eulerian methods. They are particularly suitable for combustion simulation because of the wide range of scales that need to be resolved, and the range of physics that must be modeled. Progress has been made in construction of fast Lagrangian algorithms for solving the vorticity transport equation, and extending these algorithms to the convection-diffusion-reaction equations governing reacting flows. The former has been demonstrated successfully in predicting flows at intermediate Reynolds numbers. Extension to higher Reynolds numbers will require more memory per processor and longer computational time. Simulations the reacting flows are currently limited to either simplified chemical reaction models in three dimensions, or extended chemical models in two dimensions. With NFACS facilities, we will be able to perform three-dimensional simulations with an extended-chemistry model. Another area in which computations will be enabled is in the area of unsteady boundary conditions in which active control strategies are required to improve the overall performance of a process, e.g., efficiency, without compromising other performance measure, that is emissions and noise. Simulations at higher Reynolds numbers, closer to where actual systems operate, will also be enabled, leading to better reduced models and practical engineering codes for design.

Large-Eddy Simulations (LES)

In recent years LES of turbulent combustion has received great attention as an attractive method for modeling of full-scale technical combustion processes. LES has been particularly useful where high power densities desired for efficiency involve intense turbulent mixing, which in turn is controlled by scales involving the entire system. Device optimization and the design of novel devices with high combustion efficiency depend critically on an understanding of the interaction of fluid mechanics and chemistry. The NFACS facility will enable a broad range of accurate LES simulations based on validated theory and detailed chemical kinetics that can predict performance and emissions, and are at the same time suitable and simple enough to be applied in engineering simulations. Applications include simulations of lifted turbulent diffusion flames, bluff-body stabilization, gas turbines in aircraft and stationary applications, autoignition in a non-premixed environment, and large-scale pool fires. LES will also be of major interest for applications in reciprocating engines to investigate phenomena such as turbulence production by liquid jets, and auto-ignition in direct injected (DI) diesel and homogeneous charge compression ignition (HCCI) engines. Also for the investigation of early flame development and

cycle-to-cycle variations in spark ignition (SI) engines, and the dynamic behavior of engines under variable load conditions, LES will be particularly attractive. Besides these technical devices, LES has also been recognized to be particularly useful for numerical simulations of accidental fires, where large-scale mixing processes are known to be of great importance.

Database Assessment and Data Analysis

A major research issue intimately linked to these research opportunities is the generation of chemical mechanisms to describe the reactions occurring in combustion systems. Combustion simulations can involve hundreds to thousands of dependent variables coupled in highly nonlinear ways. Extracting the relationships between these variables, determining the parametric sensitivity of solutions, and exploring reduced chemical mechanisms require sophisticated mathematical tools. Of particular importance in this area are tools such as computational singular perturbation, proper orthogonal decomposition, uncertainty quantification, sensitivity analysis, and techniques that focus on the complexity of coupled transport-chemical processes in multidimensional reacting-flow simulations. These tools are mature for a limited range of combustion applications but would become considerably more valuable were facilities such as the NFACS applied to more physically relevant systems with realistic fuels and multiple dimensions.

Detailed combustion simulations are critically dependent on large databases that characterize the chemical kinetics, thermodynamics and transport of the many species involved in combustion. Validations of these databases are computationally extremely demanding, particularly in multidimensional, time-dependent simulations, yet are critical to the success of the simulations and models. The primary databases in the combustion community for standard hydrogen and hydrocarbon fuels have been optimized with simplified low-dimensional configurations but can, with more capable computing hardware, be updated to incorporate more relevant multidimensional or time-dependent scenarios.

Technological Barriers

Combustion simulations incorporating detailed chemical kinetics are particularly demanding, as they involve hundreds or thousands of chemical species and a multiscale response spanning many orders of magnitude in space and time. However, this set of applications constitutes a unique category where projected resources can make possible a significant step towards the reasonable and useful simulation goals of more tightly coupling detailed simulations with experimental data, developing modeling strategies that allow a broad range of full-device design and optimization and of improving the lower-level understanding of detailed chemical interactions with turbulence. The NFACS proposal outlines a hardware configuration superior in many respects to existing Department of Energy “capacity” based computing hardware and is certainly capable of providing a platform for dramatic forward progress in combustion science.

Many combustion-related applications are implemented with multidimensional structured grid strategies that include block-structured approaches to adaptive mesh refinement. While these types of calculations inherently have the potential of being well suited to vector-processing computer hardware, current high-end “capacity” systems do not allow this key observation to be capitalized on effectively. The multiphysics nature of the problem implies the use of a broad variety of numerical kernels. The poor cache performance associated with sparse, multidimensional data access patterns and extremely demanding parallel communication requirements limit the peak achievable performance of these simulations with existing hardware. Moreover, as simulations with detailed chemical fidelity generate a massive volume of data to be analyzed, parallel I/O issues are paramount. Although the capability-class machines currently available do support a broad range of Department of Energy research, the proposed NFACS configuration represents a solution much more tightly coupled to the needs of combustion simulations with detailed chemistry.

Historically, advances in computing hardware are most advantageous when coupled with software and algorithm enhancements. Low Mach number simulation capabilities represent an example of a class of useful problems where an increase of more than five orders of magnitude in problem size was realized through the combined use of asymptotic analysis and adaptive mesh refinement. Although simulation methodologies are available for many of the computational problems posed here, additional development is required to harness the power of both existing and new computer architectures for these problems. One critical area of research is scalable algorithms for multiphysics reacting-flow problems. Particular issues in this area include the development of scalable solver techniques for variable coefficient and nonlinear implicit systems for grid-based discretizations and the development of improved load-balancing strategies for heterogeneous workloads. Substantial capability increases can also be achieved by developing improved discretization procedures that not only provide improved representations of the basic physical processes but also improve the coupling between these processes. Performance analysis tools, usable in complex multiphysics applications, are vitally important for achieving significant percentages of peak hardware performance. Finally, petabytes of storage for both cached and archived data will also be needed for the massive amounts of resulting simulation data to be analyzed.

Fusion Energy

Collaborators: S. Jardin and W.-L. Lee, Princeton Plasma Physics Laboratory; R. Cohen, LLNL; C. Sovinec, University of Wisconsin.

Fusion energy research is a crucial part of the DOE mission. In last year's report on DOE facility priorities for the next twenty years [D26], the International Thermonuclear Experimental Reactor (ITER) fusion experiment was the number one near-term priority, and three other fusion facilities appeared among the mid- and long-term priorities. Furthermore, the report states that computational simulation capability is expected to be vital to the success of the ITER experiment. Computational simulation supports experimental fusion research not only in the design of experimental facilities, but also in analysis of the resulting data and the development and validation of theory.

Some significant problems in magnetic-confinement-fusion reactor simulation for which major advances can be made with the next-generation of computing capability are:

- global stability of plasma confinement
- plasma microturbulence
- pellet injection
- edge plasma physics.

The computational methodologies that are used for these problems have been well developed in the fusion community. For many problems, future progress is limited by insufficient resources to perform the desired calculations in a reasonable time. In some areas, the requirements are driven by the need to integrate multiple models to produce accurate simulations of more complex physics. (An integrated simulation is usually more computationally expensive than the sum of its components.) In other areas, model and algorithm development is the key requirement. These needs call for a future computational capability that is balanced between raw performance and ease of use.

The ultimate goal of a complete, integrated, computational simulation of a fusion reactor is beyond the capability of any feasible computer system in the near term, so progress in fusion research will focus on isolated component problems. Some of these problems are described below.

Global stability. Global stability of the plasma confinement is an essential ingredient of a working reactor. Simulating the onset and evolution of instabilities and predicting confinement failure is particularly difficult because of the large range of time and length scales and high anisotropy of the plasma. Extended magnetohydrodynamic (MHD) models with coupled transport, multiple fluids, and

kinetic turbulence are required. Future computational requirements are driven by the need to increase spatial and temporal resolutions, simulate longer time periods, replace ad hoc approximations with models, and perform more simulations with varying parameter values. Algorithmic improvements in adaptive mesh refinement methods and advanced nonlinear solvers (particularly for stiff equations) are needed to mitigate the cost of these increases. The state of the art in extended MHD simulation is represented by the codes M3D and NIMROD. Typical solution times at the limit of current resources are on the order of 10^4 processor hours. A reasonable goal for future performance on LCS-1 would be to increase per-processor performance by a factor of 10–15 and increase parallel efficiency by 10–20% on 2–3 times as many processors, for a total increase of 20–125 times current performance. This would allow, for example, for doubling the resolution in three grid dimensions and time, tripling the length of simulated time, and running twice as many parameter values, at the same time as improving model fidelity.

Plasma microturbulence. The behavior of a fusion reactor is strongly affected by the balance between the heat generation of the burning plasma and the heat loss from electromagnetic turbulence. This behavior impacts the design and cost of the reactor, as it affects both material strength and cooling. Current practice is to extrapolate properties from existing reactors to newer, larger reactors; one of the goals here would be to simulate large experiments such as ITER directly.

One of the principal codes for performing such simulations is the GTC code, a 3D particle-in-cell (PIC) method using an electrostatic model capable of simulations with 10^9 particles, 10^8 grid points, and 10^5 time steps on available computers. It scales well to 2,000 processors and achieves 10–15% of peak. Current computational resources constrain simulations to have less than the desired resolution, simplified geometry, reduced physics, and fewer parameters. Increased capabilities are needed for all these constraints. For example, improved models will need to include electromagnetic dynamics, global turbulence, and kinetic equation dynamics in shaped plasmas. In addition, efficient finite-element and / or finite difference elliptic solvers are needed. Roughly 10 peak teraflop/s are required to perform such simulations at the device scale.

Pellet injection. The fuel in a fusion reactor must be replenished as it is consumed by the fusion reaction. Thus the mechanism for injecting fuel into the reactor is of major importance in the design and operation of the reactor. For example, it has been observed that the angle of fuel injection has a strong effect. Current simulation capabilities do not satisfactorily predict the behavior of the injected fuel, specifically the density distribution after the pellet ablates in the high-temperature plasma.

Solutions to this problem depend on the ability of the MHD simulation to accurately model the mass distribution along and across the magnetic field lines and flux surfaces. The adaptive mesh refinement methodology is extremely advantageous to this problem, providing over two orders of magnitude in performance gain over nonadaptive meshes. Necessary improvements in MHD simulation capability involve high-fidelity treatment of the plasma geometry, implicit treatment of stiff anisotropy of heat conduction, hybrid fluid/kinetic and multifluid models. These will require some improvement in spatial resolution and much longer simulation times, which will increase computational requirements by several orders of magnitude.

Edge plasma physics. The edge plasma model is vitally important in simulating a burning plasma. It affects the physical wall of the confinement vessel, the production and transport of impurities from the wall into the plasma, and the inventory of tritium in the reaction core. Progress in this area requires improved coupling of constituent models, especially in kinetics, atomic physics, and complex geometry. The wide range of length and time scales that must be modeled presents a significant demand for computational resources. A key accomplishment in this area will be the prediction of the structure of the so-called "pedestal" (a region of rapid radial variation of temperature and density) in the plasma edge. The pedestal height has a large impact on reactor performance and is the source of significant uncertainty

in current reactor models. With current computational resources, solution times exceed 2,000 processor hours. Future requirements are similar to those for global stability and plasma turbulence described above.

Climate Modeling

Collaborators: W. Collins, NCAR; I. Fung, UC Berkeley; DOE SciDAC Climate Team.

Understanding the effect of human-induced changes to the composition of the atmosphere on the climate system is one of the most compelling scientific missions of our time. Largely a result of energy production, these compositional changes have been demonstrated to have a detectable influence on many aspects of the climate system. Computer modeling is an important tool in the detection and attribution of recent climate change and is the principal vehicle to make predictions regarding future climate change.

Current state-of-the-art climate models are a complex coupling of models of the major climate subsystems. Present-day submodels describe the atmosphere, ocean, sea ice, and land systems. The large computational burden required to adequately describe any of these subsystems forces compromises to be made both in sophistication and fidelity. Advances in computing technology permit better simulation of the climate system by allowing the inclusion of more processes relevant to climate behavior, as well as more highly resolved discretizations of the climate subsystems. Trends in both of these aspects of climate modeling increase our understanding of human-induced effects on the recent past climate and reduce the uncertainty in our predictions of future climate change.

Uncertainty in climate change prediction has many sources, each of which may be better addressed with more powerful computers. The naturally chaotic behavior of the coupled climate system implies an internal uncertainty that may be quantified by performing ensembles of statistically independent simulations. As computer power is increased, larger ensembles may be performed to increase the statistical significance of predictions. Future human behavior is also uncertain, as many different energy usage scenarios are plausible. Investigation of many different future forcing scenarios by ensemble integration can provide valuable information to policymakers as to how they differ. Finally, the deficiencies in climate models' abilities to reproduce the recent past climate are a sizable contributor to the uncertainty in predictions of the future. Addition of important processes, as well as significant increases in spatial resolution, are necessary to increase confidence of future climate change on regional scales.

Climate codes solve equations of hydrodynamics, radiation transfer, thermodynamics, and chemical reaction rates. Current approaches are marked by finite difference calculations acting on fairly regular spatial grids, requiring high main memory bandwidth. Fast Fourier transforms (FFTs) of a short length are also used in current models, although these may be replaced in the future. Scalability of individual simulations tends to be poor relative to other advanced scientific applications, although the ensemble requirement provides an embarrassingly parallel dimension to increase scientific throughput. It should be noted that existing sea ice and land codes are difficult to vectorize.

The DOE Office of Science's Climate Change Prediction Program (CCPP) is a highly coordinated activity. Many of the CCPP-funded DOE laboratory personnel support the Community Climate System Model (CCSM3) developed at the National Center for Atmospheric Research (NCAR) in Boulder, Colorado. The latest version of this code, scheduled for public release in June 2004, will be the principal model used to produce the United States' contribution of simulations of past and future climate to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. These simulations will be the most significant consumer of compute cycles allocated to the U.S. climate modeling community at the unclassified DOE computing centers.

The development and release of CCSM3 has been an intense major effort for several years. It is expected that such activities over the next several years will be of a more evolutionary nature. Certain areas targeted by the CCSM Scientific Steering Committee are of particular interest to DOE programmatic objectives and will benefit from a sustained 50 Tflop/s machine. The interaction of

atmospheric aerosols with the climate system is a particular strength of CCSM. Multiple ensemble simulations will allow the effects of different aerosol forcings to be isolated and quantified. The next generation of climate models, including CCSM, must include a full simulation of the carbon cycle. The complexity of the model will be increased substantially by the representation of biogeochemical and ecosystem processes. Corresponding increases in compute demands will follow.

Current-generation climate models are quite good at characterizing large-scale features of the atmosphere and ocean down to the continental or basin scale. However, assessment of the consequences of potential future climate change requires confidence in the model at much finer scales. Significant increases in resolution are required for this to be achieved. The tentative release version of CCSM3 will consist of an atmospheric model with a horizontal T85 spectral truncation (approximately 1.4 degrees at the equator) and an ocean model with an approximately 1 degree mesh. Strong cases can be made for increases of a factor of 10 in each of these horizontal resolutions, resulting in an increase in the computational burden by a factor of 1,000. While the machine proposed here will not take the model this far, significant new physics can still be addressed. For instance, using a scaling model that adequately describes an increase in spectral truncation between T42 and T85 and the transition from Power3 to Power4 architectures, it is estimated that on the Power6+ the truncation may be increased to T340 (approximately 0.35 degrees) with a turnaround time of five simulated years per compute day using 512 processors. Such resolution captures features important to regional scales and provides a much better representation of individual storms and other extreme events.

Although the scalability of single runs of climate models to a large number of processors is limited, large compute facilities such as that proposed here are necessary to the DOE climate change research mission. In a given configuration of model and forcing sets, the current ensemble size of four simulations must be increased to ten or more to adequately quantify the output statistics. As model complexity increases, the number of individual forcing sets (currently five: greenhouse gas, sulfate aerosol, ozone, solar, and volcanic) will increase greatly, furthering the need for more individual and combination simulations of the recent past. Ensemble integrations of future scenarios have been even more limited and further the demand for compute cycles. Inclusion of a predictive model of the carbon cycle in CCSM3 furthers the need for more ensemble simulations. As demand for climate model output increases from the analysis community, it is clear that "capability" in this field is synonymous with "capacity" in that many more runs than presently performed are required to quantify and reduce prediction uncertainty.

Astrophysics

Collaborators: M. White and R. Klein, UC Berkeley; E. Baron, University of Oklahoma; T. Mezzacappa, Oak Ridge National Laboratory; J. Borrill and P. Nugent, Berkeley Lab; D. Swesty and E. Myra, SUNY Stony Brook.

The study of astronomy — the study of the Universe as a whole and of its component parts, past, present, and future — is surely one of the earliest sciences pursued by mankind. Its origins are intimately tied to our search for understanding who we are and what our existence means — whence, astronomy's age-old links to philosophy and religion. But more recently (within the past half millennium) astronomy has played a central role in the rise of science as an experimental and deductive activity and, in the hands of luminaries such as Galileo and Newton, in the rise of physics as the fundamental physical science. This evolution is also marked by the words used to describe the field today: "Astronomy" tends to refer to the more descriptive aspects of the subject, while "astrophysics" is used to describe activities related to the use of physical sciences (including both physics and chemistry) as explanatory tools for what astronomers observe. Furthermore, astronomy is now intimately linked to virtually all other sciences. For example, physicists study the nature of fundamental interactions by looking at the evolution of the very early Universe and by studying the properties of highly evolved stars — exploding stars (e.g., supernovae), white dwarfs, and neutron stars. Biologists and chemists are examining the origins of life by considering the organic chemistry of the interstellar medium (ISM). Geoscientists interested in the origins of the

planets are collaborating with astronomers who are finding numerous planetary systems orbiting nearby stars. The profound connections between astronomy and astrophysics and some of the deepest questions faced by mankind — What is the origin of all matter and energy? What is the fate of the Universe? What is the nature of space and of time? — continue to this day. The very recent discovery of dark matter, dark energy, and the accelerating Universe is but one example.

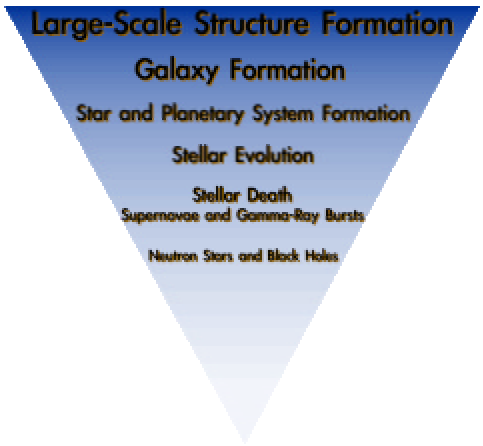
Impact on Science and Society

Certainly no physical science has succeeded in attracting the enthusiasm and interest of the public to the degree that astronomy and astrophysics have. Astronomy clubs filled with young and old enthusiasts are found everywhere in the United States, and astronomy is commonly discussed in the mass media, from newspapers and news magazines to radio and television. It is the only science to have spawned its own literature genre, science fiction, and its own federal agency (NASA) via its intimate links to the space sciences. And it is unique among the sciences in playing a role in virtually every federal agency supporting scientific research. Many of the most important areas of modern physical sciences — special relativity theory, gravity and the general theory of relativity, quantum mechanics (including nucleosynthesis and spectroscopy), plasma physics — grew out of research motivated by or related to astronomical questions. Even in the computational realm, astronomy stands out. It was the first physical science to demand computation. The ability to predict the seasons, notable events such as lunar and solar eclipses, and the motion of the planets, hinged on the ability to compute, and the ambition and scope of some of the ongoing and planned astrophysical simulations are arguably unequaled in science. Indeed, as was recognized by both the DOE/NNSA ASCI Advanced Simulation and Computing and DOE/SC SciDAC programs, some of the most important problems in modern astrophysics, such as the establishment of the Universe's distance scale and the nucleosynthesis of the iron peak and heavier elements, can only be broached via "grand challenge" simulation capabilities.

One of the practical consequences of this deep connection between astronomy and the popular imagination is that astronomy and astrophysics have proven to be a strong recruiting tool for attracting students into the physical sciences. This is an essential point in a time when the physical sciences are finding it increasingly difficult to attract "the best and the brightest" of the youth of the United States.

Scientific Opportunities

Astronomy and astrophysics can be said to suffer from an abundance of research opportunities. As we probe the Universe using more and more sophisticated technology, the number of profound (and as yet unanswered) questions has actually increased rather than decreased. In the following, traversing all scales in the Universe, we illustrate by example the richness of the questions and problems faced by modern-day astrophysicists. In all cases, simulations have played a central role in the past. This continues in the present, and is sure to be the case in the future.



Large-Scale Structure and Cosmology. Largely as a result of a new generation of technologically advanced ground-based and space-based optical and microwave telescopes, study of the large-scale properties of the Universe, and especially of its formation, have made enormous advances over the past decade. We have now entered the age of "precision cosmology." Most important, virtually all phases of the Universe, from its earliest moments to the present, are now thought to be susceptible to modeling and simulation, whose aims are to connect what is observed in the distant past to what is observed now in our corner of the Universe. Furthermore, important new cross-disciplinary areas of science, such as particle astrophysics and the physics of quark-gluon plasmas, have led to entirely new sets of questions to be addressed, for which simulation will play an increasing and ever more essential role. These are areas in which the frontiers of astronomy and of physics coincide, and where we are as yet uncertain about the most basic laws of Nature. Much of what is of interest occurs under highly nonlinear circumstances, and therefore, simulation is an essential means by which theoretical progress can be made.

Galaxy Formation and Interactions. Over the past decade, we have increasingly realized that the nonlinear stages of the formation of large-scale structure in the Universe, while seeded at the time of creation, followed in time the formation of much smaller-scale structures, namely galaxies. Studies of individual galaxies, as well as their interaction in clusters, will continue to be carried out in concert with a major revolution in observations of these systems, which now use x-rays (to trace hot cluster and interstellar medium gas), optical emissions (to trace the stars), IR (to trace the cold "baryonic" matter, composed of protons and neutrons, in ISM clouds), and radio (to trace the interactions of the cosmic microwave background photons with the hot electrons in the cluster gas). These interactions between theory and observations now demand much of theory, well beyond the simple models of just a few years ago.

Star Formation. Simulations have played a central role in driving our modern understanding of how stars and planetary systems are formed. Using modern x-ray, infrared, and radio telescopes, astronomers have been able to penetrate the interstellar dust clouds that have long hidden from view the physical processes leading to gravitational collapse of interstellar gas and star formation. Simulation has moved us to the brink of being able to put the observations into a unified physical theory of star and planet formation that will allow us to predict the variety of planetary and stellar systems to be found in the Universe. One of the most exciting areas in which computations play an important role is in understanding the variety of evolutionary paths for planetary systems. Why do the gas giant planets in our solar system sit in well-behaved orbits far outside the orbit of the Earth, while in many observed extrasolar planetary systems, gas giant planets are found at distances from the parent star even less than the distance between the Earth and the Sun, or in sweeping elliptical orbits? Ultimately, these models should allow us to predict the frequency of potentially life-bearing planets in our galaxy, as well as to understand the origins of our own Earth and Sun.

Stellar Evolution. The evolution of stars is marked by a gradual consumption of the interior nuclear fuel and, in rotating stars that have internal convection layers, by a constant level of transient energy release mediated by internal magnetic fields: stellar "activity." Our understanding of these processes relied in the past largely on one-dimensional (spherically symmetric) evolutionary models of stars. It is only recently that state-of-the-art, large-scale, multidimensional stellar evolution simulations have been attempted, elucidating new physics. For example, convection is known to play a crucial role in regulating both the internal distribution of angular momentum in rotating stars and the generation of internal magnetic fields as part of a stellar magnetic "dynamo." Large-scale simulations are at the heart of trying to understand these processes. The ultimate release of magnetic energy (leading directly to transient optical, ultraviolet (UV), and x-ray emissions, and the acceleration of high-energy cosmic rays) also remains to be understood. And simulations of processes such as magnetic reconnection form an important bridge between astrophysics and the plasma sciences. This subject area has particular relevance to us for immediate, practical reasons. The magnetic activity of the star closest to us, our Sun, is known to have

consequences for the Earth's environment ("space weather") and is strongly suspected to influence global climate change.

Stellar Death. The death of stars through spectacular stellar explosions known as supernovae produces many of the elements in the Universe necessary for life and serves as "standard candles," illuminating fundamental and profound aspects of our Universe, such as its geometry, content, and ultimate fate. Most recently, through gamma-ray and x-ray observations of the long-puzzling gamma-ray bursts (extremely bright and energetic bursts seen at cosmological distances throughout the sky), an indisputable association between these bursts and supernovae has been made. In addition to their place in the cosmic hierarchy, the extremes of density, temperature, and composition encountered in supernovae provide an opportunity to explore fundamental nuclear and particle physics that would otherwise be inaccessible in terrestrial experiment: Supernovae serve as cosmic laboratories, and supernova models are the bridge between observations (bringing us information about these explosions) and the fundamental physics we seek. In addition, proposed large-scale terrestrial experiments such as the Rare Isotope Accelerator, a priority for the U.S. nuclear physics community, and the proposed National Underground Science and Engineering Laboratory are both significantly motivated by supernova science. Much work remains to elucidate the mechanisms for stellar death via explosion. With the advent of robotic telescopes designed to maximize the success rate of finding supernovae, coupled to the use of large-aperture telescopes to measure the detailed spectra of the exploding stars, theorists are faced with new opportunities for detailed testing of supernova models. The consequent demands on simulation will be severe. And we now stand at a threshold. The Laser Interferometric Gravitational Wave Observatory (LIGO), an NSF-funded gravitational wave detector, is on line, along with other detectors around the Globe. Galactic supernovae are among the sources expected to generate gravitational waves in LIGO's bandpass. A detection by LIGO would be the first direct evidence of gravitational waves and the dynamic nature of spacetime as an active, physical fabric and participant in Universal phenomena (and not simply a void in which phenomena occur).

Numerical Relativity. Relativity has long had an intimate connection with astronomy and astrophysics — consider Eddington's classic observation of light from stars bent by the gravitational field of the Sun, the first experimental test of Einstein's theory of general relativity (GR), carried out near the beginning of the 20th century. Much of the experimental data relevant to general relativity could, until recently, be captured by relatively simple approximations of the full Einstein field equations. However, with the (indirect) discovery of gravitational radiation from binary pulsars (leading to a Nobel prize for its discoverers), the realization that highly nonlinear aspects of GR may have astronomical verification has led to a major experimental effort to detect transient gravitational waves (e.g., LIGO). Key to success will be a firm understanding of the physics leading to the gravitational radiation in the first place, and simulations of promising events such as black hole mergers and neutron star mergers are proceeding apace. Thus, much as cosmology has done over the past decade, relativity seems poised for a similar advance to "precision GR."

Research Issues

There are several cross-cutting issues that recur in virtually all of our subfields of astronomy:

- **Multi-Scale Phenomena.** As alluded to earlier, the dynamic range in both time and space for typical astrophysical problems can be enormous, and for this reason the practicalities of effective computing require the development of subgrid models that correctly describe the physical processes not directly simulated. Subgrid modeling is a science in its own right, and requires a judicious combination of theoretical work (in both physics and applied mathematics) and experimental studies that allow one to validate the model. Specific subgrid models that have to be developed to make progress in "cornerstone areas" are discussed below.
- **Multi-Physics Phenomena.** Many astrophysical problems involve a broad range of physical processes, not all of which can be captured by a single closed set of evolution equations. The

successful coupling of distinct evolution equations, each describing a particular physical phenomenon or process, is still an art rather than a science. It is a forefront research area in its own right. Examples include (1) the coupling of N-body particle and single- or multi-component fluid equations (in cosmology and galaxy formation/interaction studies); (2) the coupling of photons or neutrinos to hydrodynamics (i.e., radiation hydrodynamics) in a completely self-consistent manner (in core collapse supernovae and gamma-ray burst modeling);, and, at an even greater separation of scales, (3) the coupling of neutrinos to stellar core nuclei in supernovae, which requires both state-of-the-art macroscopic neutrino transport and microscopic nuclear structure modeling.

Large-Scale Structure and Cosmology. Key to progress is the development of a subgrid model for star formation. (In this case, the term "subgrid model" is obviously much more encompassing than is usually meant by the term in fluid dynamics.) At present, large-scale structure simulations treat stars as point masses and do not make any attempts at modeling the details of their formation. However, issues related to details of star formation, including the "birth mass function" of stars, are essential for correct prediction of the large-scale distribution of galaxies by luminosity, morphology, etc. Here, observational data (e.g., from the Sloan Digital Sky Survey) will be of great help in building these requisite models.

Galaxy Formation and Interactions. This research area requires the development of a subgrid model for stellar evolution, from stellar birth through stellar death via supernovae. Challenges include (1) the incorporation of feedbacks to the interstellar and intergalactic media (via stellar winds and supernovae and their ejecta, and the radiation fields that accompany these phenomena), (2) the correct treatment of magnetic fields (in terms of their influence on both dynamics and energetics), and (3) energetic particles (in terms of both their origins and their dynamical consequences).

Star Formation. The challenge for this area of research is to span the dynamic range from star-forming clouds of interstellar gas covering many light years, to stars and planets of sizes of thousands or tens of thousands of kilometers. The problem involves hypersonic turbulence, magnetohydrodynamics, self-gravity (solution of the multidimensional Poisson equation), chemical networks, and multidimensional radiation transport, as well as "dusty" plasmas, coupled to the ionization structure of the interstellar gas. Because plasma conditions (temperature, density, ionization state, magnetic field intensity) vary enormously in the physical regions of interest, it is not likely that a single evolution equation that correctly describes the physics in all regimes can be constructed. Instead, the challenge is to couple correctly, distinct evolution equations operating in distinct physical regions. Current models treat cloud formation, dense-core formation, star formation, and planet formation independently and with major approximations to the physics. Future hardware and future software developments must allow coupling of the different scales and improvement of the physics at each scale.

Stellar Evolution. The challenge here is to develop 3D models incorporating convection, interior rotation, pulsation, (nuclear) chemistry, radiation (both photon and neutrino), and magnetic fields, and to integrate the resulting equations on time scales comparable to a star's lifetime. The dynamical time scales (for convection, for example) are enormously smaller than the evolutionary time scales, but at times (e.g., during "shell flashes"), these time scales can become comparable. Hence, a single scheme for integrating the stellar evolution equations is unlikely to be successful. Hybrid schemes need to be developed.

Stellar Death. In the case of Type Ia supernovae, the key missing ingredients for correctly describing the explosion mechanism are a subgrid model for the deflagration (flame front) to detonation (shock front) transition (now "subgrid" is meant in the traditional sense used by fluid dynamicists) and the proper use of mesh refinement to make the simulations feasible. In the case of core-collapse supernovae, the key issue is to develop a 3D, multifrequency, multiangle radiation (neutrino) transport capability. The analysis of both the simulation data and the observational data (in order to allow comparison of time-resolved spectra and light curves) are computational challenges in their own right. Three-dimensional, multifrequency, multiangle radiation transport is also needed here in order to connect simulation data with

observational data and to remove systematic errors in supernova "standard candle" determinations of cosmological parameters.

Numerical Relativity. The challenge is to follow numerically black hole mergers and neutron star mergers sufficiently long to understand the geometrodynamics of spacetime around such mergers and to predict the gravitational wave emission through all inspiral phases. On a technical level, these simulations will present many of the challenges presented by supernova simulations, with the added complexity that the state of applied mathematics for numerical solution of the Einstein field equations seriously lags that for the PDEs that describe radiation magnetohydrodynamics.

Astrophysical Data. The benefits of the digital revolution have also led to significant costs, mostly driven by the fact that the increased data-capturing capabilities (from both telescopes and simulations) have led to a flood of new data, challenging both our ability to distribute and store/archive the data, and to analyze them effectively. Thus, the Sloan Digital Sky Survey has already produced a multi-terabyte data set. The Large-aperture Synoptic Survey Telescope (an 8-m class telescope with a 2.3 gigapixel CCD) will yield yet larger data sets. And the anticipated very-large-area optical telescopes (with diameters of 30 meters or more) will trump even these data sets. In all of these cases, it is widely appreciated that without paying sufficient attention to data handling, effective utilization of the new data (whether observational or computational) will not occur. In part, this is an issue of computational infrastructure (e.g., storage, networking, and analysis engines and displays), but in part it relates to the "middleware" required to make the data usable over long periods of time — well-defined data-interface and data-structure standards and the tools for making all this available over the network (e.g., Grid technology).

Technology Barriers

Capacity versus Capability Computing. The state-of-the-art supercomputers available to astronomers today are largely oversubscribed, especially for large simulations that require full use of the entire machine. Thus, queues for large simulations can be discouragingly long, leading to very long turn-around times for runs. Doing physics — exploring the control parameter space of models by repeated simulations — becomes essentially impossible. While capability computing hardware must certainly be developed to address the target science discussed above, the present situation clearly indicates that future plans for compute engine resources must be designed for both capability and capacity.

Memory Size and Bandwidth. Significantly increased memory bandwidth and, more important, a balance between processor speed and memory bandwidth, is the single most desired characteristic of future architectures across our science subareas. Moreover, for many forefront astronomy simulations, memory size is a paramount issue as well. For example, for many cosmological N-body as well as supernova simulations (the latter using the adaptive mesh method), current limits on the "in-core" memory are the principal constraint on the size of the problems that can be addressed.

Communication Bandwidth. Global reduction operations are at the heart of many of the solution algorithms we use. For example, these operations are required to perform the inner product computations in iterative Krylov subspace methods for the solution of the large, sparse linear systems of equations that arise in radiation transport applications. Large communication bandwidth will significantly reduce the wall clock time needed to perform such global reduction operations.

Algorithms. In astrophysics simulation, spatial and temporal dynamic ranges of 10 to 15 orders of magnitude are possible, and raw compute power often buys only a factor of 5. Thus, algorithm developments, much more so than leaps in hardware capabilities, will be key to performing successfully the ambitious simulations outlined here. For example, the development of (a) efficient MPP-based multigrid solvers (e.g., for the solution of the Poisson equation for the gravitational potential) and implicit solvers (e.g., for the solution of our radiation transport equations), both on AMR meshes, (b) methods for adaptively varying the time integration scheme (from fully explicit to fully implicit) as the situation

demands, and (c) scalable methods to perform global reduction operations are examples of algorithmic advances that would have profound effects on the efficacy of astrophysics simulations.

Parallel I/O. There are substantial efforts underway to improve the performance of parallel I/O, which has been one of the major bottlenecks in degrading "wall clock" performance on massively parallel machines (MPPs) (MPMs?). But these efforts have not yet succeeded in widening the bottleneck.

Validation. Because of limited resources, much of the astrophysical code development in the open computing community does not receive the proper level of code validation. By "code validation" we do not mean bug checking ("verification") but, rather, an assurance program that a given code accurately reproduces experimental results for values of the dimensionless control parameters that coincide with the expected regime of validity of the code. A key scientific goal of laboratory astrophysics is indeed to provide the experimental data for validating astrophysical codes.

Resources Required

"Hard" Resources. It is relatively straightforward to describe the key characteristics of the next-generation of astrophysics simulations, virtually independent of the particular subfield: There will need to be orders of magnitude increases in resolution and (spatial and temporal) scales covered (relative to existing 3D simulations). The new 3D simulations will include significantly more physics (e.g., sophisticated radiation transport, cosmic magnetic fields). The use of adaptive meshes will see extensive use. And there will be a move towards mixed or fully implicit time-integration schemes (in order to handle both longer physical time scales and physical processes such as radiation, which cannot be effectively treated with explicit schemes). All of these attributes of the next-generation simulations have implications for future computing demands. Compute engines with (a) 50 Tflop/s to 10 Pflop/s sustained speeds, (b) large memory per processor, (c) large total memory (> 10 TB), (d) significantly increased memory bandwidth, and (e) significantly increased communication bandwidth (increased performance for global communications in particular) will be needed. Petabytes of storage for both cached and archived data will also be needed, as well as network throughputs > 100 Gbps on all networks (both local- and wide-area networks). Moreover, dedicated paths and bandwidth on demand will be essential for effective interactive and collaborative visualization, particularly involving researchers at geographically distributed sites.

Metrics of Success

The metrics of success are straightforward: In each of the subfields of astronomy, the aims of simulation are well-defined. "Success" will be measured first and foremost by the extent to which the simulations reproduce the myriad observations and the extent to which they provide a theoretical framework within which all observed phenomena can be understood. In addition, it can be strongly argued that the extent to which validation of astrophysical codes is successful can be used as a further metric of success of the computational astrophysics program. The extent to which applications in computational astrophysics drive the development of "fundamental" application areas, such as computational fluid dynamics and magnetohydrodynamics, radiation transport and radiation hydrodynamics, etc., which are of relevance to many application areas across the DOE Office of Science and across federal agencies, will also be a key measure of success. Finally, the extent to which computational astrophysics as an application area helps "bring up" the next generation of computational scientists in the U.S. will be another key measure of success.

D.3 Letters of Support from Collaborators

Berkeley Lab's proposal to establish NFACS has generated enthusiastic support within the national scientific research community. Below is a list of researchers who have contributed letters of support. [This edition of the proposal omits the letters to save space.]

Leadership Computing Consortium

Nicholas M. Donofrio, Senior Vice President, Technology and Manufacturing, IBM

Dona L. Crawford, Associate Director, Computation, LLNL

Rob Pennington, Interim Director, National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign

Francine Berman, Director, San Diego Supercomputer Center and NPACI, Professor and High Performance Computing Endowed Chair, University of California, San Diego

Rick L. Stevens, Director Mathematics and Computer Science Division, Argonne National Laboratory, Professor, Department of Computer Science, University of Chicago

Rick L. Stevens, Distributed Terascale Facility, Project Director, Department of Computer Science Research Institute, The University of Chicago

Edward Seidel, Floating Point Systems Professor of Physics and Computer Science, Louisiana State University

Ruzena Bajcsy, Director, Center for Information Technology Research, In the Interest of Society (CITRIS), University of California, Berkeley

David H. Bailey, Leader, SciDAC Performance Evaluation Research Center, Computational Research Department, LBNL

Moe Khaleel, Director of Computational Sciences and Mathematics, PNNL

Jonathan Ormes, Director of Space Sciences, NASA Goddard Space Flight Center

Leadership Class Application Teams

Nanoscience

Roberto Car, Professor of Chemistry, Princeton University

David Ceperley, Professor of Physics, University of Illinois-Urbana Champaign

Giulia Galli, Quantum Simulations Group Leader, Physics & Advanced Technologies, LLNL

Steven G. Louie, Professor of Physics, University of California, Berkeley

Alex Zunger, Ph.D. Research Fellow, National Renewable Energy Laboratory

Combustion

John Bell, Group Leader, Center for Computational Sciences and Engineering, LBNL

Ahmed F. Ghoniem, Professor of Mechanical Engineering, Massachusetts Institute of Technology

Heinz Pitsch, Assistant Professor, Department of Aeronautics and Astronautics and Mechanical Engineering, Flow Physics and Computation Division, Stanford University

Larry A. Rahn, Senior Scientist, Combustion and Physical Sciences, Sandia National Laboratories, Livermore

John N. Shadid, Computational Sciences Department, Sandia National Laboratories, Albuquerque

Arnaud Trouvé, Associate Professor, SciDAC Project Leader, University of Maryland

Fusion

Stephen C. Jardin, Director, SciDAC Center for Extended Magnetohydrodynamic Modeling, Princeton Plasma Physics Laboratory

William M. Tang, Chief Scientist, Princeton Plasma Physics Laboratory

Ronald Cohen, Associate Program Leader, Fusion Energy Program, LLNL

Carl Sovinec, Assistant Professor, Department of Engineering Physics, University of Wisconsin-Madison

Climate

William D. Collins, Chair, CCSM Scientific Steering Committee, NCAR
Inez Fung, Director, Center for Atmospheric Sciences, University of California, Berkeley

Astrophysics

Eddie Baron, Professor of Physics and Astronomy, The University of Oklahoma

Richard I. Klein, Professor of Astronomy, University of California, Berkeley

Edward Seidel, Professor of Physics and Computer Science, Louisiana State University

Michael L. Norman, Director, Laboratory for Computational Astrophysics, Professor of Physics, University of California, San Diego

Jonathan Ormes, Director of Space Sciences, NASA Goddard Space Flight Center

F. Douglas Swesty, Professor of Physics and Astronomy, SUNY at Stony Brook

Stan E. Woosley, Professor of Astronomy and Astrophysics, University of California, Santa Cruz

Office of the Senior Vice President
Technology and Manufacturing

New Orchard Road
Armonk, NY 10504

March 31, 2004

Dr. Horst D. Simon
Associate Laboratory Director for Computing Sciences, and
Director, National Energy Research Scientific Computing (NERSC) Center
Lawrence Berkeley National Laboratory
One Cyclotron Road, MS-50B-4230
Berkeley, CA 94720

Dear Horst,

DOE's Leadership Class Computing System is a strategic step forward for the United States to provide the most capable computing system to tackle the most challenging science problems. We are pleased the DOE Office of Science has taken the lead to provide this level of computing to the national science community.

As the vendor of the Leadership Class Computing System, IBM is committed to provide an innovative, extremely effective, high performance computing system of unprecedented performance and capability. The proposed systems not only provide outstanding price/performance for capability problems, but also provide excellent absolute performance for science.

A wide range of IBM technology will be brought to bear to assure the LCS systems are an outstanding success. The scale, schedule and requirements for LCS mean the standard roadmap of technology is not sufficient. IBM has already adopted the concepts of "Science Drive Architecture Design" in redesigning the Power 5/6 node to focus on the balance of Flop/s and memory bandwidth. This new 8-CPU single core node is now accepted for ASCI Purple and is the basis of the LCS-1 system. We will continue the Science Drive Design approach as the details of the LCS-2 system are defined and implemented.

IBM will continue innovation to reduce latency and to improve the efficiency of science codes. The concept of virtual vectorization - accelerators that will improve efficiency of codes while still leveraging the cost effectiveness and balance of our high volume CPU cores - is something IBM will assess. We will work with LBNL and the application areas to complete the design of this effort and create effective methods to explicit the new functionality. Further, IBM is committed to improve the interconnect of the LCS systems, particularly for latency sensitive algorithms. There are multiple approaches we are exploring, from collective off load functions to enhanced adaptors for Power 5/6.

In summary, IBM views contributing to the LCS effort as a strategic step forward. We are committed to work with LBNL and the Office of Science to accomplish the aggressive goals for performance and cost effectiveness. IBM will devote significant research, design and development resource to delivering the LCS for the DOE science users.

Sincerely,

Nicholas M. Donofrio

D.4 Compliance with Section 307 of the Consolidated Appropriations Resolution, 2003

Compliance of this proposal with Section 307 of the Consolidated Appropriations Resolution, 2003, is described in the letter of the Laboratory Counsel reproduced on the following page.

Office of the Laboratory Counsel
EXT. 7025 ▪ FAX 6498 ▪ MS: 50A-4112

April 8, 2004

IN STRICT CONFIDENCE
ATTORNEY CLIENT PROTECTED INFORMATION

To: Horst Simon
Associate Laboratory Director for Computing Sciences

From: Glenn R. Woods
Laboratory Counsel

Re: LBNL Proposal to the DOE Office of Science – National Facility for Advanced Computational Science (NFACS)

This is written in regard to our meeting today concerning the Laboratory's NFACS proposal and Section 307 of the Consolidated Appropriations Resolution, 2003. You have asked me to review the proposal in view of the constraints set out in Section 307. My analysis and conclusions are set forth below.

In regard to the statements in the LBNL proposal concerning Applications Partnerships, it is clear that there is no promise of any privileged access to applications scientists and that a competitive peer reviewed access and use process will be used. In addition, in regard to the statements in the proposal concerning the Leadership Computing Consortium Members, it is clear that membership is open to the computational science community and that charter members obtain no privileges not available to others who may wish to join later. Finally, in regard to the planned relationships with the San Diego Supercomputing Center (UC San Diego) and the National Center for Supercomputing Applications (University of Illinois, Urbana-Champaign), it is my opinion that the unique nature of these facilities and their compatible technology with NFACS, provides a strong basis to partner with these NSF funded national facilities in full compliance with Section 307.

In view of the above, and the revisions which have been made to the proposal, it is my opinion that the NFACS proposal is in compliance with Section 307.

Please let me know if you have any questions.

APPENDIX E
Facilities and Resources

E.1 Berkeley Lab Support Strategy

Berkeley Lab senior management is committed to providing cost-effective infrastructure for the national user facilities and scientific resources with which it is entrusted. In 2000, when NERSC required additional space and connectivity to high-speed data links, Berkeley Lab worked with a third-party developer to remodel an available off-site building into the Oakland Scientific Facility (OSF).

It is a Laboratory strategic goal to have high performance computing at a central location on the main site. The site identified for this is the Bevatron site, a central location and the major focus of the Laboratory's principle research activities during the 1950s–1980s. Demolition of the External Particle Beam structure at the Bevatron site this year creates an optimal building site for a modern facility to house the LSC-2 system and consolidate all high performance computing within Berkeley Lab proper.

Recently, the DOE has encouraged third-party financing approaches to facilities construction, and these approaches will enable Berkeley Lab to provide the requisite NFACS building by 2007. Because Berkeley Lab is located on University of California (UC)-owned land, this process is actually less complicated for Berkeley Lab than for those national laboratories situated on federal land, which must be transferred via a quitclaim deed to a development entity. The University can simply enter into a long-term ground lease with a developer at a nominal cost. When the building is complete, DOE approves a UC lease of the facility a year at a time over the life of the building. Berkeley Lab and the University are currently developing a research office building on the main site targeted for completion in 2006 via such a third- party development. The experience and knowledge gained through this procurement give us every confidence that the NFACS building can be completed on time.

Construction of a building in this way is cost effective for the program. As is the case with the Oakland Scientific Facility, the annual lease costs would appear as part of an overhead "space charge," which includes all Berkeley Lab lease costs and is distributed across all Laboratory programs.

E.2 Building and Physical Infrastructure

Berkeley Lab's Oakland Scientific Facility includes a 20,000-square-foot computer floor and associated utilities for cooling and power. It houses a number of diverse computational resources, including two major clusters (Alvarez and PDSF, a High Performance Storage System, institutional systems specific to Berkeley Lab information technology (IT) infrastructure, and the 6,656-processor IBM SP of NERSC. Additional computer equipment for NERSC is anticipated this fiscal year. Currently, there are 5,000 square feet of computer floor available for an additional system. The LCS-1, described in this proposal, will require 2,400 net square feet and will readily fit in the existing OSF.

The follow-on system, LCS-2, will be housed in a new dedicated computer building. The site is located in the center of the Berkeley Lab campus, on part of the site of the decommissioned Berkeley Lab Bevatron, which was recently demolished. The building has been designed after the new computer facility at Sandia National Laboratory, Albuquerque, that was built and occupied in less than two years. The Berkeley Lab building will be ready to accept LCS-2. The 20,000-gross-square-foot building will contain a computer room and utility support space. The building will be on a slab, with insulated metal wall and roof panels. The clear span of the building structure will provide maximum operational flexibility. It will have a 3 foot raised floor with mechanical and electrical distribution systems. The ceiling will be approximately 20 feet high. The building will have state-of-the-art fire protection systems. Mechanical/ electrical support equipment will be located adjacent to the building. This site has all major utilities in place and the availability of up to 12 megawatts of electrical power, from the days when the Bevatron was

operative. This site also provides for the ability to expand into a second adjacent 20,000 square feet of computer floor, yielding a 40,000-square-foot computer complex. Site plans and conceptual building renditions are shown in Figures E-1 to E-3.

The contingency plan for housing LCS-2 is expansion of the OSF to gain another 20,000 square feet. The OSF was designed for such a contingency, and it can be exercised in time for LCS-2.

Berkeley Lab, therefore, has existing and committed space for NFACS.

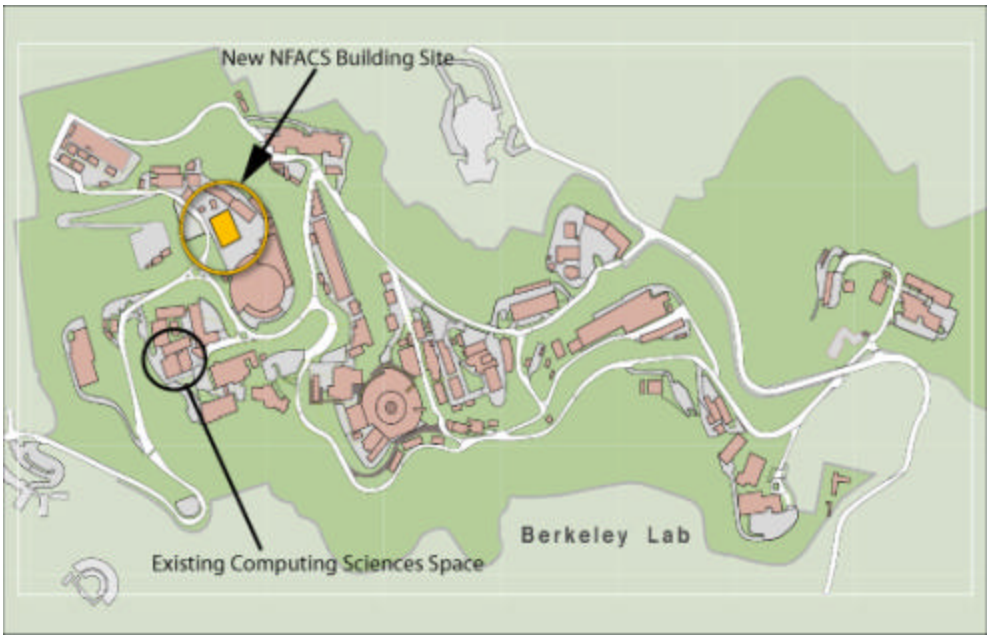


Figure E-1. Site plan locating the NFACS building on the LBNL campus, near the existing research space and offices of Computing Sciences.

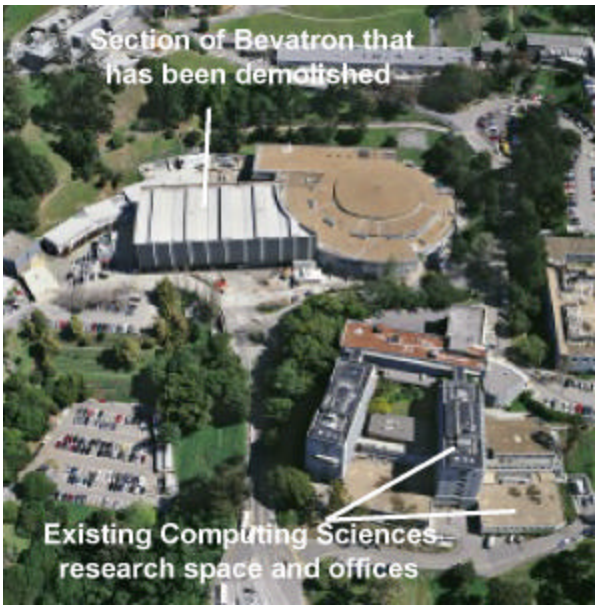


Figure E-2: Site before demolition of portion of Bevatron

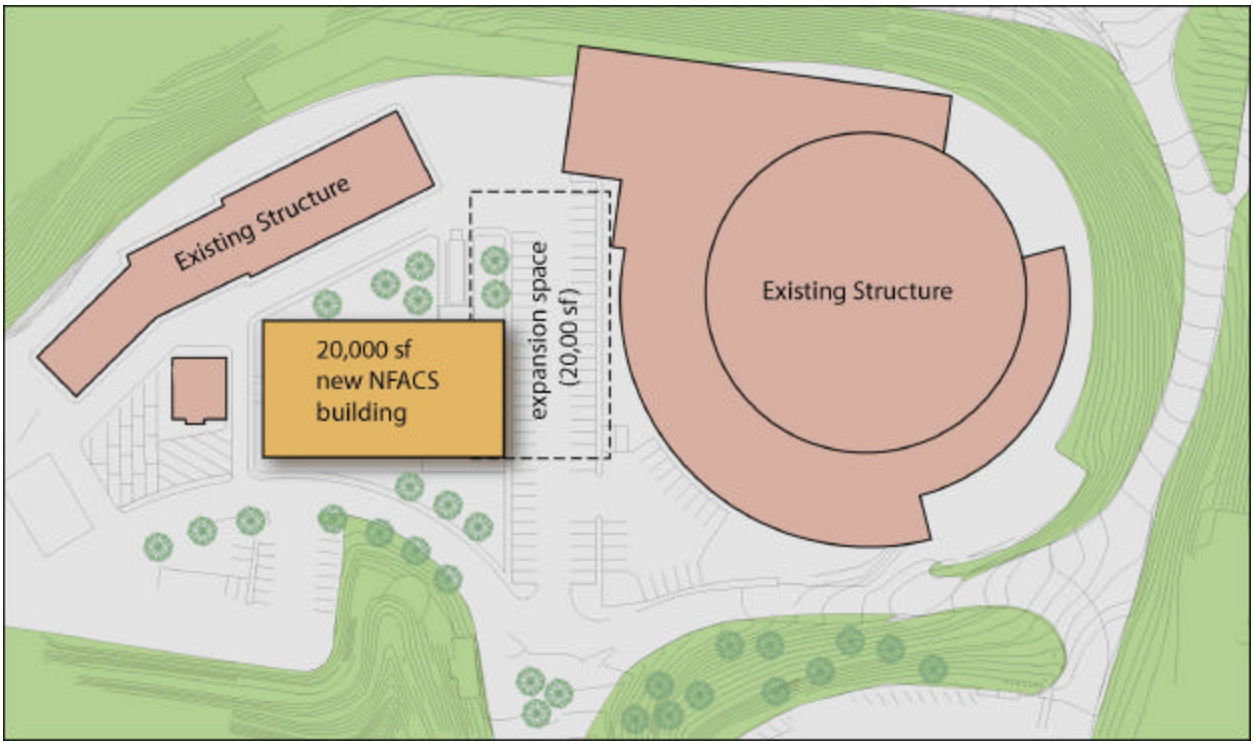


Figure E-3. Building site plan, showing available space for the 20,000 sf NFACS building and adjacent expansion space. Demolition to release this site was completed in March 2004.

Figure E-4. Conceptual drawing of the NFACS building.

E.3 Networking, Data Storage and Archives, and Security

The process of large-scale science is changing. Very large data sets from experiments, simulations, and sensor arrays are critical resources for today's science, and those resources must be provided efficiently and securely to remote high-performance computing and storage systems. At the same time, a new science paradigm is emerging in which multiple national assets — computational, data, and experimental—are employed simultaneously in the process of scientific discovery. Projects such as the DOE Science Grid are constructing the infrastructure of software and services that will automate the task of using these systems in concert for large-scale scientific problem solving.

As the home of the Energy Sciences Network (ESnet) and NERSC, the lead site for the DOE Science Grid, and one of the original six development sites for the High Performance Storage System (HPSS), Berkeley Lab has already made significant progress in integrating high-end computing, storage, and data management into the Grid environment. We will do the same for NFACS, thereby facilitating large-scale science for DOE and the nation.

Networking

As a national facility whose users routinely transfer massive quantities of data, NFACS is geared towards providing wide-area-networking connectivity at the highest possible performance. Berkeley Lab and NFACS are located near the primary switching point for national networks in Northern California at Sunnyvale — home to both the Qwest and Level3 networking hubs. The Qwest hub is the transit point for the backbones of major production networks such as DOE's ESnet, the National Science Foundation's (NSF's) Abilene, the National Aeronautics and Space Administration's (NASA's) NREN (NASA Research and Education Network), and the NSF TeraGrid, while the Level3 hub carries experimental dark-fiber networks such as the National Lambda Rail, the DOE Ultranet, and the Corporation for Education Network Initiatives in California (CENIC)/Pacific Light Rail. The proximity allows NFACS easy and cost-effective access to each of these networks. In order to promote interaction with and outreach to scientists in industry, academia, and other federal programs, Berkeley Lab will work closely with ESnet to create network peering arrangements that will maximize the effective remote access to NFACS users regardless of their institutional affiliation and facility location.

In order to ensure the highest-performance-available production network access to NFACS, Berkeley Lab will immediately upgrade its connection to Sunnyvale to optical cable OC-192 in order to match the existing backbone bandwidth of the ESnet and Abilene production networks. In order to provide more effective access to the NSF user community, ESnet and Abilene are finalizing an arrangement to peer their networks at each of these co-located hubs at Sunnyvale, Chicago, New York, and Atlanta to create a common network backplane that provides full connectivity between the labs and universities comparable to what either backbone alone can provide. ESnet and the Abilene will also deploy a monitoring infrastructure to ensure that the quality of service is maintained. These upgrades coincide with NERSC's movement to a 10-Gigabit internal network infrastructure, which is already under way. Both the upgraded internal network and wide area network infrastructure will be immediately available to the first-generation NFACS system and will continue to be expanded to match the scale of successive systems and continuously match the performance improvements of the production-network backbones.

In the first year of NFACS operation, ESnet will deploy a multiprotocol label switching (MPLS)-based quality of service (QoS) service that operates initially between ESnet border routers. This service will support guaranteed fast data paths between selected sites as a schedulable service. Within the next two years, ESnet and Abilene will deploy a system that will allow dynamic provisioning of circuits across both networks as envisioned by the Internet2 Hybrid Optical/Packet Infrastructure (HOPI) working group. This will enable dedicated "bandwidth corridors" that support high-speed transfers between sites using efficient fixed data rate protocols. This will support NFACS storage peering arrangements between other laboratories and NSF-Partnership for Advanced Computational Infrastructure (PACI) supercomputing centers in order to support our vision of a nationwide supercomputing infrastructure.

In addition to its support of production network infrastructure, the NFACS system will tie in to major experimental and dark-fiber networks, such as the TeraGrid, DOE Ultranet, and National Lambda Rail, in order to add its capabilities to a vibrant research community that combines sensors, archival data, and supercomputers to accomplish large multidisciplinary scientific projects. Berkeley Lab proposes connecting NFACS to the NSF TeraGrid with three OC-192 Packet over Sonet (PoS) lambdas to the Los Angeles TeraGrid hub. This link would provide NFACS users with the same level of connectivity as the currently established TeraGrid sites. Details of the proposed TeraGrid connection are provided in following section and the letter of support from the TeraGrid Consortium in Appendix D. In addition, ESnet and CENIC are in the process of defining a 10-Gigabit dark-fiber metro area network (MAN) that will link together Bay Area universities and research laboratories. The MAN will support a mixture of experimental and production network activities and will offer the capacity to dynamically expand to accommodate demand through dense wave division multiplexing of signals over the fibers. The customer-owned dark-fiber MAN can therefore accommodate a variety of network peering options available via Sunnyvale in a cost-effective manner. As part of the project, we will review high-performance networking options on a periodic basis and will transparently shift to lower-cost technologies as they are available.

In collaboration with ESnet and other sites, Berkeley Lab will be active in deploying the latest enhancements in local and wide-area networking systems and protocols, such as 10-Gigabit Ethernet and InfiniBand, to enable NFACS clients to move data and to provide system access for enhanced services such as remote steering, Grids, and visualization. The Laboratory will continue its emphasis on eliminating bottlenecks in the wide area network–local area network (WAN-LAN) boundary, since this is the key to success in many bandwidth-intensive applications. To deliver the full capability of NFACS system to its users, Berkeley Lab is committed to continuing its role as a center of excellence in network engineering.

Connecting NFACS to the TeraGrid

To promote interaction and outreach to scientists in industry, academia, and other federal programs, Berkeley Lab proposes connecting NFACS to the NSF TeraGrid with three OC-192 PoS lambdas to the Los Angeles TeraGrid hub. Consequently, the TeraGrid Consortium has extended its invitation to establish this connection if NFACS is funded. This link would provide NFACS users with the same level of connectivity as the currently established TeraGrid sites.

Dark fiber, dim fiber, and managed lambda connections to Los Angeles using several vendors were evaluated. The proposed solution, purchasing managed lambdas from the not-for-profit CENIC, was selected not only because it was the least expensive but also because it may be possible to obtain a multiyear irrevocable right of use (IRU). Having CENIC as the transport provider is by far the most cost-effective solution and also has other benefits. The L.A. TeraGrid hub is co-located in a CENIC rack, so the L.A. buildout is simplified to adding additional OC-192 interfaces to the existing Distributed Terascale Facility (DTF) Juniper hub router and connecting three fiber jumpers to the CENIC wavegear within the rack.

CENIC will multiplex the three OC-192 PoS lambdas onto their existing fiber and transport them to Emeryville, the closest peering point, about two miles from Berkeley Lab's OSF, the site of the NFACS and NERSC computing center. There they will be demultiplexed and cross-connected to three pairs of Metromedia Fiber Network Inc. (MFN) dark fibers, which will terminate at the 415 20th Street, Oakland, CA, PoP (point of presence). This is located in the basement of the OSF.

The three OC-192 circuits will be cross-connected to a Juniper T640 at the OSF, which will serve as the dedicated TeraGrid site router for NFACS's presence on the TeraGrid. After considering several alternatives, Berkeley Lab chose a Juniper router for maximum compatibility with the existing TeraGrid infrastructure and for the advanced features available on the Juniper T-series platform. Although the connection to the TeraGrid could be accommodated with a Juniper T320 router, Berkeley Lab proposes utilizing a T640. For a 4% higher cost than a T320, the T640 will provide a cost-effective upgrade path to

OC-768 and will also allow for significant growth space if other Northern California sites are authorized to connect to the TeraGrid through the OSF router. In addition to the OC-192 interfaces, the T640 would be configured with three 10-Gigabit Ethernet interfaces and one multiport 1-Gigabit Ethernet interface card. The three 10-Gigabit Ethernet interfaces would be connected as a single Ether Channel to the OSF TeraGrid switch.

After evaluating 10-Gigabit network switches from multiple vendors, Berkeley Lab proposes a Foundry Networks BigIron MG8 switch (in layer 2 mode only) as the dedicated TeraGrid switch. Consisting of one four-port 10-Gigabit Ethernet line card and multiple 1-Gigabit Ethernet cards, the Foundry switch will provide for gigabit aggregation in connecting with NERSC and NFACS computing and storage systems. The MG8 switch was chosen over its competitors for several reasons. First, the design and performance capabilities of the MG8 qualify it as an outstanding candidate. This, coupled with a cost savings of over $130k compared to a Force10 switch, makes it the most cost-effective choice. Finally, Berkeley Lab's excellent relationship with Foundry Networks and the proximity of their headquarters complete the equation for a successful solution. Foundry Networks is a market leader in 10-Gigabit port shipments, was among the first vendors to release a 10Gig-E product, and is a profitable company no longer dependent on venture capital.

Berkeley Lab proposes connecting NERSC's HPSS to the TeraGrid switch by adding a dedicated 10-Gigabit interface to each of two HPSS nodes running on IBM p655 AIX systems. With IBM's 8 Gb/s disk subsystem (four 2 Gb/s Fibre Channel) matching the network interface performance, the HPSS system has the underlying hardware to sustain much greater than 40–50% efficiency.

HPSS has the capability of striping data transfers, not just across interfaces in a given node, but across multiple nodes. This would allow our two HPSS nodes connected to the TeraGrid to work in concert on a single transfer, enabling sustained data rates above 10 Gb/s. NERSC's role in the HPSS development community will enable functional requirements to be efficiently added to HPSS and will enable the TeraGrid as a testbed for further enhancements, as well as providing strong production capabilities.

Data Storage and Archives

NERSC's HPSS has enough capacity to serve both NERSC and NFACS clients. NERSC currently stores approximately 1,050 terabytes (TB) of data (30 million files) and handles between 3 and 6 TB of I/O per day. The current maximum capacity of NERSC's archive is 8.8 petabytes (PB) at current tape densities; the buffer (disk) cache is 35 TB; and the maximum transfer rate is 2.8 gigabytes per second. NFACS will require large amounts of archival storage and will invest in new tape technology. For LCS-1, 500 GB tape drives and cartridges will be added to the NERSC HPSS, giving a total maximum capacity of 4.5 PB just for LCS-1. For LCS-2, 1 TB cartridges will be deployed, adding 5 PB a year (10 PB total for the time period of the proposal) to the potential NFACS storage capability, for a total of 15 PB of storage.

Berkeley Lab is one of the original six HPSS development partners and thus serves on both the HPSS technical and executive committees. Berkeley Lab and NERSC demonstrated the first Grid-enabled HPSS service at SC2001; at SC2002, Berkeley Lab demonstrated a Web-based portal to HPSS that allowed users to drag and drop files between GridFTP servers. Currently, NERSC operates DOE's first Grid-enabled production HPSS gateway (garchive.nersc.gov), which consists of Grid-enabled HPSS FTP servers that can talk to standard GridFTP clients. We also have special Grid-enabled HPSS parallel FTP clients that are interactive and allow for high-speed parallel data transfers directly to and from HPSS movers. Our code to enable Grid access to HPSS has been accepted by the HPSS consortium and will be merged into the main code releases. In addition, we have been instrumental in defining a vision for HPSS's role on the Grid. We are also in the early stages of testing a Grid-enabled Hierarchical Storage Interface (HSI) service, and are working on a production Web portal for HPSS using GridFTP, as well as a reliable file transfer service.

Berkeley Lab is also very involved in developing and providing the infrastructure for high-speed WAN access to archival storage. The NERSC HPSS contains a wealth of climate, physics, and astronomy data, routinely accessed by participants in the DOE Earth System Grid via two high-data-rate access mechanisms, GridFTP and the HRM (Hierarchical Resource Manager), which are running on a Net100 host as well as directly from HPSS. Net100 is a Linux kernel that has been modified to include a suite of techniques to enable high-speed, wide-area data transfer: fast-start TCP, auto-tuning of window size, auto-determination of the appropriate number of parallel streams to use, built-in monitoring for end-to-end performance debugging, etc. HRM is a tertiary storage system interface that optimizes multifile access to tape-based systems like HPSS, provides caching, and can participate in Grid file replication. HRM is widely used in the high energy physics community, as well as in the Earth System Grid. GridFTP provides parallel stream file transfer, third party transfers, Grid security, etc. HRM can provide a back end for GridFTP to optimize access to the tertiary storage system. This combination provides a significant increase in throughput in very high performance networks, and provides ease of use through the automated management of multi-file transfers and network and host failures.

Shared storage will be important for NFACS clients, as wider collaborations require transparent access to a common code base, data, and metadata. The NFACS HPSS will be federated with archival storage systems (both HPSS and Unitree) across all sites involved in the NFACS Leadership Computing Consortium (LCC). Users of the LCS systems will have equal access to archival data across NFACS, NERSC, and PACI facilities through the LCC storage federation. Close coordination of certificate management between DOE Science Grid, TeraGrid, and PACI sites will enable single-sign-on access across facilities and seamless transfer of data between archival storage systems. And the "bandwidth corridors" described above will support dedicated high-speed data transfers between the sites for efficient mirroring and staging of massive datasets between their respective storage systems.

In addition to archival storage systems, NFACS will be part of a wide-area shared file system that will link together all LCC partner sites including the NSF PACI supercomputing centers, NERSC, and Louisiana State University (LSU). The file system will be based initially on a WAN Global Parallel File System (GPFS) that is being developed through a partnership between IBM Research and the San Diego Supercomputer Center (SDSC), and will be usable across both Linux and IBMSP supercomputing infrastructure at LCC partner sites. In demonstrations conducted by SDSC this past year, GPFS sustained well over 900 MB/sec over a wide area 10-Gigabit link. The shared file system will enable more flexible migration between the systems for users who have shared accounts, and will help the LCC consortium form a well-integrated computing environment that better serves a national scientific user community.

Grids

As the home of ESnet and NERSC, the lead site for the DOE Science Grid, and one of the original six development sites for HPSS, Berkeley Lab has already made significant progress in integrating high-end computing, storage, and data management into the Grid environment. We will do the same for NFACS, thereby facilitating large-scale science for DOE and the nation. NERSC has established ties with all major Grid efforts in DOE and NSF and is closely collaborating with the DOE Science Grid and all its partners. The NFACS center staff will leverage the NERSC Center staff's broad experience with Grid software and services. We will work in close coordination with the LCC members to establish the peering of Certificate Authorities and trust relationships necessary to support coordinated access to Grid services. An interface to the NERSC Information Management (NIM) system will make it easy for NFACS users to get Grid authentication certificates and will form the basis for coordinated management of Grid certificates that will support single-sign-on access to Grid services across all LCC partner sites, including the PACI centers, NERSC, and the TeraGrid Consortium.

Visualization and Data Analysis

High-end visualization and data analysis tools will be essential to turn raw simulation data into scientific discoveries. NFACS will work closely with its LCC partners to apply technologies developed across the coalition available to the user community. In particular, we will work closely with Lawrence Livermore National Laboratory (LLNL) to share, test, debug and deploy together the latest ASCI tools for visualization of massive datasets, including utilization of commercial technologies to achieve new levels of graphics performance, the LLNL/Stanford-developed distributed parallel rendering software (Chromium), and proven parallel, scalable end-user applications (like VISIT and Blockbuster movie player), and the Terascale Browser. The Berkeley Lab/NERSC visualization group will also provide LCS users and LCC members with access to the VisPortal, which automates complex workflows like the distributed generation of MPEG movies or scheduling of file transfers, mediates access to limited hardware resources like off-screen graphics pipes, and controls the launching of complex multicomponent distributed visualization applications like Berkeley Lab's Visapult — an application used for remote and distributed, high performance interactive volume rendering of massive remotely located datasets. All of these tools will be tightly coupled with the high-speed networks, coordinated Grid services, storage federation and WAN GPFS capabilities deployed across the LCC sites.

Security

It is the policy of Berkeley Lab to provide a safe and secure work environment. To sustain its scientific mission, it is important that the Laboratory protect its resources and assets, both intellectual and material. Only necessary technical staff have access to computer rooms and computer facilities. The general staff and the public do not have physical access to these computer resources. All laboratory assets are tracked and protected by laboratory security services. NFACS users will access the system remotely, subject to all Berkeley Lab cyber security policies, controls, and restrictions. At the same time, as a multi-purpose, open, unclassified laboratory, it is essential that Berkeley Lab remain an open environment that promotes free intellectual exchanges and collaborative efforts within the international scientific and technical community. As an unclassified facility, Berkeley Lab makes its facilities available for use by investigators from institutions throughout the nation and the world.

Berkeley Lab utilizes a state-of-the-art network intrusion detection system (IDS) called Bro, developed by Vern Paxson of ACIRI/LBNL. Bro is usually connected to network lines via passive taps and is capable of monitoring network traffic in excess of 1 gigabit per second. Berkeley Lab and Juniper have partnered to combine Bro with the unique capabilities of a Juniper router to passively monitor network traffic and detect anomalies at speeds in excess of 40 gigabits per second with no impact on the network, as demonstrated at SC2002. Berkeley Lab proposes to connect a Bro IDS to the T640 Juniper router. This experiment in high-speed intrusion detection will be of great benefit to the TeraGrid, both as a research tool and for anomaly detection, especially as the TeraGrid evolves with more sites coming online. Connection of the Bro IDS will not impact the performance of the router or the network.

Berkeley Lab uses a best-practices approach to cyber security, ensuring that known security problems are fixed and that systems and networking are proactively managed to reduce exposure to risk while simultaneously maintaining an open environment. In order to maximize our ability to conduct science and mitigate the effects of computer security incidents, the Laboratory provides noninvasive advanced monitoring and automatic reactive tools using components that are embedded in the network as well as in every computational and storage system. Berkeley Lab's active security infrastructure is able to detect cyber attacks, detect vulnerable or compromised hosts, and initiate a large-scale coordinated response to cyber-security incidents without resorting to methods that impede legitimate system access. For example, firewalls are creating significant roadblocks to pervasive deployment of Grids. Berkeley Lab uses an active intrusion detection system that offers a compelling alternative to standard firewalls as a means to defend against cyber attacks. DOE and the Department of Homeland Security are funding efforts to extend this system to sites other than Berkeley Lab. The Laboratory will continue to use and improve

these advanced monitoring tools to provide NFACS with the best level of security with minimal impact on performance and function.

Berkeley Lab has an outstanding security record and is recognized as a leader in cyber security within DOE and beyond. This expertise will make NFACS both secure and easily accessible.

APPENDIX F
Resources and Expertise at UC Berkeley, Berkeley Lab, and NERSC

Berkeley Lab enjoys an open and unrestricted intellectual environment that is easily accessible to scientists and visitors worldwide. Berkeley Lab is located in the heart of the San Francisco Bay Area, which is home to a large number of universities, laboratories, major facilities, and a vibrant scientific and research community. The Bay Area offers the critical mass of resources and intellectual leadership necessary to assemble an institution of international prominence like the National Facility for Advanced Computational Science (NFACS).

F.1 University of California, Berkeley

Berkeley Lab's location, only a short five minute walk or shuttle bus ride away from the campus of the University of California at Berkeley (UC Berkeley), facilitates numerous formal and informal collaborations. Currently there are seven joint appointments of faculty from the Electronic Engineering and Computer Science (EECS) and Math Departments at UC Berkeley with Berkeley Lab Computing Sciences: David Culler, James Demmel, Susan Graham, Ming Gu, Arie Segev, Jonathan Shewchuck, and Katherine Yelick. The combination of NERSC facilities, combined with Berkeley Lab and campus computing efforts, creates a vibrant community for cross-institution and cross-discipline efforts in research in algorithms, architectures, and applications, and in training of future computational scientists. Several of these joint projects have had a significant impact on NERSC.

Collaboration with David Culler and the Millennium project on campus led to funding for the first cluster at NERSC (sponsored by an Intel grant), and subsequently additional joint cluster development. This work continues on campus (the Ganglia distributed monitoring and execution system for cluster is used on 500 clusters worldwide, and Millennium is being upgraded to a 1 Teraflop facility with CITRIS contributions from HP and Intel), with continued opportunities for collaboration.

Jim Demmel collaborates with Berkeley Lab in areas of numerical algorithms research, One ongoing example is the SuperLU software, which was used in a computation at NERCS that appeared in cover article in Science, and which led to a 5x speedup in the NIMROD fusion reaction plasma simulation code. Other examples include eigenvalue algorithms and cosmic microwave background radiation analysis algorithms.

Katherine Yelick leads the Berkeley Unified Parallel C (UPC) team, a collaborative effort centered at LBNL, which may produce more efficient and productive programming models for NFACS platforms. The UPC group is working on applications of the model, including a parallel mesh generation algorithm building on Jonathan Shewhuck's Triangle system, and Adaptive Mesh Refinement algorithms building on the Colella's Chombo effort. Graham leads the Harmonia project, which is building programming environments for sequential and parallel languages, in which program analysis is done dynamically as users edit their programs. Future plans include support for detection of race conditions and the integration of information for performance and debugging.

The Berkeley UPC project includes research into mechanisms for lightweight communication in the GASNet communication layer, which is designed for portability and performance. It includes a novel technique for efficiently pinning memory for direct memory access, which is used on current Infiniband and Myrinet systems and will be of direct use in LCS-2. The UPC team is also working on fast collectives that take advantage of network processor offload, and, in collaboration with Yelick and Demmel's BeBOP group, into automatic performance tuning of collectives. The group will extend the set of optimizations to include the fast hardware-based collectives in the LCS systems.

The BeBOP group has developed several optimizations for sparse matrix kernels on single-processor modern memory systems and techniques to automatically select optimizations based on the architecture and the matrix structure. They are working with Berkeley Lab scientists to extend these ideas to parallel sparse matrix kernels, and through the SciDAC TOPS effort, integrating their software into applications and libraries like PETSc. Future plans include an exploration of optimizations for vector architectures, both physical and virtual.

Kathy Yelick is also working with LBNL scientists on the evaluation of advanced architectures for scientific computing, including processor-in-memory, streams, VLIW (very long instruction word), and vectors. This LBNL architecture evaluation team worked closely with IBM in the early stages of VIVA design to understand the benefits and limitations of vectors, and what type of memory system was needed to support the more challenging U.S. Department of Energy (DOE) applications.

Jim Demmel is also the Chief Scientist for CITRIS, the Center for Information Technology Research in the Interest of Society, a four-campus, 200+ faculty research institute centered at Berkeley. Groundbreaking for a new $100M building to house CITRIS at Berkeley will occur in Fall 2004. A central CITRIS activity is the design and deployment of sensor networks, geographically distributed networks of thousands of tiny, inexpensive wireless micro-electro-mechanical system (MEMS) sensors that can collect enormous amounts of data for societal applications, ranging from energy efficiency in buildings, to disaster response and homeland security, to biomedical monitoring, to transportation network monitoring, to environmental monitoring. Many of these research projects at CITRIS have direct relevance to future NFACS goals (e.g., the handling of large data sets), so NFACS is looking forward to continued collaboration with UC Berkeley for the benefit of its user community.

F.2 National Facilities Managed by Berkeley Lab

Berkeley Lab has been a leader in science and engineering research for more than 70 years. Located adjacent to the Berkeley campus of the University of California, Berkeley Lab is a DOE National Laboratory managed by the University of California. Many of its scientific staff have joint faculty appointments with the University.

Berkeley Lab conducts only unclassified research across a wide range of scientific disciplines, with key efforts in fundamental studies of the universe; quantitative biology; nanoscience; new energy systems and environmental solutions; and the use of integrated computing as a tool for discovery. In addition to its 17 scientific divisions, Berkeley Lab hosts four DOE national user facilities. The focus of Berkeley Lab in managing these large, leading-edge scientific facilities is to provide the best resource, service, and support to the general scientific community, both university- and national laboratory–based. Berkeley Lab will draw on this extensive experience to manage NFACS.

The user facilities at Berkeley Lab support thousands of users throughout the country. In addition to the Advanced Light Source (ALS, third generation synchrotron light source) and the National Center for Electron Microscopy, Berkeley Lab manages two networking and computational facilities directly applicable to the management of the Leadership Class Computational Facility.

The Energy Sciences Network, or ESnet, is a high-speed network serving thousands of DOE scientists and collaborators worldwide. A pioneer in providing high-bandwidth, reliable connections, ESnet enables researchers at national laboratories, universities, and other institutions to communicate with each other using the collaborative capabilities needed to address some of the world's most important scientific challenges. Managed and operated by Berkeley Lab, ESnet provides direct connections to all major DOE sites with high- performance speeds, as well as fast interconnections to more than 100 other networks. Funded principally by DOE's Office of Science, ESnet services allow scientists to make effective use of unique DOE research facilities and computing resources, independent of time and geographic location.

Berkeley Lab also manages the National Energy Research Scientific Computing (NERSC) Center, a world leader in accelerating scientific discovery through computation. NERSC provides high-performance computing tools and expertise that enable computational science of scale, in which large, interdisciplinary teams of scientists attack fundamental problems in science and engineering that require massive calculations and have broad scientific and economic impacts. NERSC is the foremost resource for large-scale computation within DOE's Office of Science. At the heart of NERSC's current computer hardware capability are a 6,656-processor IBM RS/6000 SP with a peak performance of 10 teraflops and a High-Performance Storage System (HPSS) mass storage system with a 35-terabyte disk cache and an archival storage capacity of 8.8 petabytes.

Since it was established 30 years ago, NERSC has built an impressive record of technological leadership, unparalleled user support, and scientific achievement. Since its early days, institutions across the country and around the world have tapped NERSC's expertise and followed its model as they work to establish their own scientific computing centers.

As part of its commitment to provide the best systems and services to its users, NERSC conducts an annual survey to gauge satisfaction — and ask for ways to improve — in areas such as hardware, software, training, communications, and support. Asked to rate services on a scale of 1 to 7, users in 2003 rated NERSC's user support services at 6.55. Additionally, when responses show concern about certain aspects of center management, adjustments are made to try to improve the service or system. The 2003 survey results, which can be found at http://hpcf.nersc.gov/about/survey/2003/first.php, also include open-ended comments, including the following:

> *"NERSC simply is the best run centralized computer center on the planet. I have interacted with many central computer centers and none are as responsive, have people with the technical knowledge available to answer questions and have the system/software as well configured as does NERSC." —2003 NERSC User Survey Respondent*

The high ratings given by users reflects both the commitment and responsiveness of the center staff to manage systems for the best overall use. This commitment was demonstrated during the 1996 NERSC move from LLNL to LBNL, during which at least one computing system was always available. Also, when new computing systems are purchased and installed, their installation is scheduled so that users have complete access to existing systems while the new ones are tested and deployed for early use. This ensures that users continuously have access to the critical high performance computing (HPC) resources. NERSC also has a history of working with vendors to improve overall performance of these systems. For example, in 1998, NERSC was the first center to achieve a checkpoint/restart capability on the Cray T3E supercomputer. Working with IBM, NERSC's staff has been able to achieve performance runs of scientific applications at up to 68 percent of the system's theoretical peak performance, a vast improvement over the 5–10 percent of peak realized on similar systems at other centers. Another key measure of effectiveness is the overall utilization of each processor in a high-performance system. Over the past year, NERSC's IBM has been managed so that processors are utilized more than 90 percent of the time. This extreme utilization, however, must be balanced with good job-turnaround times. While filling the system with lots of small fast jobs might make for good statistics, DOE and NERSC are committed to capability computing, or running jobs requiring 512 processors of more. Balancing jobs of varying processor requirement and length is critical to making the most efficient use of the system — and keeping users satisfied. At users' suggestions, NERSC staff have continued to fine-tune the scheduling system to achieve the most beneficial turnaround times.

NERSC has been a leader in capability computing for years. In 1998, NERSC staff were part of the team that was the first to achieve true 1-teraflop/s performance for a scientific application. More recently, when the DOE Office of Science decided to launch a new program to select a small number of computationally intensive, large-scale research projects that can make high-impact scientific advances through the use of a substantial allocation of computer time and data storage, NERSC was selected as the

center best able to provide this capability. Announced last December, all of the first three INCITE awards were made to university scientists. Because of robust networking connectivity and cybersecurity, Berkeley Lab has made the interface of the university and national laboratory environments seamless. This experience and tradition is necessary for support of NFACS if it is to be available to all meritorious unclassified computational research.

Berkeley Lab's Infrastructure and Operations Plan for NFACS will leverage the existing NERSC Center's physical, organizational, intellectual, and support infrastructure. While NFACS and NERSC will operate separate computing systems and will preserve their distinct missions, the shared infrastructure will maximize the DOE's return on investment from both user facilities. As a standalone facility, NFACS would require a staff of 25 to 30; but integrated into the NERSC Center's operations, NFACS will need only 12 additional staff members to provide the same high level of services and support that NERSC users are accustomed to. Five of those staff are dedicated to the science areas. Just as Berkeley Lab's computer scientists, computational scientists, and applied mathematicians have contributed to the NERSC Center's track record of scientific breakthroughs and accomplishments, we can expect the same kinds of fruitful collaborations with researchers using NFACS.

F.3 Large Scale System Management at NERSC

NERSC provides both high-end system resources and a comprehensive support structure for over 180 projects. Much information about NERSC can be found at the NERSC Web site (www.nersc.gov), including annual reports of scientific achievement.

NERSC is one of the national user facilities operated by Berkeley Lab. Others include ESnet, the Advanced Light Source, and the National Center for Electron Microscopy. NERSC systems are located in the 20,000 sf Oakland Scientific Facility — a state-of-the-art computer complex that houses NERSC and other LBNL systems.

NERSC provides exceptional support the entire DOE computational community. The major physical resources are the 10 Tflop/s IBM SP with 6,656 Power3 processors, almost 8 terabytes of main memory, and more than 70 terabytes of storage. NERSC's HPSS system provides 9 petabytes of archive capacity. It provides high bandwidth, capability if sustained parallel transfers of more than 90 megabytes per second over jumbo frame Gigabit Ethernet. NERSC also operates an advanced visualization server, a 700-CPU Linux cluster devoted to high-energy and nuclear physics data analysis, and a variety of smaller systems.

Just having physical resources does not produce effective science. NERSC provides a full range of services starting with 24 by 7 by 365 system operations and support. Computer and Network operators monitor NERSC systems continuously, making sure the systems operate at full performance. They handle a range of issues and are coordinated with other support staff to provide timely response. Operations is backed by a professional staff of system administrators who are always available to respond to system problems and user issues. NERSC systems, storage and network staff combine standard system administration with deploying and improving advanced technology. Some accomplishments of recent years are:

- NERSC systems — even the low serial number ones — provide highly reliable service, in excess of 99% of scheduled availability.
- Since 1997, the first year of production of the massively parallel processing system, NERSC parallel systems have operated at 90–95% utilization. A very large percentage of time goes to large jobs. On the IBM SP, these are jobs that run between 512 and 4,096 processors on a day-in/day-out basis.
- NERSC staff have improved networking rates for users. A number of projects have shown a factor of 5 to 20 time improvement in transfer rates.

- NERSC storage staff provide a highly scalable storage system, capable of getting any level of performance.
- NERSC fielded the first large-scale production Linux cluster in 1997. It is still in operation, with continuously improving technology.
- NERSC is leading DOE and a number of sites by having all of its resources already on the Grid.
- NERSC is noted for its very well protected, yet open and flexible cyber security.

NERSC is the world leader at managing large-scale, massively parallel systems. Starting with the T3E and running through the 10 TF IBM SP, NERSC has shown it is possible to accomplish both high utilization and very effective capability processing at the same. In addition to the base projects, NERSC supports most of the SciDAC project teams — many with very large calculations. In FY04, DOE introduced the INCITE program, which has 10% of the entire resources of NERSC allocated to three projects. Capability computing jobs — more than 512 CPUs — dominate the NERSC workload and are the focus of the queue priority systems. NERSC-3 regularly runs jobs that use 4,096 processors.

NERSC system managers have a great deal of experience in all aspects of managing systems. NERSC has manages of the systems on call 24 x 7 to insure the be, most effective. These include:

Hardware and software configuration management. Large systems require strong hardware and software configuration management to assure consistency and reduce variation. NERSC has developed system management processes (on many operating systems, including AIX, UNICOS/mk, and Linux) to install and configure software, distribute operating images to all nodes, and perform many other tasks. NERSC's methods have been exported to a number of other sites.

System upgrades have to be tested and planned. The implications of each change are assessed. Often, there is a complicated interdependency between upgrades that has to be analyzed.

A similar methodology applies to hardware. Configuration management has to be applied, and upgrades to firmware and other hard components are planned.

Job Scheduling. The workflow through a system is one of the most important things to the user community. Users want to know when they can expect their jobs to run, and that they have a fair chance to have their work proceed in the system.

Much system management is spent setting up batch systems and then monitoring the work flow on the system. Constant adjustments are made based on what types of jobs are submitted, where the hot spots are, special requests for priority, expanded limits and other reasons, and of course providing feedback to the users on job status.

NERSC staff are highly successful at getting the right jobs (for example, capability jobs and high-priority jobs) to run at the right time. They balance conflicting priorities and keep the system busy, but at the same time make sure the system continues to run efficiently.

NERSC has even developed metrics and benchmarks for this area. They provide feedback to the users through a sophisticated, real time Web interface that also provides a history of jobs and metrics. That allows users to decide how to submit jobs to optimize their allocation as well their productivity.

Parallel file system management. One of the most complex and important parts of parallel systems is the Global Parallel File System. File systems require constant monitoring and proactive repair before problems become obvious to the users. There are specialized file system nodes that manage the data, and software layers that allow direct, high performance access.

Usage patterns are constantly changing. Some projects are active, while other are not. Disk space is too limited to provide unlimited amounts of storage, so systems use quotas, assignments to different file

systems for load balancing, and other methods to (a) provide the space that each active project needs at the time they need it, and (b) cost-effectively use the resources

Permanent file systems have to be backed up. Today there are few efficient backup tools, so optimization is important.

Interconnect management. The other complex component of the parallel systems is the interconnect technology. Nodes have multiple paths into the "switch," and there are complex interactions necessary for load balancing. Also, there is a need to constantly monitor and proactively test the interconnects to detect slow points and other anomalies.

The interaction of the interconnect and the file system is also important. Configuration, tuning, and job management are necessary to prevent or mitigate interference.

Problem solving. There are always bugs and problems – 24 by 7. Much of the system manager's time is spent resolving problems detected and reported. Some problems are easy, and some are extremely difficult. Often there must be diagnostic analysis and data collection. Detailed interactions are required with vendor support and development staff. Multiple patches must be applied and tested.

System tuning. Large systems are in constant need of assessment and tuning. As work loads shift, so do the bottlenecks and limitations. Systems staff have to identify the limits and develop ways to improve.

There is proactive testing. For example, at NERSC we run a set of tests periodically to assure that the system doe not regress in performance. For example, using this method, we were able to detect that one system was gradually slowing down 5% per month after a reboot. After a long series of bug analyses and fixes, this was eventually solved. Interestingly, no other site reported these problems, because they were not doing proactive performance testing.

System security. Well-managed systems are the primary protection from intrusion. Proper configuration, IPsec security, and monitoring and scanning are parts of providing the level of security that is appropriate for the risk and recovery time. New software functions, such as Grid Middleware, require new approaches to provide flexibility as well as protection.

System programming. Different systems — especially open systems — require occasional systems programming development to provide improvements, bug fixes (if there is not a vendor fix), and other reasons. System programming no longer means just a kernel. There are now many levels (kernel, device drivers, communication layers, schedulers, etc.) of software interacting in complex ways to provide system level services.

Account management and accounting. NERSC has an automated system to maintain all account information, and automatically install and de-install accounts on the right systems. The system, called the NERSC Information Management system (NIM), also records all usage data and places it in central database. NIM, developed at NERSC, automates much of the account management process, so that once the code (a "finger") is ported to a new system, it is more an administrative task than a technical one to install and manage accounts. It is also provides much more consistency.

Nonetheless, each system manager also keeps a set of activity logs to show what happened on the system, and who used what resources. These are used for capacity planning, analysis of security events, and many other reasons.

F.4 Leveraging NERSC Operations

Operating and maintaining NFACS can easily be integrated with the NERSC Center's current monitoring and operations support for all computing, networking, and storage systems on a 24 × 7 × 365 schedule. Routine tasks involve system administration and monitoring, initial system troubleshooting, system backup, scheduling outages, remedial maintenance, and managing the near-line and off-line storage media. Three computing system engineers will be required to maintain the LCS over the five-year

span of this proposal. Vendor personnel will also be available, sometimes on site, to ensure that the LCS system operates well with high reliability.

NERSC has a long history of managing, tuning, improving, and developing new functionality for leadership class systems. For example, by aggressively using advanced scheduler functions, NERSC has been able to double the amount of computational capability delivered by some systems, compared to that delivered by the standard vendor software. NERSC systems managers know how to balance high utilization and fast turnaround for a diverse set of clients and disciplines, and will share that expertise with LCS system managers.

NFACS/NERSC staff will provide advanced training in the use of LCS systems and technologies. Most training will be presented live over the Access Grid. Some may also be broadcast live using the Real Networks streaming video format. Presentation slides will be available online before and after the presentations, and presentation materials from past classes will be available for browsing on the NFACS Web site. Online tutorials on a variety of topics will be developed and made available on the Web, along with documentation provided by vendors.

The NERSC help desk will provide direct assistance to NFACS clients, as well as managing and resolving client problem reports. The client community will be able to ask for assistance in the way most effective for them—not just what is most efficient for the help desk. Thus telephone, e-mail, and Web interactions will be supported with timely acknowledgement and response resolution. Once a client reports a problem, the help desk will manage it until it is resolved, not just send the client to another group or have the client manage the problem.

The NERSC Center's accounts and allocation support staff and tools will help NFACS clients manage their project resources. The Web-based NIM will manage all accounts and projects for NFACS systems, automatically accumulating usage data of clients and projects, summarizing it, and implementing resource restriction if a project or client exceeds its allocation. Weekly, monthly, and yearly account summaries will be distributed to the client community and DOE. NIM will also manage user credentials and enable users to access Globus/Grid services on NFACS resources.

F.5 Comprehensive Scientific Support

Many of the important breakthroughs in computational science are expected to come from large, multidisciplinary, multi-institutional collaborations working with advanced codes and large datasets, such as the SciDAC and INCITE collaborations. These teams are in the best position to take advantage of terascale computers and petascale storage, and NERSC provides its highest level of support to these researchers. This support includes specialized consulting support; special service coordination for queues, throughput, increased limits, etc.; specialized algorithmic support; special software support; visualization support; conference, and workshop support; Web sever and Web content support for some projects; and Concurrent Version System (CVS) servers and support for community code management.

Building on the NERSC Center's experience working with the special requirements of high-end users, a Leadership Computing Applications Team (LCAT) member will be assigned to each scientific discipline represented at NFACS to help define project requirements, obtain resources, tune and optimize codes, and coordinate services. Five LCAT members will support NFACS clients. Scientific computing specialists will work directly with clients to develop and improve algorithms, enhance performance, and contribute to software development. In these collaborations, the staff member will typically be a scientist experienced in the field of study who is also knowledgeable in the computing needs of the project.

Consulting staff will solve and manage client problem reports and requests for assistance, particularly with regard to programming and application development. Three consultants will be dedicated to working with NFACS clients. They will introduce new techniques, systems, and technologies; help analyze and debug problems with user codes, as well as with systems and applications software; report problems to

vendors; and track problems so they will be corrected in a timely manner. Consulting staff will provide software support for a complex set of tools, libraries, and environments, such as MPI, TotalView, and performance-analysis tools. Software packages, including visualization and client interface software, will be supported.

Berkeley Lab's integrated hardware/software environment for remote visualization support will be fully integrated into the NFACS infrastructure. High-end visualization and data analysis tools will be essential to turn raw simulation data into scientific discoveries. NFACS will work closely with its LCC partners to apply technologies developed across the coalition available to the user community. In particular, we will work closely with LLNL to share, test, debug, and deploy the latest ASCI tools for visualization of massive datasets, including utilization of COTS technologies to achieve new levels of graphics performance, the LLNL/Stanford-developed distributed parallel rendering software (Chromium), proven parallel, scalable end-user applications (like VISIT and Blockbuster movie player), and the Terascale Browser. All of these tools will be tightly coupled with the high-speed networks, coordinated Grid services, storage federation, and WAN GPFS capabilities deployed across the LCC sites. This represents a powerful set of tools and services that will enable LCS users across the nation to rapidly understand the enormous amount of data they generate at NFACS. Without tools of this caliber and computer scientists available to support these tools, the huge data generation engines that NFACS will be deploying would be virtually useless.

The Berkeley Lab/NERSC VisPortal project is exploring ways to deliver Grid-based advanced visualization and data analysis capabilities through a Web portal interface. The portal automates complex workflows like the distributed generation of MPEG movies or scheduling of file transfers, mediates access to limited hardware resources, and controls the launching of complex multicomponent distributed visualization applications like Visapult—an application used for remote and distributed, high performance interactive volume rendering of massive remotely located datasets. The image-based rendering methods employed by Visapult are able to hide much of the latency of the intervening network. Visapult's highly tuned network implementation has enabled it to win the annual SCinet Bandwidth Challenge competition three years in a row, and is well positioned to take full advantage of DOE Science Grid and NSF TeraGrid distributed networking and computing resources. Over time, the VisPortal interface will be used to closely integrate storage resource management (SRM) systems like Hierarchical Resource Manager (HRM) with visualization and data analysis environments. The Berkeley Lab visualization group will work with the Grid community on ways to integrate these services across all Grid facilities as well as ways to deploy Grid-based data analysis technologies developed at other sites on our own systems wherever possible.

APPENDIX G
NERSC Policy Board

Daniel Reed (Chair)
The University of North Carolina at Chapel Hill
147 Sitterson Hall, Campus Box 3175
Chapel Hill, NC 27599

Albert Narath (Retired)
1534 Eagle Ridge Drive, NE
Albuquerque, NM 87122

Robert J. Goldston
Director, Princeton Plasma Physics Laboratory
P.O. Box 451, Mail Stop 37
Princeton, NJ 08543-0451

Robert D. Ryne
Lawrence Berkeley National Laboratory
One Cyclotron Road, MS-71J
Berkeley, CA 94720

Stephen Jardin
Princeton Plasma Physics Laboratory
P.O. Box 451, Mail Stop 27
Princeton, NJ 08543-0451

Tetsuya Sato
Earth Simulator Center Director-General
Japan Marine Science & Technology Center
3173-25, Showa-machi, Kanazawa-ku
Yokohama-City, Japan 236001

Sid Karin
Professor of Computer Science and Engineering
University of California, San Diego
9500 Gilman Drive, MC 0505
La Jolla, CA 92093-0505

Stephen Squires
Vice President and Chief Science Officer
Hewlett-Packard Laboratories
1501 Page Mill Road, MS 3U-10
Palo Alto, CA 94304-1126

William J. Madia
Executive Vice President of Laboratory Operations
Battelle
505 King Avenue
Columbus, OH 43201

Michael Witherell, Director
Fermi National Accelerator Laboratory
P.O. Box 500
Mail Stop 105
Batavia, IL 60510

Paul C. Messina
Argonne National Laboratory
Building 221
9700 South Cass Avenue
Argonne, IL 60439

APPENDIX H
NFACS Staff and Biographical Sketches

As an element of the high-performance computing resources at Berkeley Lab, NFACS will be closely aligned with the NERSC Center at Berkeley Lab. The High Performance Computing Facilities (HPCF) Division coordination through the Director and the General Manager will minimize the need for additional staff and maximize technology leverage. However, within the HPCF Division, NERSC will remain a separate program focused on its mission of capability computing for computational science in the service of the DOE mission. NERSC will not be diminished by NFACS. Rather, we see NERSC enhancing the operation and success of NFACS. NFACS will also draw on talent across the Computing Sciences organization at Berkeley Lab by leveraging skill and talents from the Computational Research Division and the Information Technologies and Services Division.

H.1 NFACS Staffing

The NFACS staff will consist of 13 FTEs who are full-time career employees, and in most cases are 100% assigned to NFACS tasks. NFACS will take advantage of the experience of high-level professional staff at NERSC to assure consistent and high-quality service and support for NFACS scientists. Only full-time staff with a long-term commitment to the project and significant related experience can be expected to deliver the technical expertise required to manage highly complex and unique systems such as the LCS computing platforms.

Berkeley Lab values teamwork and the continued professional development of its staff. In order to address rapidly changing requirements, work groups are assembled that span multiple groups in order to bring the best skills to a task, but also to give staff an opportunity to get engaged in new projects. The NFACS Director, Horst D. Simon, will form an LCS Support Team that is an integrated team for the full life cycle (design, deployment, testing, and operation) of LCS resources. This team will consist of systems managers, performance analysts, and science area support staff, all working as one integrated unit. Thus, the proposed structure is flexible and dynamic — exactly what is needed for such a complex, far-reaching project.

Normally, staffing a facility like NFACS from scratch requires 45 to 60 staff. However, because of heavy leveraging of NERSC infrastructure and expertise, NFACS will use just 13 staff for direct support of LCS. The NFACS Director will be supported by a management team composed of a General Manager, the LCS Team Leader, the LCS Lead System Analyst, the LCS Lead Performance Analyst, and the LCS Lead Scientific Support Analyst. The Leads report to the General Manager and will work with other LCS staff to carry out their responsibilities.

The General Manager, William Kramer, reporting to the NFACS Director, is accountable for the NFACS facility, with management responsibility for planning, budgets, enhancements, personnel, vendor and user relations, physical resources, and program and operational integration.

The LCS Team Leader, William Saphir, reporting to the General Manager, is responsible for the development, management, and operations of computing, storage, and networking resources as well as the support needed by the user community.

The LCS Lead System Analyst, Nicholas Cardo, working with the LCS Team Leader, is responsible for the deployment, management, and operations of computing resources.

The LCS Lead Performance Analyst, David Skinner, working with the LCS Team Leader, ensures that NFACS application and systems performance improvements are determined and implemented.

The LCS Lead Scientific Support Analyst, John Shalf, working with the LCS Team Leader, ensures that NFACS scientific support meets the needs of the user community. He coordinates the Leadership Computing Application Team points of contact.

The NFACS staff will also include:

- Two additional Large Scale System Analysts who focus on managing and improving NFACS's computing infrastructure, including systems, operating systems, and software utilities.
- Two additional Performance Analysts for problem management; user code optimization, and debugging; documentation; online, remote, and classroom training; and third-party applications and library support. User Services also maintains collaborations with groups of researchers at other computing centers and manages user allocations.
- Four additional Scientific Support Analysts. Each Science Area Analyst is responsible for one of the Leadership Computing Applications Team (LCAT) areas. While these five staff will be part of the LCS Team, they may be co-located with the staff and activities of the five leadership computing applications areas, possibly at sites other than Berkeley Lab. Science Area Analysts will be computational science experts and will connect NFACS to the application area teams, providing focused support. They will be the point of contact (POC) to expedite any problems or concerns. Each application area will be provided with a single POC. The POC will handle requests by the computational scientists from the application area, and will facilitate special requests. They will also provide focused algorithm help, and work as partners with the applications area scientists to tune performance of application codes for the LCS systems.

H.2 Current Support of Key Personnel

The current funding of NFACS key personnel is listed below. Saphir, Shalf, Skinner, and Cardo will move to 100% NFACS funding in the first year. Their positions at NERSC will be back-filled with new hires. Kramer will move to 50% NFACS funding, relinquishing his SciDAC and Organizational Burden roles.

Simon, Horst

Project: LBNL Directors Organizational Burden
Percent Support: 100%
Duration: October 2003–September 2004
Annual Funding: N/A

Kramer, William

Project: National Energy Research Scientific Computing Center (NERSC)
Percent Support: 50%
Duration: October 2003–September 2004
Annual Funding: $28.2M base (operating and equipment)

Project: SciDAC: DOE Science Grid: Enabling and Deploying the SciDAC Collaboratory Software Environment
Percent Support: 20% (Per SciDAC renewal budget and YTD effort actuals)
Duration: October 2003–September 2004
Annual Funding: $225K (FY04 $150K rec'd to date with additional $75K expected per SciDAC renewal request)

Kramer, William (cont.)
Project: LBNL Computing Sciences Organizational Burden
Percent Support: 30%
Duration: October 2003–September 2004
Annual Funding: N/A

Saphir, William
Project: National Energy Research Scientific Computing Center (NERSC)
Percent Support: 100%
Duration: October 2003–September 2004
Annual Funding: $28.2M base (operating and equipment)

Shalf, John
Project: National Energy Research Scientific Computing Center (NERSC)
Percent Support: 50%
Duration: October 2003–September 2004
Annual Funding: $28.2M base (operating and equipment)

Project: Adaptive Mesh Refinement Visualization
Percent Support: 45%
Duration: October 2003–September 2004
Annual Funding: $460K

Project: Laboratory Directed Research and Development (LDRD) Project: Architectural Alternatives
Percent Support: 5%
Duration: October 2003–September 2004
Annual Funding: $175K

Cardo, Nicholas
Project: National Energy Research Scientific Computing Center (NERSC)
Percent Support: 100%
Duration: October 2003–September 2004
Annual Funding: $28.2M base (operating and equipment)

Skinner, David
Project: National Energy Research Scientific Computing Center (NERSC)
Percent Support: 100%
Duration: October 2003–September 2004
Annual Funding: $28.2M base (operating and equipment)

H.3 Biographical Sketches

Biographical sketches of key staff are included on the following pages.

Horst D. Simon

NFACS Director
Associate Laboratory Director for Computing Sciences
Lawrence Berkeley National Laboratory

Tel: (510) 486-7377
Fax: (510) 486-4300
Email: simon@nersc.gov
http://www.nersc.gov/~simon/

Education

- Ph.D., Mathematics, University of California, Berkeley, 1982.
- Diplom in Mathematik, Technische Universität Berlin, 1978.
-

Position History

Lawrence Berkeley National Laboratory, (1996–present)

- Associate Laboratory Director for Computing Sciences, 2004– present.
- Division Director, National Energy Research Scientific Computing (NERSC) Center, 1996– present.
- Division Director, Computational Research Division (CRD), 2002–present.

Silicon Graphics, Inc. (1994–1996)

- Manager of Research Marketing Development, 1994–1996.

Computer Sciences Corporation (1989–1994)

- Manager, Research Department, Contract to Numerical Aerodynamic Simulation (NAS) at NASA Ames Research Center, Moffett Field, 1989–1994.

Boeing Computer Services (1983–1989)

- Manager, Research Department, Moffett Field, 1987–1989.
- Boeing Technical Marketing Manager, 1987–1989.
- Manager, Computational Mathematics, Bellevue, Washington, 1986–1987.
- Project Manager, Boeing Research Program, 1986–1987.
- Project Manager, NSF Supercomputer Initiative, 1986–1987.
- Technical Staff Member, Computational Mathematics Group, 1983–1986.

State University of New York (SUNY), Stony Brook (1982–1983)

- Assistant Professor, Department of Applied Mathematics, 1982–1983.

Summary of Qualifications

Horst Simon was recently named Associate Laboratory Director for Computing Sciences at Berkeley Lab. In addition, he continues his responsibilities as Division Director for both the National Energy Research Scientific Computing (NERSC) Division and the Computational Research Division. In his new role as the Associate Laboratory Director for Computing Sciences, Horst represents the interests of the three computing divisions — NERSC, Computational Research, and Information Technologies and Services — in the formulation of Laboratory policy, and in communicating Laboratory actions on policy and procedures as appropriate. He also coordinates constructive interactions within the Computing Sciences divisions to seek coupling with other scientific programs.

Horst joined Berkeley Lab in early 1996, as director of the newly formed NERSC Division, and was one of the key architects in establishing NERSC at its new location in Berkeley. The NERSC Center is DOE's flagship supercomputing facility for unclassified research, funded by DOE's Office of Science, and is used by 2,276 users at 312 institutions. Under Horst's leadership, NERSC has enabled important discoveries in fields ranging from global climate modeling to astrophysics. Horst is also the founding director of Berkeley Lab's Computational Research Division, which conducts applied research and development in computer science, computational science, and applied mathematics. His research interests are in the development of sparse matrix algorithms, algorithms for large-scale eigenvalue problems, and domain decomposition algorithms for unstructured domains for parallel processing. His recursive spectral bisection algorithm is regarded as a breakthrough in parallel algorithms for unstructured computations, and his algorithm research efforts were honored with the 1988 Gordon Bell Prize for parallel processing research. Horst was a member of the NASA team that developed the NAS Parallel Benchmarks, a widely

used standard for evaluating the performance of massively parallel systems. He is also one of four editors of the twice-yearly TOP500 list of the world's most powerful computing systems.

Awards and Honors

- H. Julian Allen Award (jointly with the NAS Parallel Benchmarks Team) for notable scientific papers written by authors at NASA Ames Research Center, for the NAS Parallel Benchmarks (1995).
- Gordon Bell Prize (jointly with group from Cray and Boeing) in recognition of superior effort in parallel processing research (1988).
- University Award, SUNY Stony Brook (1983).

Selected Publications

Books (editor)

- *Scientific Applications of the Connection Machine*, Conference Proceedings, World Scientific Publishing Company, Teaneck, New Jersey, July 1989; second edition 1992.
- Parallel Computational Fluid Dynamics, MIT Press, Cambridge, Mass., 1992.
- *Parallel Processing for Scientific Computing* (with D. Bailey, P. Bjorstad, J. Gilbert, M. Mascagni, R. Schreiber, V. Torczon, and L. Watson), Proceedings of the 7th SIAM Conference, SIAM, Philadelphia, 1995.
- *Solving Irregularly Structured Problems in Parallel* (with A. Fereira, J. Rolim, and Shang-hua Teng), Proceedings Irregular 98, Springer Lecture Notes in Computer Science No. 1457, August 1998.

Papers

- Dr. Horst Simon has a total of more than 166 refereed papers in journals, book chapters, and proceedings.

Professional Activities and Organizations

- Chair, SIAM Activity Group on Supercomputing (1994–1996).
- Member, Society for Industrial and Applied Mathematics, SIAM Activity Groups on Linear Algebra and on Supercomputing, IEEE.
- Member, IEEE Computer Society.
- Member, Association of Computing Machinery (ACM).
- Member, IEEE Gordon Bell Prize Committee (1990–1994).
- Associate, Foresight Institute.
- Institutional Representative to Coalition for Academic Scientific Computing (CASC).
- Member of Editorial Board of five scientific journals (IJHPCA, Scientific Programming, IJCSE, Adv. In Eng. Software, Journal of the Earth Simulator)

Boards

- Wissenschaftlicher Beirat, Konrad Zuse Zentrum, Berlin (ZIB), Germany.
- Member, Scientific Advisory Board, CSCS (Swiss National Supercomputer Center), Manno, Switzerland.
- Member, International Advisory Panel for the Institute of HPC (iHPC), Singapore.
- Member, Industrial Advisory Board, Department. of Computer Science, UC Davis.
- Member Advisory Board, iPARK, San Jose, CA (1999-2000).
- Member Board of Directors, Pumpkin Networks, Sunnyvale, CA (1999-2001).

Partial list of Collaborators

C. Ding, Hongyuan Zha, Z. Zheng, J. Demmel, A. Sohn, R. Biswas, Shang-Hua Teng, S. Barnard, A. Pothen, Erich Strohmaier, C. Farhat, S. Lateri, S. Barnard, A. Karp, D. Heller, J. Lewis, R. Grimes.

Teaching

CS267 is a one-semester graduate class in *Applications of Parallel Computers*, University of California, Berkeley. Spring 1997, (jointly with Professor David Culler). Fall 2002, (jointly with Professor James Demmel).

William T.C. Kramer

General Manager
NFACS
Lawrence Berkeley National Laboratory

Tel: (510) 486-7577
Fax: (510) 486-4300
Email: kramer@nersc.gov

Education

- M.E., Electrical Engineering, University of Delaware, 1986
- M.S. and B.S., Computer Science, Purdue University, 1976 and 1975
- PhD Candidate, Computer Science, University of California, Berkeley, 1999–present
- Position History
- PI for the DOE Science Grid – 2003 to present
- General Manager of the NERSC Facility, 2000- present
- Deputy Division Director and Head, High Performance Computing Department, NERSC Division, 1996–present
- Director, Advanced Air Transportation Technologies Program, NASA AATT Program Office, 1995–1996
- High Speed Processor (Cray) Manager; Branch Chief, Computational Services, Numerical Aerodynamic Simulation (NAS) Systems Division, NASA Ames Research Center, 1986–1994

Summary of Qualifications

Mr. Kramer's strong, innovative management and technical skills and his proven ability to quickly establish excellent services and support large-scale, advanced systems has been key to NERSC's success at LBNL. As one of the first employees of the relocated NERSC, he led the reimplementation of NERSC at LBNL with an expanded mission and 20% reduction in funds, including reinstallation of all systems and hiring over 60 technical staff on a rapid schedule. Mr. Kramer led the NERSC-3 and NERSC-4 supercomputer procurement. He has played a key role in deploying the IBM SP, a fifth-generation, early delivery supercomputer, which is the world's largest unclassified supercomputer. Mr. Kramer also placed the first UNIX supercomputer, the Cray–2, into production in 1986 while at NASA. Innovations from the NERSC procurement process are now being used at several other centers.

Mr. Kramer is responsible for all aspects of NERSC, and shares division-wide management responsibilities. He introduced project planning and metrics for the division, leading the development and implementation of a new 28,000 sf computing facility from advocacy and requirements to implementation. He has twice directed the moves of the entire NERSC center. He led the installation, testing, and introduction of the early, very large IBM SP, the first large T3E, early J-90s, and HPSS. He was first to use checkpoint/restart in a MPP production environment, and first to demonstrate the ability to manage very large MPPs with utilization over 90-95%, among other innovations. He was instrumental in managing the paradigm shift for NERSC from traditional vector computing to massively parallel.

Mr. Kramer has taught numerous classes, seminars, and tutorials on computing topics, ranging from computer graphics and visualization to high performance computing. He has taught at two universities, has worked in private industry, and has been active in computer security, at times assisting in federal investigations. Mr. Kramer's research interests include large-scale system management, scheduling, performance evaluation, and the integration of high performance networks and computers.

Professional Activities

SC 2005 General Chair, GGF-58 Organizing Committee SC 2002 Xnet Chair, GGF-5 Organizing Committee, SC 2001 High Performance Bandwidth Chair, SC 2000 Vice Chair for Information Architecture and Chair of SCinet 2000; Served as SC 2001 Network Bandwidth Challenge Leader and part of the Network Measurement team; Served as the SC 'XY consultant on future networking for site selection for future conferences; SC 98 Exhibits Chair; SC 96 HPC Challenge Chair; Invited member of the Arctic Region Supercomputer Center Advanced Technology Panel; Invited member of the Cray Customer Advisory Board; Member of the IBM *e-server* Customer Advisory Board

Awards and Honors

Four (4) LBNL Outstanding Performance Awards, NASA Group Achievement Award, DECUS Board of Director's Medal, DECUS Outstanding Contributor Award, DECUS Outstanding Unit Award, Certified Private Pilot, Open Water Scuba Instructor, Emergency Medical Technician, Certified Fire Fighter, Eagle Scout, USMA (West Point) appointment

Selected Publications

- Kramer, William T. C., Deborah A. Agarwal, Arie Shoshani, Brent R. Draney, Guojun Jin, Gregory F. Butler, and John A. Hules, "Deep Scientific Computing Requires Deep Data," accepted for publication in The IBM Journal on Research and Development.
- Paxson, Vern, Steven Lau, James Rothfuss and William Kramer; Trends in Computer Security for Open Scientific Facilities, Tutorial at SC 2003, November 2003 in Phoenix, AZ, and SC 2002, November 16-22 in Baltimore, MD
- William Kramer, "Building Network Testbeds for High Performance Applications," IEEE Computer Magazine, June 2002 (cover article).
- Kramer, William and Steven Lau; Trends in Computer Security for Open Scientific Facilities, Tutorial at the Global Grid Forum – 8and the High Performance Distributed Computing Conference, June 22-27 in Seattle, WA
- Kramer, William and Clint Ryan, "Performance Variability on Highly Parallel Architectures," accepted for publication and presentation – the International Conference on Computational Science 2003, Melbourne Australia and St. Petersburg Russia, June 2-4, 2003.
- Kramer, William, "NERSC and Blue Planet," invited presentation at the 7th SciCOMP Conference, Gottingen, Germany, March 7, 2003.
- Bair, Raymond, et. al., "High-Performance Networks for High-Impact Science - Report of the High-Performance Network Planning Workshop Conducted by the Office of Science, U.S. Department of Energy," August 13-15, 2002, http://doecollaboratory.pnl.gov/meetings/hpnpw/finalreport/high-performance_networks.pdf.
- Paxson, Vern, Steven Lau, James Rothfuss and William Kramer; Trends in Computer Security for Open Scientific Facilities, a full day Tutorial at SC 2002, November 16-22 in Baltimore, MD.
- Kramer, William, "Accelerating Scientific Discovery Through Advanced Computation," and invited keynote presentation at 10th ECMWF (European Center for Mid-range Metrological Forecasting) Workshop on Terascale Computing, - Reading England, November 2002.
- McCurdy, C. William, Rick Stevens, Horst Simon, William Kramer, et al, Creating a Science Driven Computer Architecture: A Path to Scientific Leadership, LBNL Technical Publication, Number 5483, October 2002
- Wong, Adrian T., Leonid Oliker, William T. C. Kramer, Teresa L. Kaltz, and David H. Bailey, "Evaluating System Effectiveness in High Performance Computing Systems," Proceedings of SC2000, November 2000.
- Wong, Adrian T., Leonid Oliker, William T. C. Kramer, Teresa L. Kaltz, and David H. Bailey, "System Utilization Benchmark on the Cray T3E and IBM SP," presented at the 5th Workshop on Job Scheduling Strategies for Parallel Processing, May 2000, Cancun Mexico.
- Bailey, David, et al. "Valuation of Ultra-Scale Computing Systems: A White Paper," published as an official report of the Department of Energy, December 22, 1999.
- Kramer William, Francesca Verdier, Keith Fitzgerald, James Craw, and Tammy Welcome, "High Performance Computing Facilities for the Next Millennium," presented at SC99, Portland, OR, and published as part of the Tutorials Program, November 1999.
- Simon, Horst D., William T. C. Kramer, and Robert F. Lucas, "Building the Teraflops/Petabytes Production Supercomputing Center," EuroPar '99 in Toulouse, France, September 1999.
- "How Are We Doing? A Self Assessment of the Quality of Services and Systems at NERSC (Oct. 1, 1997- Dec 31, 1998," William T.C. Kramer, LBNL Technical Report LBNL-43131, May 1999.
- Shoshani, Arie, Craig Tull, Brian Tierney, Harvard Holmes, Robert Lucas, and William T.C. Kramer, "Large Scale, Data Intensive Computing," tutorial at the SC '98 Conference, November 8, 1998, Orlando, FL.
- Kramer, William T.C., "So, now that you have your Teraflops computer, what do you do," invited presentation at the LBNL Science and Technology Seminars Series, September 11, 1998 at the LBNL Washington D.C. office.

William C. Saphir

LCS Team Leader
NFACS
Lawrence Berkeley National Laboratory

Tel: (510) 486-4373
Fax: (510) 486-4004
Email: WCSaphir@lbl.gov

Education

- Ph.D., Physics, University of Texas at Austin, 1992.
- B.S., Physics, Massachusetts Institute of Technology, 1986.

Position History

- Chief Architect, NERSC Division, Lawrence Berkeley National Laboratory, 2003–present.
- Director of Engineering, Scale Eight, 2000–2002.
- Senior Director of Operations and Customer Service, Scale Eight, 2002–2003.
- Staff Scientist and Group Leader, Future Technologies Group, NERSC, 1996–2000.
- Parallel Systems Consultant and Research Scientist, Numerical Aerodynamic Simulation (NAS) Division, NASA Ames Research Center, 1992–1996.
- Systems Programmer and Consultant, Project Athena, MIT, 1984–1985.

Summary of Qualifications

Bill Saphir is an internationally recognized expert on parallel computing technology, and is well known for his work on high performance communication and performance analysis for parallel computers. As Chief Architect for NERSC, he is responsible for setting the technology direction for the NERSC facility. He made significant contributions to the MPI-2 standard and is the original developer of MVICH, the implementation of MPI over Infiniband that runs on Virginia Tech "Big Mac" supercomputer. As group leader for the NERSC Future Technologies group he spearheaded Berkeley Lab's entrance into cluster computing. At NASA, he was a member of the team that developed the NAS Parallel Benchmarks, one of the most effective tools for measuring the performance of parallel computers. He has presented tutorials on cluster computing, MPI, and other topics in parallel computing.

Dr. Saphir was responsible for software development for Scale Eight's Global Storage Service, which stored and served dozens of terabytes of data for Microsoft, MTV, and other customers from using several Linux clusters each composed of hundreds of servers. He later assumed responsibility for the 24x7 operation of this service, and for all user support. Under his direction, service availability exceeded 99.97%.

Selected Publications

Book

1. MPI: The Complete Reference, Volume 2 (with W. Gropp, et al.) MIT Press, 1998.

Articles

Parallel Computation

1. The Effect of Message Buffering on the Communication Performance of Parallel Computers NAS TR RNS-94-004, April, 1994.
2. NAS Experiences With a Prototype Cluster of Workstations/ Supercomputing '94 Proceedings, November, 1994 (With William Kramer, et al.).
3. Job management Requirements for NAS Parallel Systems and Clusters. Workshop on Job Scheduling for Parallel Processing, International Parallel Processing Symposium, 1995.
4. JSD: Parallel Job Accounting on the NAS SP2. NAS TR 95-16, July, 1995.
5. The NAS Parallel Benchmarks 2.0. NAS TR 95-020, December, 1995 (with David Bailey, et al.).
6. MPI-2: Extending the Message Passing Interface. Proceedings of Euro-Par 96 Parallel Processing in Lecture Notes in Computer Science 1123, 1996 (with Al Geist, et al.).

7. The NAS Parallel Benchmarks 2 Results. NAS TR 96-010, August, 1995 (with Alex Woo, Maurice Yarrow) and NASA HPCCP Computational Aerosciences Conference, August 1995 and Supercomputing 96 (poster), November 1996.
8. The NAS Parallel Benchmarks 2.1 RESULTS. NAS Technical Report 96-010, August, 1995 (with Alex Woo, Maurice Yarrow) and NASA HPCCP Computational Aerosciences Conference, August 1995 and Supercomputing 96 (poster), November 1996.
9. New implementations and results for the NAS parallel Benchmarks 2 (with R. F. Van der Wijngaart, A. C. Woo, M. Yarrow). 8th SIAM Conference on Parallel Processing for Scientific Computing, Minneapolis, MN, March 14-17 1997.
10. Implementing the MPICH ADI on an Unreliable Transport Layer. (with B. Pfrommer and P. Bozeman). LBNL Report LBNL-40343. May 1997.
11. A Survey of MPI Implementations. NHSE Review 2(1) 1997 and LBNL Report LBNL-41025/UC-405 November 1997.
12. The Design and Evolution of the MPI-2 C++ Interface. (w/J. Squyres and A. Lumsdaine). Proceedings of the International Scientific Computing in Object-Oriented Parallel Environments Conference 1997.
13. On the Efficacy of Code Optimizations for Cache-based Microprocessors (with R. Van der Wijngaart). LBNL-40345. May 1997.
14. Performance and Scalability of the NAS Parallel Benchmarks on the Cray T3E (w/S. Caruso). LBNL-403444. May 1997.
15. M-VIA: A Modular High-Performance Implementation of the Virtual Interface Architecture for Linux (with P. Bozeman).
16. On the Development of Open Source System Software for High Performance Computing. LBNL-45479. March 2000.

Physics

1. The Kinetic Theory of the Standard Map. Proceedings of "Aspects of Nonlinear Dynamics: Solitons and Chaos," Brussels, December, 1990, edited by I. Antoniou and F. Lambert, (Springer, Berlin, 1991) (with Hiroshi Hasegawa).
2. Decaying Eigenstates for Simple Chaotic Systems. Physics Letters A **162** (1992) 471 (with Hiroshi Hasegawa.)
3. The Non-Equilibrium Statistical Mechanics of the Baker Map. Physics Letters A **162** (1992) 477 (with Hiroshi Hasegawa).
4. The Nonequilibrium Statistical Mechanics of Chaotic Maps. Dissertation, The University of Texas at Austin, May, 1992.
5. Unitarity and Irreversibility in Chaotic Systems. Physical Review A **46** (1992) 7401 (with Hiroshi Hasegawa).

Seminars and Tutorials, Panels (selected)

1. Collective Communication in PVM 3. Workshop on Distributed Computing in Aeroscience Applications, October 19, 1993.
2. A Comparison of Communication Libraries: NX, CMMD, MPI, PVM. NAS, November 30, 1993.
3. Devil's Advocate: Reasons *Not* to Use PVM. PVM User Group Meeting, May 20, 1994.
4. Sorting Out Communication Libraries, a Comparison of NX, CMMD, MPI and PVM. Tutorial at Supercomputing '94, November 14, 1994.
5. An Intensive and Practical Introduction to MPI. Supercomputing '96, '97, '98. Full and half day tutorial.
6. MPI-2. Supercomputing '97, '98.
7. M-VIA: Virtual Interface Architecture for Linux. Linuxexpo, June 1999.
8. Linux for Scientific Computing. O'Reilly Linux Conference, August 1999
9. M-VIA and MPI for VIA. Joint PC Cluster Computing Conference (JPC4-5). September 1999.
10. Production Linux Clusters. SC '99 (full day tutorial). With P. Bozeman, R. Evard and P. Beckman.LBNL-43738 (and SC'03 with R. Evard, P. Beckman, and S. Coghlan).

Nicholas P. Cardo

LCS Lead System Analyst
NFACS
Lawrence Berkeley National Laboratory

Tel: (510) 486-4765
Fax: (510) 486-4316
Email: cardo@nersc.gov

Education

- M.S., Computer Engineering, San Jose State University, 1994.
- B.S., Computer Science, New Jersey Institute of Technology, 1987.

Position History

- Project Lead, Lawrence Berkeley National Laboratory, 1999–present.
- Senior Systems Programmer/Analyst, Sterling Software, Inc., NASA Ames Research Center, Moffett Field, California, 1989–1998.
- Systems Analyst, Sterling Software ZeroOne Systems, John von Neumann National Supercomputer Center, Princeton, New Jersey, 1987–1989.

Accolades

- Certified AIX Support Specialist, 2000
- SP-XXL President, 2002–2004
- SP-XXL President, 2004–2006

Summary of Qualifications

I have developed utilities for the management of Cray Data Migration Facility (DMF) for UNICOS. These include a handle to filename search utility, inode recreation capability for soft deleted files, the ability to move an inode between file systems without recalling the data back to disk, and the ability to search for files by DMF handle. I designed and developed several kernel modifications including a disk high water mark feature. I designed and developed a fast inode scan algorithm. This evolved into utilities for locating files quickly on a file system as well as a fast way to estimate the number of tapes required for performing system backups. I designed and developed one of the first full production schedulers for the Portable Batch System (PBS). This scheduler was designed to work on Cray parallel vector processor systems and included the capability of performing resource management. I also participated as one of the early testers of PBS for production systems.

I am currently the project lead for the IBM RS/6000 SP at the National Energy Research Scientific Computing Center (NERSC). My responsibilities include configuration management, software management to maintain a highly reliable and stable system, hardware problem determination, and root cause analysis for system problems. I am also involved in software projects to improve the usability of the system as well as the administration of the system.

I have established working relationships with IBM developers and support specialists. Through these relationships I provide valuable information used in their development efforts. The end result is a feature-rich product that meets the customers' needs.

I have been President of the SP-XXL organization since 2002 and re-elected in 2004 for another two year term. This organization represents the largest Scientific and Technical Computing institutions world-wide that utilize IBM hardware/software.

I bring with me over thirteen years of experience in System Management, System Programming, and System Administration. Utilizing my System Administration and System Programming knowledge allows me to acquire a very detailed understanding of system internals, and with my Computer Engineering education, I am able to gain a very detailed understanding of hardware and software interactions.

David Skinner

LCS Lead Performance Analyst
NFACS
Lawrence Berkeley National Laboratory

Tel: (510) 486-4746
Fax: (510) 486-7202
Email: deskinner@lbl.gov

Education

- Ph.D. Theoretical Chemistry, Hertz Fellow, University of California Berkeley, May 2000.
- B.S. Chemistry, University of Kansas, May 1995.

Position History

- NERSC User Services Scientific Consultant, Lawrence Berkeley National Laboratory, August 2000–present.

Project Lead for IBM SP user support. Responsible for tracking and resolving user problems on the SP. Provides in depth consulting for strategic user projects. Expertise in MPP support in the areas of programming models, message passing, I/O, and performance. Expertise in chemistry software support. Installs and maintains scientific and UNIX software, including several molecular dynamics codes. Oversaw the transition of users and scientific applications from the phase I SP cluster (gseaborg) to the new phase II cluster (seaborg).

Provides technical support for architectural and performance studies. Key member of the IBM Blue Planet project. Writes and ran benchmarks and kemels to test Blue Planet ideas. Brings a user perspective to the project: programmability and usability considerations.

Developed and gave advanced user training talks: "Effectively Addressing Memory with Loop Constructs"; "MPI Scaling on the IBM SP: Techniques and Pitfalls"; "Effective Memory Use on the SP"; "Scaling I/O and MPI".

Programming projects include expanding the capabilities of IBM's parallel performance data collection tool and porting and writing the memory profiling code Xtream which provides the scientist/programmer with clear-cut answers to questions about how programming language, loop constructs, and data locality impact application performance on a given architecture. Working with code developers at IBM's Watson Research Center, replaced the previous performance data collection method (hundreds or thousands of Unix processes computing their HPM performance data records separately) with a single HPM file which aggregates the performance data using file locking. The new method is six times faster than the old method for one sample program running on 1,000 processors. In addition to accumulating the total HPM statistics across tasks into a concise report the new version reports the minimum, maximum, and average counts, which is useful in detecting load imbalance, and includes a new –mpi option which allows NERSC and its users to collect for the first time MPI performance statistics.

Received Outstanding Performance Awards for writing a from-scratch implementation of multiple compiler support for IBM compilers that predate IBM's official support for "NDI" compilers; for writing benchmarks and performing acceptance and full configuration testing for the NERSC-4 procurement; and for transitioning users from the phase 1 to the phase 2 SP for the NERSC-3 procurement.

- Graduate Research Assistant, 1995–2000.

Publications

- "Evaluation of Cache-based Superscalar and Cacheless Vector Architectures for Scientific Computations", Supercomputing 2003. L. Oliker, J. Carter, J. Shalf, D. Skinner, S. Ethier, R. Biswas, J. Djomehri, R. Van der Wijngaart.
- "Scaling Up Parallel Scientific Applications on the IBM SP," D. E. Skinner, LBNL Technical Report LBNL-54254, D. Skinner. http://hpcf.nersc.gov/computers/SP/scaling/.
- "A Performance Evaluation of the Cray X1 for Scientific Applications," presentation at VECPAR'2004.
- "Quantum and semiclassical approaches to chemical reaction dynamics," D. E. Skinner, LBNL Technical Report LBNL-47146, May 2000.
- "Application of the forward-backward initial value representation to molecular energy transfer," D. E. Skinner and W. H. Miller , LBNL Technical Report LBNL-44187 and Journal of Chemical Physics, August 1999.

- "Application of the semiclassical initial value representation and its linearized approximation to inelastic scattering," D. E. Skinner and W. H. Miller, LBNL Technical Report LBNL-42302 and Chemical Physics Letters Vol. 300 1-2, Jan 29. 1999.
- "Quantum mechanical rate constants O+OH {reversible reaction} H + O{sub2} f or Total Angular Momentum J > 0," D. E. Skinner, T. C. Germann, W. H. Miller, LBNL Technical Report LBNL-41297 and Journal of Physical Chemistry A Vol. 102 21, May 21 1998.

Summary of Experience

- Science
  - Energy Transfer via Semiclassical Dynamics
    - Calculation of Collisional Energy Transfer in Molecular Systems
    - Approximate Quantum Mechanics via many (1.0e6) trajectories
    - Distributed Trajectory Code in C++ with interfaces to LAPACK
  - Thermal Rate Constants via Exact Quantum Dynamics
    - Time Dependent Quantum Mechanics of Long Lived Complexes
    - Quantum Dynamics on Grids via Parallel FFT/Matrix Methods
    - Large Scale T3D Computation, HPF and MPI
  - Semiconducting Nanoclusters
    - Development and Validation of a Kinetic Theory of Charge
    - Transfer Integration of ODE's, master equations (matrix ODE's
- Programmer of Parallel and Distributed Programs for Chemical Physics
  - Experience in Coding C, C++, Fortran, and Fortran90
  - Parallel Quantum Dynamics on Grids (T3D, RS6000, Origin 2000, IBM SP)
  - Distributed Trajectories in C++ Using MPI and Pthreads (Linux)
  - Author of IVRpack, a C++ trajectory library
  - Perl scripting
- Experienced in Maintaining a Diverse Network of Machines
  - Windows PC, Sparc Solaris, SMP Linux, Dec Alpha, Origin 2000, AIX RS6000 and Mac Sysadmin Experience
  - Data Acquisition for Femtosecond Spectroscopy using PC A/D boards
  - NFS, NIS, Packet Filtering, Netatalk, ssh and rdist
  - Deployed ipchains Firewall, QOS and SSL Webmail
  - Linux kernel configuration and builds

John Shalf

Lead Scientific Support Analyst
NCFAS
Lawrence Berkeley National Laboratory

Phone: (510) 486-4508
Fax: (510) 486-5812
Email: jshalf@lbl.gov

Research

Performance Evaluation for High Performance Computing Applications, Grid Computing, High Performance Networking, Visualization Systems, Computer Architecture.

Education

Virginia Polytechnic Institute and State University, Blacksburg Virginia.

- Completed all coursework for Masters Degree in Electrical Engineering 1992-1995.
- Graduated in May of 1991 with BS in Electrical Engineering.

Honors/Affiliations

- Visiting Scientist Max-Planck-Institut fuer Gravitationphysick / Albert Einstein Institute in Potsdam Germany 1997.
- Senior Staff for $2.2M/year 3 year NSF-KDI Grant for "An Astrophysics Simulation Collaboratory" 1999-2002.
- Technical Advisory Board Member for $6M euro/yr EU GridLab Project (http://www.gridlab.org).
- R&D100 Award for LBNL/NERSC RAGE Robot 2002.
- Bandwidth Challenge Team Leader for LBNL team 2001, 2002.
- Supercomputing 2002: Taught tutorial class on parallel and distributed computing.
- Invited participant for in 2001 NSF/ANIR Grand Challenges in E-Science workshops: credited as major contributor to the final document that led to 2002 NSF Experimental Networks CFP. (http://www.evl.uic.edu/activity/NSF/index.html).
- Participated in Award-Winning Supercomputing HPC Challenge Efforts 1995,1997,1998,2001,2002.

Position History

Staff Scientist, *December 1999–present*
Lawrence Berkeley National Laboratory
Berkeley CA 94720

Research Programmer, *January 1995–December 2000*
HPC Consulting Group / General Relativity Group / Visualization and Virtual Environments / StarTAP.
National Center for Supercomputing Applications, Urbana IL.

- HPC consultant: analyzed, optimized, and developed HPC codes for NSF supercomputing center applications
- Developed Chesapeake Bay VR/Visualization tool for CEWES/Army Corp of Engineers. Currently included as an example program for all versions of CaveLIB shipped by VRCO.
- StarTap International Networking Access Point applications team.
- General Relativity Group/Laboratory for Computational Astrophysics: Developed portions of Cactus code framework and AMR visualization and data management infrastructure.
- Project development leader for Simulated Cluster Archive portal (http://sca.ncsa.uiuc.edu/) and LCAVision AMR visualization tool (http://zeus.ncsa.uiuc.edu/~miksa/LCAVision.html).
- NCSA project lead,Portal and Visualization Tool development for NSF-KDI Astrophysics Simulation Collaboratory project (http://www.ascportal.org).

Selected Publications

- L. Oliker, A. Canning, J. Carter, J. Shalf, D. Skinner, S. Ethier, R. Biswas, J. Djomehri, R. Van der Wijngaart, "Evaluation of Cache-based Superscalar and Cacheless Vector Architectures for Scientific Computations," *Proceedings of Supercomputing 2003*, November 15-21, Phoenix Arizona. (*LBNL-53117*).
- J. Shalf and W. Bethel, "How the Grid Will Affect the Architecture of Future Visualization Systems," IEEE Computer Graphics and Applications, Visualization Viewpoints column, May/June 2003. *(LBNL-51723)*.
- J. Shalf and E. W. Bethel, "Cactus and Visapult: An Ultra-High Performance Grid-Distributed **V**isualization Architecture Using Connectionless Protocols," IEEE CG&A special issue on Grid Visualization, Mar/April 2003 *(LBNL-51564)*.
- Terry J. Ligocki, Brian Van Straalen, John M. Shalf, Gunther H. Weber, Bernd Hamann, "A Framework for Visualizing Hierarchical Computations" in *Hierarchical and Geometrical Methods in Scientific Visualization*, pp 19–40, 2003.
- Gabrielle Allen, Tom Goodale, Michael Russell, Edward Seidel and John Shalf, *Classifying and Enabling Grid Applications*, Chapter 23, "Grid Computing: Making Global Infrastructure a Reality," editors, Fran Berman, Geoffrey Fox, Tony Hey, published by Wiley & Sons, 2003 (http://www.grid2002.org).
- Gregor von Laszewski, Michael Russell, Ian Foster, John Shalf, Gabrie lle Allen, Greg Daues, Jason Novotny, Edward Seidel, "Community Software Development with the Astrophysics Simulation Collaboratory," Concurrency and Computation: Practice and Experience, 2002.
- Gabrielle Allen, Tom Goodale, Gerd Lanfermann, Thomas Radke, Edward Seidel, Werner Benger, Hans-Christian Hege, Andre Merzky, Joan Massó and John Shalf "Solving Einstein's Equations on Supercomputers," *IEEE Computer*, *32*, (1999) [cover story].
- Werner Benger, Ian Foster, Jason Novotny, Edward Seidel, John Shalf, Warren Smith and Paul Walker. "Numerical Relativity in a Distributed Framework" *Proceedings of the Ninth SIAM Conference on Parallel Processing for Scientific Computing*, 1999.
- Mike Norman, John Shalf, Stuart Levy, Greg Daues, "Diving Deep: Data Management and Visualization Strategies for Adaptive Mesh Refinement Simulations," IEEE Computing in Science and Engineering, Special issue on Massive Data Visualization, July-August, 1999, pp 22–32.
- M.L. Norman, P. Beckman, G. Bryan, J. Dubinski, D. Gannon, L. Hernquist, K. Keahey, J.P. Ostriker, J. Shalf, J. Welling, S. Yang, "Galaxies Collide on the I-WAY: An Example of Wide-Area Collaborative Supercomputing," *International Journal of Supercomputing Applications and High Performance Computing*, Vol. 10, No. 2/3 Summer/Fall, 1996.

APPENDIX I
Bibliography

1. C. William McCurdy et al., "Creating Science-Driven Computer Architecture: A New Path to Scientific Leadership," Lawrence Berkeley National Laboratory report LBNL/PUB-5483, October 2002; http://www.nersc.gov/news/blueplanet.html.

2. "Facts on ASCI Purple," Lawrence Livermore National Laboratory report UCRL-TB-150327 (2002); http://www.sandia.gov/supercomp/sc2002/flyers/SC02ASCIPurplev4.pdf.

3. Daniel A. Reed, ed., "Workshop on the Roadmap for the Revitalization of High-End Computing," June 16–18, 2003 (Washington, D.C.: Computing Research Association).

4. Phillip Colella, Thom H. Dunning, Jr., William D. Gropp, and David E. Keyes, eds., "A Science-Based Case for Large-Scale Simulation" (Washington, D.C.: DOE Office of Science, July 30, 2003).

5. "Red Storm System Raises Bar on Supercomputer Scalability" (Seattle: Cray Inc., 2003), http://www.cray.com/company/RedStorm_flyer.pdf.

A1 M. Brehm, et. al, "Pseudo vectorization, SMP, and Message Passing on the Hitachi SR8000-F1" Euro-Par 2000 - Parallel Processing: 6th International Euro-Par Conference, Munich, Germany, August/September 2000.

A2 H. Nishiyama, et. al, "Pseudo-vectorizing Compiler for the SR8000", Euro-Par 2000 - Parallel Processing: 6th International Euro-Par Conference, Munich, Germany, August/September 2000.

A3 R. Bader, et. al. "TeraFlops Computing with the Hitachi SR8000-F1", High Performance Computing in Science and Engineering. Transactions of the First Joint HLRB and KONWIHR Status and Result Workshop 2002.

C1 David H. Bailey et al, "The NAS Parallel Benchmarks", Intl. Journal of Supercomputer Applications, vol. 5, no. 3 (Fall 1991), pg. 66-73.

C2 P. A. Agarwal, et. al, "Cray X1 Evaluation Status Report", ORNL Technical Report ORNL/TM-2004/13, January, 2004.

C3 L. Oliker, et. al, "A Performance Evaluation of the Cray X1 for Scientific Applications," High Performance Computing for Computational Science VECPAR 2004, to appear.

D1 http://www.pnl.gov/scales/

D2 Kendall, R. A., E. Aprà, D. E. Bernholdt, E. J. Bylaska, M. Dupuis, G. I. Fann, R. J. Harrison, J. L. Ju, J. A. Nichols, J. Nieplocha, T. P. Straatsma, T. L. Windus and A. T. Wong, Comput. Phys. Commun., 128:260-283, 2000 [b] http://www.emsl.pnl.gov/docs/nwchem.

D3 R. Car and M. Parrinello, Phys. Rev. Lett., 55, 2471 (1985).

D4 W. M. C. Foulkes, L. Mitas, R. J. Needs, and G. Rajagopal, Rev. Mod. Phys., 73, 33 (2001).

D5 R. Q. Hood, M. Y. Chou, A. J. Williamson, G. Rajagopal, R. J. Needs, and W. M. C. Foulkes. Phys. Rev. B, 57, 8972 (1998).

D6 J. C. Grossman, L. Mitas, and K. Raghavachari. Phys. Rev. Lett., 75, 3870 (1995).

D7 J. C. Grossman and L. Mitas. Phys. Rev. B, 52, 16735 (1995).

D8 V. Natoli, R. M. Martin, and D. M. Ceperley. Phys. Rev. Lett., 70, 1952 (1993).

D9 L. Mitas and R. M. Martin. Phys. Rev. Lett., 72, 2438 (1994).

D10 N. W. Ashcroft, Physics World, 8, 43 (1995).

D11 K. Matsuishi, H. K. Mao E. Gregoryanz, and R. J. Hemley. J. Chem. Phys., 118, 10683 (2003).

D12 M. Sprik. J. Phys.: Condens. Matter, 12, A161 (2000).

D13 M. Reed, "Nanostructured Systems" in "Semiconductors and Semimetals," Vol. 35, Academic Press, Boston (1992).

D14 M. Grundmann, D. Bimberg and N. Ledentsov, "Quantum Dot Heterostructures," Wiley & Sons, New York (1998).

D15 V.I. Klimov et al., Science 290, 314 (2000).

D16 L. Pavesi et al., Nature 408, 440 (2000).

D17 W.V. Schoenfeld, T. Lundstrom, P.M. Petroff and D. Gershoni, Appl. Phys. Lett. 70, 2194-2196 (1999).

D18 D.L. Klein, R. Roth, A.K.L. Lim, A.P. Alivisatos and P.L. McEuen, Nature 389, 699-701 (1997).

D19 D. Loss, D.P. DiVincenzo, Phys. Rev. A 57, 120-126 (1998).

D20 K.W. Barnham and G. Duggan, J. Appl. Phys. 67, 3490 (1990).

D21 L.W. Wang, and A. Zunger, J. Chem. Phys. 100, 2394-2397 (1994). A. Canning, L.W. Wang, A. Williamson and A. Zunger, J. Comput. Phys. 160, 29 (2000).

D22 A. Franceschetti, H. Fu, L.W. Wang and A. Zunger, Phys. Rev. B. 60, 1819 (1999).

D23 H.-J. Choi, M. L. Cohen, and S. G. Louie, to be published.

D24 M. Rohlfing and S.G. Louie, Phys. Rev. Lett. 81, 2312 (1998).

D25 S. Ismail-Beigi and S.G. Louie, Phys. Rev. Lett. 90, 076401 (2003).

D26 Facilities for the Future of Science: A Twenty-Year Outlook, US DOE, Office of Science, Nov 2003.

APPENDIX J
Acronyms and Abbreviations

ADM Arnowitt-Deser-Misner (computational relativity)
ALS Advanced Light Source (LBNL experimental facility)
AMR Adaptive mesh refinement (numerical technique)
ANL Argonne National Laboratory (DOE laboratory)
APDEC Applied Partial Differential Equations Center (SciDAC project)
ASCI Advanced Simulation and Computing (DOE/NNSA program)
BeBOP Berkeley Benchmarking and Optimization (UC Berkeley project)
BGK Bhatnagar, Gross and Krook (fusion)
BGL Blue Gene/L (IBM-LLNL research system)
BLAS Basic linear algebra software
BSSN Baumgarte-Shapiro-Shibata-Nakamura (computational relativity)
BW Bandwidth
CCPP Climate Change Prediction Program (DOE/OS program)
CCSM Community Climate System Model (climate modeling)
CEIMC Coupled Electronic-Ionic Monte Carlo (nanoscience)
CEIMD Coupled Electronic-Ionic Molecular Dynamics (nanoscience)
CENIC Corporation for Education Network Initiatives in California
CGD Climate and Global Dynamics
CITRIS Center for Information Technology Research in Interest of Society
COTS Commercial off-the-shelf
CP Car-Parrinello (chemical dynamics algorithm)
CPU Central processing unit
CTSS Commercial TeraGrid Software Stack
CVS Concurrent Version System (software product)
DARPA Defense Advanced Research Projects Agency
DCA Dynamical cluster approximation (computational chemistry)
DEISA Distributed European Infrastructure for Supercomputing Applications
DFT Density functional theory (computational chemistry)
DHS Department of Homeland Security
DI Direct injected (diesel engine)
DIMM Double inline memory module
DNA Deoxyribonucleic acid (biology)
DNS Direct numerical simulation (turbulence computation)
DOE U.S. Department of Energy
DPCS Distributed Production Control System (LLNL software)
DSD Distributed Systems Department
DTF Distributed Terascale Facility
EECS Department of Electrical Engineering and Computer Science (U.C. Berkeley)
ERCAP Energy Research Computing Allocations Process
ES Earth Simulator
ESG Earth Systems Grid (DOE program)
ESnet Energy Sciences Network
ESSL Engineering and Scientific Software Library (IBM product)
ETF Extensible TeraGrid Facility
FE Finite element
Flop/s Floating-point operations per second
FFT Fast Fourier transform
FFTW (self-tuning FFT software)
FPMD First principles molecular dynamics (computational chemistry)
FPR Floating point register
FTC Fusion Technology Committee
FTE Full-time equivalent
FTP File transfer protocol
FY Fiscal year
GB Gigabyte
Gb/s Gigabits per second
Gflop/s Gigaflop/s (billion floating-point operations per second)
GHz Gigahertz
GPFS Global Parallel File System (IBM product)
GR General relativity
GSFC Goddard Space Flight Center (NASA center)
GSI Grid Security Infrastructure
GTC Stellarator Monte Carlo Transport (fusion code)
GYRO (gyrokinetic fusion code)
HCA Hardware custom accelerators (IBM design)
HCCI Homogeneous charge compression ignition (turbulence)

HECRTF High-End Computing Revitalization Task Force (multi-agency working group)
HOPI Hybrid Optical/Packet Infrastructure (Internet2 Working Group)
HPC High performance computing
HPCF High Performance Computing Facilities Division (Berkeley Lab)
HPCRD High Performance Computing Research Department (Berkeley Lab)
HPCS High Productivity Computing Systems
HPSS High Performance Storage System (IBM storage system)
HRM Hierarchical Resource Manager
HSI Hierarchical Storage Interface
IB InfiniBand (network technology)
IDS Intrusion detection system
IEEE Institute for Electrical and Electronic Engineers
INCITE Innovative and Novel Computational Impact on Theory and Experiment
I/O Input/output
IRU Irrevocable right-of-use (networking)
ISM Interstellar matter
ITER International Thermonuclear Experimental Reactor (fusion program)
JAMSTEC Japanese Marine Science and Technology Center
KDI Knowledge Discovery Initiative (NSF program)
LAN Local area network
LBNL Lawrence Berkeley National Laboratory
LCAT Leadership Computing Applications Team
LCC Leadership Computing Consortium
LCRM Livermore Computing Resource Management (LLNL software)
LCS Leadership Class System
LDA Local density approximation (computational chemistry)
LED Light-emitting diode
LES Large eddy simulation (turbulence computation)
LIGO Laser Interferometer Gravitational Wave Observatory
LLNL Lawrence Livermore National Laboratory (DOE laboratory)
LONI Louisiana Optical Network Initiative
LSMS Locally Self-consistent Multiple Scattering (computational chemistry)
LSST Large-aperture Synoptic Survey Telescope
LSU-CCT Louisiana State University Center for Computing and Technology
LU Lower-upper diagonal (numerical linear algebra technique)
MAN Metro Area Network
MASS Mathematical Acceleration SubSystem (a math and science library from IBM)
MB Megabyte
MC Monte-Carlo (numerical technique)
MD Molecular dynamics
MEMS Micro electro-mechanical systems
MFN Metromedia Fibre Network, Inc.
MHD Magneto-hydrodynamics (plasma physics)
MICS Mathematics, Information and Computer Science (DOE program)
MM Molecular mechanics (computational chemistry)
MPEG Moving Pictures Experts Group (data compression standard)
MPI Message Passing Interface (parallel computing software)
MPLS Multiprotocol Label Switching (network technology)
MPP Massively parallel processing
MSP Multi-Streaming Processor (Cray design)
NAS Numerical Aerospace Simulation (NASA Ames computer facility)
NASA National Aeronautics and Space Administration
NCAR National Center for Atmospheric Research (NSF facility)
NCSA National Center for Supercomputer Applications (NSF facility)
NERSC National Energy Research Scientific Computing Center
NFACS National Facility for Advanced Computational Science
NIM NERSC Information Management (account and allocation software)
NIMROD Non-Ideal Magnetohydrodynamics with Rotation, Open Discussion (fusion)
NLR National Lambda Rail (network)
NNSA National Nuclear Security Administration (DOE program)
NPACI National Partnership for Advanced Computational Infrastructure (NSF)
NPB NAS Parallel Benchmarks
NREL National Renewable Energy Laboratory (DOE laboratory)
NREN NASA Research and Education Network (network)
NSF National Science Foundation
NWCHEM Northwest Chemistry (PNNL software)
OASCR Office of Advanced Scientific Computing Research (DOE program)
OC Optical cable (networking standard)
ODE Ordinary differential equation
ORNL Oak Ridge National Laboratory (DOE laboratory)
OSF Oakland Scientific Facility (LBNL computer center)

PACI Partnership for Advanced Computational Infrastructure (NSF program)
PB Petabyte
PDE Partial differential equation (numerical approach)
PDSF Parallel Distributed Systems Facility (NERSC computer system)
PERC Performance Evaluation Research Center (SciDAC program)
PERCS Productive, Easy-to-use, Reliable Computing System (IBM project)
PETSc Portable, Extensible Tool for Scientific Computing (numerical library)
Pflop/s Petaflop/s (quadrillion floating-point operations per second)
PIC Particle in cell (computational technique)
PIMD Path integral molecular dynamics (computational chemistry)
PNNL Pacific Northwest National Laboratory (DOE laboratory)
POC Point of contact
PoP Point of presence (network technology)
POP Parallel Ocean Program (climate modeling code)
PoS Packet over SONET (network technology)
PPDG Particle Physics Data Grid (DOE program)
QFS (Sun Microsystems file system product)
QM Quantum mechanics
QMC Quantum Monte Carlo
QoS Quality of service (network technology)
RDMA Remote direct memory access
RFI Request for Information
SAN Storage area network
SC DOE Office of Science
SCaLeS Science Case for Large-scale Simulation (DOE working group)
SciDAC Scientific Discovery through Advanced Computing
SDSC San Diego Supercomputer Center (NSF facility)
SI Spark ignition (engine)
SLURM Simple Linux Utility for Resource Management (LLNL software)
SMP Symmetric multiprocessor
SONET Synchronous Optical Network
SP Scalable Parallel (IBM parallel computer product)
SPAI Sparse approximate inverse (numerical technique)
SRM Storage Resource Manager (LBNL research project)
SSP Single-Streaming Processors (Cray design)
SuperLU (numerical software product for sparse LU factorization)
TB Terabyte
TCP Transmission Control Protocol
TeraGrid (NSF distributed facility)
Tflop/s Teraflop/s (trillion floating-point operations per second)
TLBE Thermal Lattice Boltzmann Equation (fusion code)
TOPS Terascale Optimal PDE simulation (SciDAC project)
UC University of California
UID User identification
UPC Unified Parallel C (programming language)
ViVA Virtual Vector Architecture (LBNL-IBM project)
VLIW Very long instruction word (computer architecture)
VVP ViVA Virtual Processor
WAN Wide area network
WRF Weather Research and Forecasting Model

APPENDIX K
Budget

[Proprietary and confidential information deleted.]