

Creating Science-Driven Computer Architecture: A New Path to Scientific Leadership

C. William McCurdy,¹ Rick Stevens,² Horst Simon,¹ William Kramer,¹ David Bailey,¹
William Johnston,¹ Charlie Catlett,² Rusty Lusk,² Thomas Morgan,² Juan Meza,¹
Michael Banda,¹ James Leighton,¹ and John Hules¹

¹Computing Sciences Directorate
Ernest Orlando Lawrence Berkeley National Laboratory
University of California
Berkeley, California 94720

²Argonne National Laboratory
9700 S. Cass Avenue
Argonne, Illinois 60439

October 2002

Contents

Creating Science-Driven Computer Architecture: A New Path to Scientific Leadership.....	1
<i>The Challenge.....</i>	<i>1</i>
<i>The Context of This Proposal: A Mission for the DOE to Restore American Leadership in Scientific Computing..</i>	<i>1</i>
1. Introduction: A Strategy for Creating a New Class of Computer Architectures for Scientific Computing ..	2
1.1 Sustained Cooperative Development of New Computer Architectures	2
1.2 A Focus on Sustained Performance of Scientific Applications	3
1.3 A Strategy to Pursue Several Architectures at Multiple Sites	4
1.4 Embedding Ultrascale Capabilities in a National Grid Infrastructure.....	5
1.5 A New Investment in the Computer Science Research and Scientific Research Communities.....	5
2. A Focus on Sustained Performance of Scientific Applications.....	6
2.1 Application Teams to Drive the Design of New Architectures	6
2.1.1 Combustion Simulation and Adaptive Methods (J. Bell, P. Colella, J. Grcar, LBNL)	7
2.1.2 Computational Astrophysics (R. Rosner, U. of Chicago; J. Borrill, P. Nugent, J. Shalf, M. White, LBNL).....	8
2.1.3 Nanoscience (A. Canning, L. W. Wang, LBNL; L. Curtiss, P. Zapol, M. Sternberg, D. Wolf, S. Zygmunt, ANL; S. Louie, U.C. Berkeley; M. Stocks, ORNL)	9
2.1.4 Biochemical and Biosystems Simulations (M. Colvin, G. Galli, F. Gygi, LLNL; D. Dixon, T. Straatsma, PNNL; T. Head-Gordon, LBNL)	11
2.1.5 Computational Environmental Molecular Science (D. Dixon, T. Windus, W. De Jong, PNNL)	14
2.1.6 Climate Modeling (R. Malone, LANL; W. Washington, NCAR; C. Ding, M. Wehner, LBNL)	15
2.1.7 Accelerator Modeling (R. Ryne, LBNL; K. Ko, SLAC).....	16
2.1.8 Lattice Quantum Chromodynamics (R. Sugar, UCSB)	17
2.1.9 Advanced Simulations of Plasma Microturbulence (W. Tang, PPPL).....	19
2.1.10 Quantum Monte Carlo Calculations of Nuclei (S. Pieper and R. Wiringa, ANL)	20
2.1.11 High Energy / Elementary Particle Physics (L. Price, ANL)	21
2.2 A New Architecture Strategy: Beyond Evaluation to Cooperative Development.....	23
An Early Indication of the Feasibility of This Approach: The IBM/ANL/LBNL Workshop of September 3–4, 2002	24
2.3 Key Machine Parameters	25
3. Strategy to Leverage Major U.S. Computing Capability.....	25
3.1 Blue Gene/L	26
3.1.1 Blue Gene Architecture Overview	28
3.1.2 Collaborative Development with IBM	29
3.1.3 Milestones.....	30
3.2 Blue Planet: Extending IBM Power Technology and Virtual Vector Processing	31
3.2.1 Blue Planet System Architecture	32
3.2.2 The Virtual Vector Architecture	33
3.2.3 Designs with Future Enhancements for Scientific Computing	34
3.2.4 Milestones.....	36

4. Possible Hardware Collaborations with Other Laboratories and Universities	36
4.1 <i>Technology Exploration in VLIW and Circuit-Board-Level Optical Interconnects with Hewlett-Packard</i>	37
4.2 <i>Expansion of a NERSC Production Computational Resource with a Cray XI</i>	38
4.3 <i>Tracking the DARPA HPCS Program</i>	39
4.4 <i>Alternative Technologies</i>	39
5. Computer Science Research Strategy for Petascale Applications.....	39
5.1 <i>Research Team and Goals</i>	40
5.2 <i>Fault Tolerance</i>	41
5.3 <i>Numerical and Communication Libraries</i>	41
5.4 <i>Compilation Techniques</i>	41
5.5 <i>Parallel I/O</i>	41
5.6 <i>Performance Analysis and Tuning</i>	41
6. Advanced Grid and Networking Infrastructure As a Science Enabler	42
6.1 <i>Grid Middleware: Creating an Integrated Science Apparatus</i>	44
6.2 <i>Ultrascale Networks: Removing the Barrier of Location</i>	47
6.3 <i>Workplan</i>	48
6.4 <i>Milestones</i>	49
7 Sustaining Science-Driven Computing.....	50
7.1 <i>Managing Multiple Paths to New Computing Architectures</i>	50
7.2 <i>Estimated Costs Associated with Each Technology Option</i>	51
7.3 <i>Decisions in the ANL and LBNL Partnership</i>	51
8. Conclusion.....	52

Creating Science-Driven Computer Architecture: A New Path to Scientific Leadership

The Challenge

The purpose of the new program proposed here is to bring into existence a new class of computational capability in the United States that is optimal for science. In recent years scientific computing has increasingly become dependent on hardware that is designed and optimized for commercial applications. However, while science in this country has benefited from the improvements in computers that derive from advances in microprocessors following Moore's Law, a strategy of relying on machines optimized primarily for business applications has handicapped American science. The performance of the recently completed Earth Simulator in Japan exposed the seriousness of this problem in the critical arena of American scientific supercomputing.

Typical scientific applications are now able to extract only 5 to 10 percent of the power of supercomputers essentially built from commercial web and data servers. By contrast, the design of the Earth Simulator makes 30 to 50 percent of its power accessible to the majority of types of scientific calculations. The combination of its 40 trillion operations per second and its optimization for scientific computing gives researchers using it a clear advantage over those without access to such capability.

The United States must undertake a program that will result in scientific computing capability that durably returns the advantage to American science, because doing so is crucial to the country's future. A new class of computer designs will revolutionize the power of supercomputing for science, but it also will affect scientific computing at all scales. What is called for is the opening of a new frontier of scientific capability that will ensure that American science is not handicapped in its pursuit of research in critical areas such as nanoscience, climate prediction, combustion, and fusion energy.

The Context of This Proposal: A Mission for the DOE to Restore American Leadership in Scientific Computing

The Department of Energy must create a new class of computing capability for the country by undertaking the research and development necessary to build supercomputers optimized for science in partnership with the American computer industry. The combined strengths of the national laboratories and the university computer science community must be brought to bear on this problem to ensure its success within five years or less. The DOE must make these computers available to the *entire* U.S. scientific community, especially in the development stages, so as to expose these machines to the rigorous testing that American scientists must be able to give them if there is to be broad-based confidence in their full-scale deployment. The Department must take on the role of stewardship of computer science research in this class of computer design for science, because the computer architecture research community is critical to the success of its larger mission in science and national security.

We believe that the Department of Energy must enlarge its vision of its responsibility for establishing and maintaining the superiority of scientific computing in this country, because if it restricts its attention only to its national laboratories and university grantees, it cannot argue that its investments are being made in the larger best interest of the U. S. scientific community. While that certainly does not mean that the DOE can or should provide supercomputing access for the entire country, it does mean that the Department must make the new class of machines available to the academic scientific research community until they can be deployed by other scientific agencies.

In this document we propose a multi-site strategy for accomplishing this mission, initially pursuing three different directions of hardware development and deployment, and establishing highly capably networking and grid infrastructure connecting them and their partners to the broad research community.

1. INTRODUCTION: A STRATEGY FOR CREATING A NEW CLASS OF COMPUTER ARCHITECTURES FOR SCIENTIFIC COMPUTING

In the 1980s and early 1990s there were more than twenty U.S. companies producing supercomputers that were designed for scientific and technical applications. Among them were Alliant, BBN, Convex, Cray Research, Cray Computer, IBM, Intel Supercomputer Division, Kendall Square, and Thinking Machines. The primary role of national laboratories and universities at that time was to evaluate their offerings and select the best for use in scientific research. The market for high-performance scientific and technical computing was a significant focus of the computer industry.

Today the situation is radically different because the market for commercial Web and data servers has grown to completely overshadow the market for high performance scientific computing. Supercomputers in use now largely consist of commercial servers, entirely unmodified except for being connected by networks that have not increased in capability at the same rate as the processors they connect.

Today's situation calls for a strategy that creates a new class of supercomputing machinery by leveraging Moore's Law and the technology that underlies commercially viable computers and the microprocessors they are based on, instead of merely utilizing its existing implementations. Most crucially, this strategy must provide a new way to couple scientific applications to development of computer architectures, thereby opening a sustainable path to petaflops/s-level performance and beyond.

1.1 Sustained Cooperative Development of New Computer Architectures

The Argonne and Berkeley Laboratories propose a new type of development partnership with computer vendors that goes beyond the mere evaluation of the offerings that those vendors are currently planning for the next decade. We propose and describe in this paper a comprehensive strategy which includes development partnerships with multiple vendors. Those partnerships will bring to bear:

1. teams of scientists and computational mathematicians who will modify and optimize their applications for future systems through the use of simulators and prototypes of new hardware
2. teams of computer architects from major U.S. computer vendors who will interact directly with the scientific applications teams, and
3. teams of computer scientists who will work with both applications scientists and computer architects to analyze and abstract the requirements of scientific applications so that they can be addressed in hardware and to develop the software environments that will allow scientists to extract the maximum performance and capability from that hardware.

This strategy is directed at challenging and partnering with vendors to create architectures that perform to a target level on a specific suite of scientific applications. Unlike the current approach that has become standard for both scientific and commercial computing, this new approach does not merely abstract requirements of example applications as the primary mechanism for making performance-related design decisions. That strategy fails for scientific computing because the core algorithms for scientific applications are vastly more diverse than those in commercial applications and because they continue to be combined in new ways as scientific understanding and problem solving approaches progress.

This new strategy requires sustained partnerships because the development time for new computer architectures is from three to six years and one may need to iterate over several generations of new hardware to achieve the desired outcome. Because of the long duration of the typical microprocessor development cycle, from design to commercial manufacture, the general purpose microprocessors for 2005 are already designed and in the stages of implementing their manufacture. U.S. vendors cannot commit to significant changes in their development plans without a sustained commitment from their partners such as those in the national laboratories. They cannot succeed with those changes if they are not able to constantly reevaluate and modify their plans in partnership with the scientific community. The national laboratories are a natural home for such partnerships, which will also include extensive participation of the university computer science and applications communities. The laboratories can sustain and manage the partnerships and deliver both software and in-depth analysis of applications performance to meet development milestones.

1.2 A Focus on Sustained Performance of Scientific Applications

Within a decade we believe it is possible to provide scientific computing in the U.S. with orders of magnitude increase in sustained performance. The development partnerships we will implement are motivated by the simple goal of achieving maximum sustained performance of scientific applications. The most successful architectures will enable a range of scientific applications to reach those sustained performance levels, but optimization for specific scientific applications is also a benefit of the strategy.

This strategy is not the same as simply “standing up” very large computers. It is intended to change the offerings of computer vendors at all levels, from the laboratory scale, single-principal-investigator scale, to the scale of supercomputers. We propose to change the building blocks of high performance scientific computing and the communications fabrics that enable their integration into large-scale systems. If we can improve applications performance of the building blocks, the nodes that consist of a few processors or tens of processors, to routinely

reach sustained performances of 30 to 50 percent of their peak capability, the core productivity of American computational science will increase dramatically at all scales.

The current measure of the upper levels of scientific computing is a peak speed of tens of teraflop/s (trillions of floating point operations per second), of which typical applications extract 5 or 10 percent (or less). The strategy we propose here is directed toward the ultimate goal of providing peak capabilities at a petaflop/s (1000 teraflop/s) by 2007–2009, of which 30 to 50 percent will be accessible to scientific users.

Doing so will require between 50,000 and 100,000 processors. No computer architecture known today will scale to that size and provide high levels of sustained performance. We propose that the Department of Energy step up to this challenge, and the strategy we offer can meet it.

1.3 A Strategy to Pursue Several Architectures at Multiple Sites

We propose a comprehensive strategy that makes use of the entire complex of the Office of Science national laboratories and investment in university scientific and computer science communities. The proposed initial expression of the program will be in three of the multipurpose DOE laboratories: Argonne, Berkeley Lab, and Oak Ridge. Even in this initial implementation, we propose a collaboration between these three laboratories, together with the Princeton Plasma Physics Laboratory, Pacific Northwest National Laboratory, Brookhaven National Laboratory, and a collaboration of the special-purpose laboratories who lead programs in accelerator design and quantum chromodynamics calculations.

The cost of scientific supercomputing is an issue of national strategic importance. The strategy we propose to implement will pursue *at least* three options:

1. At the highest cost per peak teraflop/s, the first option will involve custom components at all levels in an architecture known to be successful in scientific applications, *parallel vector processing*. The initial stages of this effort have been announced with the evaluation of a beta-test version of the Cray X1 at Oak Ridge National Laboratory.
2. At half this price and with the promise of sustainably high cost-effectiveness, the second option will involve commercial microprocessors in a new architecture that will be programmable in the same way as the first option, *ViVA* or *Virtual Vector Architecture*. IBM will partner with Lawrence Berkeley National Laboratory to implement early versions of this architecture and deliver Blue Planet, a 160 teraflop/s mature implementation in the second half of 2005.
3. At the lowest cost per peak teraflop/s, the third option will be based on “system-on-a-chip” architecture that is being explored most visibly in the IBM Blue Gene project. This architecture is arguably the most promising for reaching the petaflop/s goal of this proposal; however, its suitability for general scientific use has not yet been demonstrated. But at half the cost of option 2 and one quarter the cost of option 1, this path is extremely cost-effective to pursue and provides the best long-term bet currently known to the scientific community. IBM will partner with Argonne National Laboratory to develop new expressions of this architecture and deliver a 180 teraflop/s implementation appropriate for general scientific exploitation in 2005.

Option 2 will provide twice the sustained capability of the Earth Simulator at half the cost.

Option 3 will provide a new architecture family for scientific computing and one that makes a

definitive step towards cost-effective petaflop/s computers with high sustained levels of performance.

This proposal focuses on options 2 and 3 and the significant development partnership with IBM that will bring both of these new architectures into existence in a form, at a scale, and on a schedule that the scientific community can exploit successfully. But other options for the nation also exist, and we have begun to explore them as well. Among them are:

- A partnership with Hewlett-Packard to develop new architectures based on microprocessors that company is developing with Intel.
- A production implementation of the Cray X1 in 2004 at the NERSC Center to expose that architecture to the scientific community with the full production support of the NERSC staff, already experienced in vector architectures.
- Exploration of alternative technologies that are in development at universities today, and leveraging the new DARPA High Productivity Computing Systems program.

1.4 Embedding Ultrascale Capabilities in a National Grid Infrastructure

Regardless of which architectures are deployed in the next three years to enable the U.S. scientific community to maintain international leadership in computational science, it will be critical to integrate these new computing platforms with the emerging computational infrastructure that provides integrated access to large-scale data, computing platforms, instruments and users. It is critical that DOE deploy these systems in a manner that creates a national footprint and fully integrates them with the emerging research Grid.

A number of the scientific disciplines that DOE must address require a new scale of networking connectivity to exploit this new frontier of computing. The strength of American science is due in no small part to the infrastructure that supports it, and networking technology is rapidly reaching a level of performance that will change fundamentally our expectations of accessibility of our computational resources and our ability to move tasks and data between them transparently. We describe therefore in this proposal a dramatic step forward in the deployment of a national high-performance networking infrastructure that complements the advanced architectures that will be developed.

1.5 A New Investment in the Computer Science Research and Scientific Research Communities

Two other classes of investment are critical to the success of this proposed new program at DOE. They must not be neglected in the initial description of the initiative to ensure leadership in computing for science.

- A long-term program of computer science research in computer architecture for science that reestablishes a viable national university research community in this discipline as well as a core capability in this area in the DOE national laboratories.

- An investment across the entire range of the research portfolio of the DOE Office of Science to enable researchers funded by ASCR, BER, BES, FES, and HENP to make use of new architectures as they become available as prototypes and production installations.

In the 1980s and early 1990s there were roughly 50 groups of computer scientists in universities across the United States investigating new computer architectures specifically directed at high performance scientific computing. They were primarily funded by DARPA as part of the High Performance Computing Initiative. Today there are only a handful of such groups, and their interest, like that of the larger computer science community, is turning primarily to new technologies that exploit Web and Grid computing as well as a pervasive digital infrastructure for the country.

DOE must accept the responsibility of reestablishing a national research community in this discipline if it is to provide a sustainable path to leadership in scientific computing. University researchers must be able to depend on stable interest from a funding agency in this arena if they are to devote their careers to it. The DOE is capable of managing such a program, but it will be a new effort requiring new investments and program structures. No other U.S. funding agency has the capability and mission requirements to assume this leadership role for advanced scientific computing.

The first steps in investing across the research portfolio of the Office of Science were taken in the SciDAC initiative. Those were only small steps by comparison with what is required to bind the scientific research communities in BER, BES, FES, and HENP, together with the applied mathematics community in ASCR, to the project of creating new computing architectures for science. Those communities are ready for this challenge, but the effort required to explore multiple architectures and programming models is well beyond their present resources. This initiative must be driven by scientific applications and managed to use them to drive the creation of new computing resources.

2. A FOCUS ON SUSTAINED PERFORMANCE OF SCIENTIFIC APPLICATIONS

2.1 Application Teams to Drive the Design of New Architectures

The goal of this proposal is to develop petaflop/s platforms before the end of the decade that allow a sustained performance of 30 to 40 percent of peak on major DOE computational science applications. In the Power X track, the near-term goal for 2005 is to develop platforms that deliver twice the sustained applications performance of the Earth Simulator—about 40–50 Tflop/s—at about half the system price. The Blue Gene/L track has more ambitious performance goals of a full petaflop/s peak system by 2006 and a 3 Pflop/s peak system in late 2007. Because of the novel architecture employed in Blue Gene, high sustained-to-peak ratios will be a big challenge.

In order to achieve these goals, we have selected a preliminary set of mission-critical computational science applications from across the DOE complex. In all applications areas, we have engaged leading scientists from across DOE labs and universities to become applications partners. In order to achieve the high sustained performance goals, we need to engage these applications partners and the hardware vendors in a cooperative platform development strategy

described in section 2.2. The near-term goal for 2005 is to demonstrate 50 Tflop/s sustained performance or higher on at least three of the applications developed in cooperation with the applications partners.

From the long list of potential applications partners who will benefit in their computational science goals from the new architectures proposed here, we have assembled a small but high quality and highly motivated team of applications scientists and computational mathematicians. The goal of this applications team is to interact in ongoing workshops with computer scientists from ANL and LBNL, selected university-based computer science researchers, and vendor developers. This core group will not be larger than a single roomful of people. The goal of this activity is to nucleate a sustained partnership that will work together for the next five years to achieve a fusion of architecture, applications, and systems software.

The following sections on the individual applications describe science goals that can be accomplished at three levels of sustained performance: 50 Tflop/s, 250 Tflop/s, and 1,000 Tflop/s. They coincide with the major milestones of the architecture track of 180 Tflop/s, 1,000 Tflop/s, and 3,000 Tflop/s peak machines proposed in sections 3.1 and 3.2.

2.1.1 Combustion Simulation and Adaptive Methods (J. Bell, P. Colella, J. Grcar, LBNL)

One of the critical components of U.S. energy policy is the development of methods for burning hydrocarbon fuels that are both efficient and low in emissions. Computer simulations of combustion processes have the potential for providing fundamental scientific insight into combustion processes, as well as being tools to aid in the engineering design of specific combustion devices.

There are two critical algorithmic components required to successfully simulate these problems on computers capable of sustained performance in the 50–1,000 Tflop/s range. One is the use of low-Mach-number formulations to eliminate time step constraints due to stiff acoustics, while retaining bulk compressibility effects. The second is the use of adaptive mesh refinement and other adaptive techniques to concentrate computational effort in localized regions where it is most needed, such as a thin flame zone. The combination of these two approaches leads to calculations involving discretizations of classical partial differential equations (elliptic, parabolic, and hyperbolic) on locally structured but globally irregular grids. Since this class of mathematical problems can be used to represent a broad range of problems in the DOE science portfolio, success in the development these technologies for combustion would also provide new and improved tools for other areas as well.

The use of the approach described here decreases the computational cost by a factor of 10,000 over conventional resolved-acoustic approaches discretized on uniform grids. For that reason, it would be possible to simulate a broad range of combustion problems with full-fidelity models. A set of steps along the way to full-fidelity models is clear.

Scientific goals achievable with an increase in computational capability include:

- 50 Tflop/s: Simulation of turbulent natural gas combustion for laboratory-scale systems.
- 250 Tflop/s: Simulation of natural gas combustion for engineering-scale systems.
- 1,000 Tflop/s: High-fidelity computations for problems involving gasoline-like fuels, such as n-heptane.

2.1.2 Computational Astrophysics (R. Rosner, U. of Chicago; J. Borrill, P. Nugent, J. Shalf, M. White, LBNL)

Astrophysics spans a wide range of problems of central concern to DOE:

- The early universe is the ultimate high energy accelerator, and most likely holds the keys to understanding grand unification.
- Cosmology has revealed a new type of force field, related to the ad hoc repulsive term first considered (and rejected) by Einstein.
- Astronomical nucleosynthesis during the first few seconds of the universe and subsequently within both normal stars and novae and supernovae (as well as spallation occurring in places such as the interstellar medium) is at the heart of virtually all element generation in the universe.
- The evolution of stars such as the Sun is thought to be a key to understanding the evolution of the Earth's climate.

Thus, there are deep connections between the physics studied by nuclear astrophysicists and stellar astronomers and the kinds of physics and chemistry problems studied by DOE scientists in the Office of Science's nuclear, high energy, and energy technologies efforts, as well as in the DOE/NNSA's Stockpile Stewardship program.

Modern massively parallel computing systems and software have evolved to the point where a number of significant physics problems can be attacked by direct numerical simulations, in which the calculations are used analogously to laboratory experiments to explore and understand physical processes. Such exploration and understanding occurs in two distinct ways: first, by extensive exploration of the control parameter space by large numbers of otherwise identical calculations; second, by "hero" calculations, which stretch the capabilities of existing computers by simulating physical processes for which no conceivable laboratory experiment or direct astrophysical observation can be carried out (a typical example being the modern generation of general relativistic fluid simulations of neutron star or black hole mergers).

Other types of astrophysics and plasma physics calculations are beyond current capabilities for direct numerical simulations in the U.S., but can be envisaged as solvable by computers that conceivably can be built, based on our current knowledge and extrapolation of technology. An example is the study of fully turbulent jets, for which a key criterion is the Reynolds number (Re), which scales with the ratio of the largest to smallest resolved spatial scale as $Re \sim 2 (L_{\max}/L_{\min})^{4/3} \sim 2 \times 10^4$ for $L_{\max}/L_{\min} \sim 10^3$ (which can be achieved today for occasional "hero" calculations). Carrying out such calculations at modestly larger Reynolds numbers (i.e., further into the well-developed turbulent regime) and over a range of control parameters¹ is the next key step in carrying out detailed physics studies; such studies are clearly going to be possible within the next five years.

Among the largest 2D and 3D simulations carried out on any extant computer are simulations of Type Ia (exploding white dwarf) and Type II (core collapse) supernovae (typically using high-resolution Godunov-based schemes such as PPM and wENO); simulation of stellar interior convection, using high-resolution pseudo-spectral schemes solving the Boussinesq or anelastic

¹ In addition to the Reynolds number, important control parameters include the jet-to-ambient fluid density ratio and the ratios of governing diffusivities such as material diffusion, momentum diffusion (viscosity), etc.

equations; simulations of structure formation in the early universe, especially those linking N-body codes (typically based on some sort of tree approach) to hydro codes (again, usually based on some type of Godunov scheme); and simulations of cataclysmic events such as black hole mergers, based on the use of newly developed schemes for numerically solving Einstein's equations. With this step, the scope of numerical simulations as a proxy for laboratory experimentation will become significantly larger than it is today, and will begin to seriously encompass a wide variety of complex physical systems.

Another aspect pertains to developments in detector technologies that have driven an exponential increase in data-gathering capacity. This trend is transforming astrophysics from a data-starved to a data-swamped discipline, where the challenges include building and managing extremely large astronomical databases (tens to hundreds of terabytes). From satellite missions providing high-resolution images of the Big Bang to dedicated telescopes surveying millions of galaxies, from neutrino detectors buried deep in the Antarctic ice to huge interferometers detecting gravity waves for the very first time, this data deluge has the capacity to transform our understanding of the cosmos, addressing such fundamental questions as the nature of the mysterious dark energy and dark matter that comprise 95 percent of the Universe.

Exploiting the full scientific potential of the data will require heroic efforts in both data analysis and numerical simulation. First we need to derive a set of statistics characterizing our Universe from the petabytes of raw data, and then we need to compare these reference points with the results of detailed simulations of candidate astrophysical models—quantum fields evolving in the very early Universe, the formation of superclusters of galaxies, supernovae explosions, or colliding black holes. The computational constraints we face in applying these methods derive from the rapidly increasing data size and power law algorithmic scaling for the data analyses, and from the enormous dynamic range needed—from quantum to cosmic scales—in the simulations.

Scientific goals that could be achieved with an increase in computational capability include:

- 50 Tflop/s: We could perform precision analysis of the PLANCK satellite's cosmic microwave background radiation data to determine the fundamental parameters of cosmology to the percent level or better.
- 250 Tflop/s: We could perform a full hydrodynamical simulation of structure formation in a cosmological volume, resolving the largest individual galaxies.
- 1,000 Tflop/s: We could simulate a supernova in 3D and follow it with time evolving spectrum synthesis including all the microphysics, and, if appropriate, begin to follow its core collapse into a black hole.

2.1.3 Nanoscience (A. Canning, L. W. Wang, LBNL; L. Curtiss, P. Zapol, M. Sternberg, D. Wolf, S. Zygmunt, ANL; S. Louie, U.C. Berkeley; M. Stocks, ORNL)

Fabrication and integration of nanoscale devices promises to revolutionize science and technology, provided the unique properties of nanostructures can be understood through modeling and simulations. Nanoscience has potential applications from targeted drug delivery to single or few electron devices for computers that will take us well beyond Moore's law. The DOE is one of the major players in the National Nanoscience Initiative. Computational modeling will play a central role in the realization of the potential of nanoscience. In nanoscience much of

the arsenal of traditional materials science of analytical/theoretical methods cannot be applied to complex nanoscale systems. Realistic, quantitatively accurate calculations on nanoscale systems will require new human and computer resources that are not presently available to the scientific community.

The simulation of self-assembly processes has been identified as one of the grand challenges for computational nanoscience. Self-assembly is the coordinated action of independent entities under distributed (i.e., non-central) control to produce a well-defined larger structure that may exhibit some desired collective effect. The independent building blocks self-assemble to form complex structures having properties that differ dramatically from those of the constituents. Self-assembly is becoming an increasingly important strategy for the fabrication of nanostructured materials that will be used in energy conversion, catalysis, biomolecular materials, and nanoelectronics.

A substantial investment in the development of new, robust computational tools and theoretical models is required to accurately simulate the complexity and variety of self-assembly processes. The tools and models must be able to cover a large range of time and length scales, and a wide range of chemical interactions from van der Waals interactions to covalent bonds. In addition, computational studies of self-assembly will require substantial enhancements in DOE computer capabilities.

The computational methodologies in nanoscience range from classical dynamics to *ab initio* electronic structure calculations usually based on a density functional theory approach to solving Schrödinger's equation. Generally speaking, there are two goals in computational nanoscience:

1. To simulate the synthesis process of nanostructures.
2. To simulate the electronic, optical, and mechanical properties of a synthesized nanostructure or nanostructure complex.

Goal 1 often involves molecular dynamics or Monte Carlo simulations but has also been tackled by finite element methods. Extension of all the methods used to describe assembly of nanostructures to fully 3D systems is critical and necessary in order to realistically incorporate the physics and faithfully capture the statistical mechanics of a self-assembling system.

Goal 2 is often addressed by electronic structure calculations. The challenge of nanostructure calculations is their size, often ranging from a few thousand atoms to a few million atoms. In the simulation of the synthesis, the total simulation time is beyond what can be done with accurate *ab initio* methods. Therefore at present we are limited by both length and timescales in our calculations. In electronic structure calculations, linear scaling algorithms that exploit some locality in the system show promise to handle larger systems. Currently, conventional *ab initio* methods scale as the cube of the size of the system. The density functional-based tight binding (DFTB) method is an example of a promising technique that is being developed for the nanoscale regime. This is an approximate electronic structure method that potentially can handle systems of a very large numbers of atoms with explicit treatment of electrons. The DFTB method is also used for exploiting electronic structure information directly in molecular dynamics (MD) simulations.

Computational studies of self-assembly processes will require multiscale simulations based on a variety of algorithms covering different length and time scales. In the nanoscale regime it will be necessary to carry out computations on nano building blocks including the interfaces between them, which control the thermodynamics and kinetics of self-assembly. The accuracy required in

the nanoscale regime requires methods that explicitly include electrons in the simulations (such as density functional theory). The information from the electronic structure regime is needed as input to the mesoscale or coarse-grained regime, where the simulations are performed on thousands or millions of the nano building blocks represented by unified particles. Multiscale simulations that combine both the nanoscale and mesoscale regimes will be able to simulate self-assembly and will require significant expansion or current computer capabilities.

The scalability of the electronic structure methods with system size presents significant challenges. Various standard methods scale as N^2 to N^7 , where N is the number of atoms. Development of algorithms with improved scaling is an important field of current research, with the common goal to achieve linear scaling of computational effort with system size. With the most powerful computers we have at present of a few teraflop/s, even with linear scaling methods we can only perform calculations for about one thousand atoms in electronic structure calculations, and a few picoseconds timescales for about a hundred atoms in an *ab initio* molecular dynamics simulation. This means we can only simulate the smallest nanoscale systems such as fullerenes, and we are unable to accurately simulate the synthesis of even these simple systems. An *ab initio* electronic structure calculation for a 10,000 atom system, on the other hand, would allow us to study some of the technologically important nanostructures such as quantum dots, wires etc., as well as single building blocks in complex nanostructure devices. At present these systems can only be simulated using less accurate semi-empirical methods. With a 250 Tflop/s computer, we could perform *ab initio* calculations of up to 50,000 atoms. This would allow us to simulate the interactions between a few different building blocks of a nanoscale device as well as simple interfaces between “soft” biologically or organically based structures and “hard” inorganic ones. We would also be able to simulate with less accurate methods the synthesis of small (a thousand atoms) nanostructures. With a 1,000 Tflop/s computer, we would be able to simulate a complete nanostructure device with *ab initio* accuracy. A 1,000 Tflop/s computer would also allow us to do a synthesis simulation of a few nanoseconds with *ab initio* accuracy. We would also be able to perform excited state and transport calculations for large nanostructures.

Scientific goals that could be achieved with an increase in computational capability include:

- 50 Tflop/s: *Ab initio* electronic structure calculation for a 10,000 atom system. This is sufficient to investigate interaction of two nanoparticle (5 nm in diameter) building blocks that are used in the self-assembly of nanoparticle arrays on surfaces.
- 250 Tflop/s: *Ab initio* electronic structure calculation for a 50,000 atom system. This would allow the investigation of interactions in clusters of nanoparticles with varying surface termination, stoichiometry, and other properties that might be influencing the assembly process.
- 1,000 Tflop/s: Simulate a complete nanostructure device using *ab initio* methods. This would allow simulations of self-assembly of small arrays of nanoparticles on a surface and other phenomena such as self-assembly of DNA strands.

2.1.4 Biochemical and Biosystems Simulations (M. Colvin, G. Galli, F. Gygi, LLNL; D. Dixon, T. Straatsma, PNNL; T. Head-Gordon, LBNL)

The emerging, detailed understanding of biology is one of the great triumphs of science, holding vast potential for improving human well being. One important consequence is that a growing

fraction of biology can now, in principle, be simulated computationally. Indeed, no field offers a greater combination of challenge and promise for computer modeling. For example, all biological processes are eventually determined by molecular interactions and conformational changes. However, these processes typically involve a complex of a large number of individual molecules interacting with assemblies of very large molecules. A detailed understanding of cell signaling pathways, such as the epidermal growth factor (EGF) or tumor necrosis factor (TNF α) pathways that control cell differentiation, growth, and death and inflammatory response requires that complex protein-protein interactions be studied. A thorough understanding of this pathway is of crucial importance for the development of cancer treatments as well as understanding how cancer is initiated. Because computer simulations can provide atomic-level detail that is difficult or impossible to obtain from experimental studies, computational studies are essential. However, this requires the modeling of an extremely large complex of biomolecules, including bilayer lipid membranes, transmembrane proteins, a complex of many intercellular kinases, and, of course, thousands of waters of solvation. This leads to systems on the order of 5+ million atoms, and a simulation on such a system would provide the first molecular view of the initiation of the signal transduction pathway.

The challenges involved in modeling biological processes arise from the range of timescales, from femtosecond enzymatic reactions to multisecond protein folding pathways, and size scales, from small organic molecules to multi-protein complexes with millions of atoms and ultimately very large scale biological pathways and systems. Moreover, the subtle energetics of biochemical reactions and the complex environments in which they occur push the limits of available simulation methods. Overcoming these challenges to allow predictive biological simulations will lead to breakthroughs with profound impacts on human health, energy production, environmental protection, and materials engineering. The enormous complexity of biological systems and the difficulty of using information from small model systems to address complex, collective phenomena at large scales requires significant advances in theories, algorithms, software, and hardware. Currently available computing resources allow computer simulations of biomolecular systems to be routinely carried out for about 100,000 atoms for tens of nanoseconds. Computer resources will need to increase in power by at least 3 orders of magnitude in order to be able to allow microsecond simulations of systems with several million atoms. This is an ambitious goal, but absolutely necessary if we want to make a significant impact on the nation's key scientific problems in health, environmental protection, and national security.

Over the past four decades, a wide variety of chemical simulation methods have been developed. The choice of methods involves trade-offs between accuracy, size of the chemical system, and computational cost. At one extreme are *ab initio* quantum chemical (QC) methods that can in principle predict any chemical property to high accuracy but are in practice computationally limited to relatively small molecular systems and non-dynamic simulations. At the other extreme is classical molecular dynamics (MD), which simulates the motions of atoms in their chemical context but usually uses molecular mechanical ("ball and spring") force fields that have limited accuracy. While these methods can be used to understand some biological phenomena, a number of biological simulations fall in between these two methods, such as elucidating enzyme mechanisms and modeling DNA mutation. The enormous improvements in chemical simulation algorithms and parallel supercomputer speeds promise to bridge the current gap between quantum chemistry and molecular dynamics to allow so-called *first principles molecular dynamics* (FPMD) simulations of biomolecular systems in the near future.

FPMD simulations are extremely demanding computationally. The FPMD applications have been highly optimized for current teraflop/s-speed computers on which they are capable of simulating up to a few hundred atoms for a few picoseconds. However, these methods also constitute a nearly exact simulation of nature, and even within these computational limitations, it is becoming an important tool for studying fundamental biochemical processes. For example, recent publications have demonstrated that FPMD yields excellent results for the structure of pure water, the solvation of ions, the energetics of simple biochemical reactions, and the conformations of biochemical building blocks such as the phosphodiester backbone of DNA. Ongoing research projects involve using FPMD to simulate the activation of anticancer drugs and the biochemical proton transfer processes. These results for small biochemical systems provide tantalizing glimpses of the value of longer time and larger system size simulations that will be made possible with faster computers. The following list gives examples of some biochemical simulations that would be made possible with successive tenfold increases in computer speeds over our current ~ 1 Tflop/s capabilities.

Performance Increase	Example Biological Application
10 \times	Chelation processes of radioactive ions
100 \times	Biochemical reactions catalyzed by a few amino acids
1,000 \times	Enzyme-catalyzed reactions in model active site
10,000 \times	Receptor-ligand binding (excluding large-scale protein conformational changes)
100,000 \times	Ligand-induced conformational changes in proteins
1,000,000 \times	Folding simulations for small protein regions Ion transport through transmembrane pores

Using Earth Simulator (ES)-class computers, we expect major progress in the predictive power of FPMD techniques for simulations of biological systems. We also expect major progress in the development and use of coupled techniques linking various quantum mechanical approaches, such as FPMD and quantum Monte Carlo methods, and linking quantum mechanical and classical atomistic approaches. These coupled techniques will provide codes that both experimentalists and theorists can use to guide, interpret, and predict new phenomena and future experiments. Within the biology arena, the use of ES-computers will effectively bring quantum simulations into the realm of biology.

The computer simulation of large biomolecular systems on parallel architectures presents a number of significant challenges, in terms of memory management, processor speed, inter-processor communication, I/O bandwidth, and storage requirements. First, the size of these systems requires the use of distributed data models. Second, even the least complex description of intermolecular interactions in terms of effective pair potentials requires a significant amount of inter-processor communication of data. There will be an even greater demand for interconnect data traffic for simulations using interaction functions beyond pair-wise additive contributions, such as electronic polarization. Third, each molecular dynamics time step requires several synchronizations of all processors, which leads to significant load balancing obstacles. Fourth, simulation trajectories need to be stored for subsequent analysis calculations and visualization, presenting additional I/O bandwidth, network bandwidth, and data storage challenges. With the

increasing scale and complexity of new computer architectures, these issues become even more important, and will need to be (and are being) addressed by novel software designs. For example, we have begun to explore computational techniques that address some of these efficiency issues by designing our simulations such that multiple regions of phase space are simultaneously sampled in a single calculation. A number of applications can significantly benefit from such an approach, including calculation of thermodynamic quantities.

The development of computer architectures beyond teraflop/s performance, combined with new advances in computational sciences and new advances in experimental biological research, will accelerate our understanding of biological processes in microbial systems all the way to humans, and will have significant impact on developments of new treatments and techniques in health care, environmental stewardship, and protection against biological threats.

Scientific goals that could be achieved with an increase in computational capability include:

- 50 Tflop/s: Simulation of biomolecular systems from the current 1 to 10 nanosecond scale to the microsecond time scale; FPMD simulations of systems comprising 3,000–4,000 atoms for several picoseconds; FPMD simulations of systems comprising 200–300 atoms in the nanosecond range.
- 250 Tflop/s: Simulation of large biomolecular complexes in which slow conformational changes determine biological function; simulation of enzyme catalysis reactions using hybrid classical and quantum mechanical methods; microscopic modeling of DNA repair mechanisms and drug/DNA interactions.
- 1,000 Tflop/s: Simulation of processes in the millisecond time scale, such as protein folding and membrane transport and membrane fusion processes.

2.1.5 Computational Environmental Molecular Science (D. Dixon, T. Windus, W. De Jong, PNNL)

Computational molecular science is a key technology for addressing the complex environmental cleanup problems facing the Department of Energy's nuclear production sites as well as other polluted sites in the nation. To support the development of innovative technologies for remediating these sites, we will need to develop reliable models to investigate the impact of the technology and the appropriate level of risk of using the technology or of doing nothing. Reliable models of the risk of proposed remediation technologies will be critical to developing the safest and most cost-effective approaches to site cleanup and to public acceptance of the cleanup process and results. The underlying data used in the models which can be provided by computational molecular science must be of high quality for the models to be accepted.

Computational molecular science involves solving equations derived from quantum mechanics, including the effects of relativity, classical mechanics, and statistical mechanics. The accurate solution of these equations leads to very expensive calculations whose cost increases very rapidly with the number of atoms. For instance, doubling the number of particles increases the time for the calculation by a factor of at least 32. The development of new theoretical methods, new algorithms, and new software which incorporates these advances on new high performance computer architectures promises to lead to the reliable prediction at an unprecedented level of the molecular behavior of environmental contaminants.

The science goals for next generation computer architectures are:

- 50 Tflop/s: Accurate calculations for realistic models of lanthanides and actinides on complex mineral surfaces (<1.0 kcal/mol thermodynamics) to develop parameters for reactive transport models of the vadose zone.
- 250 Tflop/s: Accurate calculations for realistic models of lanthanides and actinides on complex mineral surfaces interacting with aqueous solutions (<1.0 kcal/mol thermodynamics) to develop parameters for reactive transport models of the vadose zone.
- 1,000 Tflop/s: Accurate calculations for realistic models of lanthanides and actinides on complex mineral surfaces interacting with aqueous solutions (<50% error in reaction rates) to develop parameters for reactive transport models of the vadose zone.

2.1.6 Climate Modeling (R. Malone, LANL; W. Washington, NCAR; C. Ding, M. Wehner, LBNL)

Future climate change as a result of human activities is an issue of great social and political importance. Government policymakers require significantly more accurate information about the effects of fossil fuel consumption to responsibly formulate energy usage policy. Observations of recent changes to the earth's climate system are limited. Hence, comprehensive numerical models of the coupled climate subsystems are our most important tool to detect and attribute changes in the present climate as well as make reliable predictions of future climate change. Key to the responsible use of these predictions is a far greater quantification of their uncertainties, especially on regional scales. Higher performance computer technology can be exploited to accomplish this by refining resolution and enlarging simulation ensemble size.

The climate system is a complicated coupled system of distinct components. The atmosphere, ocean, cryosphere, land, and biosphere all interact in a complex manner. The enormous ranges of spatial and temporal scales imply very large numbers of grid points and time steps in numerical simulation models. The most sophisticated climate models, known as "coupled general circulation models," are composed of individual submodels of each of these systems. Each of these submodels, complex in their own right, discretize the planet into separate grids. The equations of motion for the atmosphere, ocean, and sea ice all must be solved on these grids. Unresolved and/or local phenomena (e.g., turbulence, cloud processes, radiation, chemistry, etc.) are treated as source terms to these hydrodynamics equations. Limitations in present computer speed restrict production simulations in the U.S. to a horizontal resolution of about 300 km in the atmospheric submodel and about 60 km in the ocean. Useful regional climate change prediction will require an atmospheric resolution of 20 km or less. Resolution of ocean mesoscale eddies, believed to be important to energy transport in the seas, will require an oceanic resolution of 10 km.

Climate change research within the Department of Energy coordinates the efforts of multiple DOE labs with other agency projects, principally those of NSF, NOAA, and NASA. A DOE SciDAC consortium has a broad mandate to consider all future architectures. The most recently developed U.S. coupled general circulation model is the Community Climate System Model (CCSM2), developed at the National Center for Atmospheric Research (NCAR). At current resolutions, this model requires 5.5 wall-clock hours per simulated year on 128 processors of the current NERSC Power 3 machine (Seaborg). Extrapolating the performance of the CCSM2 simulation on Seaborg at current resolution and utilizing the performance characteristics of the proposed computer design suggests that multi-century integrations at half of the target resolution

(40 km in the atmosphere and 20 km in the ocean) can be performed on a 50 Tflop/s (sustained) machine with a two or three month turnaround. At 250 Tflop/s, integrations at the target resolution (20 km in the atmosphere and 10 km in the ocean) can be made in about a year. In order to perform ensemble simulations at this resolution needed to quantify simulated internal variability, a machine exceeding 1,000 Tflop/s is required.

- 50 Tflop/s: Multi-century CCSM2 simulation using 40 km resolution in the atmosphere and 20 km resolution in the ocean.
- 250 Tflop/s: Multi-century CCSM2 simulation using 20 km resolution in the atmosphere and 10 km resolution in the ocean.
- 1,000 Tflop/s: Ensemble simulations of multi-century CCSM2 simulation using 20 km resolution in the atmosphere and 10 km resolution in the ocean.

2.1.7 Accelerator Modeling (R. Ryne, LBNL; K. Ko, SLAC)

Particle accelerators have enabled remarkable scientific discoveries and important technological advances that span many important programs in the DOE. In the High Energy and Nuclear Physics (HENP) program, experiments associated with high-energy accelerators have led to important discoveries about elementary particles and the fundamental forces of nature, quark dynamics, and nuclear structure. In the Basic Energy Sciences (BES) program, experiments with synchrotron light sources and spallation neutron sources have been crucial to advances in the materials and biological sciences. In the Fusion Energy Sciences (FES) program, great strides have been made in developing heavy-ion particle accelerators as drivers for inertial fusion energy. Beyond impacts to basic and applied science, accelerators have also been proposed that address environmental, energy-related, and national security issues through projects such as the accelerator transmutation of waste, accelerator-driven fission and fusion energy production, and radiography for stockpile stewardship.

Computational accelerator science (CAS) is a key ingredient, along with theory and experimentation, needed to develop the next generation of particle accelerators and new accelerator technologies. Applications of CAS fall mainly into three areas: (1) simulation of beam dynamics in accelerators, (2) electromagnetic modeling of accelerator structures and components, and (3) the simulation of high gradient and high power devices involving a combination of intense particle beams, lasers, plasmas, and electromagnetic fields. The main equations governing these three areas are Hamilton's equations, Maxwell's equations, and the combined Vlasov/Maxwell equations. The 3D solution of these equations is computationally challenging because of the size of the problems and number of floating point operations required. For example, the simulations may require billions of particles and mesh points in order to achieve the accuracy needed for design purposes and to achieve convergence of the numerical predictions. Multi-physics simulations, involving additional phenomena such as particle collisions, ionization, multi-species effects, and radiation, place even greater demands on the computational platforms. These computational challenges must be met because these complex computer models have important consequences for cost, risk, and ultimately the success of major accelerator projects.

The scientific community is now faced with several key accelerator modeling problems whose solution will require the use of "ultrascale" computers. With a capability of 100 Tflop/s, it will be possible to perform high-accuracy simulation and computer-aided optimization of very large

electromagnetic structures for next-generation accelerators. The ability to design high-efficiency structures has significant cost ramifications for future accelerator projects. Also, it will be possible to perform end-to-end design and simulation of a heavy ion fusion (HIF) accelerator complex. The ability to accurately model HIF systems is crucial for the successful development of this technology, which could lead to U.S. energy independence. Using a 100 Tflop/s computer it will also be possible to perform end-to-end design and simulation of fourth-generation light sources, which are important for imaging and studying ultrafast physical processes. With a capability of 250 Tflop/s it will be possible to perform self-consistent simulations of high power klystrons and RF sources. Such sources are critical for the design of next-generation high energy accelerators, and furthermore have significant industrial applications. It will also be possible to perform a complete design of a plasma “afterburner,” a concept that may make it possible some day to reach the very highest energies in an extremely cost-effective manner. It will also be possible to model, including multiple physical phenomena, high intensity proton accumulator rings and electron-based recirculating systems, where the beams may travel for thousands of turns and where collective effects are important. With a capability of 1,000 Tflop/s it will be possible to model, over very long simulated time scales, the dynamics of intense beams and colliding beams in circular accelerators, and to design electron cooling systems. This will make it possible to expand the operational envelopes and luminosities of colliders, thereby maximizing the physics output of these most unique and complex devices for studying the fundamental forces of nature.

- 50 Tflop/s: High-accuracy simulation and computer-aided optimization of large electromagnetic structures for next-generation accelerators; end-to-end design and simulation of fourth-generation light sources.
- 250 Tflop/s: Self-consistent simulations of high power klystrons and RF sources; complete design of a plasma “afterburner”; model high intensity proton accumulator rings and electron-based recirculating systems.
- 1,000 Tflop/s: Model (over very long simulated time scales) the dynamics of intense beams and colliding beams in circular accelerators; design electron cooling systems.

2.1.8 Lattice Quantum Chromodynamics (R. Sugar, UCSB)

Quantum chromodynamics (QCD) is the component of the Standard Model of high energy physics that describes the strong interactions. Although the Standard Model has been enormously successful in describing a wealth of phenomena in high energy and nuclear physics, our knowledge of it is incomplete because it has proven to be very difficult to extract many of the predictions of QCD. Large-scale numerical simulations within the framework of lattice gauge theory are required to fully understand the physical phenomena encompassed by QCD and to make precise calculations of the theory’s predictions. Such simulations are necessary to solve the fundamental problems in high energy and nuclear physics that are at the heart of the Department of Energy’s large experimental efforts in these fields.

Major goals of the DOE’s experimental program in high energy and nuclear physics are (1) to verify the Standard Model or discover its limits, (2) to determine the properties of hadronic matter under extreme conditions, and (3) to understand the structure of nucleons and other hadrons. Lattice QCD calculations are essential to research in all of these areas.

Recent refinements of numerical algorithms, coupled with major increases in the capabilities of massively parallel computers, have enabled the accurate calculation of a few crucial quantities, such as the strong coupling constant and the masses of the b and c quarks. Furthermore, the experience gained to date allows confident predictions of the computing resources required for the accurate determination of a broad range of fundamental quantities. A number of these quantities will be obtained in the next generation of lattice QCD simulations, which will require computational facilities capable of sustaining tens of teraflop/s. These include matrix elements of the electroweak operators needed to determine some of the least well known parameters of the Standard Model, the masses of glueballs and particles with exotic quantum numbers, and the equation of state of strongly interacting matter at high temperatures. The U.S. lattice QCD community must carry out these simulations in the next few years if it is to provide timely support for the experimental programs in high energy and nuclear physics, and if it is to keep pace with theoretical physicists in Europe and Japan.

In the longer term, there are calculations of great importance which will require extended use of facilities capable of sustaining hundreds of teraflop/s or even petaflop/s. Examples include matrix elements required for a quantitative understanding of the violation of CP symmetry, the phase structure and equation of state of strongly interacting matter at finite baryon density, and a detailed understanding of the internal structure of nucleons.

Lattice gauge theorists have been using high performance computers in their research for over twenty years. They have extensive experience with a wide range of architectures, including array processors, commodity clusters, vector supercomputers, massively parallel computers, and machines specially designed for the study of QCD. They currently make extensive use of NERSC, the Center for Computational Science at ORNL, and other national supercomputer facilities. Lattice gauge theorists have usually been among the first to develop code for new architectures, and typically obtain some of the best performances on them. The U.S. lattice QCD community has a grant from the DOE's SciDAC program to develop the software infrastructure that it needs to carry out terascale simulations. For all these reasons, it is in an excellent position to participate in the hardware evaluation efforts proposed in this document.

Lattice gauge theorists at Columbia University have pioneered the construction of special purpose computers for the study of QCD. In doing so they have taken advantage of simplifying features of QCD calculations, such as regular grids, uniform, predictable communications between processors, and low memory requirements to design economical, high performance machines. The current generation of Columbia computers, the QCDSF, won the Gordon Bell prize for price/performance at SC98. The Columbia group has formed a partnership with IBM to design and construct its next special purpose computer, the QCD On a Chip (QCDOC). This machine, which targets a price/performance of \$1/sustained Mflop/s, plays a major role in the U.S. lattice QCD community's plans to build terascale computing facilities. The first multi-teraflop/s QCDOCs are planned for deployment in 2003. This computer is in the direct design line of IBM's Blue Gene/L series. The QCDOC proposed by the U.S. lattice gauge theory community for 2003 will be made available to the architecture development partnership proposed in this document for testing and benchmarking, enabling them to investigate this architecture at an early date, while giving the U.S. lattice gauge theorists an early start on terascale computing.

The most computationally intensive component of any lattice QCD calculation is the repeated inversion of large sparse matrices, whose dimensions are in the millions in current applications.

These inversions are typically done with the conjugate gradient, stabilized biconjugate gradient, or minimal residual algorithms. Codes are highly scalable. At present, individual production runs are being carried out on as many as 512 processors on commercial supercomputers, and on thousands of processors on the QCDSF. Naturally, performance depends on the computer and on the problem being studied. As an example, lattice QCD codes typically sustain 200 to 300 Mflop/s per processor in production runs on the NERSC IBM SP, Seaborg.

An executive committee consisting of R. Brower (Boston U.), N. Christ (Columbia U.), M. Creutz (BNL), P. Mackenzie (Fermilab), J. Negele (MIT), C. Rebbi (Boston U.), S. Sharpe (U. Washington), R. Sugar (UCSB, Chair), and W. Watson III (JLab) has been formed to plan the computational infrastructure needed by the U.S. lattice QCD community. (Planning documents can be found at www.lqcd.org). The chair of the committee expressed that they would be delighted to work with ANL and LBNL on the evaluation of future IBM Power and IBM Blue Gene/L families of parallel architectures.

2.1.9 Advanced Simulations of Plasma Microturbulence (W. Tang, PPPL)

This is a subject of great importance in DOE's Fusion Energy Science program. The key issue, which impacts the cost of a magnetic fusion reactor, involves optimizing the balance between the self-heating of the plasma from the fusion reactions and the heat leakage caused by electromagnetic turbulence. Simulations can provide the insights and guidance needed for the cost-effective design of facilities and the subsequent interpretation of experimental results addressing this complex problem. Fusion reactor design studies today rely on empirical extrapolations of turbulent transport properties from present-day, smaller-scale experiments to future larger devices. The scientific understanding needed to improve upon such estimates and provide a path forward to actually control the turbulence demands first-principles direct numerical simulations.

A good example of progress is the full utilization of the terascale computer at NERSC to produce three-dimensional particle-in-cell simulations of turbulence suppression. Valuable new insights into how plasma confinement might be significantly improved in a reactor-scale plasma were gained from more recent electrostatic simulations, which utilized over a billion particles and showed full scalability for 2,000 processors. This indication of a clear path forward to efficient future use of more powerful parallel computers is most encouraging, since more realistic physics models, which must include electromagnetic dynamics, will require about a factor of 50 to 100 increase in computing power. The Earth Simulator class of supercomputers can meet these petascale computational demands. Access to such computational resources will enable U.S. fusion researchers to revolutionize future advances by accelerating the pace towards greatly enhanced simulation and modeling capabilities. The U.S. fusion program would accordingly reassert its internationally recognized leadership in this area.

Direct numerical simulations utilizing advanced particle-in-cell methods involve explicit, finite-difference, unstructured mesh methodologies. The largest simulations to date have typically been carried out using 1 billion guiding centers, 125 million mesh points, and 7,000 time-steps, taking up 72 hours of CPU time of 1,024 processors on the IBM SP supercomputer at NERSC. On the scalar processors of the IBM SP, the relevant code (GTC) currently achieves about 10 to 15 percent efficiency with near-perfect scalability up to 2,000 processors to date.

There are programming and algorithmic challenges to be overcome to address these scientific problems. They include multi-dimensional domain decomposition in toroidal geometry and

mixed distributed/shared memory programming. Other problems that we will have to tackle in this collaboration are load balancing on computers with large numbers of processors, optimization of fundamental gather-scatter operation in particle-in-cell codes, and scalable parallel I/O for petascale range of data sets.

Science goals for 50, 250, and 1,000 Tflop/s systems proposed in the future include:

- 50 Tflop/s: Electromagnetic physics description of a thermonuclear plasma with 3D global turbulence simulations including kinetic electron dynamics.
- 250 Tflop/s: Spatial extension of 3D global turbulence simulations to be able to treat key fine-scale dynamics, such as the formation of internal transport barriers, in addition to large-scale nonlocal physical processes.
- 1,000 Tflop/s: Temporal extension of 3D global turbulence simulations to the longer time-scales of actual plasma discharges, including the evolution of density, temperature, and current profiles.

2.1.10 Quantum Monte Carlo Calculations of Nuclei (S. Pieper and R. Wiringa, ANL)

A major goal of nuclear physics is to understand the stability, structure, and reactions of nuclei as a consequence of the interactions between individual nucleons. This goal can be achieved by carrying out calculations of light nuclei (up to ten nucleons) using realistic two- and three-nucleon potentials. One approach is to use Green's function Monte Carlo (GFMC) calculations, which are one type of quantum Monte Carlo (QMC) methods. These calculations are severely dependent on the growth of computer power; each additional nucleon requires almost an order of magnitude increase in computer resources. Besides demonstrating the ability of modern nuclear theory to model the properties of light nuclei, these calculations provide predictions for astrophysically interesting reactions that cannot be measured in the laboratory and for neutron-rich systems that can be used to model heavier nuclei that will be produced at the Rare-Isotope Accelerator.

At present these calculations are computationally limited to $A = 10$ (we use about 1 percent of the NERSC Seaborg computer per year). With substantially more computing power it would be possible to progress to $A = 12$ systems, in particular ^{12}C , which is of great interest because of the many experiments that use it as a target and because of its importance in nuclear astrophysics. Examples are the recent nuclear transparency experiment at Jefferson Lab, longitudinal and transverse spin response measurements, and understanding the triple-alpha burning reaction. In addition, these computers would enable calculations of neutron-rich matter such as is found in neutron stars.

The time-consuming parts of these codes can be logically expressed as repeated, extremely sparse complex matrix operations on a large vector. The sparse matrix has a structure based on 4×4 and 8×8 blocks, except that the blocks are not necessarily contiguous. The elements of these blocks are not all independent. Therefore specialized routines that apply one block to the vector have been written, making use of the relations between the elements of the block. The programs have been well tuned, both to the RISC architecture of the individual nodes and to achieving high scalability on parallel processors. As an example of this, a 6.7-hour, 128-node run (2,048 MPI processes) on NERSC's Seaborg IBM SP achieved a sustained speed of 0.55 Tflop/s. The per-processor speed was 298 Mflop/s, and the parallelization efficiency was 92.1 percent.

This was a calculation of a negative-parity state in ^{10}Be that involves long vectors; simpler cases can run at higher per-processor speeds, but it would not be reasonable to use so many processors for such quick calculations.

The $A = 10$ calculations allow us to reliably estimate the needs for ^{12}C : one calculation requires 325×10^{15} floating point operations, or 20 hours of a 50 Tflop/s (actually delivered) machine. Each processor would need 2 GB of memory. The results above show that a few thousand processors could be used with good efficiency; however, using 10,000 might be a problem. Each state of ^{12}C is a separate calculation, and one would want to use several different interactions, so several weeks of a 50 Tflop/s machine would be needed for a complete study. With current computers we can simulate neutron matter by doing GFMC for 14 neutrons in a box with periodic boundary conditions. A calculation with 14 neutrons and one proton in the box would simulate the matter actually found in parts of neutron stars; such a calculation would be about four times as hard as the ^{12}C case and thus require about a 200 Tflop/s computer. Once ^{12}C is in hand, it would be extremely interesting to compute the triple-alpha burning reaction rate, knowledge of which is needed for understanding nucleosynthesis in massive stars. This involves computing the overlap of a $^4\text{He} + ^8\text{Be}$ scattering state and ^{12}C . A single calculation would be an order of magnitude bigger than just the ^{12}C calculation and hence would require more like a 500 Tflop/s computer.

Science goals for future systems include:

- 50 Tflop/s: GFMC simulation of ^{12}C .
- 250 Tflop/s: GFMC simulation of 14 neutrons and one proton in the box (simulates the matter found in parts of neutron stars).
- 1, 000 Tflop/s: Computing the overlap of a $^4\text{He} + ^8\text{Be}$ scattering state and ^{12}C to understand nucleosynthesis in massive stars.

2.1.11 High Energy / Elementary Particle Physics (L. Price, ANL)

High energy physics (elementary particle physics) is the study of the basic constituents of matter and the forces and interactions between them. The study is carried out using all available tools. Although naturally occurring particles from cosmic rays and radioactive decay of nuclei are used for appropriate studies, the need for higher energies and intensities of particles to use in collisions has driven the development of increasingly powerful and clever accelerator systems as a major thrust in enabling new generations of experiments. Experiments to exploit more powerful accelerators and to answer more basic questions about particle physics have become extremely complex, consisting of millions of channels of detector readout, generating very large quantities of data to be sifted and analyzed. Results from experiments are systematized and explained through development of theories that provide a conceptual framework and mathematical structure for describing current data and predicting the results of measurements not yet made.

This cycle of advancing the science of high energy physics relies on powerful computing in at least three major ways. Design of accelerators relies on the well-known science of electromagnetism and is therefore well suited to the use of detailed simulation to design future accelerators. Experimental high energy physics requires computing at several steps in the data chain from acquisition to reconstruction to repeated analysis of the huge amounts of data created, in addition to requiring simulation and analysis of large quantities of what-if data to compare

with actual experiments. Finally, some apparently correct theoretical descriptions of nature are cast in a form that cannot be readily made to provide numbers to compare with experiments in analytic fashion. In those cases, computational techniques permit numbers to be obtained through lengthy calculations. The current prime example of the need for extensive computation to extract theoretical results is the use of mesh or lattice techniques to calculate quantum chromodynamics and other non-Abelian gauge theories of elementary particles.

Although all three of these applications require and can benefit strongly from a large increase in available computational power, we will focus on the benefits to experimental high energy physics after the data is initially recorded for analysis. The HEP analysis problem has become the paradigm of data-intensive science, especially as access to analysis tools, including computation, is needed by large numbers of distributed users for a single large dataset. The analysis task, as constrained by present computing systems, is a complex set of tasks that is typically carried out in different places at different times. The world's largest database was created to meet the current needs of the BaBar experiment at SLAC, and the future needs of the Large Hadron Collider experiments are the major drivers of key U.S. Grid projects. Ideally, the analysis task for experimental data starts with a bright idea by a researcher for a way of searching the very large quantity of experimental data from an experiment for a pattern that will reveal a new particle or process. The complexity of the system and resource limitations mean that analysis stages are presently organized to streamline the search paths that have been identified and prioritized in advance. Those streamlined analyses can typically be carried out in hours, already not really interactive. New ideas that require changes to upstream parts of the analysis chain can incur delays of months before they are run and results evaluated for the next stage of optimization. Many good ideas are not pursued because of resource limitations.

HEP experimental analysis, with limitations noted above, is now carried out in dedicated systems of order 1 Tflop/s per major experiment (of which about five exist worldwide) and with disk storage systems of order 200 TB. These numbers are growing steadily with time. HEP will benefit substantially from a coordinated system providing computing power, data storage, and network communications increase by an order of magnitude in the next one to two years and by two orders of magnitude within three to four years.

The availability of substantially more powerful computing, storage, and communication systems will not by itself enable the radically new ways to carry out analysis that are envisioned to provide much stronger and shorter connections between new ideas and viewing results. The new computational systems will need to be matched by software and middleware tools that integrate the multiple steps now required and mask the complexity of database systems, storage locations, etc. Much expertise exists within the HEP and computer science communities that is currently devoted to developing computation and data Grid tools to solve some of the problems described here. This effort will need to be expanded and refocused on exploiting the step in computational power and enabling the new approaches that will be possible.

As the total computational power of a new shared system becomes available, we can identify stages in utilization. With an initial (early) deployment of a total 50 Tflop/s system, one HEP experiment could make the transition to use a portion of the system as its main interactive analysis engine, developing the new paradigm for integrated analysis with substantially increased resources. This would allow substantial advances in the science represented by that experiment. If this were applied to the BaBar experiment, for example, pursuit of more bright ideas could result in additional understanding of the origins of matter-antimatter asymmetry and therefore of

the evolution of the early universe. By the time 250 Tflop/s is available, the experience of the initial collaboration using the system can be extended to perhaps two others in an appropriate stage of development. The 1,000 Tflop/s system might be properly timed to enable the experiments now under construction for the Large Hadron Collider, strongly supported by both DOE and NSF, to start their physics analysis with these important new capabilities. The result will be rapid progress and the ability to explore the more difficult areas that might otherwise be neglected in understanding the major problems of understanding the fundamental nature of matter: the origin of electroweak symmetry breaking, the underlying symmetries of the universe, and the question of the existence of new forces between particles.

- 50 Tflop/s: Data analysis of BaBar experiment to understand the origins of matter-antimatter asymmetry and therefore of the evolution of the early universe.
- 250 Tflop/s: Extension of integrated data analysis system to two other experiments.
- 1,000 Tflop/s: Physics analysis of experiments from Large Hadron Collider, e.g., the origin of electroweak symmetry breaking, the underlying symmetries of the universe, and the question of the existence of new forces between particles.

2.2 A New Architecture Strategy: Beyond Evaluation to Cooperative Development

In order to reach our overall goal of developing platforms that deliver petaflop/s peak capability and allow sustained performance of 300–400 Tflop/s for scientific applications by the end of the decade, we must aim at radically changing the relationship of hardware vendors and computational scientists. The historical relationship of vendors and computational scientists has been one of evaluation and feedback. While there has been a tremendous change in the hardware platforms over the last two decades—from parallel-vector machines in the early 1980s, to microprocessor-based massively parallel platforms in the late 1980s to mid-1990s, to commodity server-based clusters of SMPs in the late 1990s—there has been no fundamental change in the way these platforms have been designed and built. The main criteria were technological, architectural, and economic feasibility. Applications performance did not enter as a primary design criterion. Machines were designed by computer architects, built by commercial vendors, and the role of applications was to evaluate different platforms in comparison studies and benchmarking *after* several competing platforms had been built. Results of the evaluations were reported back to the vendor. However, processor and architecture design cycles are three to six years. Therefore, in most cases any applications feedback could not influence the next two generations of processor and architecture design, and was largely ignored.

Staying on this path of largely excluding scientific applications from the design considerations for high performance computing (HPC) platforms has led to the current situation in the U.S.: the current HPC platforms developed by U.S. vendors are largely targeted towards commercial applications and are not very well suited to meet the demands of scientific computing applications. This situation leads to the very low sustained performance rate of only 5 to 10 percent of peak that is typical for most applications on today's cluster of SMP architectures developed by U.S. vendors.

We are proposing to change this situation by fundamentally altering the role of applications scientists in the processor and architecture development process. Applications scientists will

become integral members of the development teams. In collaboration with computer scientists, the applications requirements will be translated into clearly quantifiable measurements of architectural parameters. In ongoing development meetings, the hardware designers will use these requirements to explore design trade-offs and will propose prototype architectures back to the applications teams. Iterating this process will lead to defining prototype designs. Vendors will then deliver small-scale early prototypes for actual applications testing. Further iterations and refinements will then lead to the final, large-scale production machine.

***An Early Indication of the Feasibility of This Approach:
The IBM/ANL/LBNL Workshop of September 3–4, 2002***

This new approach was tried in a workshop held in early September 2002 at Berkeley Lab, involving applications and computer scientists from the Argonne and Berkeley laboratories as well as the Power 5/6 and Federation switch design team from IBM. The requirements from four applications areas—astrophysics, climate, nanoscience, and combustion (adaptive mesh refinement)—were analyzed and discussed. Derived from this analysis was Table 1, which lists common desirable architectural features across different applications.

**Table 1
Application Requirements**

	AMR	Coupled Climate	Astrophysics		Nanoscience	
			MADCAP	Cactus	FLAPW	LSMS
Sensitive to global bisection	X	X	X		X	
Sensitive to processor-to-memory latency	X	X				
Sensitive to network latency	X	X	X	X	X	
Sensitive to point-to-point communication	X	X				X
Sensitive to OS interference in frequent barriers				X	X	
Benefits from deep CPU pipelining	X	X	X	X	X	X
Benefits from large SMP nodes	X					

In further analysis, the qualitative dependencies in Table 1 were reduced to the following set of parameters, which then guided the discussion of potential modifications to the Power X/ Federation architecture:

- Ratio memory bandwidth/Flop/s

- Ratio interconnect bandwidth/Flop/s
- Latency: memory and network

The resulting ViVA architecture is described in detail in section 3.2.2. While only a first step, the workshop demonstrated that cooperative development is possible and may lead to highly productive new improvements in computer architecture for scientific applications.

2.3 Key Machine Parameters

The consistent, significant discrepancy between the theoretical peak performance of a computer system and the actual sustained application performance seen by scientific and technical codes mandates better metrics to provide a better indication of the performance future systems may deliver. Our team has been examining the proposed architectures as well as existing systems using a set of composite metrics that take into account the fundamental importance of memory bandwidth relative to processor performance. This type of evaluation recognizes the basic fact that the achieved performance rate relative to the theoretical peak is actually a measure of how much time the processor worked on the application, and how much time it waited for data to finish loading from or storing to memory. Thus ratios of data movement rates and data processing rates give a much more realistic sense for how productive a given computer system will be, and how much of its theoretical peak speed can be expected for typical applications. We have begun to analyze these ratios both within a node and in an overall system between nodes (where a node is defined as the smallest grouping of processors with shared memory).

As with any complex system, there are many factors that influence even the most carefully chosen metrics. For example, inter-node performance is also affected by the node interconnect network topology, link bandwidth, and overall bisection bandwidth (a measure of aggregate capacity of the network, with implications regarding bottlenecks that might slow traffic between nodes). Finally, any interconnection architecture (within or between nodes) will have inflection points whereby performance will vary non-linearly as the size of the system grows.

As useful as these metrics are in comparing architectures from the application perspective, critical decisions regarding national-scale investments require a more comprehensive comparison—one that takes into account cost. Our strategy is based on a simultaneous focus on both performance and cost. Our preliminary analysis indicates that the architectures we have proposed to pursue have the potential to not only deliver substantially more capability than is available anywhere today, but to do so at substantially better price/performance.

3. STRATEGY TO LEVERAGE MAJOR U.S. COMPUTING CAPABILITY

In the next two subsections of this preproposal we describe two distinct but related architecture development activities we propose to conduct jointly with IBM. Our plan is to build one integrated team that spans ANL/LBNL, our science applications partners, and university-based computer science researchers, and that will involve other DOE laboratories such as ORNL, PNNL, and BNL.

The first activity described below is aimed at accelerating and enhancing the development of the IBM Blue Gene family of computers currently under development in IBM research and to deploy one or more such systems into production use for the open scientific community. ANL will lead

the Blue Gene effort, with LBNL providing a more supporting role. The Blue Gene path is important because it opens up a new class of machines to potential production status and because it will firmly explore the feasibility of exploiting the scalability of systems-on-a-chip technology, one of the key technologies for large-scale systems development in the 21st century. If this path is successful, we believe it may offer substantial performance capability to a range of large-scale applications at a price/performance ratio unachievable by any other technology in the pipeline.

The second activity is focused on enhancing and extending the IBM Power architecture and SP line of parallel supercomputers (the current largest scale systems in production use by DOE at NERSC). LBNL will lead the effort to enhance Power, with ANL providing a more supporting role. The enhanced Power path is significant because it represents the single largest installed base of high-performance computing in the U.S. and billions of dollars of software investment on applications and systems software. This path will aggressively explore the option space of improving the effectiveness of Power and the SP on scientific applications, while preserving the software investments of the last decade. If this path is successful, we believe it will enable the scientific computing community to continue focusing on production science for the next decade.

Both of these paths require close collaboration of computer scientists, computer designers, and applications and algorithms developers. Our plan is to form one integrated team of technical experts that will work on both architecture paths to provide for rapid information flow and to enable a comprehensive assessment of the progress and viability of each path. This means that we will not initially partition the applications space by architecture path until we have learned for certain that one path or the other is more appropriate.

3.1 Blue Gene/L

IBM announced a significant new effort in December 1999 to develop a computer system capable of sustaining near petaflop/s on molecular modeling problems critical to life science research. This project, called Blue Gene, has evolved in the last three years to include a family of more general-purpose computer systems being developed in IBM research in collaboration with DOE laboratories and university groups.

The Blue Gene family (BG/X) of computers is a departure from existing IBM large-scale systems in several ways. First, it is aimed at extreme scalability, with the first system in the series called BG/L targeting 65,536 nodes at full scale and future systems envisioning even larger scale. Second, the BG/X series is based on extreme levels of system integration leveraging systems-on-a-chip (SoC) technology, which enables the collocation of CPUs, memory, and network interfaces to be integrated on a single die. SoC technology enables high performance and low power it also reduces the number of discrete component types in the system, lowering manufacturing cost and improving overall reliability. Finally, the system uses a new software strategy that has minimal operating system functionality at each compute node, with most operating systems functions moved to a separate set of processors specially configured to provide OS services and I/O services. Each of these new directions contributes to put the BG/X family on the path to sustained petaflop/s capabilities for much lower cost than present products. Current indications are that the general approach represented by BG/X has the potential to achieve sustained petaflop/s performance on some applications by end of 2007 for about \$200 M. Sustained price/performance for applications on BG/L is likely to be one-half to one-fourth that of other approaches.

The first system (i.e., BG/L) to be developed in this family has a planned peak performance of 180 teraflop/s and is scheduled to have first availability in late 2004. The five year roadmap for the Blue Gene project has two follow-on machines to BG/L under consideration. BG/P with a target peak performance range of 1000 Tflop/s (300 Tflop/s sustained) and provisional availability in late 2006/ early 2007 and BG/Q with a target peak performance range of 3000 Tflops (1000 Tflops/s sustained) and availability targeted for late 2007/early 2008. There is also a near term variant of the BG/L system planned called BG/D which incorporates denser memory packaging to permit configurations with more external memory per node. ANL has established a preliminary plan with IBM to co-develop and deploy a full-scale BG/L system at ANL in late 2004/early 2005 to the DOE Office of Science community, potentially followed by a BG/P system in 2006/2007 deployed at LBNL for the production user community followed by a BG/Q system deployed at ANL in late 2007/2008.

The Blue Gene project is leveraging experience gained with the DOE-supported joint project with IBM, Columbia University and BNL to develop and deploy the 10 Tflops/s QCDOC machine. Blue Gene differs from the QCDOC machine in several important ways. BG/L uses a “double” floating-point unit, capable of two fused multiply-adds per cycle compared to QCDOC’s single floating-point unit. The BG/L network is faster and more sophisticated as it supports point-to-point routing (via a 3D torus) where as the QCDOC network is strictly nearest neighbor. The BG/L also has dramatically improved internal memory bandwidth (22 GB/s vs. 8 GB/s) from the embedded DRAM to the CPU. And it has faster network connections and implements a more advanced communications stack. The QCDOC project and ANL have agreed to work together to directly leverage the relevant experience gain that that project.

While the original application focus of the Blue Gene project was to address protein folding problems, the current design point is truly general purpose with industry standard PowerPC instruction sets, standard product-level compilers, support for MPI-based message passing and remote memory access using put/get operations, high-bandwidth access to on-chip memory, and high interconnect bisection bandwidth (Table 2 above). All of these features make the machine architecture much more in line with the directions recently taken by commodity clusters. Indeed, we believe that the BG/X series of machines can be viewed as one potential evolutionary successor to such clusters, enabling in many cases well written application software developed for commodity clusters to be ported with few changes.

Each of BG/L’s 64K compute nodes have a relatively slow clock rate (700 MHz), which contributes both to lower cost and low power consumption. This low power consumption enables the systems to be air cooled and permits dense packaging, which together enable 1,024 computing nodes per rack. The current design point for BG/L requires about ~1 MW of power for 180 Tflop/s peak systems, ~300 tons of cooling, and less than 4,000 sq ft of floor space. These numbers are in many cases an order of magnitude better than competing proposals.

Since all inter-node networking functions are integrated onto the same chip as the processor and embedded DRAM, there is no need for a separate switch and switching fabric, which also contributes to improving reliability and lower cost. The BG/L node has two identical processors. In normal operation, one processor is fully dedicated to message passing, and the other is reserved for the computational work. Under software control, both can be used for processing, effectively doubling the amount of available computational power. For our estimates, we assume the standard mode of processing and only count the performance of one processor in computing peak performance numbers.

3.1.1 Blue Gene Architecture Overview

BG/L is a scalable system with the maximum size of 65,536 compute nodes; the system is configured as three-dimensional torus ($64 \times 32 \times 32$). Each node is implemented on a single CMOS chip (which contains two processors, caches, 4 MB of EDRAM, and multiple network interfaces) plus external memory. This chip is implemented with IBM's system-on-a-chip technology and exploits a number of pre-existing core intellectual properties (IPs). The node chip is small compared to today's commercial microprocessors, with an estimated 11.1 mm square die size when implemented in 0.13 micron technology. The maximum external memory supported is 2 GB per node; however the current design point for BG/L specifies 256 MB of DDR-SDRAM per node. There is the possibility of altering this specification, and that is one of the design alternatives we propose to evaluate during our assessment of the design for DOE SC applications.

The systems-level design puts two nodes per card, 16 cards per node board, and eight node boards per 512-node midplane (Figure 1). Two midplanes fit into a rack. Each processor is capable of four floating point operations per cycle (i.e., two multiply-adds per cycle). This yields 2.8 Tflop/s peak performance per rack. A complete BG/L system would be 64 racks.

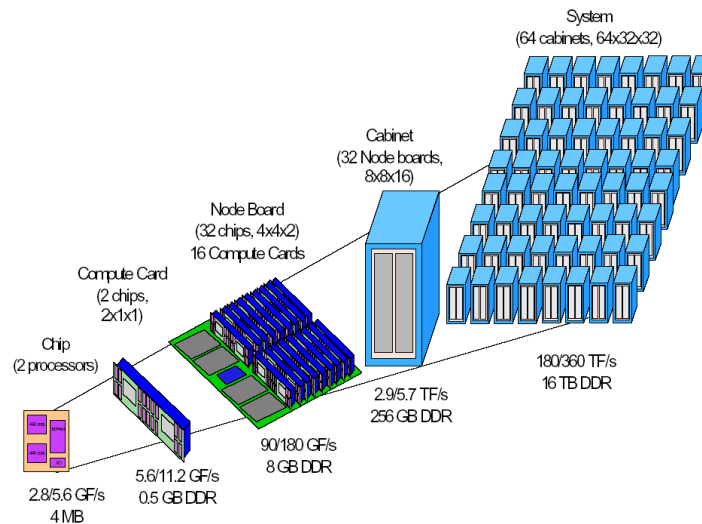


Figure 1. Blue Gene system design.

In addition to each group of 64 compute nodes, there is a dual processor I/O node which performs external I/O and higher-level operating system functions. There is also a host system, separate from the core BG/L complex, which performs certain supervisory, control, monitoring, and maintenance functions. The I/O nodes are configured with extra memory and additional external I/O interfaces. Our current plan is to run a full Linux environment on these I/O nodes to fully exploit open-source operating systems software and related scalable systems software. The more specialized operating system on the compute nodes provides basic services with very low system overhead, and depends on the I/O nodes for external communications and advanced OS functions. Process management and scheduling, for example, would be cooperative tasks involving both the compute nodes and OS services nodes. Development of these OS services can precede the availability of the BG/L hardware by utilizing an off-the-shelf 1,000-node Linux cluster, with each node having a separate set of processes emulating the demands of compute nodes. ANL has previously proposed to DOE the need for a large-scale Linux cluster for scalable

systems software develop and that system would be very suitable to support the development of these OS services as well.

BG/L nodes are interconnected via five networks: a 3D torus network for high-performance point-to-point message passing among the compute nodes, a global broadcast/combine network for collective operations, fast network for global barrier and interrupts, and two Gigabit Ethernets, one for machine control and diagnostics and the other for host and fileserver connections

3.1.2 Collaborative Development with IBM

We have developed with IBM a comprehensive strategy for collaborative development that addresses the critical needs of moving the Blue Gene project from a research project to a project that could provide high-level production support to critical applications. We outline that plan here.

Simulation software framework. A set architecturally accurate simulators have been developed for the BG/L project to enable porting and tuning of applications prior to the availability of hardware. We plan to deploy these simulators and make them available to our application partners and work with them to have BG/L ready codes and libraries well in advance of the systems. We also plan to work closely with IBM to extend these simulators to incorporate changes to the architecture over time, and specifically to track the evolution of the architecture for BG/P and BG/Q. The simulations run on Linux clusters, and we propose to deploy a series of large-scale Linux systems to support the use of these simulators by the community. These large-scale Linux systems will also enable the development of OS services and I/O frameworks for the BG/L system in advance of hardware.

Systems software development. Argonne has already been working closely with IBM on the development of two critical items of systems software for BG/L, the message passing system based on MPICH/ADI-3 and process management. We plan to put in place a much more comprehensive systems software development effort that will target the OS functions needed to support BG/L codes that today run on Linux clusters. To do this we will need to develop mechanisms to offload OS functions from the compute node to the OS services nodes. These developments will be pursued in conjunction with the broader community and will leverage the SciDAC systems software projects and existing open-source activities. The Argonne/Berkeley collaboration will also provide the community with open access to development testbeds already being deployed at ANL for this purpose.

Applications development. There are several major challenges for effective use of BG/X for scientific applications. Applications have to parallelize to a degree substantially beyond the current levels of control concurrency. Most current high-end applications do not scale much beyond 1,000 nodes. To effectively use BG/L, applications will need to scale to 64K nodes. Thus significant work is needed to improve applications scalability. Applications need to be optimized for the multi-level memory hierarchy of BG/L. BG/L has very high internal memory bandwidth, ~11GB/s between the on-chip RAM and the register set of each of the two CPUs (22 GB/s aggregate). Given the rather modest 2.8 Gflop/s peak performance of the CPU, this is almost 4 bytes per flop, which is the among the highest relative burst memory bandwidth of any existing system (Table 2 above). However that chip memory is relatively small at 4 MB. Off chip memory bandwidth is still relatively high at 5.5 GB/s, but to achieve high sustained node performance, applications will have to effectively manage this memory. Finally, applications will

have to be modified to fit in the relatively small memory footprint of the node (256 MB) and to effectively exploit the distributed OS services model. One promising path to addressing the applications needs on BG/L is to build highly tuned numerical libraries such as PETSc (which is used by Jardin's fusion code) that have been tuned to BG/L that address the memory size and memory hierarchy issues. This approach can shorten the time to get many applications up on BG/L and provide a faster path to BG/P and BG/Q. For many of the application projects we have selected for initial development, we believe both of these challenges can be met, but it will be require sustained cooperation between IBM, our laboratory based computer scientists, algorithm developers, university partners, and applications scientists with access to significant scalability testbeds, simulators, and early hardware platforms.

Performance analysis and tuning. As mentioned earlier, determining and enhancing the expected performance of real applications on BG/L and follow-on platforms will be a major activity of this project. We intend to leverage the SciDAC Performance Evaluation Research Center and other projects to develop predictive performance models of our selected applications and to use these models to influence design decisions for both software and future hardware platforms. By focusing on a multi-year development trajectory rather than a single platform, we can feed forward lessons and insights from previous generations into the next systems. Our strategy depends on being able to use BG/L as a testbed for the development of applications and insight for BG/P, and BG/L and BG/P for BG/Q. We expect there will be considerable effort required to both tune applications on the real hardware to achieve maximal performance and to debug the process and methods used in simulation and performance estimation. A critical element of the strategy is to move towards a future architecture design methodology that is more quantitatively informed by real results from real applications rather than rely on general rules of thumb, as historically been the case. This approach requires the sustained participation of computer scientists, applications scientists and computer designers over a period of many years.

Projections forward into BG/P and BG/Q. We believe that the proposed BG/X approach is one of the most viable for achieving cost-effective, sustained petaflop/s by the end of the decade. A critical factor in the approach is that we have a multi-year commitment to deploy a series of machines in the same family that permit us to leverage investments made in applications, systems software, analysis and simulation tools, and development methods. If our work with BG/L proves successful, we would anticipate deploying a BG/P class machine (peak petaflop/s with a goal of sustaining several hundred teraflop/s) in the production computing infrastructure of DOE Office of Science, while working towards the development of a machine capable of sustained petaflop/s.

3.1.3 Milestones

- 4QCY02 – Workshop with IBM and ANL/LBNL team
- 1QCY03 – Establish initial BG/L modeling and simulation capability
- 2QCY03 – Install architecture simulation platform at ANL
- 3QCY03 – Develop performance models of BG/L for applications
- 4QCY03 – Install 512 node BG test system (~1.4 TF) at ANL
- 1QCY04 – Complete initial applications performance validation

- 2QCY04 – Complete BG/P design requirements
- 3QCY04 – Install additional (512 node) BG test systems at LBNL and NCSA
- 4QCY04 – Upgrade architecture simulation and development platform at ANL
- 1QCY05 – Install 64K node ANL BG/L system (180 TF peak, 60 TF sustained)
- 2QCY05 – Establish BG/P simulation and modeling capability
- 3QCY05 – ANL BG/L system in experimental production
- 4QCY05 – Install 1024 node BG/P test systems (10 TF) at ANL and LBNL
- 1QCY06 – BG/Q design requirements completed
- 2QCY06 – Upgrade architecture simulation and development platform at ANL
- 3QCY06 – Establish BG/Q simulation and modeling capability
- 4QCY06 – Install 128K node LBNL BG/P system (1000 TF peak, 300 TF sustained)
- 1QCY07 – Initial BG/Q software environment completed
- 3QCY07 – LBNL BG/P systems in production

3.2 Blue Planet: Extending IBM Power Technology and Virtual Vector Processing

Currently the most dominant architecture for high end computing is the IBM Power series, which by itself accounts for 55% of the Top20 computers in the world. IBM has a large and diverse installation base for the Power line, giving it a tremendous advantage in funding research and development of new functionality. However, this diversity contributes to the problem that the Power architecture is not specifically tuned to the needs of the scientific market. A key component of this proposal is to enhance the Power architecture for scientific computing while in parallel implementing a system by the end of 2005 that has at least twice the sustained performance of the Earth Simulator at roughly half the hardware cost! The details of this approach are:

1. Blue Planet System: Work with IBM to enhance IBM's current Power 5 plan in order to deploy a system with approximately 150 Tflop/s peak performance that will be able to sustain 40 to 50 Tflop/s sustained performance on at least several real scientific codes. Each of the 2,048 nodes will consist of eight "single core" CPUs that provide double the memory bandwidth of the standard Power 5 CPUs and will have their own dedicated L1, L2, and L3 caches. The CPUs will have a peak performance of roughly 8–10 Gflop/s.

The Blue Planet system will have 16,384 CPUs, the maximum main memory bandwidth possible, 8,192 switch links, and 2.5 petabytes (PB) of shared, parallel storage in FY 2006. This system will have more memory bandwidth, more interconnections, and lower interconnect latency than IBM had previously planned.

2. Work with IBM to develop a new capability of Virtual Vector Architecture (ViVA) which harnesses the eight individual Power 5 CPUs in a node into a single 60–80 Gflop/s vector

unit. This is the equivalent of what the Cray X1 does using four individual single-streaming processors (SSPs) within one multi-streaming processor (MSP). ViVA will be implemented on the Blue Planet System and has the potential to further improve the performance of codes.

3. Building on experiences with Blue Planet and ViVA, cooperatively work with IBM to further enhance the Power 6/7 and other future processor designs. Teams of DOE computational and computer specialists will work with IBM processor designers toward the goal of further improving memory and interconnect bandwidth in the Power series. The goal will be to field a petaflop/s (peak) computer capable of 20–25% sustained rates on diverse application by 2009.

The details of the three-pronged parallel approach are described in the following three subsections.

3.2.1 Blue Planet System Architecture

Blue Planet will take the IBM technology due for implementation in the second half of CY 2005 and expand its capability in multiple ways as an immediate and highly reliable enhancement to the scientific computational power of the DOE science community. Specifically, the system will be built on Power 5+ CPUs, which will run at approximately 2+ GHz, and the Federation switch. At that rate, each CPU is theoretically capable of 8–10 Gflop/s. However, using new, previously unplanned functionality as well as special packaging will allow the system to achieve a much higher sustained percent of peak on true scientific codes from multiple disciplines. The new features are:

- A new packaging of eight “single core” modules per node so that each CPU has its own dedicated L1, L2, and L3 caches (Figure 2). This configuration provides 60–80 Gflop/s nodes by using the single core chips. Each node will have twice as many memory buses—one GX bus per CPU—as IBM’s standard offering for 8-way nodes, which will enhance both memory and interconnect performance. Unlike the currently planned Power 4 based eight-way dual-core CPU nodes, these CPUs will run at the maximum clock speed achievable. The nodes will have twice the memory bandwidth to main memory and a three-tier cache system.

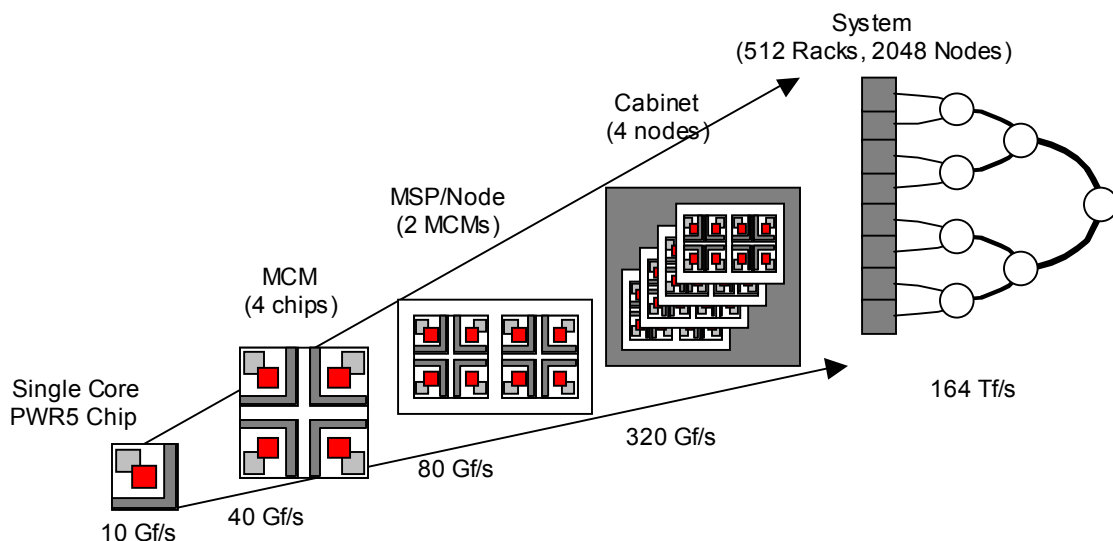


Figure 2. Blue Planet architecture.

- Each Federation network will be expanded by a factor of 4 from 1,024 links (512 nodes) to 4,096 links (2,048 nodes). This requires adding a third stage to the switch and improving the entire software to scale to at least 2,048 nodes. The system will have two IBM Federation networks, so it will have a total of 8,192 switch links, yielding considerable improvement in network bandwidth.
- At the same time the Federation switch scales by a factor of 4, IBM will combine improvements in the hardware and software to decrease MPI latency. The scaling and latency improvements of the switch define Federation+, a midlife improvement to IBM's switch technology. IBM has made midlife improvements to processors but never before to switch technology.
- Operating system, compiler, and library technology will take full advantage of the increased scale and performance of the Blue Planet system.

The resulting system will be able to sustain 40 to 50 Tflop/s on applications from several disciplines. The Blue Planet system will have 16,384 CPUs, each with 8–10 Gflop/s; 2,048 eight-processor SMP nodes; 8,192 switch links; and 2.5 PB of shared parallel disk storage. The amount of memory on the system will be configured to assure maximum main memory bandwidth. Currently memory technology trends indicate this has to be 256 TB of memory to attain this maximum bandwidth configuration—a capacity that is very expensive and not required by the applications (many of the applications scale in memory use by N^2 while they scale in computation by N^3). If memory bandwidth can be maintained at 64 TB or 128 TB of memory, the system will be delivered with the most cost-effective amount.

Blue Planet will have more memory bandwidth, more interconnections, and lower interconnect latency than was previously planned by IBM. It will be delivered in FY 2006, possibly in two phases. The need for phasing may come from component availability, since all aspects of the system will be new and production lines will be ramping up for the expected demand. The first phase will be delivered in the second half of CY 05, with the second phase following within nine months.

Blue Planet requires approximately 6 MW of power for the computer and peripherals and requires 1,700 to 2,000 tons of cooling. The entire system will fit within 12,000 square feet of computer space and can be housed in the existing LBNL Oakland Scientific Facility without constructing additional space.

3.2.2 The Virtual Vector Architecture

The basic intent of the Virtual Vector Architecture (ViVA) facility is to allow customers/applications to run high performance parallel/vector style code on a traditional high data bandwidth SMP. What is described below is currently an unsupported function within IBM, involving compilers, operating systems, Hypervisors, firmware, processor/systems, and productization. On the positive side, the Power 5 processor/system design does have basic functionality to support ViVA.

ViVA would further enhance the Blue Planet system and take IBM in a new direction. Power 5 chips have the ability to synchronize the CPUs using a hardware communication link for barrier synchronization. This hardware feature is currently not planned for exploitation because there is not an identified requirement within the existing markets. However, the synchronization feature can be used to harness individual CPUs into a “virtual vector” unit. This is the same concept

implemented in the Cray X1² MSP CPU, which has four separate 3.2 Gflop/s SSPs that synchronize for vector processing.

The basic intent of ViVA is to allow customers/applications to run high performance parallel/vector style code on a traditional high data bandwidth SMP. Initially, ViVA's goal is not to improve memory bandwidth per se. It does greatly enhance the ability of compilers (and programmers) to exploit fine-grained parallelism automatically, as is done with compilers that run on existing vector systems. The result should be to increase the proportion of applications that have higher sustained performance, thereby making the system much more cost effective for a wider range of scientific applications.

ViVA will be implemented on the Blue Planet system through software that uses the Power 5 architectural features. It will be evaluated and available to the applications that benefit from it. If the evaluation and use of ViVA shows benefit, not only will it enhance application performance of Blue Planet beyond what is described above, but it is conceivable that further vector-like support will be possible in future generations of the IBM Power architecture. If the ViVA experiment is not as successful, Blue Planet will perform with no lower performance than that described above.

3.2.3 Designs with Future Enhancements for Scientific Computing

The third part of the effort is the longest term. Cooperative IBM-DOE design of future Power CPUs will be initiated. The design cycle of a complex chip like the Power 5 takes five to six years. Year one is typically the high-level design and feature selection, followed by two or more years of implementation and then two or more years of testing, prototyping, and assembly design.

Building on the experiences with the Blue Planet and ViVA experiment, DOE Lab staff will cooperatively work with IBM to further enhance the Power 6/7 and other future system component designs. Teams of DOE computational and computer specialists will work with IBM hardware and software designers to further improve memory and interconnect bandwidth in the Power series. The goal will be to field a petaflop/s (peak) computer capable of 20–25% sustained rates on diverse application by 2009.

The initial work will be done in a series of “lockdown” meetings with designers, where DOE's scientific application requirements will be analyzed and understood. Design alternatives will be developed and evaluated for their potential to better meet the requirements. The result will be a series of more detailed meetings to review and resolve the design details.

DOE applications and representative code kernels will be evaluated by instrumenting and understanding the behavior of the codes on existing hardware. Special performance profiling tools—some existing only in IBM labs—will be used to gain an improved understanding of the codes. Based on the results of these studies, models of the codes will be developed and run on software simulators for the proposed hardware design. The outcome should be to identify and hopefully resolve performance bottlenecks at the design stage of future-generation Power processors rather than after delivery, as a traditional evaluation does.

² In fact, the MSS/SSP implementation was first done on the Cray SV1 systems and has been in operation for since 1999.

Examples of further candidate improvements that will be evaluated and in some cases selected for implementation are:

- VMX2 external connections
- I/O decentralization
- FFT breakthroughs—libraries and possibly hardware acceleration engines
- Ability for an application to use 100% of bisection bandwidth
- A striped-down version of MPI (MPI lite) that will have some restrictions for major improvements in performance
- Microkernel OS running on the compute nodes, as exists on the T3E
- Improved daemon control that synchronizes their running to have far less impact on applications
- Hardware collectives (e.g., reduce all) supported with improved hardware and software
- Synchronized time-outs of tasks to have far less impact on applications
- Unified Parallel C
- Advanced cooling
- Non-segmented addressing
- Improved OS for scientific programming (AIX or Linux)

In addition to the issues involved with design and development of future Power CPUs and improvements, significant software challenges existing in order to make Blue Planet operate well at the scale proposed. These challenges will have to be addressed in the same cooperative manner as outlined above for the long term in order to assure the implemented solutions are the most effective possible for the applications.

One example of these issues is that the communication software is very sensitive to interference from interrupts and the asynchronous nature of dispatching of application threads by individual operating systems on each node, especially in a large-scale system. This interference is especially disruptive for global operations across large numbers of processors such as MPI collective communications. New solutions and programming models to help synchronize dispatch cycles and eliminate the overhead of collective operations will need to be developed. This may occur through hardware accelerators that attach directly to the switch in addition to software. Better synchronized individual OS activity across the nodes of the system will need to be made more robust and usable. New programming models (e.g., UPC) will need to take direct advantage of such adapter hardware functions, which provide improved memory access across nodes. The zero copy transport protocols will have to be made more efficient and robust to eliminate the memory bandwidth bottleneck during transport. Very low latency communication will need processor assist in terms of fast synchronization lock instructions and for barrier synchronization, which are being developed/considered.

Likewise, scaling a shared file system to 2.5 PB and 2,048 nodes will require significant redesign and enhancement. The target of 1 GB/s of IO performance per IO server will need significant enhancements to zero copy transport, distributed locking design (for parallel access), and the

disk allocation manager. Efficient metadata serving will be a significant challenge at such scale. It will be critical to ensure that robust fault tolerance from disk and node failures is built into the system, which requires major enhancements to the high availability features. The ability to efficiently back up a system with 2.5 TB of disk needs to be addressed. External access to the cluster file system data at high speed is another critical problem that needs to be addressed, particularly in the Grid applications.

Overall, there is a wide range of areas that must be addressed initially with the Blue Planet system and then improved for the petaflop/s system. These issues cannot be done at small scale, so a system the size of Blue Planet is the only way to gain confidence and experience that the improvements will work on the petaflop/s system, regardless of which processor type is used.

3.2.4 Milestones

- 3QCY02 — Initial Power 6 design lockdown meeting
- 4QCY02 — Power 4, single-core, four-way, 16 nodes (32 by end of year) with enhanced performance profile capability
- 4QCY02 — Study to decide if pseudo vector (minus barrier) is possible on Power 4 for a prototype
- 4QCY02 — Study to assure the single-core high bandwidth concept shows benefits over standard 32 node
- 1QCY03 — DOE application studies concluded and shared with IBM before the design cutoff for Power 6
- 3QCY03–4QCY04 — Power 4+, single-core, four-way, Federation switch, 32 nodes (64 by end of year)
- 3QCY03–4QCY04 — If rudimentary ViVA is possible (minus barrier), experiment with implementations without compiler technology
- 3QCY03–4QCY04 — Single-core high bandwidth concept shows benefits over standard 32 node study
- 4QCY05 — Installation and demonstration of Blue Planet
- 2QCY09 — Petaflop/s system with Power architecture and the following switch to Federation

4. POSSIBLE HARDWARE COLLABORATIONS WITH OTHER LABORATORIES AND UNIVERSITIES

The principal technology options that ANL and LBNL will pursue in this proposal are described above in Section 3. However, over the trajectory of the computing initiative we propose here, and even in its first three years, we expect that new options will appear to which the Argonne and Berkeley Labs can and should contribute. Those options will have to be the primary responsibility of other institutions, however. We describe a few of them here because there are already specific opportunities for collaboration that we can envision today, but other technical

directions that may not yet be as well developed as the ones we describe here should not be neglected.

4.1 Technology Exploration in VLIW and Circuit-Board-Level Optical Interconnects with Hewlett-Packard

Recent discussions with Hewlett-Packard's technology and science staff revealed new and innovative approaches that could have high payoff for future scientific applications. HP has demonstrated the ability to compete and win HPC systems with their PNNL and ASCI awards. Furthermore, they have more systems on the TOP500 list than any other vendor.

The possibilities that were described to us by Hewlett-Packard immediately suggest that PNNL, if it is willing to accept primary responsibility for investigating these directions, could lead an effort to which ANL and LBNL could contribute substantially. The opportunities that arise immediately originate from features of the IA-64 microprocessor that are not being exploited in any current architecture.

HP's basic strategy is to compete in the mid to high end of technical computing with SMPs and CC-NUMA SMPs in the Fortune 1,000 companies. HP will cluster nodes together to build bigger systems between 1–10 Tflop/s for universities, laboratories, and large commercial needs, and further cluster the SMP nodes together for very large systems (ASCI and Earth Simulator class) for national centers.

The CPU technology is the IA-64 or Itanium Processor Family (IPF) architecture, of which HP is a co-developer with Intel. The IPF architecture is capable of a number of features that HP would attempt to exploit under this strategy. Currently HP has no plans to use these features, since there is no incentive in the commercial marketplace. Some of the possible ideas are:

- Increase the use and length of very long instruction word (VLIW) architectures. VLIW provides increased concurrency by executing multiple standard instructions at one time as one longer complex (or very long) instruction. This technique requires architectural as well as advanced compiler support to be effective.
- HP is investigating the use of optical pin-outs on their application-specific integrated circuit (ASIC) chips beginning in a limited manner in the 2006 timeframe. This will enable the use of optical technology on circuit boards, increasing the transmission rate between components on a chip as well as allowing for innovative cooling and system designs, since components will no longer be limited in their physical separation distance.
- Implementing hardware accelerators alongside the main CPUs. Accelerators can be a combination of IPF cores and field-programmable gate arrays (FPGAs). Examples of the type of accelerators HP might produce are FFT engines, sequence assembly engines, and sparse array engines.
- IPF explicitly supports high-end security by implementing a protection ring structure similar to that done in the Multics research system, and by executing the RSA encryption kernel faster than any other processor.

HP is also partnering with the University of Southern California's Information Sciences Institute (ISI) to explore processor-in-memory (PIM) technology under a DARPA contract.

HP has a strong history of innovation and implementation in the scientific and technical market. They hold more patents on nanotechnology switching than anyone else. The combination of corporate interest and technology expertise has high potential.

4.2 Expansion of a NERSC Production Computational Resource with a Cray X1

We expect and hope that the evaluation of the Cray X1 at Oak Ridge will be successful and that Cray Inc. will succeed in fielding an effective vector architecture. If early evaluations of the X1 are positive, the NERSC Center is in a unique position in the DOE complex to place an early system of modest size in full production and provide access to it for the entire DOE scientific community. NERSC has extensive experience in supporting vector architectures, and its systems programmers have contributed to the evolution of the Cray UNICOS operating system upon which many of the parallel computing features of the X1 operating system are based. NERSC can make such a system available and give its users the full range of consulting and programming services that the NERSC user community has come to expect from the Center for its other computing resources.

The X1 consists of building blocks of CPUs, called multi-streaming processors (MSPs). Four MSPs are clustered onto a node board. The proposed system would have 192 MSPs (~180 for computation), each operating at a peak rate of 12.8 Gflop/s per CPU. The aggregate system-wide memory would be 1.5 TB (8 GB per MSP) and 50 TB disk. The X1 would run a single system image (SSI) that enhances the usability and manageability of the system. Table 3 shows a possible phased deployment of production systems beyond the beta test version being tested at ORNL.

Table 3
Two-Phase Deployment of Cray X1

Phase	Date	Number of CPUs	Aggregate Memory
Phase 1	Q3/CY03	64 liquid-cooled MSPs	0.256 TB
Phase 2	Q4/CY03	Add 128 MSPs = 192	1.5 TB

Initial performance projections combine information from pen-and-paper analysis, running kernels of codes on Cray's SV2 simulator, and early performance from a partially functional prototype. Based on this data, it is reasonable to expect the final system to have significant performance on scientific codes:

- Sustained System Performance (SSP-2) value of 0.75 Tflop/s (~30% of peak).
- The POP climate model will performance at approximately twice the speed of the SX-6 on a per CPU basis.
- Projected Linpack of 1.98 Tflop/s (86% of peak), effectively matching the percentage of peak performance achieved by the Earth Simulator.

A NERSC-4-like contract would be used to award a firm, fixed-price contract to Cray with focused requirements from the overall DOE scientific workload, a true acceptance test designed to make the system capable for production use within the NERSC Center.

4.3 Tracking the DARPA HPCS Program

In 2002 DARPA started a new program in High Productivity Computing Systems (HPCS). One-year contracts were awarded to Cray Inc., Hewlett-Packard, IBM, Silicon Graphics Inc., and Sun Microsystems. A second 36-month research and development phase will be followed by selection of a contractor for a 48-month full-scale development phase. This program is intended to be a focused research and development program, creating new generations of high end programming environments, software tools, architectures, and hardware components in order to realize a new vision of high end computing.

The goals of HPCS are to address the issues of low efficiency, scalability, software tools and environments, and growing physical constraints. These goals are closely related to some of the objectives of this proposal, and consequently the ANL/LBNL project team will monitor and interact with the HPCS architecture teams, by participating in the program workshops and communicating with the vendor teams. LBNL is already subcontractor on two of the projects (HP and Sun). However, early pilot platforms from the program are planned to be available only in 2008 and beyond. Hence this program, which is of considerable long-term interest for DOE, will not offer the immediate sustained performance opportunities that we are seeking in this proposal.

4.4 Alternative Technologies

In addition to working with established vendors on current and future architectures, we believe it is important for the DOE laboratories and this project in particular to participate in the research and development of over-the-horizon technologies and architectures that may provide additional options for improving applications performance and scientific computing capability.

Examples of these technologies include massively parallel ensembles of programmable logic devices, reversible logic based systems (which may address the power and thermal issues associated with pushing beyond petaflop/s configurations), hardware implementations of cellular automata, and all optical devices. We propose that several small and modest-scale technology tracking and long-term research activities be coupled to our primary architecture development strategy; these activities will provide connections to innovative work being done in universities and laboratories. The motivation for these projects and connections is to encourage the applications and computer science community to continuously think about alternative ways of achieving capability that may be feasible in the five to ten year time frame.

5. COMPUTER SCIENCE RESEARCH STRATEGY FOR PETASCALE APPLICATIONS

As a key ingredient of the new strategy proposed here, DOE must reestablish a national research community in computer science research, focused on high performance computing questions, including computer architecture, software, tools, algorithms, and libraries. In order to maintain a sustainable path to leadership in scientific computing, DOE must ensure an environment where university researchers can depend on stable funding if they are to devote their careers to it. A comparable core capability of computer science research in the DOE labs must be maintained.

The DOE is very capable of managing such a program, as the success of SciDAC shows, but it will be a new effort requiring new investments and program structures.

Historically, developing applications that achieve high performance on any sequential or parallel system has required multiple iterations of performance analysis and software refinement. In each cycle, one first identifies the key program components responsible for the bulk of the program's execution time and then modifies the program to improve its performance. Petascale systems containing thousands of processors pose new and larger challenges for software development and optimization, including memory hierarchy management, communication latency tolerance and avoidance, fault-tolerance, I/O, adaptation and load balancing, all realized by scalable software and libraries.

Although one could launch a broad-based effort to develop all the requisite scalable software for petascale systems, experience has shown that demonstrable progress is best achieved by focusing software research and development on the problems exposed by specific applications. Driving software research and development with a set of six to eight applications that are of national importance focuses effort on the most critical software challenges. Moreover, it is the shortest path to demonstrated scalability and high performance.

Hence, a major element of our strategy is to create a small but high quality and highly motivated team of applications scientists, computer scientists from ANL and LBNL, and selected university-based computer science researchers. The goal of this activity is to nucleate a sustained partnership that will work together for the next five years to achieve a fusion of architecture, applications, and systems software.

5.1 Research Team and Goals

We have recruited to this effort Dan Reed from the University of Illinois, Kathy Yelick from the University of California Berkeley, Kai Li from Princeton University, Tom Sterling from Caltech/JPL, and Bob Lucas from ISI/USC to join us in this endeavor. This core team, coupled with the core applications partners discussed above, will work closely with the computer science and applied mathematicians at ANL and LBNL to develop a detailed program plan for this project.

We expect significant computer science research will be required to sustain our strategy. Important areas include the development of system software for the BG/X machine series, with particular focus on scalable process management and programming models, but with a focus also on scalable I/O, memory hierarchy management, fault tolerance, and software libraries. On the enhanced Power architecture path, work is needed in compilation technology and performance analysis for future systems. The team members represent the highest quality computer scientists in the nation; all have a demonstrated track record of success in enabling high performance computing for real applications.

We expect that one outcome from this long-term partnership will be a new cohort of graduate students and related research projects that will help nurture a new generation of computer scientists who have a primary research interest in the enablement of scientific applications. Below, we outline the key challenges in fault tolerance, numerical and communication libraries, compilation techniques, parallel I/O, and performance analysis and tuning.

5.2 Fault Tolerance

Any system with tens of thousands of nodes will experience node and component failures. To maximize effective system use, one must create nimble applications that can adapt to failures—checkpointing alone is unlikely to be sufficient. Hence, we must explore algorithm-based fault tolerance, redundant computation techniques, and flexible communication library support for fault detection and recovery, all driven by the needs of our target application suite.

5.3 Numerical and Communication Libraries

High performance parallel numerical libraries provide application developers access to state-of-the-art algorithms without requiring extensive code development or use of low-level parallel constructs. We will develop a robust suite of libraries tailored to the needs of the target application suite. This will include extension of scalable message passing libraries like ANL's MPICH to operate in the petascale regime and optimization of numerical libraries like PETSc for the Blue Gene memory hierarchy. The SciDAC Integrated Software Infrastructure Centers focused on numerical algorithms and libraries, such as the Applied Partial Differential Equations Center (APDEC), Terascale Optimal PDE Simulations (TOPS), and Terascale Simulation Tools and Technologies (TSTT), should be augmented to address the challenges of the new architectures.

5.4 Compilation Techniques

The virtual vector architecture described in section 3.2.2 raises a number of research questions in compilation techniques. From a programming perspective, there are several ways to exploit the architecture; for example, the compilers could be enhanced to support automatic parallelization across several threads and use the GPPR for thread synchronization. Other models are possible and need to be investigated and evaluated. More generally, the low-latency interconnects are a perfect match for alternative programming models, e.g., models based on global memory and UPC.

5.5 Parallel I/O

To understand I/O needs in the petascale regime, we will leverage the team's application and system I/O characterization tools. These insights will guide tuning of the MPI-IO support in MPICH as well as extensions to open source I/O libraries like PVFS. In addition, we will work with IBM to configure and optimize GPFS and Linux I/O.

5.6 Performance Analysis and Tuning

Efficient instruction scheduling, cached data reuse, memory hierarchy management, and code generation are all critical to achieving high performance on modern microprocessors. Understanding application and system interactions will require hardware performance measurement, code quality assessment, and resource assessment, all at unprecedented scale. This work will leverage skills and software from the SciDAC Performance Evaluation Research Center (PERC), which includes several members of the proposing team at ANL, LBNL, and

Illinois. As such, it will extend existing tools and develop new offline and online techniques for scalable performance measurement and analysis of systems with tens of thousands of processors.

6. ADVANCED GRID AND NETWORKING INFRASTRUCTURE AS A SCIENCE ENABLER

The current state of high performance computational science is at the level of teraflop/s computers and terabyte data sets—often termed *terascale* environments. Networking capabilities and software to enable the integration of distributed resources (termed *middleware*, or *Grid* software) in terascale environments today involves high-end networks running at up to tens of gigabits per second and very basic Grid capabilities. To utilize these for science requires such significant application effort and cooperation among laboratories that they are routinely used by only a small handful of advanced application teams.

In a recent science requirements workshop³, several science disciplines were asked to provide information on how they currently use networking and Grid computing, and what they saw as the future process of their science that would require, or be enabled by, significant increases in capacity and functionality. Several general observations and conclusions may be made after analyzing the science application scenarios, and the fundamental observations that emerged from this workshop are becoming common themes in discussions about very large-scale science.

First, science is a distributed endeavor. The participants are distributed, and the science apparatus that they use is distributed. Experimental science requires specialized, often immensely expensive instruments whose data output requires supercomputers to process and large-scale archives to store. Computational science employs the supercomputer as the data-generating instrument as well as for data analysis, and nearly all fields of science require advanced visualization capabilities.

Second, the integration of this science apparatus requires a more sophisticated level of middleware than is currently available, because today's Grid systems are so highly dependent upon manual intervention and coordinated planning that it can take days or weeks to arrange for and debug an experiment.

Third, there is a tremendous amount of commonality among science disciplines in terms of their requirements for Grid middleware and network capabilities.

Yet even for terascale environments, there are also fundamental challenges to providing adequate network capacity, and improvements in Grid middleware will only further expose the limitations. A scientist at a university should be able to retrieve a terabytes-sized dataset from a laboratory for analysis. This will be a typical size for both instrumentation and simulation datasets in just a few years, and already is in some fields today. Yet in today's ESnet environment, where 100 Mb/s would be considered excellent end-to-end performance, the 10 TB data set would take on the order of two weeks to transfer. Between major laboratories the situation is better; however even on the fastest portions of the ESnet backbone, this 10 TB data set will require one to two days to transfer. Today's fastest research testbeds, like the one between Argonne and NCSA

³DOE Office of Science, High Performance Network Planning Workshop, August 13-15, 2002, Reston, Virginia (<http://doecollaboratory.pnl.gov/meetings/hpnpw>)

using the I-WIRE optical infrastructure, are beginning to demonstrate 20–40 Gb/s capabilities, in which the 10 TB would move in 45–90 *minutes*.

Even so, moving to ultrascale science will require 1,000 TB files. Even at 40 Gb/s, this would take nearly three days to transfer (and would take over two *years* using ESnet). Clearly ultrascale science will require networks running not in the tens but in the hundreds of Gb/s.

In short, all of the science areas need high-speed networks and advanced middleware to couple, manage, and access the widely distributed, high performance computing systems, the many medium-scale systems of the scientific collaborations, high data-rate instruments, and the massive data archives that, together, are critical to next generation science, and to support highly interactive, large-scale collaboration. That is, all of these elements are required in order to produce an advanced distributed computing, data, and collaboration infrastructure for science that will enable paradigm shifts in how science is conducted. Paradigm shifts resulting from increasing the scale and productivity of science depend completely on such an *integrated advanced infrastructure* that is substantially beyond what we have today (Figure 3).

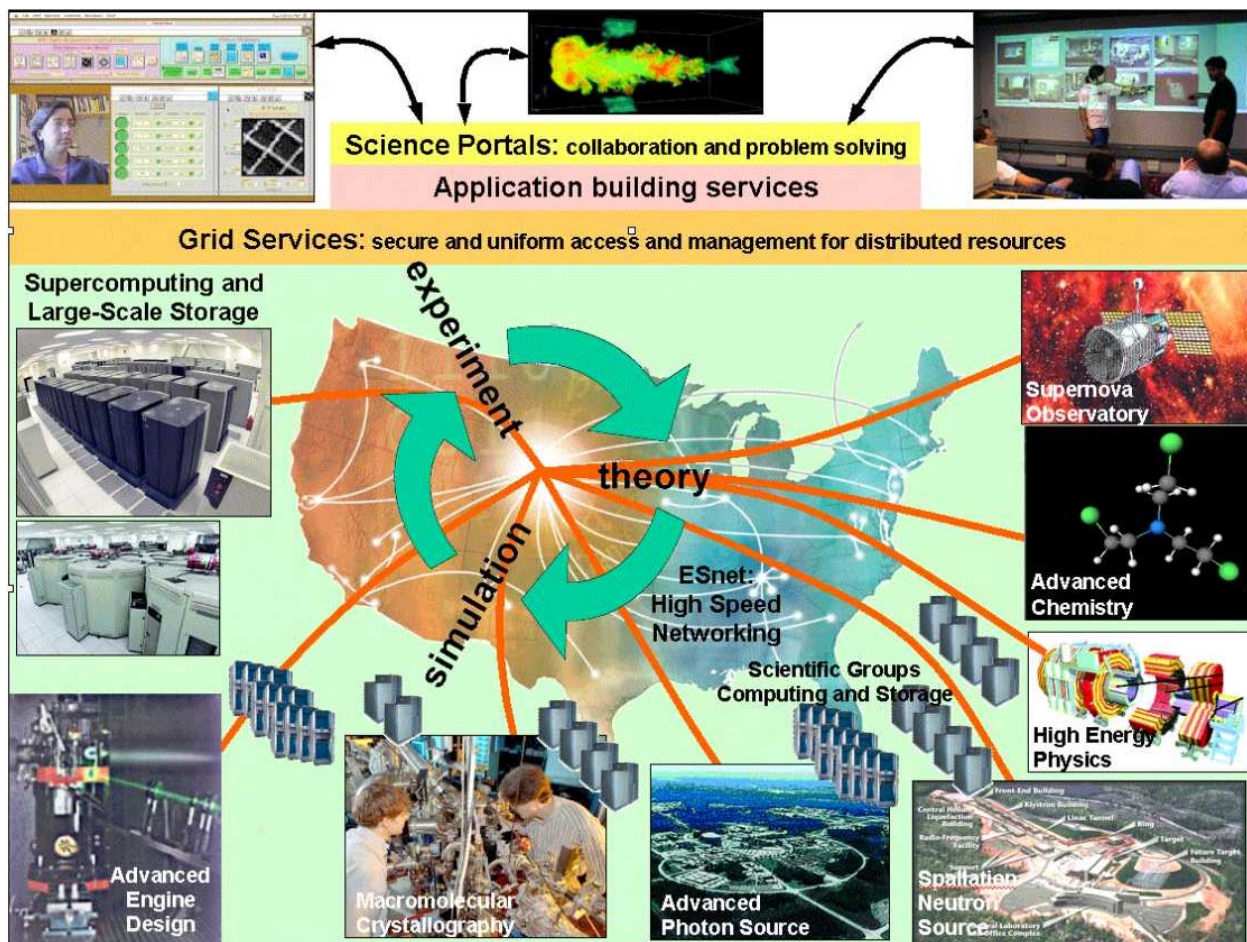


Figure 3. Integrated cyber-infrastructure enables advanced science: (1) Provide the DOE science community with advanced distributed computing infrastructure based on large-scale computing, high speed networking, and Grid middleware. (2) Enable the collaborative and interactive use of the next generation of massive data producing scientific instruments. (3) Facilitate large-scale scientific collaborations that integrate the federal labs and universities.

The need for such paradigm shifts is not speculative, but rather is evident from observing the challenges to scientists *today*. Several areas of science are pushing the existing infrastructure to its limits trying to move to the next generation of science. Examples of this include high energy physics with its worldwide collaborations analyzing petabytes of data (described below), and the data-driven astronomy and astrophysics community that is trying to federate the huge databases being generated by a new generation of observing instruments so that entirely new science can be done by looking at all of the observations simultaneously. (E.g., the National Virtual Observatory⁴ illustrates this point very well. Specifically see “New Science: Rare Object Searches.”⁵)

The task of the integrated advanced infrastructure is to deliver an overall computing, data, and collaboration quality of service to scientific projects. That is:

- Computing capacity adequate for a task is provided at the time the task is needed by the science.
- Data capacity sufficient for the science task is provided independent of location, and in a transparently managed, global name space.
- Communication capacity sufficient to support all of the aforementioned is provided transparently to both systems and users.
- Software services supporting a rich environment that lets scientists focus on the science simulation and analysis aspects of software and problem solving systems, rather than on the details of managing the underlying computing, data, and communication resources.

The clear message from all of the science application areas is that the paradigm shifts in how science is done will come about from a well integrated, widely deployed, highly capable distributed computing and data infrastructure, and not just any one element of it.

The requirements for the highly capable distributed science environments needed to support the sorts of science described above include a range of technologies, all of which must be integrated and persistent. The technologies that we discuss here are either being deployed today, or are in development. This is not a list of things that will require a decade of computer science research before we can deploy them; on the other hand, there is a good deal of development and deployment to be done in order to make these technologies into a highly capable infrastructure.

Two years ago we did not have the systems, communications, tools, or experience to do this. Today, we are at a point where building and deploying is doable in the three- to five-year timeframe in all of the technology areas, given adequate support.

6.1 Grid Middleware: Creating an Integrated Science Apparatus

The evolution of middleware and distributed systems in the scientific computing environment is currently embodied in the endeavor called computing and data Grids.^{6,78}

⁴ National Virtual Observatory, NVO. <http://www.us-vo.org/>

⁵ Astro-IT challenges and big UK surveys, A. Lawrence. In *Virtual Observatories of the Future*. 2000. Caltech. <http://www.roe.ac.uk/wfau/nvo/index.htm>

⁶ Computational and Data Grids in Large-Scale Science and Engineering, W. Johnston. Future Generation Computer Systems, 2002.

The role of Grid middleware is to greatly simplify the construction and use of widely distributed and/or large-scale collaborative problem solving systems. Grid-managed resources are the geographically distributed, architecturally and administratively heterogeneous computing, data, and instrument systems of the scientific milieu.

Grid middleware provides services for uniform access, management, control, monitoring, communication, and security to application developers using these distributed resources.

The international group working on defining and standardizing Grid middleware is the Global Grid Forum⁹ (GGF), which now consists of about 700 people from some 130 academic, scientific, and commercial organizations in about 30 countries.

GGF involves both scientific and commercial computing interests. It also involves an evolving understanding of the issues that must be addressed in order to facilitate the expeditious construction of the complex distributed systems of science from a very dynamic pool of resources.

There is now enough experience in building Grids that the basic access and management functions noted above are fairly well understood, and reference implementations are available for most of these through the Globus toolkit.¹⁰

However, as our experience with Grids grows, more issues arise that must be addressed in order to meet the goals of easily building effective distributed science systems.

In order to be effective, the Grid middleware must be widely deployed. This involves two things: first, a recognition on the part of the funding agencies that Grids represent an essential new aspect of the infrastructure of science, and thus must be supported as persistent infrastructure; second, an educational process that addresses the critical sociological issues involved in changing operational procedures, inter-site cooperation and sharing, homogenizing security policy, etc. Many of these issues have been addressed in the building and operation of networks, and now have to be addressed in the operation of computing, data, and instrumentation facilities.

The type of Grid middleware described thus far provides the essential and basic functions for resource access and management. As we deploy these services and gain experience with them, it has also become clear that higher-level services are also required in order to make effective use of distributed resources. These higher-level services include, e.g., functionality such as brokering to automate building application-specific virtual systems from large pools of resources. Another high-level service is collective scheduling of resources so that they may operate in a coordinated fashion, so that, for example, a high performance computing system could do the real-time data analysis that would enable a scientist to interact with experiments involving on-line instruments or to allow simulations from several different disciplines to exchange data and cooperate to do a

⁷ The Anatomy of the Grid: Enabling Scalable Virtual Organizations, I. Foster, C. Kesselman and S. Tuecke. International J. Supercomputer Applications, 2001. 15(3). <http://www.globus.org/research/papers.html#anatomy>

⁸ *The Grid: Blueprint for a New Computing Infrastructure*, I. Foster and C. Kesselman, eds. 1998, Morgan Kaufmann. http://www.mkp.com/books_catalog/1-55860-475-8.asp

⁹ The Global Grid Forum is an informal consortium of institutions and individuals working on wide area computing and computational Grids: the technologies that underlie such activities as the NCSA Alliance's National Technology Grid, NPACI's Metasystems efforts, NASA's Information Power Grid, DOE ASCI's DISCOM program, and other activities worldwide. www.gridforum.org

¹⁰ The Globus Project. <http://www.globus.org>

whole system simulation, as is increasingly needed to study real, complex physical and biological systems. These services are currently being developed and/or designed.

Higher level services also provide functionality that aids in componentizing and composing different software functions so that complex software systems may be built in a plug-and-play fashion. These services are being approached by leveraging large industry efforts in XML-based Web services by integrating Web services and Grid services. This will allow the use of commercial and public domain tools such as Web interface builders, problem solving environment framework builders, etc., to build the complex application systems that provide the rich functionality needed for maximizing human productivity in the practice of science. This area is a recent activity of the GGF (e.g., the Open Grid Services Architecture¹¹), and much work remains, but the potential payoff for science is considerable.

The final aspect of the middleware is the support that is needed on the resource platforms themselves.

Computing systems must have schedulers that enable co-scheduling with other, independent resources. Data archive systems must have access servers that allow for reliable, high-speed, wide-area network data transfer. Networks must provide capabilities for quality-of-service (usually in the form of bandwidth guarantees) that let distributed resources communicate at high during critical times in coupled simulation or on-line instrument data analysis. All of the storage, computing, and network resources must have support for the detailed monitoring that is essential for debugging and fault detection and recovery in widely distributed systems.

These services must be developed, installed, and integrated into the operational environments of all of the individual systems that make up the resource pools of science.

In summary, the goal of Grids is to provide significant new capabilities to scientists and engineers by facilitating routine construction of information-based and collaboration-based problem solving environments that are built on demand from large pools of shared resources.

Functionally, the goal of Grids is to provide tools, middleware, and services for:

- building the application frameworks that allow discipline scientists to express and manage the simulation, analysis, and data management aspects of overall problem solving
- providing a uniform description of computing, data, and communication resources and the policies associated with them
- providing resource management services to deliver end-to-end quality of service
- providing a uniform interface to a wide variety of distributed computing and data resources
- supporting construction, management, and use of geographically and administratively distributed virtual systems
- facilitating human collaboration through common security, resource, and data sharing services
- providing remote access to, and operation of, scientific and engineering instrumentation systems

¹¹ The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration, I. Foster, C. Kesselman, J. Nick and S. Tuecke. <http://www.globus.org/research/papers.html#OGSA>

- managing and securing the access to the computing and data infrastructure as a persistent service.

To provide that advanced infrastructure that will facilitate the next generation of science, Grids must be combined with the next generation of ultrascale computers and storage systems and with widely deployed, very high bandwidth networks that provide the underlying capacity in order to knit together the many physical resources into a coherent and managed whole that provides the integrated advanced infrastructure needed for science.

The combined expertise at LBNL and ANL represents perhaps the strongest team of Grid experts in the world, as evidenced by the fact that Grid research at ANL is among the most frequently cited work in the field, the de facto Grid middleware software (used by over 90 percent of Grid projects around the world) is jointly led by ANL, and several of the most highly recognized, successful Grid deployment efforts have been (or are being) led by leaders from LBNL (DOE Science Grid, NASA IPG, etc.) and ANL (TeraGrid, GriPhyN, etc.). Finally, leaders from LBNL (Johnston, Jackson, Genovese, et al.) and ANL (Foster, Catlett, Stevens, Messina, Tuecke, Schopf, et al.) are directly responsible for both establishing and leading today's Global Grid Forum, where Grid standards are being developed.

6.2 Ultrascale Networks: Removing the Barrier of Location

As with petaflop/s computing, the challenges presented in reaching ultrascale network performance cannot be addressed by simply purchasing more or larger systems from vendors and suppliers. Making the capability usable end-to-end for the scientist involves a combination of carefully phased investments; acceleration of current engineering, development, and deployment activities; investment in a strategically selected set of research objectives aimed at directing the engineering and development efforts three to five years down the road; and involving industry partners who can leverage the commercial market to support the transition into production capabilities.

As with middleware, there are a number of fundamental areas in which network research and development is essential to moving to ultrascale capabilities. Today's network protocols (TCP/IP) have evolved from the worldwide Internet, where the data transport protocols and embedded algorithms for congestion control are optimized to adapt to network load (or perceived load) in a widely shared environment and to sacrifice individual performance in favor of overall network stability. Simply put, TCP ensures that no application gets more of its "fair share" of the network, and all applications are created equal. This is necessary for large, shared networks, and ultrascale science will rely on such networks. However, these optimizations present serious limitations to mission-critical capabilities ranging from those that require large bulk transfer (TCP algorithms statistically limit individual application throughput to 50 percent of available capacity) to those requiring priority response.

Similarly, today's network protocols were initially designed when 1 MB was considered to be a very large total number of bytes in a single network transfer. Consequently the error detection and correction algorithms are no more capable of detecting errors in a petabyte transfer than were old software systems at distinguishing between the dates January 1, 2000, and January 1, 1900.

Lastly, advances in optical networking technology have summarily overtaken local area networking and host interface technology, such that the basic assumptions used for the entire history of the Internet in architecting wide area networks—balancing investments in computers, equipment, and bandwidth—have been overturned. We are presently in a transition period where, while wide area bandwidth prices have undergone dramatic improvements in price per unit bandwidth, networking equipment (routers, interfaces, switches) and protocols continue to assume a large, shared, static backbone network as has been the case from the early dedicated 9,600 bps lines of the ARPANET to today’s leased 10 Gb/s single optical wavelength path “lambdas.” Further, the cost of moving data from optical to electronic domains (i.e., host and router interfaces) now vastly dominates the cost of the overall network.

The result is that emerging optical switching and routing technologies have the potential to harness the bandwidth necessary for ultrascale capabilities at substantially lower costs than are presently involved in building networks capable of tens of Gb/s. This is an area of particular importance to ultrascale computing because it sits at the intersection of the architecture of the network and the function of both networking protocols and middleware. DOE has already invested in a number of pioneering efforts in the middleware protocols necessary to exploit these changes, resulting in currently experimental functions and protocols in the Globus toolkit for controlling network bandwidth from the middleware level.

6.3 Workplan

These challenges will be addressed by a cross-institutional team representing a broad spectrum of networking expertise including:

- ANL — high-performance networking, optical communications architecture and engineering, and middleware
- LBNL — production networking support and services, measurements
- ORNL — protocol development
- Sandia-CA — network research.

2003–05 Network

Planning and initial 10 Gb/s network testbed

- A detailed study phase in 2003 to examine alternative approaches to providing the underlying communications infrastructure to exploit emerging optical technologies and leverage experience in today’s 10 Gb/s networks
- A 10 Gbps testbed linking three or four sites to be used for protocol and middleware development

2003–04 Grid, Net, and Blue Planet Integration

Initial integration of Blue Planet with the network and the Grid

- The engineering issues of high-speed connection of the Blue Planet system to the initial network testbed will be determined, and an approach developed. The initial high-speed connection will be established.

- The operational issues of integrating Blue Planet with the DOE Science Grid will be determined, and the issues of Grid accounting, Grid compatible firewall and security policies will be resolved.
- The Open Grid Services Architecture Grid services, together with the basic Unix-style Grid services, will be installed and tested in order to make Blue Planet available to the science community.
- User application support teams for developing large-scale, distributed science applications will be established.

2004–05 Network

Implement a 40–60 Gb/s experimental infrastructure, using multiple 10 Gbps channels (“lambdas”) to begin to integrate computational, middleware, and networking components (hardware and software) at scale. It is expected that five to ten site resources will be connected.

2005–06 Grid, Net, and Base System Distributed Capabilities

Capabilities specifically requested by the science community will be added to the Grid and network middleware.

- Transparent and high-speed access to the tertiary storage associated with Blue Planet from remote science sites, and Blue Planet transparent and high-speed access to data managed at remote science sites.
- Grid data caching at or near the Blue Planet system will be established for remote file replication and management.
- Overall system Quality of Service capabilities will be designed and integrated into Blue Planet (e.g., co-scheduling) and the network (e.g., bandwidth reservation)

2005–08 Network

Phased deployment of 160–240 Gb/s core optical network supporting dynamic bandwidth allocation per site, ranging from 10–160 Gb/s based on application and resource requirements (timescale of minutes to hours for allocation of bandwidth in units of 1–10 Gb/s). Experimental terabit/second core network.

6.4 Milestones

- 2QCY03 Initial 10 Gb/s testbed (ANL/LBNL/ORNL)
- 2QCY03 Integration of Blue Planet with network and Grid (ANL/LBNL)
- 4QCY03 Blue Planet available on the Grid for science community (ANL/LBNL)
- 4QCY03 Detailed terabit/s network plan and budget
- 2QCY04 Initial 60 Gb/s core network (ANL/LBNL/ORNL)
- 2QCY04 Integration of Power 4+ Blue Planet with network and Grid (ANL/LBNL)
- 3QCY04 Additional laboratories added to core network at 10–40 Gb/s
- 2QCY05 High-speed, transparent integration of local and remote tertiary storage

2QCY05 160–240 Gb/s core network

2QCY05 Initial Grid and net Quality of Service functions on Blue Planet

2QCY05 Integration of Power 5 Blue Planet with network and Grid (ANL/LBNL)

3QCY05 Strategy, technology, and initial deployment of network data caching

1QCY07 Experimental Tb/s core network

7 SUSTAINING SCIENCE-DRIVEN COMPUTING

7.1 Managing Multiple Paths to New Computing Architectures

While it is beyond the scope of this preproposal to provide a complete management plan for DOE’s new scientific computing initiative (currently with the working title Ultrascale Scientific Computing), we can articulate some important principles and challenges that should inform the management structure:

- At least three independent architectural options should be pursued by teams that have the independence to determine the detailed technical directions within them.
- Periodic major reviews of these options and the teams pursuing them should occur at intervals of one year to 18 months.
- A review by the scientific community, broadly construed, should determine which technical option or options will be chosen for large-scale implementation.

There should be several stages that precede the review that determines if one of the major technical options is to be chosen for large-scale implementation:

- Development of experimental prototypes of the hardware at modest scale.
- Development of systems software and programming environments that make those prototypes usable in early production.
- Modest-scale production implementations with software and other support that makes them usable during a testing period by a broad range of scientific users, both in the DOE laboratories and universities.

While it will be a challenge to manage the entire program as a single “construction project,” it can be done with a timeline that involves critical decision points for each technical option. It is not unprecedented to have technology selection as part of the construction of large-scale experimental facilities. Such “in vivo” technology selection has been done with detector technology for particle accelerators, for example.

A basic issue to be addressed by the project timeline is starting the development projects for future generations of hardware, beyond the first one chosen to be implemented at large scale. One or more of the technology options that are among the first three might be candidates for the second generation, for example. New technology options will appear and must be evaluated if the project is to have a sustained impact on American science. We can thus articulate another fundamental requirement of the management of this “construction project”:

- The project time line should include the initiation of new technology options and their exploration for subsequent generations of new architectures at large scale.

7.2 Estimated Costs Associated with Each Technology Option

This preproposal is not accompanied by a detailed analysis of costs. That analysis and full budgets will accompany a full proposal at a later date. The expenditures associated with each technology are different, and we have given an idea of which can be expected to be most and least expensive in section 1.3. The price of pursuing these options is not just in hardware and personnel, it also involves physical facilities to house the computers, maintenance, electricity, and cooling. However, the general scale of the costs of each technology option can be understood approximately based on the timelines and the stages of development described above.

Each of the three directions described in this preproposal has a modest startup cost. Prototype hardware for each option will cost in the vicinity of \$5 M to \$20 M, and to develop software and explore the performance of applications will cost roughly \$10 M per year.

Early production implementation and the associated software efforts for each option will be more expensive, but will leverage the investment in personnel and scientific collaborations that supported the prototype hardware. At this scale, hardware costs should be in the vicinity of \$50 M or less, and annual maintenance charges of approximately 10 percent of the hardware cost must be expected.

To scale any option to a level that significantly exceeds the capability of the Earth Simulator today will cost between \$100 M and \$400 M for the hardware alone, and maintenance, electricity, and building facilities become a significant additional annual expense.

The current progress in network hardware and landline fiber trends, and the evolution of Grid technology, indicates that an investment of \$10–15 M in the first two years, \$60 M in the next two years, and \$85 M in FY07–08, will be necessary to achieve a well-integrated petascale networking and Grids environment.

7.3 Decisions in the ANL and LBNL Partnership

The ANL/LBNL partnership does not preclude partnerships and exchange with other laboratories and universities of information in the best interest of American science and scientific computing. However, this preproposal explicitly describes a partnership between two national laboratories and IBM. It is reasonable, therefore, to ask how decisions within that partnership will be made, since the same vendor is involved in both.

The ANL/LBNL partnership is based on a long history of substantive collaboration between the two laboratories in many computing projects. The majority of their SciDAC projects, for example, involve these two labs as principal partners. The computing strategies of these two laboratories have been developed in concert for the last five years. Our well-recognized efforts in computational Grids are an example of that joint strategy.

The consortium these two laboratories propose to lead will be governed internally by an Executive Committee co-chaired by the heads of the computing efforts of both laboratories who report to the Laboratory Directors. Those two co-chairs would be Rick Stevens and Bill

McCurdy. The executive committee would include representatives of the applications teams and the computer science teams pursuing the two IBM options.

This Executive Committee would undertake quarterly reviews of progress in both technology reviews and recommend shifts of personnel and other resources as appropriate. Maximum effectiveness of the transfer of software and algorithms experience from the two technical directions would be ensured by this mechanism.

The project will also have an external Advisory Committee that will advise its managers on the choice of scientific applications and collaborators in computer science and computational mathematics that will optimize the two new architectures being developed at every stage in their evolution. The Advisory Committee's charge will be to maximize the impact of this collaboration on the entire U.S. scientific community in an open and transparent way.

8. CONCLUSION

The DOE has the opportunity to undertake an initiative to provide a fundamentally new class of computer architectures for science. That project must of necessity be of large scale and must be managed in a way that is both cost effective and ensures its success.

There is an important aspect of this project that distinguishes it from the building of experimental facilities such as particle accelerators or light and neutron sources. Those experimental facilities, once built, have a useful lifetime of more than a decade, and their scientific impact depends in large part on the fact that the scientific community can depend on their operation over an extended period. Computing machinery of one generation has a useful lifetime of at most five years. The impact of this project depends on its being able to create a new class of technology and sustain its development and construction over several subsequent generations.

This project differs from the construction of the large experimental facilities by the DOE over the past decade in a fundamental way. There are several technologies in this arena that might be pursued to completion as very large scale facilities, and it is not possible today to choose the one (or ones) that will succeed. Since computer technology continues to develop at a predictable pace, it is possible to craft a plan that regularly reviews and selects the major technical directions of the overall project and brings at least one facility into production at large scale by 2006.

The steady-state condition of American supercomputing should provide architectural diversity to its scientists, because computing technology continues to evolve and present new opportunities, and because of the diversity of scientific applications that are important to the country. Therefore we should plan a sustained program at the DOE that will create a sequence of large-scale facilities overlapping in time and integrated via Grids, networks, and advanced middleware. We must endeavor to secure the funding for such a plan and manage its multiple paths in a coordinated fashion.

“Progress toward the goal of practical fusion energy is inextricably linked to greater understanding of the physics of hot plasmas. In turn, scientific progress requires advanced simulations and detailed comparisons with experiments. Simulations serve as proxies for experiments, being easier to build and diagnose and having more scope for parameter variation. The goal is to increase the fidelity of simulations through improved physics models, improved numerics, and the use of larger computers. The ongoing benchmarks for codes are critical comparisons with experimental data. The quantity of data involved and the geographic dispersion of resources and researchers present additional challenges. To carry out this program, we will need an integrated approach to computation, data management, data transport, visualization and analysis. To complement fusion applications, a high-performance, coordinated infrastructure will be required which must include advanced computational engines, high-speed networks, and sophisticated middleware.”

Martin Greenwald (MIT)
Bruce Cohen (LLNL)
Doug McCune (PPPL)
Bill Nevins (LLNL)
David Schissel (General Atomics)

“Clearly NCAR’s Science has benefited immensely from our mantra of a balanced capable end-to-end cyber infrastructure. We see now that the challenges are scaling up to Tera and soon even Peta levels which bring new problems to the forefront. All of science can benefit from organizations like DOE, NSF and NASA as they tackle the problems associated with putting teracapable tools into the hands of scientists and if the earth Simulator serves as a stimulus to apply resources to solving these problems it will have been worth it.”

Al Kellie, Director Scientific Computing Division, NCAR

“DOE’s plans to design and implement an integrated advanced infrastructure in support of basic science certainly has the potential to substantially accelerate progress in climate science. The balanced and distributed capabilities under discussion would enable climate scientists to more rapidly explore important modeling questions, such as the role of resolution in simulation quality, and to efficiently facilitate such scientific investigations in a collaborative setting.”

James Hack, Senior Scientist,
NCAR Climate and Global Dynamics Division

“The Human-Planet relationship is ready for significant infusions of Information Technology and this has been clearly recognized by Japan in the development of their Earth Simulator System. There is an urgent need in the United States for a systematic and comprehensive augmentation to our end-to-end simulation capabilities to accelerate progress in our scientific understanding of global and regional climate change, extreme weather events, solar-terrestrial interactions, and the impacts of human activity on earth systems. The needed quantitative studies will stress all elements of modern cyberinfrastructure—high performance computation, data storage, data management and curation, networking, visualization, collaboration environments, and the appropriate middleware to support distributed work. Additional capacity is clearly warranted by the maturity of our current scientific tools and the scientific community in the United States is poised to use augmented cyberinfrastructure capabilities effectively and productively.”

Tim Killeen, Director, NCAR

“Chemistry is one of the base sciences on which many applications are built. The chemistry community is extensive and incorporates a wide range of experimental, computational, and theoretical approaches to the study of problems including advanced, efficient engine design; cleanup of the environment in the ground, water, and atmosphere; the development of new Green processes for the manufacture of products that improve the quality of life; biochemistry for biotechnology applications including improving human health and cleanup; and the use of all of these to improve Homeland Security. The advanced computing infrastructure that is being developed will revolutionize the practice of chemistry by allowing us to link high throughput experiments with the most advanced simulations. Chemical simulations taking advantage of the soon-to-come petaflop architectures will enable us to guide the choice of expensive experiments and reliably extend the experimental data into other regimes of interest. The simulations will enable us to bridge the temporal and spatial scales from the molecular up to the macroscopic and to gain novel insights into the behavior of complex systems at the most fundamental level. In order for this to happen, we will need to have an integrated infrastructure including high speed networks, vast amounts of data storage, new tools for data mining and visualization, modern problem solving environments to enable a broad range of scientists to use these tools, and, of course, the highest speed computers with software that runs efficiently on such architectures at the highest percentages of peak performance possible.”

David Dixon (PNNL)

Larry Rahn (Sandia)

Donald Thompson (Oklahoma State U.)

Piotr Piecuch (Michigan State U.)

Martin Head-Gordon (University of California, Berkeley)

David Leahy (Sandia)

Mark Gordon (Iowa State and Ames Lab)

William Green (MIT)

Bruce Bursten (Ohio State U.)

Gustavo E. Scuseria (Rice U.)