"Reasserting U.S. Leadership in Scientific Computation"

"The U.S. faces a major challenge in scientific computation, the foundation of scientific discovery in the 21st Century. A clear path, however, is available for the U.S. to regain leadership in this critical field."

-Office of Advanced Scientific Computing Research

Thomas Zacharia
Computing and Computational Sciences
Oak Ridge National Laboratory
Oak Ridge, Tennessee 37831

Scalable System Software – Al Geist, Stephen Scott
Performance Evaluation – Pat Worley, Tom Dunigan
System Installation and Management – Arthur Bland, Richard Alexander
Networking – Nagi Rao, William Wing
Data Management – Nagiza Samatova, Randy Burris
Visualization – Jim Kohl, Stewart Dickson
Application Support – Trey White, Mark Fahey
Applied Math, Tools and Algorithms – Ed D'Azevedo, Bill Shelton, David Bernholdt
Biology and Genomes to Life – Andrey Gorin, Ying Xu
Nanoscale Materials Science – Malcom Stocks, Thomas Schulthess
Climate – John Drake, David Erickson
Chemistry – Jeff Nichols, Robert Harrison
Fusion – Don Batchelor, Ed D'Azevedo
Astrophysics – Tony Mezzacappa

July 4, 2002

This white paper contains proprietary information of Cray, Inc. and is provided with permission of Cray, Inc. for the sole purpose of evaluating the white paper.

EXECUTIVE SUMMARY

This white paper describes an integrated response to the challenge of the Earth Simulator to the U.S. pre-eminence in high performance computing and computational sciences. The Center for Computational Sciences (CCS) at ORNL proposes to partner with Cray, our sister laboratories, and university partners to evaluate and drive the development of the Cray SV2 system to provide systems that can match or exceed the Earth Simulator for scientific simulations to reduce the science gap created by the Earth Simulator and to significantly enhance high-end computing. We propose a three-phase strategy wherein, in Phase 1 we propose to acquire a 32-processor Cray SV2 and evaluate the processors, memory subsystem, scalability of the architecture and software environment. In Phase 2, based on the results of Phase 1, we propose to expand the Phase 1 system to a 512-640 processor system to develop and deploy the scalable OS, file system, and other critical features. Finally in Phase 3, based on the results of Phase 2, we propose a 60 Tflops SV2e system to carry out grand challenge research at unprecedented scales of computing.

A team of physical, biological, computational and computer scientists and applied mathematicians have come together to provide a plan that addresses six strategic grand challenge areas: Coupled Climate Modeling, Chemistry, Numerical Computation of Wave-Plasma Interactions in Multi-dimensional Systems, Biology and Genome to Life, First Principles Simulation of Materials Properties, and Astrophysics. The scientific teams attacking each of these areas are made up of outstanding researchers in their respective fields. In addition, the program is organized to leverage expertise in the DOE laboratories and universities and new capabilities developed under the SciDAC ISICs to provide needed computational techniques and software tools for the grand challenge efforts.

A strong institutional commitment exists at ORNL in support of the proposed response, including construction of state-of-the-art facilities for the Center for Computational Sciences (CCS). The new facility will have a 40,000 sq.ft. computer room with the infrastructure to house an innovative, scalable, massively parallel computer capable of surpassing the Earth Simulator. In addition to being a hub for ESNET with current connectivity of OC12, the CCS will be connected to the DWDM network in Atlanta at OC192 or better allowing other DOE laboratories and universities to connect at very high bandwidth. The Center will work with DOE in making its resources available to SciDAC and other funded projects. Finally, the challenge of helping to meet the human resources shortfall in science and engineering is addressed by an innovative educational outreach program targeted at women and minorities. In this effort, high performance computers are used both as a magnet to draw students to science and engineering and as an enabling tool to assist in their understanding of technical matters.

The Japanese Earth Simulator Project, currently the largest topical computing facility in the world, presents immediate challenges to the U.S. science community to maintain its lead in understanding the complex interactions between the environment and the climate. In spite of the computational crisis brought about by the Earth Simulator, it provides the

strongest endorsement for the role that topical centers such as CCS will play in "Reasserting U.S. Leadership in Scientific Computation."

INTRODUCTION

"Japanese Computer Is World's Fastest, as U.S. Falls Back" was the title of the *New York Times* article that heralded the arrival of the "Earth Simulator" to the Top500 list, where it is likely to stay at the number 1 position for quite some time. While we have known for several years that the Japanese were building this machine, it still comes as a shock that the machine performs as well as it does on scientific applications that are strategic to the United States. It is likely to have a significant impact in the leadership role of the United States, not only in high performance computing, but more importantly in such important and strategic areas as climate change research, nanoscale science and engineering technology, biotechnology and energy technology. Just as our major scientific user facilities attract the best minds from around the world, so too will this machine galvanize the attention of the best computational scientists around the globe, including our own. Furthermore, certain classes of science can only be investigated on computers of the scale of the Earth Simulator.

What distinguishes the Earth Simulator from current U.S. offerings is that the design of the computer was optimized to meet the needs of high-performance scientific computing rather than the high-volume commercial market. Two important benefits are provided to the scientific users. First, the computer delivers a high fraction of its peak performance to science. Second, the balance of the computer makes it easier to use. As a result, it takes less time to write and to debug software for scientific applications.

The primary goal of this proposed effort is to develop and deploy an integrated response to the challenge of the Earth Simulator to the U.S. pre-eminence in high performance computing and computational sciences in six strategic grand challenge areas: ***Climate, Chemistry, Fusion, Biology and Genomes to Life, Nanoscale Materials Science, and Astrophysics***. A supporting element of this proposed effort is a computer science effort to develop software and techniques needed directly for grand challenge efforts. An important additional ingredient is the support for computational science education initially at the undergraduate and post-graduate level.

The Center for Computational Sciences (CCS) at ORNL proposes to partner with Cray to evaluate and drive the development of the Cray SV2 system to provide systems that can match or exceed the Earth Simulator (GS40) for scientific simulations to reduce the science gap created by the Earth Simulator and to significantly enhance high-end computing. We propose a three-phase strategy wherein, in Phase 1 we propose to acquire a 32-processor Cray SV2 and evaluate the processors, memory subsystem, scalability of the architecture and its software environment. In Phase 2, based on the results of Phase 1, we propose to expand the Phase 1 system to a 512-640 processor (8-10 cabinet) system, to develop and deploy the scalable OS, file system, and other critical features. Finally, in

Phase 3, we propose a 60 Tflops SV2e system to carry out grand challenge research at these unprecedented scales of computing.

This proposed effort leverages existing SciDAC projects at ORNL and other DOE laboratories and universities. Base research programs and computational resources exist at each participating institution. ***Funds for personnel requested in this white paper will be used primarily to support new personnel including postdoctoral and graduate students***. It is proposed to use existing wide area networks and computer center facilities and expertise at Oak Ridge National Laboratory until the construction of the new facilities are completed in March-June of 2003. ORNL also has many years of experience in operating and using parallel systems.

The largest fraction of the requested funding is for a scalable high-performance computing system. The hardware procurement will include an early model of a scalable supercomputer system capable of delivering 25-50 Tflops sustained performance. This system is expected to be faster than the Earth Simulator as well as current and anticipated systems at existing supercomputing centers. As the system matures, the goal will be to make the system available to additional DOE researchers. Resources will be allocated by peer review process. This proposed effort constitutes an expansion of CCS facilities existing at ORNL. The intention is to initially rely on ESNET and the new 10 Gbit/s DWDM link to Atlanta for necessary communication between the grand challenge researchers and the resources established by this proposed effort. This proposed effort will greatly benefit from a 40Gbit/s DWDM I/O network that connects the participating institutions.

Those domestic computer hardware vendors who are expected to have suitable systems available in the appropriate timeframe for the initial system have been identified. Non-disclosure presentations have been provided by IBM, HP, SGI, and Cray Inc. It is anticipated that a sole-source procurement of a research system, to be selected by DOE, will result in installation in FY03.

In addition to the participation of hardware vendors, university partners such as the University of Tennessee, Georgia Tech and Virginia Tech have pledged active participation and support to ensure the success of this endeavor. Clearly, the proposed effort will engage the best researchers from our sister laboratories and academia. It is our intent to set aside an allocation of resources to be used by our laboratory and university partners for their own research on grand challenge problems. Scientific collaboration with the academic partners will be encouraged.

The Japanese Earth Simulator Project, currently the largest topical computing facility in the world, presents immediate challenges to the U.S. science community to maintain its lead in understanding the complex interactions between the environment and the climate. In spite of the computational crisis brought about by the Earth Simulator, it provides the strongest endorsement for the role that topical centers such as CCS will play in "Reasserting U.S. Leadership in Scientific Computation." With more than 17 years of experience with both vector and parallel systems, ORNL is ready to take on the challenge of building a major new topical research center for the DOE scientific community.

The proposed effort has assembled outstanding interdisciplinary teams of application scientists, mathematicians, and computer scientists with key academic, national laboratory and vendor collaborators. We have an opportunity to restore the preeminence of the Office of Science and indeed the Nation in computing and computational sciences. The approach has already been articulated by the Office of Advanced Scientific Computing Research:

• *Partner with domestic vendors to evaluate and drive the development of promising computing architectures to provide systems that match or exceed the GS40 for science simulations.*
• *Reduce the science-gap created by the Earth Simulator by significantly enhancing the existing high end computing capability as well as networking, incorporating the best available U.S. technology.*

GRAND CHALLENGE RESEARCH

Grand challenge research areas of strategic importance to DOE and the nation have been funded separately through the SciDAC initiative and other base programs. This proposed effort leverages the developments occurring in these grand challenge applications areas. **(Additional funding will be necessary to develop specific capabilities in support of this proposed effort as well as to leverage these on-going projects)**

Collaborative Design and Development of the Community Climate System Model for Terascale Computers *(Robert Malone (Co-PI) Los Alamos National Laboratory; John Drake (Co-PI)* ***Oak Ridge National Laboratory****; Lawrence Berkeley National Laboratory; Pacific Northwest National Laboratory; Argonne National Laboratory; Lawrence Livermore National Laboratory; National Center for Atmospheric Research)*

The multi-institutional team will develop, validate, document and optimize the performance of a coupled climate model, the Community Climate System Model, using the latest software engineering approaches, computational technology and scientific knowledge.

Numerical Computation of Wave-Plasma Interactions in Multi-dimensional Systems *(Don Batchelor (PI)* ***Oak Ridge National Laboratory****; Princeton Plasma Physics Laboratory, Massachusetts Institute of Technology, Lodestar Corporation, CompX Corporation)*

The goal of this research is to use advanced terascale computing to obtain quantitatively accurate predictive understanding of electromagnetic wave processes, which support important heating, current drive, and stability and transport applications in fusion-relevant plasmas. The project will focus on increasing the speed and efficiency of two-dimensional wave solvers as well as obtaining a fully three-dimensional solution to the integral wave equation. This will allow computational studies of wave-plasma interactions in non-axisymmetric configurations. The goals of the research involve

answering the following questions: To what extent do realistic 2D and 3D equilibrium variations modify the local deposition of wave energy and momentum in the plasma? To what extent do non-Maxwellian particle velocity space distributions modify local deposition of wave energy and momentum in the plasma? What is the mechanism by which lower hybrid waves, launched with a phase velocity several times the electron thermal speed, are able to couple strongly to electrons and drive substantial currents? What is the effect of global plasma modes on the wave fields produced by launching structures?

Advanced Methods for Electronic Structure *(Robert Harrison, (PI)* ***Oak Ridge National Laboratory****; Lawrence Berkeley National Laboratory)*

This project will advance the capabilities of quantum chemical methods to describe efficiently and with controllable accuracy the electronic structure, statistical mechanics, and dynamics of atoms, molecules, and clusters. The software development effort will focus on the speed and scaling with respect to system size of electronic structure methods, the elimination of basis set error from all-electron calculations, the advancement of density functional methods, and improved descriptions for open-shell systems including excited states and electron correlation. Novel models and methods will be developed and implemented for use on massively parallel computers with the NWChem computational chemistry software package and other existing computational chemistry packages.

Theoretical Studies of Metals, Alloys, and Ceramics *(Malcolm Stocks, Thomas Schulthess, Bill Shelton,* ***Oak Ridge National Laboratory****)*

First principles quantum mechanical calculations are used to predict materials properties and to relate them to the electronic structure. Full potential linearized augmented plane wave (FLAPW), Gaussian basis cluster, and multiple scattering theory (MST) Green's function, density functional theory, electronic structure methods are developed and applied to calculating equilibrium structures, energies, elastic, thermal, magnetic and transport properties. Accurate FLAPW calculations of intermetallic compounds, energies of point defects, planar faults, and interfaces as well as thermal expansion are used to explain and predict mechanical behavior and to guide experimental alloy development. Highly precise cluster techniques are used to study the properties of intergranular phases in ceramics, the effects of impurities and dopants on interface bonding, and to guide the development of structural ceramics. Relativistic and non-relativistic MST, spin dynamics, and constrained density functional theory are developed to study magnetism in inhomogeneous systems — alloys, surfaces, interfaces, and bulk amorphous alloys. First principles theories of spin dependent transport in magnetic heterostructures are also developed. Parallel algorithms are used to significantly extend the size and complexity of systems amenable to treatment by first principles methods.

Shedding New Light on Exploding Stars *(Tony Mezzacappa (PI),* ***Oak Ridge National Laboratory****; University of Illinois NCSA, University of Tennessee at Knoxville; State University of New York at Stony Brook; University of Washington; North Carolina State*

University; Florida Atlantic University; Clemson University; University of California at San Diego)

The Terascale Supernova Initiative is a multidisciplinary collaboration of one national laboratory and eight universities to develop models for core collapse supernovae and enabling technologies in radiation transport, radiation hydrodynamics, nuclear structure, linear systems and eigenvalue solution, and collaborative visualization. The goal is a standard 3D model of core collapse supernovae with accurate multigroup neutrino transport, realistic nuclear and weak interaction physics, and the effects of general relativity.

Computational Biology *(Andrey Gorin, Ed Uberbacher, Ying Xu,* ***Oak Ridge National Laboratory)***

Computational biology research focuses on the computational, informational, and biological challenges created by the wealth of information that has resulted from large-scale genome sequencing. The group's research interests include computational analysis of the structure and function of genes and gene products, annotation of genome sequences, protein folding and structure prediction, microbial genomics and environmental biotechnology, and integrated analysis tools and databases that support fundamental biological research.

BASIC MATHEMATICS AND COMPUTER SCIENCE RESEARCH

This proposed effort leverages the developments occurring in basic mathematics and computer science research funded under the base program and the SciDAC projects through the close association of the PIs and their collaborators. **(Additional funding will be necessary to develop specific capabilities in support of this proposed effort as well as to leverage these on-going projects.)**

Terascale Optimal PDE Solvers (TOPS) ISIC *(David Keyes (PI), Old Dominion University; Argonne National Laboratory; Lawrence Berkeley National Laboratory; Lawrence Livermore National Laboratory; The University of California, Berkeley; Carnegie Mellon University, The University of Colorado; New York University;* ***The University of Tennessee)***

This ISIC focuses on developing and implementing optimal or near-optimal schemes for PDE simulations and closely related tasks, including optimization of PDE-constrained systems, eigenanalysis, and adaptive time integration. The ISIC will research, develop and deploy an integrated toolkit of open source, (nearly) optimal complexity solvers for the nonlinear partial differential equations that arise in many Office of Science application areas, including fusion, accelerator design, global climate change, and reactive chemistry. These algorithms, primarily multilevel methods, aim to reduce computational bottlenecks by one to three orders of magnitude on terascale computers, enabling scientific simulation on a scale heretofore impossible. Along with usability, robustness, and algorithmic efficiency, an important goal will be to attain the highest

possible computational performance in its implementations by accommodating to the memory bandwidth limitations of hierarchical memory architectures.

Terascale Simulation Tools and Technologies (TSTT) ISIC *(Jim Glimm (PI), Brookhaven National Laboratory; Argonne National Laboratory; Lawrence Livermore National Laboratory;* ***Oak Ridge National Laboratory****; Pacific Northwest National Laboratory; Sandia National Laboratory; Rensselaer Polytechnic Institute; State University of New York, Stony Brook)*

The primary goal of this ISIC is to develop technologies that enable application scientists to easily use multiple mesh and discretization strategies within a single simulation on terascale computers. The focus will be on high-quality hybrid mesh generation for representing complex (and possibly evolving) domains, high-order discretization techniques for improved numerical solutions, and adaptive strategies for automatically optimizing the mesh to follow moving fronts or to capture important solution features. Results of this effort will be encapsulated into software components with well-defined interfaces that enable different mesh types, discretization strategies, and adaptive techniques to interoperate in a "plug and play" fashion.

Center for Component Technology for Terascale Simulation Software ISIC *(Rob Armstrong (PI), Sandia National Laboratory; Argonne National Laboratory; Los Alamos National Laboratory; Lawrence Livermore National Laboratory;* ***Oak Ridge National Laboratory****; Pacific Northwest National Laboratory; Indiana University; The University of Utah)*

This ISIC will involve research into software component technology for high-performance parallel scientific computing to address problems of complexity, reuse, and interoperability for scientific simulation software. In particular, the research under this ISIC will address the following four areas. First, it will extend the Babel language interoperability technology to support Fortran 90 as needed by SciDAC application collaborators. Second, it will develop component schema and communication protocols between the ISIC component repository and component tools developed by collaborators. Third, it will continue to investigate parallel data redistribution approaches and integrate that support into the Babel language interoperability framework. Finally, ISIC team members will work with SciDAC application groups to use component technology in large, sophisticated simulation codes.

High-End Computer System Performance: Science and Engineering ISIC *(David Bailey (PI), Lawrence Berkeley National Laboratory; Argonne National Laboratory; Lawrence Livermore National Laboratory;* ***Oak Ridge National Laboratory****; The University of Illinois, Urbana-Champaign; The University of Maryland; The University of Tennessee; The University of California)*

The performance ISIC will focus on how one can best execute a specific application on a given platform. The research results from this effort are expected to permit the generation of realistic bounds on achievable performance, and to answer three fundamental

questions: (1) why do these limits exist; (2) how can we accelerate applications toward these limits; and (3) how can this information drive the design of future applications and high-performance computing systems.

Scalable Systems Software ISIC *(Al Geist (PI),* ***Oak Ridge National Laboratory****; Argonne National Laboratory; Lawrence Berkeley National Laboratory; Los Alamos National Laboratory; Pacific Northwest National Laboratory; Sandia National Laboratory)*

This ISIC will address the lack of software for effective management and utilization of terascale computational resources. The virtual center will be a multi-institution, multidisciplinary group composed of experts working together to develop an integrated suite of machine-independent, scalable systems software components needed for the SciDAC Program. The goal is to provide open source solutions that work from small to large-scale systems.

Scientific Data Management Enabling Technology Center *(Ari Shoshani (PI), Argonne National Laboratory; Lawrence Livermore National Laboratory;* ***Oak Ridge National Laboratory****; Georgia Institute of Technology; North Carolina State University; Northwestern University; University of California San Diego)*

The goal of this ISIC is to provide a coordinated framework for the unification, development, deployment and reuse of scientific data management software. The ISIC will target four main areas that are essential to scientific data management, emphasizing efficient management and data mining of very large, heterogeneous, distributed datasets. In addition, four tier levels of data have been identified: storage, file, dataset, and dataset federation. This "area and tier" framework is the basis for a management structure that will ensure the vertical integration of technologies in each area over these tier levels, as well as cross-utilization of area technologies. The result will be efficient, well-integrated, robust scientific data management software modules that will provide end-to-end solutions to multiple scientific applications.

DOE Science Grid: Enabling and Deploying the SciDAC Collaboratory Software Environment *(Bill Johnston (PI), Lawrence Berkeley National Laboratory; Argonne National Laboratory; Pacific Northwest National Laboratory,* ***Oak Ridge National Laboratory****)*

This collaboratory will define, integrate, deploy, support, evaluate, refine and develop the persistent Grid services needed for a scalable, robust, high-performance DOE Science Grid. It will create the underpinnings of the software environment that the SciDAC applications need to enable innovative approaches to scientific computing through secure remote access to online facilities, distance collaboration, shared petabyte datasets and large-scale distributed computation.

Earth System Grid II: Turning Climate Datasets into Community Resources *(Ian Foster (PI), Argonne National Laboratory; University Center for Atmospheric Research;* ***Oak Ridge National Laboratory****; Lawrence Livermore National Laboratory)*

This is a collaboratory for the development of a virtual collaborative environment linking distributed centers, models, data, and users. The creation of the ESG II has a potential for a very significant improvement of scientific productivity and collaboration among the investigators in the USA. Proper deployment of ESG II has an equal potential to facilitate exchange among climatologists all over the world, providing a badly needed platform for the management of the massive amounts of data that are being generated. Development of this and similar concepts is essential for rapid, precise, and convincing analysis of short- and long-term weather patterns, particularly in the period when increasing pollution introduces changes that may affect us for many generations to come.

TECHNICAL PLAN

This part of the white paper describes the high performance systems necessary to respond to the Earth Simulator challenge. The primary constraint limiting the choice of a new machine at this time is that the machine should be capable sustaining at least 25-50 Tflops on the grand challenge applications discussed above. Rather than identify a particular vendor that we find appealing and include it in our white paper, our approach has been to evaluate the products that vendors are proposing to bring on line in the future and determine how well they are suited to our proposed problems. We have worked on the premise that DOE will make the decision on the specific machines to be procured under this initiative.

There are four vendors with products that offer the possibility of the 25-50 Tflops sustained for at least some of the grand challenges. We are interested in all vendors with products in this class. We collectively heard presentations from IBM, HP, SGI, and Cray Inc. Of these four vendors, the CCS has already on the floor the current offerings from IBM and HP/Compaq. Additionally, installations of HP and SGI are underway at other laboratories. A scalable vector machine from Cray however has not been installed at any site today and would make an extremely timely and interesting machine to evaluate. Figure 1 shows a comparison of potential U.S. vendor offerings scaled to the peak performance of the Earth Simulator and plotted with the GS40. It illustrates that the SV2 has very impressive cache and interconnect bandwidths and is the only system with memory bandwidth comparable with the GS40. These metrics have been identified by science applications to be critical to getting good performance on large systems.

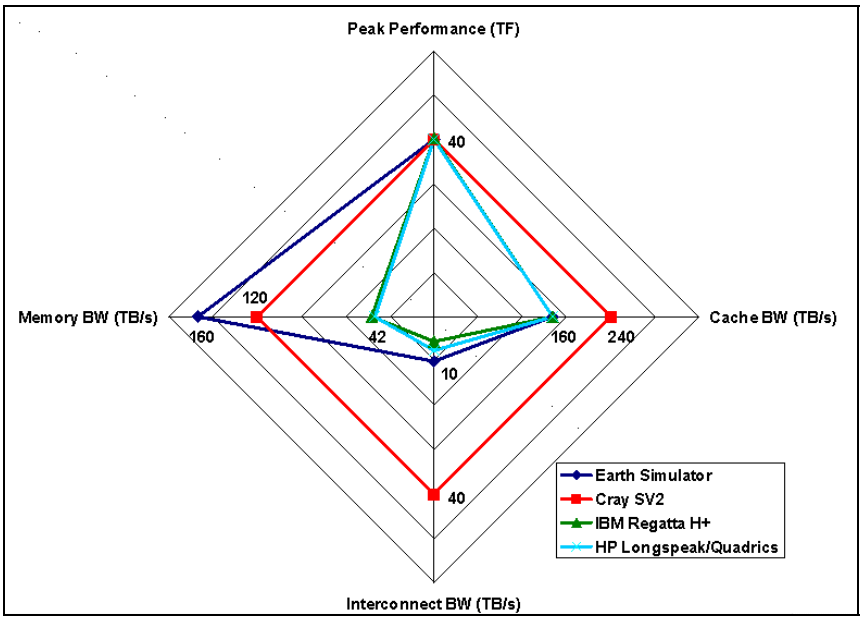

Figure 1. Comparison of U.S. machines that could potentially respond to GS40 shows that Cray SV2 has significantly better interconnect bandwidth and comparable memory bandwidth to GS40.

However, our interests are not limited to these vendors. We have had discussions with other vendors such as SRC and Key Research, both of whom have provided us with information on their future products, and are open to any ideas that may emerge between now and the time we must procure a system.

Our plan includes having in the Center a very large, powerful, scalable, high-performance computing system so that real progress can be made on the proposed problems. The CCS will provide the normal computer center functions, including user services functions such as consultation, documentation, training, numerical libraries, and utilities. The systems and operations staff will maintain and operate the hardware and system software. There will be a traveling training program that will go to each of the participating sites, as appropriate. The Center, as well as individual Grand Challenge areas, will host workshops, a visitor program, and a summer program.

EARLY EVALUATION

Parallel, high-performance computer systems for scientific and technical applications typically evolve through three phases: infancy, adolescence, and maturity. The infancy phase involves the testing of the basic architectural design, such as its ability to compute with both local and remote data without severe bottlenecks or hotspots. The emphasis is on the design, not the speed of its floating-point operations, its communications speed, nor its ease of use. The system may be delivered without a complete or stable operating system, and system software, such as compilers and debuggers, may be very “buggy,” or in some cases, non-existent.

The adolescence phase is entered when the basic design has proven successful. The vendor has typically produced its second-generation computer system with most of the major flaws in its first generation corrected. The emphasis shifts toward the scalability of the architecture, technology upgrades to improve performance and communication/computation balance, and the addition of peripherals necessary to attack real applications. The operating system is relatively stable with infrequent crashes. Basic system software exists, but not necessarily optimized for the numeric processor or communications network. Software tools are beginning to appear, but the system still requires an experienced, skillful person to achieve top performance.

The final stage of maturity is reached when the typical computer user, with the help of system software and software tools, can achieve acceptable performance. Codes may be easily ported from similar systems, or even from radically different architectures. Computers in this stage are truly production machines.

The computer discussed in the previous section, which is expected to be the showpiece of the Center for Computational Sciences, will be in the adolescent stage. In addition to evaluating computers in this intermediate stage and helping them move on to maturity, the Center can be of tremendous assistance in evaluating systems in the infancy stage. ORNL has demonstrated this ability on the many types of architectures including hypercubes, meshes, SMP clusters, and special purpose hardware. ORNL has a long history of evaluating new computer architectures. We have evaluated serial number one systems from a wide array of vendors, including the iPSC/1, iPSC/2, iPSC/860, and Touchstone Sigma from Intel, the KSR-1, SRC Prototype, Chen-1000, and Intel Pentium Pro processor. We have also participated in the evaluation of the Tera MTA at SDSC, the Swiss Tx at ETHZ, and the CPlant at SNL. We recently have had a string of firsts for the Office of Science, in installing and bringing to production the first IBM Power3 WinterHawk-I in the DOE lab system; the first DOE site to get the WinterHawk-II system, and one of only three DOE sites selected for early evaluation of NightHawk-I. The CCS was the only Office of Science site to receive a Compaq AlphaServer SC system. Most recently, the CCS received the first of IBM's latest Power4 based p690 systems and has the largest such system installed to date. The CCS has evaluated more than 30 different systems and architectures in the last 10 years.

The authors of this white paper have much expertise in the development of parallel applications and the porting and tuning of these applications on new architectures. For example, the first full application to exceed 1 Tflops was developed at CCS by Malcolm Stocks. This materials application won the 1998 Gordon Bell Prize. John Drake and Pat Worley have years of experience developing and evaluating parallel versions of atmospheric climate codes. Jeff Nichols and Robert Harrison are credited with the development of the NWChem chemistry software and also making the first large IBM SP system ever installed function effectively. Al Geist and his group are credited with the development of PVM and other critical software and are currently involved in the development of the scalable systems software suite. The Distributed Computing group received two 1999 R&D 100 awards for their recent software developments. Mark Fahey and Trey White lead a group of application engineers who helped convert

parallel vector applications into MPP applications at the DoD ERDC MSRC and are now staff at ORNL. The HPSS development team lead by Randy Burris has made several important strides in creating an efficient high-performance storage system. These are just a few of the researchers at ORNL working on high-performance applications in support of the Office of Science who will bring their expertise to the early evaluation.

Early Evaluation of Cray SV2

Of the four vendors discussed in the previous section, DOE has already on the floor the current offerings from IBM, HP/Compaq and SGI. On paper, Cray's new SV2 system offers the most potential and is the least understood. The SV2 is the first of Cray's new scalable vector systems that offer high-speed custom vector processors, very high memory bandwidth, and an exceptionally high-bandwidth interconnect linking the nodes. Cray is the only U.S. manufacturer of multi-terascale systems with vector processors. The processor performance of the Cray SV2 is anticipated to be comparable to the NEC SX-6 processor of the Earth Simulator on many Office of Science research applications. A new feature of this system is that it combines the processor performances of traditional vector systems with the scalability of more modern microprocessor based architectures. The SV2 is the first vector supercomputer designed to scale to thousands of processors in a single system image.

ORNL proposes to partner with Cray to evaluate and drive the development of the Cray SV2 system to provide systems that can match or exceed the GS40 for scientific simulations to reduce the science gap created by the Earth Simulator and to significantly enhance high-end computing. We propose a three-phase strategy wherein, in Phase 1 we propose to acquire a 32-processor Cray SV2 and evaluate the processors, memory subsystem, scalability of the architecture and software environment. In Phase 2, pending on the results of Phase 1, we propose to expand the Phase 1 system to a 512-640 processor (8-10 cabinet) system, to develop and deploy the scalable OS, file system, and other critical features. ORNL's experience with the Intel Paragon, IBM RS/6000 SP, and Compaq AlphaServer SC showed that important scalability issues in the software and hardware are not seen on small evaluation systems. Finally, in Phase 3, we propose a 60 Tflops SV2 or SV2e system. Pending the Phase 2 evaluation of SV2, we mitigate the risk of Phase 3 by including the option of a 60-100T IBM system.

As part of this collaboration, Cray will provide hardware and technical support at a significantly reduced cost. Cray will deliver a 32-processor SV2 in Q4 2002. We will use important applications codes from climate, bio-informatics, materials science, chemistry, astrophysics, and fusion to predict the expected sustained performance of multi-terascale Cray systems on large-scale simulations. Over the life of this partnership, we will continue to be one of Cray's beta-test sites for high-performance computing, providing DOE with access to the latest Cray hardware and software technology for evaluation. As an evaluation system with limited production commitments, this will be an ideal system for DOE to test new hardware and software, in particular to demonstrate scalable bandwidth in the system interconnect. Moreover ORNL, with its large IBM Power4 p690, IBM Power3, and HP AlphaServer SC systems, will be ideally positioned for head-

to-head comparisons of the latest generation IBM and Cray products for their suitability for terascale simulations.

Cray SV2 Description

The Cray SV2 is an attempt to incorporate the best aspects of previous Cray vector and MPP systems into one design. Like traditional vector systems, it is expected to have high memory bandwidth and high realizable percentage of theoretical peak. Like the Cray T3E, the design has a high-bandwidth, low-latency, scalable interconnect and scalable system software. And like the SV1, the design leverages commodity CMOS technology and incorporates non-traditional vector ideas, like vector caches and multi-streaming processors.

The SV2 multi-streaming processor (MSP) has multiple vector pipes designed to increase vector throughput without necessarily increasing the inner-loop vector length. Unlike the SX-6, the SV2 keeps the relatively short vector length of 64 elements, thus making it easier for algorithms to effectively use the pipes. The primary strategy for finding work for the eight vector pipes of a single MSP is parallelism through outer loops and pipelined operations.

Each MSP has 2 MB of cache memory, and the cache has more than enough single-stride bandwidth to saturate the vector units of the MSP. The cache is useful because the bandwidth to main memory is not enough to saturate the vector units without data reuse; memory bandwidth is roughly half the "saturation" bandwidth. This design represents a compromise between non-cache systems, like the SX-6, and cache-dependent systems, like the IBM p690, with memory bandwidths an order of magnitude less than the saturation bandwidth. Because of short cache lines and extra cache bandwidth, random-stride scatter/gather memory access on the SV2 is expected to be just a factor of two slower than stride-one access, not the factor of eight or more seen with typical cache-based systems like the IBM Power 4, HPAlpha, or Intel IA-64.

It is important to note that the SV2's cache based design deviates from the full-bandwidth design model only slightly. Each SV2 processor is designed to have more single-stride bandwidth than an SX-6 processor; it is the yet-higher peak flop/s that create the imbalance. A relatively small amount of data reuse, which most modern applications do exhibit, could enable a very high percentage of peak performance, and worst-case data access could still provide double-digit efficiencies.

Four MSPs and a flat, shared memory form an SV2 node. The memory banks of a node provide some 200 GB/s of bandwidth, enough to saturate the paths to the local MSPs and service requests from remote MSPs. Each bank of shared memory is connected to a number of banks on remote nodes, with an aggregate bandwidth of roughly 50 GB/s between nodes. This is a remarkable number; it represents a byte per flop/s of interconnect bandwidth per computation rate, compared to 0.25 bytes per flop/s on the Earth Simulator and less than 0.1 bytes per flop/s expected on an IBM p690 with a

maximum number of Federation connections. Comparisons with current DOE systems are even more lopsided.

The collected nodes of an SV2, eventually up to 4096 processors, form a single system image. A single four-processor SV2 node behaves like a traditional SMP, but each processor has the additional capability of directly addressing memory on any other node. Remote memory accesses go over the interconnect directly to the requesting processor, bypassing the local cache. This mechanism is more scalable than traditional shared memory, but it is not appropriate for shared-memory programming models, like OpenMP, outside of a given four-processor node. The mechanism is an excellent match for distributed-memory programming models, however, particularly those using one-sided put/get operations, and it is expected to provide very low latencies and unprecedented bandwidths.

In large configurations, SV2 nodes are connected in an enhanced 3D torus, or wrap-around mesh. Despite the remarkable bandwidth expected per connection, this topology has relatively low bisection bandwidth compared to crossbar-style interconnects, such as those on the SX-6 and p690. Whereas bisection bandwidth scales as the number of nodes ($O(n)$) for crossbar-style interconnects, it scales as the 2/3 root of the number of nodes ($O(n^{2/3})$) for a 3D torus. We have seen from mesh-based systems such as the Intel Paragon, the Cray T3E, and ASCI Red, however, that mesh-based topologies may be quite adequate for scaling to thousands of processors.

The primary benefit of full-bisection interconnects is more likely scheduling flexibility. A single job on a mesh-based topology usually achieves better performance if scheduled on contiguous nodes, while a job on a full-bisection network may be scheduled on any available nodes. Jobs on mesh-based systems may be unable to start despite the availability of enough processors because of fragmentation of those processors. This problem was ameliorated on the T3E through automated job migration; jobs are moved to create larger groups of contiguous processors. Additional benefits of implementing such job migration include system-initiated gang scheduling and checkpoint-restart. Like the T3E, these capabilities will be available in the SV2.

Because of the tightly coupled parallelism of vector processors, high memory bandwidth, high-bandwidth/low-latency interconnect, and scalable systems software, the SV2 has the potential to provide more capability for scientific computation in the near term than any other system we know of, including the SX-6 (see Table 1). Because of the short vector pipes, multiple vector units, memory caching, and distributed memory of the SV2 design, the software optimizations needed for efficiency on the SV2 may be virtually identical to the optimizations needed for clusters of traditional SMPs. For example, modern high-frequency commodity processors, like the IBM Power4, have multiple floating-point units with deep pipelines and aggressive memory prefetching. They require moderately large inner loops with no dependencies, just like SV2 processors. The expected efficiency and capability of the SV2 are simply much higher. Table 2 compares some of the key ratios of the SV2 and follow-on SV2e with the IBM p690 and the Earth Simulator.

Table 1. Comparison of SV2 and SX6 on several climate applications

	SX6 (MFlops)	SV2 (MFlops)
Matrix Multiply	4,000	10,000
Ocean Model		
Kernel A	3,900	4,400
Kernel B	2,500	2,700
Kernel C	1,000	700
Kernel D	1,800	1,400
FE Weather Code		
Kernel A	84	540
Kernel B	2,000	5,900
Kernel C	150	5,300
Eng. Analysis		
Kernel A	2,700	4,900

Table 2. Comparison of High Performance Systems

	p690	**SV2**	**SV2e**	**GS40**
Memory Bandwidth (GB/s) per processor	5.5 (1)*	38	19	32
Peak GFLOP/s per processor	5.2	12.8	19.2	8
Memory Bandwidth per peak FLOP/s (bytes/FLOP)	1 (0.2)*	3	1	4
Processors per node	32	4	8	8
Peak GFLOP/s per node	166.4	51.2	153.6	64
Interconnect bandwidth/node (GB/s)	16	51.2	51.2	16
Interconnect bandwidth/Peak FLOP/s (bytes/FLOP)	0.096	1	0.33	0.25

* STREAM benchmark results for the IBM p690 are much lower than the theoretical peak memory bandwidth. Values based on STREAM results are shown in parentheses. STREAM values for the GS40 and SV2/2e are expected to be much closer to peak.

In the discussions between Cray and ORNL, a plan for scaling the system to thousands of processors has been developed. During Phase 1, Cray will make an initial delivery of an 8-node, 32 processor, 410 Gflops, air-cooled system in late 2002 for evaluation. If the evaluation goes well and there is sufficient interest to go forward, in Phase 2 the system would be upgraded to a 160-node, 640 processor, 8 Tflops, liquid-cooled configuration in 4Q 2003. If the scaling of this system proves strong and SV2e results from Cray show good efficiencies, then a 3072 processor, 60 Tflops system could be built in Phase 3 (4Q 2004) using the SV2e processors. However, through collaboration with ORNL, Cray believes that we could pull in this final delivery date by as much as three months to get a 60 Tflops capability system available to the Office of Science as early as Summer 2004.

Application Performance

ORNL and Cray have developed an application performance plan that will place ORNL staff at Cray and Cray staff at ORNL to make rapid progress on four application areas. These areas are the CCSM2 coupled climate code, the bioinformatics codes FASTA and BLAST, the LSMS electronic structure code, and NWchem from the chemistry domain. Each of these codes and application areas a strategically important to both DOE and Cray and will form the basis of a long-term partnership in many scientific and engineering codes.

Scalability of the Operating System and Parallel File System

Cray has a rich suite of system software that is built on the IRIX operating system software obtained from SGI when Cray was part of that company, as well as on the T3E scalable system software. The goal of the system software is to present a single system image to the user as well as the system administrator. The SciDAC Scalable Systems group and the CCS systems software group will work with Cray to test and enhance the performance, stability, and manageability of the operating system. The SciDAC Scalable Systems Software project headed by Al Geist at ORNL will add the SV2 to the list of architectures that their software suite will support. The 32-processor SV2 system at ORNL will be used to test and validate this support. Cray engineers will be invited to participate in the system software standardization effort going on in this SciDAC project.

Large, long-running jobs, such as climate simulations, require a balanced system view that includes the processor interconnect, the full memory hierarchy, and the I/O subsystems in addition to single processor performance. While climate simulations generate huge output files on the order of terabytes of data, both fusion and chemistry codes depend on out-of-core solutions that stress all aspects of the I/O system. The I/O subsystem is often overlooked by vendors and represents a significant part of the evaluation for production readiness. The CCS has applications that can challenge even the most hardened file systems and many years of experience configuring and managing large, parallel file systems. We will work with Cray to bring their file-system solution to the required high levels of performance and reliability.

Stability and robustness of the computer system are important to many applications because of the nature of long running simulations. Cray offers a unique hardware and software solutions with the SV2 interconnect and the Cray software tools. Scalability questions of the IRIX-based operating systems on the Cray platform are critical to success. We will work with Cray to test and enhance the scalability of their OS.

Application Performance Evaluation

In order for the Cray SV2 to become a viable option for the U.S. response to the Earth Simulator, a vigorous effort must be undertaken to efficiently deploy applications software onto the SV2 platform. This proposed effort leverages the close association of

the PI's to the SciDAC science application projects and other strategic applications software. ORNL will use its expertise in climate, bio-informatics, materials science, chemistry, astrophysics, and fusion to evaluate these applications for usability, scalability and performance on the terascale SV2 system.

The SV2 research plan will leverage another SciDAC project—the Performance Evaluation Research Center (PERC). The goal of the Center is to develop a science for understanding performance of scientific applications on high-end computer systems and develop engineering strategies for improving performance on these systems. ORNL researchers Pat Worley and Jack Dongarra are PIs on PERC and will be using tools such as ATLAS and PAPI to evaluate and tune scientific applications on the SV2.

ORNL was recently awarded a prototype topical computing center for fusion energy applications, extending the scope of the Center for Computational Science. A SciDAC award in fusion has funded the development of new codes and methods that expand the ability to simulate RF plasma absorption and to simulate plasma instabilities and transport. The topical center will also evaluate Monte-Carlo particle-in-cell (PIC) transport codes that are of interest to both the fusion and accelerator communities. These calculations are important for the design of new experiments as well as the diagnosis of experimental results. But the terascale systems in place are not capable of some of the calculations that the SciDAC researchers want to perform. Multi-scale phenomena in the RF problem have been resolved to only a hundred or so waves, missing some of the crucial interactions. The transport codes only compute in 2-D and parameterize the third dimension. For some of the design regimes of proposed fusion experiments and reactors, a full 3-D model will be required. The current resources support only 50x50x160 mode 3D models and 250x250 modes in 2D. Out-of-core solvers are also considered for solving larger problems that are 5 to 10 times available memory. Scientists are working on alternative basis formulation and new physics to ultimately achieve equivalent of 250x250x250 mode resolution in 3D. Hence emerging architectures at the terascale are of great interest and must be evaluated for suitability of fusion codes.

The SV2 is expected to perform well with vector operations supporting the large matrix calculations of the RF code. There may be significant effort required to adapt codes, but several research staff members have worked with vector machines in the past. The current codes also run very well on cache-based machines, so direct comparisons will be possible.

The materials simulation of electronic structure has been a centerpiece of ORNL's scientific computation since the Gordon Bell Awards in 1991 and 1998. Theory, methods, and codes have advanced with computational capabilities. The materials codes were once optimized for vector performance and will be again for this proposed effort. Since the kernel of many of these computations is a dense matrix operation, we expect excellent performance can be obtained on the SV2 as it has for the cache-based machines.

The future versions of these fusion and materials codes being developed for larger simulations will require moving to sparse algorithms. Today's cache-dependent systems do not perform as well on sparse algorithms because the memory accesses do not follow

unit-stride patterns, therefore wasting memory bandwidth fetching pieces of cache lines that are never used. The SV2 provides a "gather/scatter" instruction that allows sparse components of a calculation to be consolidated into a vector register without wasting memory bandwidth on unused loads. Therefore, we expect that the SV2 architecture will provide much better performance on the new generation of codes than will current microprocessor based systems.

After prototyping requirements for a climate topical center, ORNL and the Center for Computational Sciences initiated a Climate and Carbon Research (CCR) focus in 2001. The new DOE projects of the SciDAC program are all using the CCS for computational support in the development of simulation studies and development of new methods. ORNL researchers have been involved with the development of new methods and parallel algorithms that scale to effectively take advantage of new computer architectures. Our collaborators in the scientific community include the major U.S. climate and weather modeling centers. Active projects supported by DOE involve development and improvement of the Community Climate System Model (CCSM2) with NCAR and a consortium of DOE laboratories, investigation of novel programming paradigms and parallel methods for climate modeling, and ensemble forecasting of IPCC scenarios using the DOE Parallel Climate Model (PCM) and the CCSM2. After evaluation of the Cray system, the machine could be useful in these model development efforts and in production runs. Evaluation of a large Cray system would provide useful information to the climate community as well as to DOE for purposes of SciDAC.

The announcement of the performance of the Japanese Earth Simulator in March 2002 has focused evaluation and benchmarking efforts at ORNL and other institutions. This proposed effort will continue this focus in the area of climate modeling, since these are the results available from the GS40. In particular, we will collect and document comparative performance data for

- the fully coupled CCSM2 at both T42 and T85 horizontal atmospheric resolution,
- a stand-alone ocean simulation at 1/10 degree using the POP2 code, and
- a stand-alone, high-resolution atmospheric code and the spectral transform kernel that it uses.

Some of the kernels used in these calculations will be optimized and collected in vendor math libraries for the SV2. This will support future work in the climate, ocean and weather communities.

Scaling the components of coupled climate models like CCSM2 is important on the SV2 as well as on other cache-based machines. Current models are limited in the number of processors that can be used by the data decomposition. Moving to a finer grain parallel decomposition will be required to take advantage of a 25-50 Tflops system. How scaling and vector performance are balanced is new territory in computer design and parallel algorithm choice. There are a number of hardware features of the SV2 that must be stressed to understand this balance. The adequacy of a small "Ecache" between the memory and vector units must be studied. The effectiveness of the torus interconnect and the bandwidth to the nodes in a distributed fashion all effect the question of the correct balance for a given application.

Evaluate Program Development Tools

Cray offers a wide set of tools for software development, including compilers, debuggers, and performance-analysis tools. Relative to other vendors, the Cray compilers have an impressive list of features. Cray's compilers have long been recognized for both excellent optimization performance and features that are important to the scientific community. All of the SV2 compilers, including C, ANSI C++, and Fortran, have a unified optimizing back-end that attempts a wide range of optimizations, including loop unrolling, software pipelining, and branch tuning based on run-time flow analysis. Both Fortran and C compilers claim full OpenMP support. One of our first goals will be to evaluate whether the advertised features actually lead to efficient and robust executables.

The SV2 compiler technology is on track for exploitation of the MSP, and we can expect good vectorization from most of the routines in the climate code CCSM2. MSPs have been available in the Cray SV1 for several years so the compilers have had time to mature. An area of concern for the compiler writers is the land model. It consists of many subroutine calls and "if tests" inside the main loop, both of which inhibit vectorization. A rough argument based on an expected percentage of peak performance at 25% for average codes suggests that the SV2 will provide a 6-10x speed up over the IBM p690 with Colony system that ORNL currently uses. This assumes that code vectorization can be achieved.

The SV2 also opens up possibilities for alternative programming paradigms, something that ORNL has pioneered with application codes. An MPI-OpenMP hybrid-programming paradigm has been used and found to be most effective for the cluster machines such as the IBM p690 system. But the Cray systems also support a single-sided, get-put style of distributed communication in the SHMEM library. This support comes from globally addressable memory and is potentially a lower-latency, lighter-weight protocol for access to distributed data. The SGI systems have been shown to support this style, and the MLP (multi-level parallelism) style of Jim Taft at NASA and the MPI-2 one-sided computation are interesting possibilities for the SV2. These protocols will be explored in the context of the climate codes and basic performance tests. Some of these new tests, analogous to the MPI ping-pong and the STREAMS benchmarks, must be developed in support of this effort.

Cray has chosen the de facto industry standards for parallel debugging and MPI performance analysis: Etnus Totalview and Pallas Vampir, respectively. No similar standard yet exists for parallel CPU-performance analysis, however. Cray offers its own graphical tools, Performance Visualizer, for application-level performance profiling, and Performance Manager, for system-level performance tuning. We will evaluate the effectiveness of these tools in light of those available on systems from other vendors.

Table 3. Plan for a three-stage approach in collaboration with CRAY

Applications optimizaiton	Vendor and Math Libraries	Operating system scalability	System robustness and production readiness	Training
Stage 1: 400 GF SV2 system, two air cooled cabinets				
a) Vectorization of kernels • Spectral (PSTSWM) • CRM physics b) CCSM2 at T85 scoping • Vector performance • Characterization of scalar performance • Lnd plan c) Early science run of 1/10degree POP ocean	Blocked 1-D FFT Legendre as matrix multiply MPI performance of collective ops	Evaluate need for specialized routing chips in SV2e configs.	Test production usefulness with POP runs Attach IO and archive systems	CRAY trains several apps personnel in effective use of the machine ORNL provides feedback on documentation
Stage 2: 8-10 Tflops	SV2 system, 10 liquid cooled cabinets = 640 processors			
a) Decomposition studies • Vector length vs. 2-D decomp • Memory optimization b) Ensemble runs of CCSM2 at T85 c) High resolution ocn runs and T170 atm	Utility layer communication routines: Halo, transposes, etc.	Develop scalable OS from 8 to 9 cabinets	Stress test and development of fault detection and recovery	General training of the user base
Stage 3: 60 Tflops	Fully configured SV2e system = 3200 Processors			
CCSM2 (T85) at 30 years/day POP 1/16 degree ocean simulations Weather model performance T341 atmospheric simulations and re-evaluation of physical parameterizations	SciDAC ISIC products and libraries made available and optimized.	Full OS scaling	Production testing Job management With checkpoint restart Maturity and stability	Innovative use seminars on computational science given by CS and apps teams

The table shows a preliminary plan for our collaboration with CRAY on climate codes. The areas for collaboration are listed as the table column headings and are taken here as task areas. We will prepare detailed tasks for other grand challenge areas that have been identified in this white paper.

Budget

The table below shows the 18month and the 30 month options based on Cray SV that would allow an interesting and timely response to the Earth Simulator.

	Phase 1 (SV2) FY02	Phase 2 (SV2) FY03	Phase3 (SV2e) FY04	Pricing
30 Month Option	2 Air Cooled Chassis (400 GF)	10 liquid Cooled Chassis (8TF)	25 liquid cooled Chassis (60TF)	$207M
18 Month Option	2 Air Cooled Chassis (400	50 liquid Cooled Chassis (40TF)		$288M

The price is based on an evaluation technology partnership with Cray. The pricing does not include the required storage area network (~$25M) and the wide area networking (~$75M) that would be needed to engage the broad scientific community. Both SV2 and/or SV2e based solutions requires immediate engagement with Cray. Detailed budgets will be prepared in consultation with DOE. Based on our study, the planning numbers for a timely response to the Earth Simulator based on Cray SV is approximately $400M for the 18 month option and $300M for the 30 month option. Either option requires aggressive engagement of Cray immediately to guide the development and deployment. The Center for Computational Sciences at ORNL has worked with Cray over the last couple of months, specifically on the Earth Simulator Response, to build a strong working partnership that is critical to ensure the success of such a response.

We have the building infrastructure in place to house the Cray SV2, we have the network connectivity in place to make the resource available broadly, we have the SciDAC projects in place as the foundation to mount such a response, we have the experience, and most importantly, we have the partnerships with Cray, our sister laboratories and universities in place. The Center for Computational Sciences and our partners stand ready if we are called upon to respond to the challenge posed by the Earth Simulator.