UCRL-JC-134806
PREPRINT

An ASCI Terascale Simulation Environment Implementation

Mark Seager

This paper was prepared for submittal to the
Mannheim Supercomputer '99 Conference
Mannheim, Germany
June 11, 1999

May 20, 1999

Lawrence Livermore National Laboratory

An ASCI Terascale Simulation Environment Implementation[1]

Dr. Mark Seager
Lawrence Livermore National Laboratory
POBOX 808, L-60
Livermore, CA 94550
seager@lnl.gov, 925-423-3141
UCRL-JC-134806

ABSTRACT

The ASCI Blue-Pacific Sustained Stewardship TeraOp/s computer demonstrated 1.2 TeraOP/s in September 1998. Two thirds of the system was delivered to the Lawrence Livermore National Laboratory in early October 1999 and the remainder in December 1998. Since that time ASCI scientists have been performing "full-system" runs of remarkable scientific value in Quantum Chemistry, Biology, Molecular Dynamics, Turbulence, Neutron Transport and Astrophysics. In addition, there has been an intensive ASCI 3D physics simulation development effort. The SST also supports a large production workload. This paper focuses on the architecture of the Blue-Pacific SST, the integration of this platform into a full simulation environment, full-system science runs and a discussion of the operational model for the SST.

1 Introduction

In support of the comprehensive test ban treaty, the United States Department of Energy (USDOE) has identified the need to move quickly from a nuclear weapons stockpile stewardship program relying on underground nuclear tests to one that has more dependence on simulation methods. Computation in conjunction with above ground experiments (AGEX) and stockpile surveillance are now major components of the Stockpile Stewardship Program (SSP). The Accelerated Strategic Computing Initiative (ASCI) was identified as the methodology for accelerating the United States High Performance technical computing marketplace to meet the demands of SSP. The ASCI program now has many facets: 1) Applications development; 2) Platforms; 3) Problem Solving Environment (PSE); 3) Alliances; 4) Numerical Environment for Weapons Simulation (NeWS); 5) Distributed and Distance Computing (DISCOM[2]). These program elements are all driven by stockpile stewardship programmatic objectives. The three dimensional, full system, refined physics applications are being constructed to meet the demanding modeling and simulation requirements coming from SSP time and technical capability objectives. From this, application ultra-scale computational requirements are derived that span 1.0 TeraOP/s (trillion of floating point operations per second) by late 1996 to 100 TeraOP/s in the 2004 timeframe. The platform strategy leverages the business plan multiple vendor partnerships to bridge the gap from 1 to 100 TeraOP/s in an astonishingly short eight years. This is roughly 2.5x the rate of performance improvement in underlying microprocessor technology! In order to accommodate this highly accelerated pace, multiple steps have been identified:

1) Option Red (1 TeraOP/s in 4QCY96) at Sandia National Laboratory;
2) Option Blue (3 TeraOP/s) at Lawrence Livermore and Los Alamos National Laboratories;
3) Option White (10 TeraOP/s in 1QCY00) at Lawrence Livermore National Laboratory;
4) 30 TeraOPs (30 TeraOP/s in 2QCY01) at Los Alamos National Laboratory;
5) Option Purple (50+ TeraOP/s in 2QCY02);
6) 100 TeraOPs (100 TeraOP/s in 2004)

The funding lines for these machines are about three to five years long. Procurement takes about 18 to 24 months from initiation of the procurement effort to initial delivery. Each of the procurements is lead by

---

[1] This document is an account of work sponsored by an agency of the US government. Neither the US government nor UC, nor any of their employees makes any warrantee or assumes any legal liability for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe on privately owned rights. Reference to any commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the US government or UC. Views and opinions of authors expressed do not necessarily reflect those of the US or UC and shall not be used for advertising or endorsement.

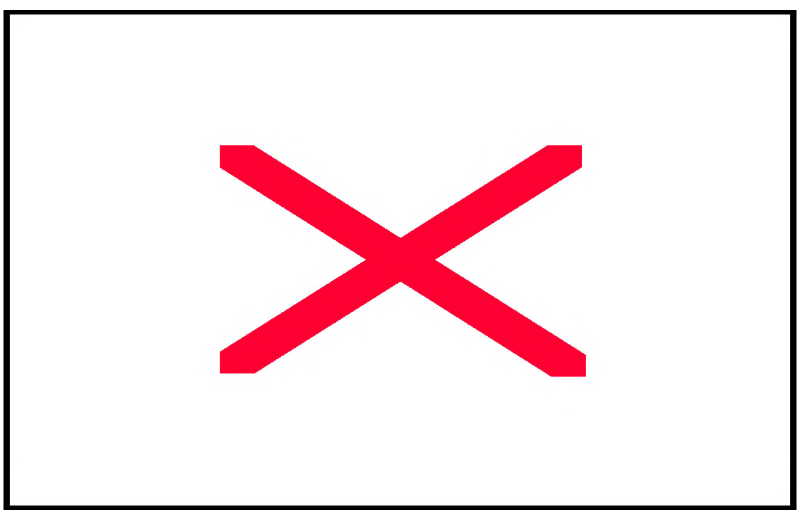
Work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

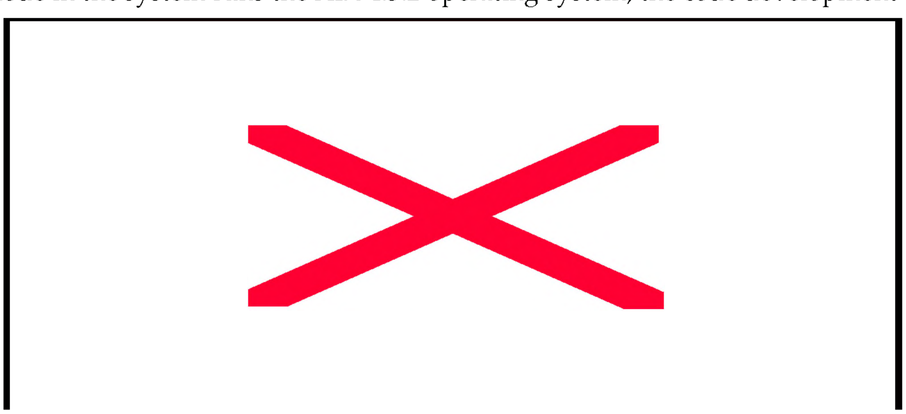
one of the three USDOE Defense Program (DP) Laboratories in the sprit of "One Program – Three Laboratories." Option Red is a partnership between Sandia National Laboratories and Intel that delivered over one TeraOP/s performance on MPLINPACK in December 1996. Option Blue led to two partnerships LLNL with IBM (Blue-Pacific) and LANL with SGI (Blue-Mountain). These partnerships demonstrated terascale performance with over 3.0 TeraOP/s peak in September-October 1998. Option White is an extension of the IBM LLNL partnership with IBM and is scheduled to deliver multiple sustained TeraOP/s performance with over 10.2 TeraOP/s peak performance in 1QCY00. The 30 TeraOPs procurement being lead by LANL just released a Request For Proposals (RFP) on May 17, 1999. Initial discussions have begun on the Option Purple procurement.

The Blue-Pacific contact with IBM has three deliveries: Initial Delivery (ID); Technology Refresh (TR) and Sustained Stewardship TeraOP/s (SST). The ID system was delivered within 90 of days receipt of contract, September 1996. It consisted of a 512-way THIN2/WIDE2 nodes that were evenly split between classified, Secure Computing Facility (SCF), and unclassified, Facility for Advanced Scalable Technology (FAST) computing. The ID had a peak of 136 GigaOP/s (counting both halves) and 98 GB of memory and 2.3 TB of global disk. Although split in two, this delivery represented a boost of 2x in capability for the SCF over the previous generation MPP, the Meiko CS-2. The ID system also extended major features (e.g., one CPU/node and operating system; high speed, low latency MPI communications; local and global disk arrays) of the CS-2 programming model (there were minor differences such as Solaris vs. AIX, PGI compilers vs. IBM compilers, etc). In March and April of 1998 the ID system was upgraded with Silver node technology. The TR delivery had 918 GigaOP/s peak, 196 GB memory and 2.3 TB of global disk. This delivery provided the single largest boost in computational power ever received in the history of LLNL: a whopping factor of 6.6.

2 ASCI Blue-Pacific SST Architecture

The ASCI Blue-Pacific Sustained Stewardship TeraOP/s ultra-computer has a peak capability of 3.889 TeraOP/s and 2.6 TB of memory, 62.5 TB of RAID5 global parallel file system space and 17 TB of local disk. The SST is a hyper-cluster of IBM SP Silver Nodes. The hyper-cluster consists of three sectors (known as S, K and Y, which together spell SKY) each of which is an IBM SP system. The SST has a two stage interconnect. The first stage (within each sector) is based on the IBM high speed (TB3) switch. The second stage (connecting the sectors) is based on High Performance Gateway Nodes (HPGN). High Performance MPI communications between any two nodes on the SST is supported. The software environment on the SST is the IBM Troutbeck release of PSSP. This contains Loadleveler for batch job control, GPFS for global file system, MPI communications and switch support and system administration tools for providing a single system image. Since each node in the system runs the AIX 4.3.2 operating system, the code development environment includes all the standard workstation development tools (F90, C, C++ compilers, CVS source code control, etc.) as well as the TotalView multi-processor debugger for MPI+Threaded applications.





Each SP sector consists of 488 Silver nodes and 24 links to the HPGNs. Each SP sector has its own global parallel file system (GPFS) served from 56 file servers. The total IO bandwidth for the SST is 6.6 GB/s to the global file system and over 11 GB/s to the local file systems.

Jobs can be scheduled on each SMP, on each SP sector and over the entire SST hyper-cluster.

Each SMP in the hyper-cluster is an IBM Silver Thin node. The Silver node has four PowerPC 604e microprocessors running at 332 MHz. The PowerPC 604e has quite good 32b and 64b numerical performance and the memory bandwidth on the node is superior. On the Livermore Fortran Kernels (LFK), the Silver node gets 114 MFLOP/s geometric mean. This benchmark is representative of Legacy 2D codes at LLNL and is typical for codes with good L2 cache utilization. Larger codes that don't have quite as good L2 cache utilization see about a 1.5x to 2.0x speed up over the POWER2 THIN2 node running at 66.7 MHz. This performance is due to the outstanding memory bandwidth of 1.3 GB/s available on each Silver node. This represents a memory to peak performance (B/F ratio) of 0.51. However, we have observed that the Silver node delivers a very high fraction of memory bandwidth to applications (about 200 MB/s to a single thread and 600 MB/s to four threads). So the delivered B/F ratio is 200/114 = 1.75.

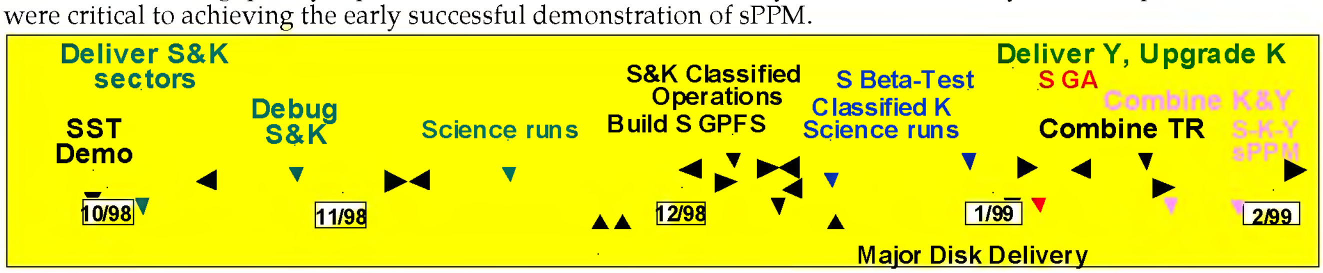
The Silver nodes also have a TB3MX adapter that connects the node to the first level (High-Speed SP) switch. This link is 150 MB/s bi-directional (300 MB/s total) and delivers about 80 MB/s MPI single task point-to-point bandwidth. The delivered MPI latency is about 23 microseconds.

In addition to the SST platform, which is in general availability (GA) status on the Secure Computing Facility network, the ASCI program enjoys a 336 Silver node Combined Technology Refresh (CTR) system on the unclassified network. The ASCI alliance partners and ASCI researchers doing unclassified code development utilize this platform.

3 Blue-Pacific SST/CTR Integration Strategy and Schedule

The Blue-Pacific SST was demonstrated at the IBM SP Laboratory in Poughkeepsie, NY on September 27, 1998. This milestone was achieved three months before the contractual requirement of late December 1998. At that time the SST delivered 1.2 TeraOP/s performance on the sPPM turbulence research code. Additionally, another run was accomplished with the largest sPPM grid to date: 70.8 billion zones. The previous sPPM record was 1.0 billion zones. This jump in delivered capability clearly demonstrates the quantum leap forward the SST brings to the ASCI program.

The strategy with demonstrating the SST at IBM's plant in Poughkeepsie was to have the IBM development environment as close to the sPPM demonstration environment as possible. This facilitated the quick resolution of problems as they were identified. In addition, recycling defective parts through IBM manufacturing quickly repaired hardware failures due to early life infant mortality. These aspects were critical to achieving the early successful demonstration of sPPM.



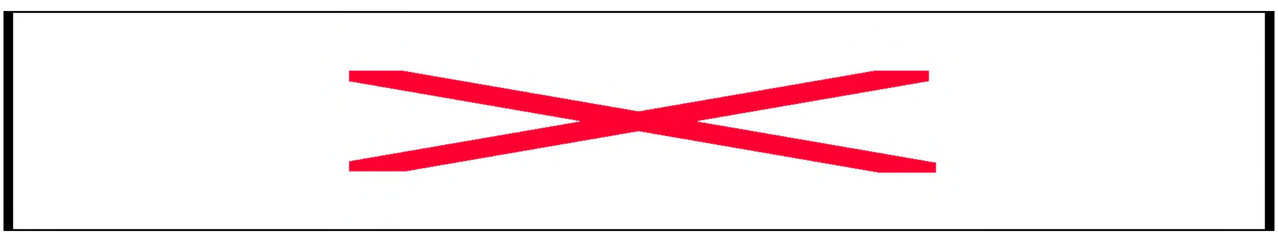
Once the sPPM demonstration was accomplished, two of the three sectors were delivered to LLNL for shakedown and programmatic full-system use. One sector, Y, was left behind in order to accelerate Troutbeck scaling work. While the S-K sectors were at LLNL it was discovered that a process change in manufacturing of PowerPC 604e parts yielded a "cache paradox" error that was only identified with LLNL applications after full system runs spanning over 8-16 hours. In addition, several severe problems were found in the LoadLeveler, GPFS and switch control software. These factors contributed to the nearly one month S&K debug period. After the S&K sectors stabilized, full-system runs were accomplished with many applications codes of interest to the ASCI program. The strategy for the full-system runs was to make as good of use of the system resources as could be managed while minimizing the user community impact of a maturing HW/SW base.

As the S&K sectors stabilized and full-system runs were ongoing the global file system was added and configured. This required intense coordination between IBM development and site engineering, LLNL systems and user groups. In addition, the machines were transitioned to the classified network and one system put into beta-test for the ASCI classified code development community. The rational for this was to incrementally increase the number of nodes available to the user community while minimizing the disruption caused by introducing large amounts of capability to the computing environment. The K sector continued its science run status and completed the largest Monte Carlo calculation ever attempted.

Just before the end of CY98, IBM finished PSSP 3.1 (Troutbeck) scaling and delivered the Y sector with Troutbeck. At this point we upgraded the K sector to Troutbeck and initiated another round of dual sector stabilization and global I/O configuration and tuning. After initial steps of K-Y on Troutbeck stabilization were accomplished, the sPPM demonstration was replicated at LLNL before the end of January 1999. This was quite a feat because the S sector was still at a previous AIX and PSSP software release.

Another strategy of the SST integration project was to always provide more nodes to the ASCI program at each step. So the TR machine was not retired until **after** Sector S went GA. The TR was decommissioned, memory upgraded from 512 MB/node to 1.5 GB/node and the local disks were doubled. After the classified portion of the TR was sanitized, it was combined with the unclassified portion of the TR. In addition, more global disk was added to bring the total to 20 TB. This operation required the complete de-install of both systems and moving the frames to the other side of the room and re-installing the frames from scratch. The machine, now dubbed the Combined Technology Refresh (CTR) machine, was then upgraded to Troutbeck. All this work was accomplished while upgrading K&Y and dealing with the fall out of a new 488 node system in GA status (Sector S). This period clearly represented the peak workload for the integration team and stretched everyone to the breaking point. However, good relations between LLNL and IBM were maintained due to the over two years of developing the partnership before the SST demonstration and delivery.

Within two weeks of GA status Sector S was running at full capacity. This user load exceeded our expectations and could be attributed to two factors: 1) the users were ready and waiting for more nodes; 2) the programming environment on Sector S was identical to the TR machine (except for the details of the global file system).



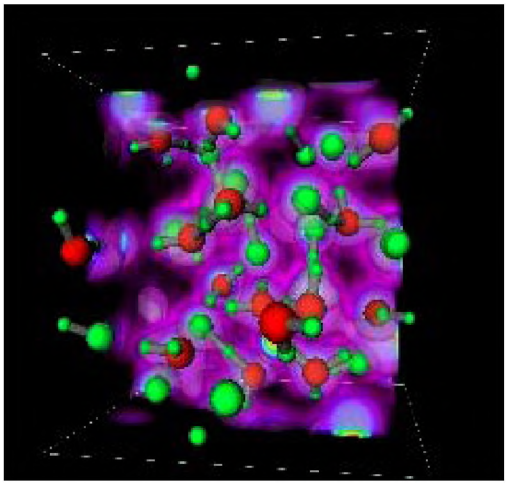
After the CTR went into GA status, K&Y were stabilized to the point they could withstand the rigors of the official acceptance test. The formal acceptance test consisted of running a synthetic workload of over 15 unclassified, but export controlled applications with a total of over 50 different data sets (problems). These sectors sailed through the acceptance test with an availability of over 99%. Over 1,900 jobs completed successfully. Only one or two jobs did not complete due to system errors. In addition, several programmatic users with time critical applications that could not get enough time on S because it was saturated, ran on K&Y during the acceptance test. Given the recent experience stabilizing Troutbeck, we were very pleasantly suprised at the swift progress through the acceptance test. Once the K&Y machines passed acceptance they were immediately made available to the ASCI program (modulo security plan approval) and sailed through a two-week beta-test approval. With K&Y in GA status, we moved over 14 TB of user data generated on Sector S to K&Y and retired Sector S from GA status and upgraded it to Troutbeck. At that point USDOE directed LLNL and LANL to stand down for security training and upgrades. That process took over a month out of the SST integration schedule. However, after we were able to stand back up again, the Sector S was ready to go and placed in GA status again.

With all three sectors in GA status with Troutbeck, we could again begin classified full-system runs. To date, several full-system applications have been run including one utilizing over 5,000 processors on 1,250 nodes that solved a problem with 364 Million Zones. Before the SST arrived the largest runs attempted had 1's to 10's of millions of zones.

4 Full-System Science Runs

The following are a few unclassified examples of applications running on the SST today in "full-system" mode. Although not yet routine, "full-system" runs are becoming common (weekly). In addition, a new paradigm at Livermore Computing has been developed to facilitate the "full-system" operational mode, this is described in the next section.

4.1 Quantum Molecular Dynamics



Francois Gygi, Giulia Galli and Francis Ree of LLNL recently completed a "full system" run of the JEEP code on the ASCI SST ultra-computer. Insensitive high explosives are preferred over sensitive explosives because of their safety and performance. They generally produce detonation products containing HF and H_2O, both of which are strong hydrogen-bonding molecules. Relevant experimental data (e.g., shock wave data) on HF are almost nonexistent because of its corrosive nature. The lack of shock-wave data, coupled with the pronounced sensitivity of detonation properties to effective HF-HF and HF-H_2O intermolecular interactions warrants an accurate theoretical treatment of HF-HF and HF-H_2O potentials. To substitute relevant shock experiments, quantum-level simulations of HF- H_2O mixtures were required. These simulations of unprecedented scale were performed on the ASCI Blue Pacific SST ultra-computer, using the JEEP first-principles molecular dynamics code. Quantum simulations are essential to correctly describe hydrogen-bonded liquids at high pressures and temperatures, where more conventional models have limited predictive power.

Quantum Molecular Dynamics methods simulate the behavior of materials at the microscopic level, the only input being the identities of the atoms and the laws of quantum mechanics. Using these methods, the computer can be turned into a virtual laboratory where one can follow the trajectories of atoms as electrons form and break chemical bonds. The simulations involved 600 atoms, with 1,920 electrons, described at a very accurate level of quantum mechanical theory.

These simulations provide crucial insight into the properties of HF/H2O mixtures at high pressures and temperatures, which were found to be complex fluids exhibiting a variety of proton transfer mechanisms and containing several ionic molecular complexes.

The simulations and ran on 3,840 processors of the SST. This research was presented at the prestigious American Physical Society's Centennial meeting in March 1999. The inset figure was produced using visualization software developed at LLNL, and shows atoms and electronic density in a fragment of the simulated sample.

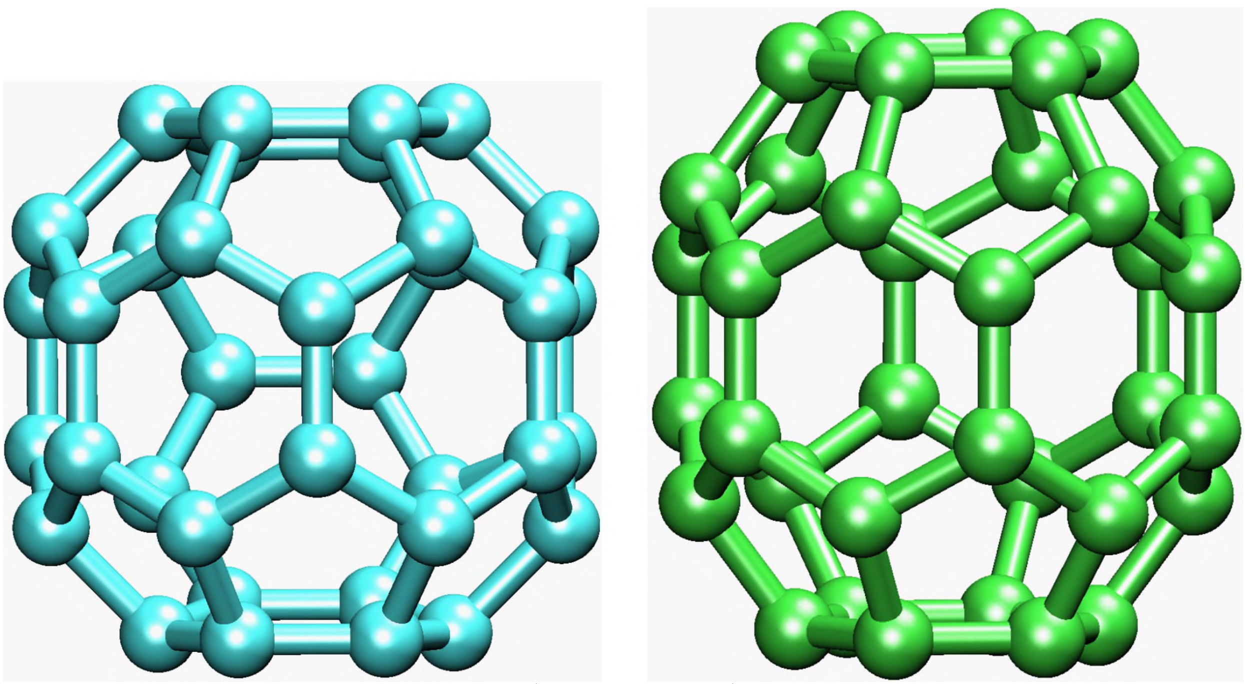
4.2 First Principles Quantum Chemistry

Mike Colvin from LLNL in collaboration with J. Grossman, University of California at Berkeley and C.L. Janssen, Sandia National Laboratory at Livermore used the ASCI Blue Pacific SST ultra-computer to perform the largest first-principles quantum chemistry calculations ever done. Quantum chemistry involves predicting chemical properties by numerically solving the Schrodinger wave equation without recourse to empirical data. The emerging role of quantum chemistry as an important tool in chemical and biological research was illustrated by the 1998 Chemistry Nobel Prize award to one of the founders of quantum chemistry, John Pople. However, to fulfill its ultimate promise, quantum chemical calculations must be made feasible for much larger chemical problems, which will require the use of the very largest parallel supercomputers. An important step towards this goal has been achieved by the implementation on ASCI's Blue Pacific SST ultra-computer of a parallelized quantum chemistry program (MPQC) developed at Sandia by Drs. Curtis Jansen, Ida Nielsen, and Ed Seidl.

The initial application of MPQC is to two chemical problems that provide important validations of less accurate chemical calculations and demonstrate the utility of large-scale quantum chemical calculations for applications in materials science, biochemistry, and environmental cleanup. These capabilities will play an important role in LLNL's new Computational Biochemistry effort led by Dr. Michael Colvin, that will apply quantum chemical simulations to improve understanding of anticancer drugs, environmental carcinogens, and the fundamental biophysics of DNA damage and repair.

Both calculations to be performed on the ASCI SST computer involve second-order Moller Plesset Perturbation Theory (MP2), a quantum chemical method that has proved very accurate in predicting the structure and energy of molecules, but has been impractical for larger chemical systems. The first application on the ASCI SST is the MP2 calculation of the binding energy of the water trimer using an unprecedented 1329 basis functions. This is the largest no-symmetry MP2 calculation ever performed and can be compared to an earlier record of 640 basis functions run on a 48-processor IBM SP2 last year in Sweden, which involved 40 times fewer floating point operations.

The second application to determine the three dimensional structure and electronic state of the C_{36} "Buckyball" one of the smallest stable members of the Buckminsterfullerene family of compounds. These compounds have promise for a vast array of applications, from high temperature superconductors to precise delivery of medicines to cancer cells. The experimental discovery of C_{36} was just announced last July and quantum chemical calculations by Dr. Jeffrey Grossman and co-workers at UC Berkeley narrowed down the possible structures to the two shown in the figure (D_{6h} and D_{2d}). Although there is tentative experimental data favoring the D_{6h} structure, the existing theoretical results indicate that the D2d and an unpaired-electron state of D_{6h} are slightly favored. The goal of these calculations is to more definitively establish which is the actual structure and electronic state of C_{36}.

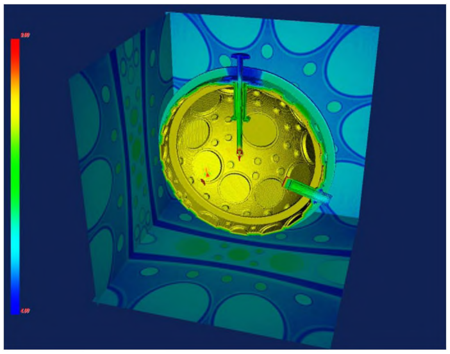


Two possible structures for the C_{36} buckyball (D_{6h} right, D_{2d} left) suggested by the experimental data and earlier calculations. Accurate quantum chemical calculations are being performed on the ASCI SST ultra-computer to determine the most stable structure and spin state.

4.3 Transport Methods

Radiation transport is central to many Laboratory physics applications, including nuclear weapons, inertial confinement fusion, plasma processing, and combustion. Many production codes in ASCI program spend upwards of 80% of their time doing radiation transport calculations. A large team lead by Peter Brown and Britton Chang worked to investigate novel parallel numerical methods for radiation transport.

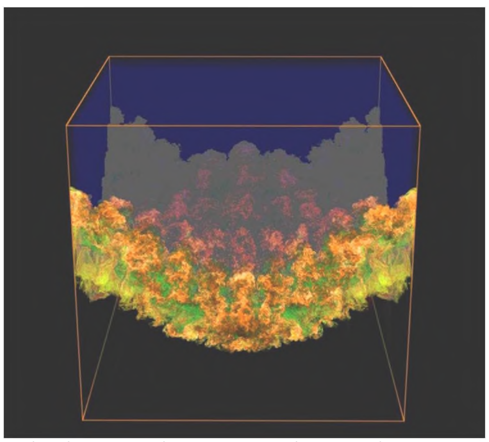
One of the highlights was the "full system" run of the Ardra neutron transport code. This code is parallel in all phase space variables and is used to explore novel algorithms for neutron and radiation transport. For example, the team is investigating the use of the above multigrid solvers to speed up the standard Diffusion Synthetic Acceleration scheme. Ardra also allows the use of new "harmonic projection" method to obtain a ray effects-free P_N solution.



The full-system SST calculation employed more than 160 million zones with over 15 billion unknowns. It took 27 hours to solve on 3,840 processors of the SST. The purpose of the calculation was to simulate the flux of fusion neutrons that comes out of the NOVA target chamber. The inset picture displays the complex geometry involved in the description of the NOVA target chamber. Isosurfaces in the figure display a cut-away view of the interior of the chamber. In particular, the target assembly is shown along with the interior geometry of the chamber. The colors on the isosurfaces represent neutron scalar flux values for the most energetic neutrons, with red denoting the regions of highest flux and blue the lowest.

This calculation also demonstrates the fact that high-fidelity simulations require complex three-dimensional (real world) geometries, span a wide range of distances (e.g., the test chamber is 20 ft in diameter and the target pellet is less than 1mm in diameter) that must be highly resolved and have a large number of materials (from Air at 10^{-5} torr to gold in this problem) that have widely varying material properties. Each of these aspects represents its own uniquely difficult computational challenge.

4.4 *Turbulence Research*



Laboratory researchers Art Mirin, Ron Cohen and others at LLNL collaborated to run the largest Richtmyer-Meshkov instability calculation ever. This instability occurs, for example, when a shock passes through an interface of two fluids of differing density. This three-dimensional simulation was carried out with the simplified Piecewise Parabolic Method Code (sPPM), a reduced implementation of the algorithm developed by Colella and Woodward, on a 2,048x2,048x1,920 mesh, using 960 nodes, arranged in an 8x8x15 domain decomposition. This 8 billion zone simulation ran for just over a week, and attained roughly half a TeraOP/s of sustained performance over that period on this highly complex simulation. Over two TeraBytes of graphics data were produced in over 300,000 files. The inset figure shows a volume rendering of the entropy at the conclusion of the calculation. The 3D simulation shows marked contrasts from corresponding 2D simulations, in which low-to-moderate–resolution runs indicate a preferred progression to larger, rather than smaller, scales as the simulation progresses. The high resolution allows elucidation of fine-scale physics; in particular, when compared with coarser resolution cases, a possible transition from a coherent to a turbulent state with increasing Reynolds number was observed.

5 TeraScale Simulation Operational Model

5.1 *Challenges and Mission*

The key challenge facing Livermore Computing in fielding the fastest ultra-computer in the world is that this scale of computing requires a new operational model in order to be successful. Thus, the organizational operation model must transform itself from what it has always been, namely a large, but fundamentally traditional, scientific computing center, into something much more like an *experimental research facility*. This new role can be called a center for scientific simulation. The change in emphasis *from computing to simulation* is essential. While the former is reasonably well understood, the latter represents a significant challenge: the development of a seamless partnership between the ability to generate data and the ability to assimilate it at the terascale rate.

Three factors are at work here. First, the number of bytes of data generated requiring analysis over the next decade is on the order of a few quintillion (10^{18}) bytes. Managing this information presents a challenge, associated with the sheer *quantity of information*.

Second, the scientific applications being developed today promise a level of physical and numerical accuracy that can be viewed to be more like a scientific experiment rather than a numerical simulation.

This requires that the supporting environment be sufficiently sophisticated to allow for adequate analysis of the data. Thus, there is a change in the *quality of data*.

Third, the size and computational demands of the scientific applications are growing faster than the rate at which the computers are growing in capability. In other words, it takes *longer to run a full-system production calculation*. To exacerbate this situation, consider that the mean time to failure of a 6,000-processor computer is between one-fifth to one-twentieth of the time it takes to complete a full system science calculation. Both of these factors necessitate sophisticated software such as gang schedulers and checkpoint-restart requirements that must be developed and implemented on these large processor machines.

Taken together, these refinements in quality and quantity of data and in the time it takes to generate and understand this information have changed "the rules of the game" completely. More of the same is not the answer, radical transformation is. In fact, the running of a three-dimensional scientific application is moving from the regime of numerical simulation to one more like physical experimentation, demanding disciplined coordination and teaming between system administrators, operators, visualization experts and other center service providers with the code developers and analysts, the traditional customers. This will be the simulation environment of the future. This transformation requires a lot of work or it will not happen. Livermore Computing, as part of the ASCI, began that transformation in 1997 and continues with it today.

5.2 *Technical Objectives*

The technical objective is to construct an environment to coordinate two complementary elements (1) the most advanced computers aggregated in configurations of such formidable capability that their footprint, power, and scaleable software requirements (O/S, tools and applications) will be unequaled outside of the DOE Stockpile Stewardship Program, and (2) Data and Visualization Corridors (DVCs) for the management, transmission and comprehension of the vast data sets generated.

At LLNL, there are a variety of complementary computing program elements. On the Stockpile Stewardship Program (SSP) side are Stockpile Computing (SC) and ASCI. There is also Multi-programmatic and Institutional computing (M&IC) consisting of multiple programs desiring to have access to high performance computing. Stockpile Computing traditionally provides support for the broad operational, integration and services base of the Center, including networking. In fact, it represents the foundation upon which ASCI and M&IC build.

It is neither possible nor advisable to consider these elements as independent of one another. In fact, they are potentially complementary, but only if they are coordinated to create a coherent environment. Here, the focus is on how they are coordinated.

5.3 *Five Strategies*

Five primary strategies are employed to coordinate contributing subprogram elements, so that the simulation environment develops according to a sensible plan and on a reasonable schedule. These strategies are

1. Enterprise Planning and Systems Engineering. In late 1997, LC began developing an enterprise planning and systems engineering approach to running the terascale computer center. As part of this process, it was recognized that some mechanism was needed to pull together the contributions from the various subprogram-funding elements. This was accomplished through a number of End-to-End (E2E) Blueprints for major services (such as Parallel I/O, Data and Visualization Corridors, Terascale Systems Integration, Shared File Services, Distributed Resource Management, etc.). These plans combined requirements gathering, architecture design and tradeoffs, procurement, integration and operations into one plan, called a Blueprint. Since they span all aspects, from requirements gathering to production operations, they are called End-to-End. This explicit kind of planning was conceived in FY98, will mature in FY99 and will become the backbone of the simulation environment in FY00. It is impossible to overestimate the importance of these plans to achieving the simulation environment. To discuss this approach in another way, this kind of planning represents a crosscut across the complementary funding elements described above, unifies their purpose, and creates coherent, coordinated, operational products.

In order to make this concept clearer, we will describe two of the Blueprints briefly. The I/O Blueprint determines the requirements, architecture, procurement schedule, integration schedule, and operational requirements of the parallel I/O network and storage devices between the (1) ASCI 3.9 TF Platform, (2) tertiary storage environment and (3) rendering engine (a triangle, see illustration below). The requirements come directly from SSP and the I/O rates determined (for instance, three independent four-way-parallel streams each at 32MB/sec sustained to tape for the storage component) meet those requirements. The Blueprint was designed in FY98 and will be largely integrated in two phases, May and August 1999. It represents a crosscut, as it pulls together hardware funding for networks and tertiary storage, Stockpile Computing support for operations and integration, and ASCI PSE support for HPSS into one stream providing one coordinated product. The figure below shows the architecture for FY99.

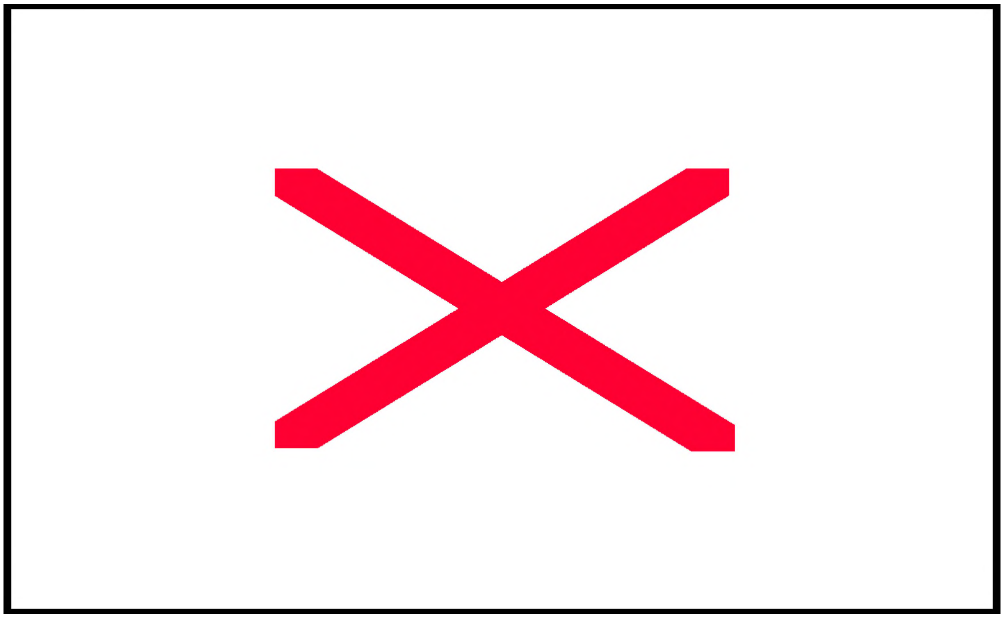


Figure 1: Interlocking Blueprints. Depiction of the Interface between three Blueprints (1) ASCI Platform Integration, (2) I/O Blueprint and (3) DVC. Bandwidth requirements explicitly called out.

As a second example, the Data and Visualization Corridor (DVC) Blueprint provides the architecture for the Assessment Theater and other elements of the visualization environment (networks, rendering engines, office displays). The Blueprint also coordinates PSE deliverables, such as tools that are being developed to interactively navigate the generated data and select subsets to analyze. The Assessment Theater represents the highest-end user DVC interface. The DVC Blueprint picks up responsibility at the rendering engine (the place where the I/O Blueprint terminates) and carries the data (both analog and digital) to user offices and to the Assessment Theater. The B132 Assessment Theater was installed in December. The theater includes a 3x3 array of state-of-the-art projectors to achieve extremely high resolution and superior image quality on a 3,840 x 3,072-pixel screen (to increase to 5x3 projectors or 6,400 x 3,072 pixels in FY 1999). The Assessment Theater is connected to the LC complex via the Laboratory's fiber optics infrastructure.

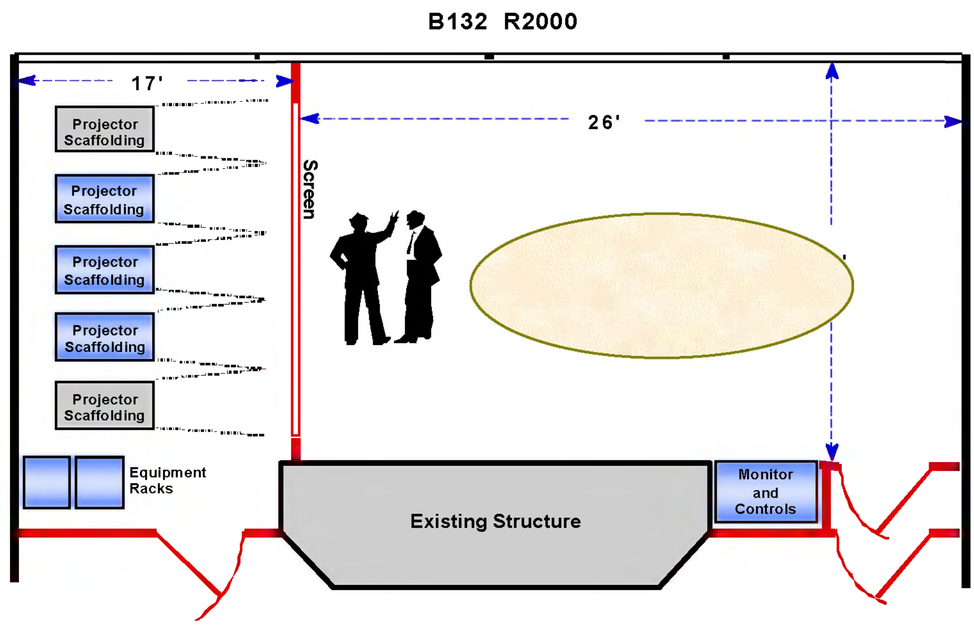
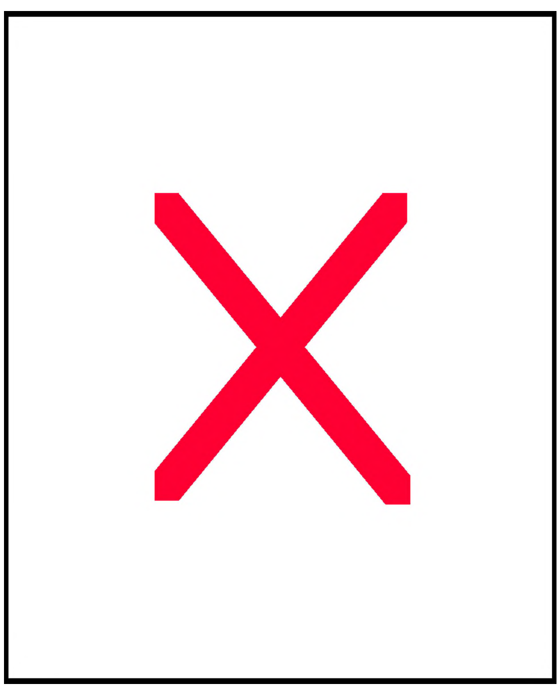


Figure 2: B132 Assessment Theater was brought into service on schedule in December 1998. Theater features a 3x3 array of projectors (expanding to 5x3 in FY99).



The Assessment Theater was used in 1998 and early in 1999 to visualize the results of major calculations completed on two sectors of the SST, including classified program applications, human genome, *ab initio* molecular dynamics and turbulence calculations.

A record shattering 8 billion-zone sPPM turbulence calculation is illustrative of the renderings viewed on the large screens. This calculation required about two weeks of machine time on two sectors of the SST system, and demanded 24x7 coverage by systems administrators and by the physical scientists running the code in order to complete it. The visualization required over 2.0 TeraBytes of data on 300,000 files. All components of the DVC were utilized to perform this calculation, storage, visualization and analysis. Substantial planning was required to get all components of the system ready in time for this calculation. As the calculation proceeded, close coordination between the LLNL systems integration, visualization, networking, storage and IBM on-site customer engineering, service and development in Poughkeepsie, were required to keep the underlying "experimental apparatus" functioning with minimal interruption.

2. Leverage Multiprogrammatic and Institutional Computing (M&IC) Program. M&IC represents a funding source both from other Laboratory programs (Energy, SSP, Environment, Biology and Biotechnology, etc.) and the Institution into the unclassified computing environment. This environment shares much of the infrastructure (and costs) of the unclassified environment with ASCI and SSP, providing a tremendous source of leverage for all partners.

3. Develop a Superb User Services Capability tailored to an "Experimental Facility". A major simulation system must provide very high-quality customer services and operational support. This is required to meet the needs of a user base that has grown dramatically, and to address the complexity of the simulation environment that is far beyond that of the "production computing" era. The User Services Group consists of a 14-persons including documentation specialists, instructors and computer scientists, plus a constellation of numerous other specialists in the tools and visualization groups who contribute daily to user support. Included is a well-staffed 9-hour Hotline (that automatically transfers calls to machine operators at night) utilizing Trouble Ticket systems, a very detailed and professional customer oriented web site with complete documentation and dynamic machine status pages, parallel computing courses, tools and graphics and numerical consulting. This posture builds on the Center's tradition of service, but is becoming increasingly more sophisticated. This improvement represents a costly but necessary step in the transformation from a Production Computer Center to a Simulation Center.

4. **Develop Teams to monitor "Experimental Shots"** For the operational and systems activities, there will be a separate user–services oriented development group. Major production calculations require weeks to complete and significant troubleshooting since nodes and software fail on time-scales more rapid than application convergence. For calculations to finish in a timely manner, calculation hangs must be kept to a minimum or eliminated completely. A day of machine time costs at least $100,000, for depreciation alone, ignoring the cost of the rest of the facility and infrastructure. The Center has initiated a planning process to grow both the systems and operations groups so that there will be the appropriate level of coverage during all full system runs. In particular, 24x7 coverage by staff with systems expertise will be required in the future. Root access will be needed to troubleshoot, debug and restart hung calculations. Whether this effort will be managed by the systems group or a highly trained subset of the operators is yet to be determined, but both these areas must grow dramatically before the middle of FY00. Similarly, teams will be formed to help the scientists use the tools necessary to manipulate the generated data before it is transmitted to the Assessment Theater. When these two steps are complete in FY00, the LC will have completed the necessary initial steps to negotiate transition from production computer center to experimental research facility.

5. **Plan and Build a Facility for the Next Decade:** The time-scales and demands coming from (ASCI require a facility of extraordinary scope. In FY98, LC successfully completed a Conceptual Design Review (CDR) and received Defense Program approval to proceed with the design and construction of a new $83.5M Terascale Simulation facility (TSF), beginning in FY00. The first computer room will be available in August 2002, in time to receive the a follow-on ASCI system. It is the unique combination of computer complex, data assessment, and networking capability (proposed for this facility) that represents the necessary ingredient for the success of ASCI. To house these systems and associated assessment infrastructure, LLNL requires a significant increase in electrical power, mechanical support, physical space and networking infrastructure. It also requires a provision for the collocation of the staff required to operate, service, and develop the computing systems and supporting assessment infrastructure. The proposed TSF will meet these requirements. To amplify, this CDR was prepared to address the five following specific project objectives

- High levels of power and cooling available as needed to support up to two systems whose power requirements are equivalent to those coming from two 2004-class 100 TeraOP/s systems.
- A guarantee that any ASCI-scale system will have sufficient high-quality floor space to be sited, and a high probability that two systems (with requirements like those coming from two 2003/4 100 TeraOP/s computers) can be sited simultaneously. This requires a total of more than 47,000 square feet of unobstructed machine area space in appropriate geometry.
- Provision for high performance Advanced Simulation Laboratories for data assessment and a Network Operations Center guaranteeing high bandwidth connectivity to simulation laboratories and customers in local and remote locations.
- Consolidation of Center staff in this complex with adequate space for vendors, tri-lab and Alliance collaborators.
- Beneficial occupancy of at least 23,500 feet of machine room area by August 2002 (major system delivery late in 2002).

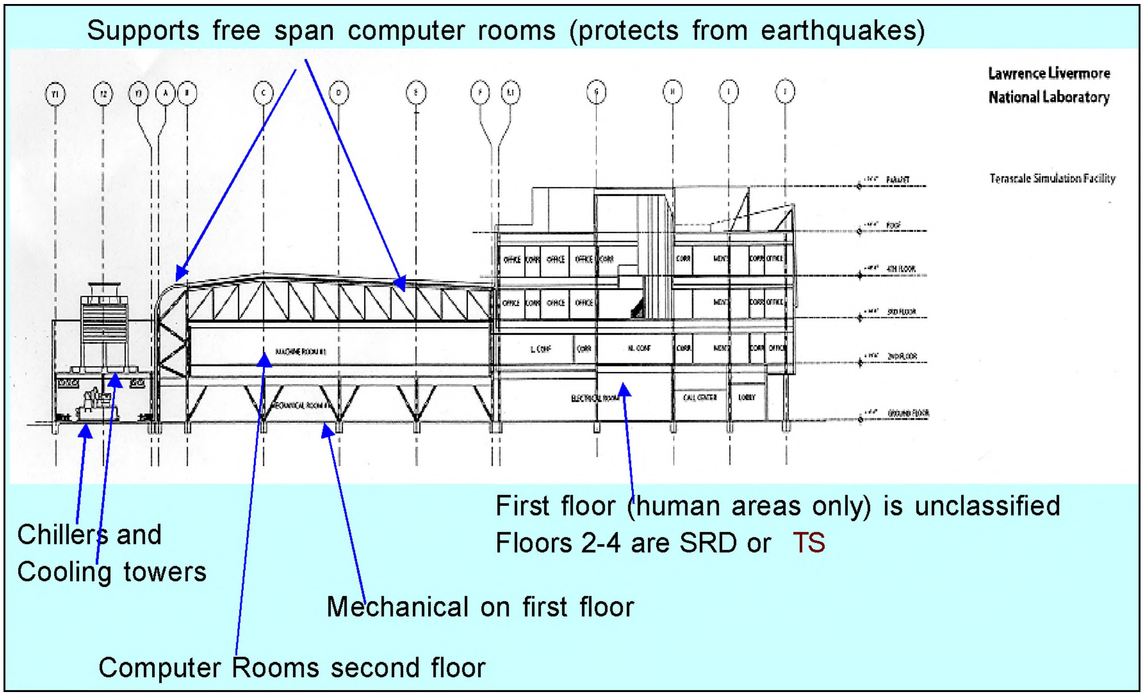


Figure 3: Terascale Simulation Facility. Cross-section of proposed Terascale Simulation Facility (TSF). Total power requirements (24 MW) will support two ASCI-scale systems simultaneously on more than 47,000 square feet of floor.

6 Summary

The ASCI program has identified a staged approach to providing 100 TeraOP/s capability to the Stockpile Stewardship Program by 2004. Leveraging the business plan of the vendors through partnerships, the program has delivered three major terascale systems on- or ahead-of-schedule. In particular, the Blue-Pacific SST demonstrated a startling 1.2 TeraOP/s on sPPM and ran a 70.8 billion zone demonstration problem on September 27, 1998, a full three months ahead of the contractual requirement. The SST hyper-cluster architecture has demonstrated "full-system" capability with representative applications covering a wide range of scientific disciplines.

Delivery to LLNL was staged in order to maximize the resources available to ASCI scientists and engineers, while at the same time minimizing the impact of cutting-edge hardware and software environments. This integration process took well over four months to accomplish and required complex planning, project execution and the dedicated efforts of hundreds of people from LLNL and IBM.

Stunning scientific results have been achieved with the Blue-Pacific SST in Quantum Molecular Dynamics; first principles quantum chemistry; transport methods; turbulence research and biological sciences. However, the operational model required at the ultra-scale is quite different from the conventional computational center approach.

The Livermore Computing organization is currently radically recasting the operational model to support the SST and future ASCI platforms delivered to LLNL. Five strategies are being employed to accomplish the transition. These strategies include detailed system engineering, different organizational models and building a new state-of-the-art Terascale Simulation Facility (TSF).

7 Acknowledgements

Work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract W-7405-Eng-48. The efforts described in this paper are the result of the hard work and dedication of hundreds of people at LLNL and IBM. The Blue-Pacific partnership should be acknowledged as the source of progress on many fronts. In particular IBM management should be highly commended for the dedication and focus they applied to this daunting effort. In addition, direct input from Michel McCoy, Steve Ashby and the full-system science teams were essential to writing this paper.