The ASCI Blue Mountain 3-TOps System

On the Road to 100 TeraOps

This leaflet describes the Blue Mountain System at Los Alamos National Laboratory (LANL). The system is a component of the ASCI program, a collaboration between DOE Defense Programs and Sandia, Lawrence Livermore, and Los Alamos national laboratories, that will create leading-edge capabilities in simulation and computations modeling that are essential for maintaining the safety, reliability, and performance of the US nuclear stockpile. In an era without nuclear testing, these computational goals are important to stockpile stewardship.

To meet the needs of stockpile stewardship in the year 2010, modeling and simulation applications must achieve validated higher-resolution, three-dimensional, complete-physics, and full-system capabilities. This level of computation requires high-performance computing (HPC) far beyond our current level of performance. LANL's Blue Mountain System is the first step in furthering these computational goals.

The complete system, mostly delivered between June and November of 1998, consists of 48 Silicon Graphics Origin 2000 shared memory multi-processor computers with 128 250-MHz processors on each machine (total of 6144 processors). These 48 machines have a composite of 1.5 Terabytes of RAM and 76 Terabytes of fiber channel disk. Jointly, the machines represent a peak capacity of 3.072 TeraOps (3 trillion floating point mathematical operations) per second, with an expected sustained performance of 1 TeraOp per second on the demonstration code, *simplified Piecewise Parabolic Method.* In its full configuration, the Blue Mountain system is one of the most powerful computers installed on-site in the world.

A considerable challenge in the deployment of the ASCI Blue Mountain system is connecting the 48 individual machines into an integrated parallel compute engine. This challenge is currently being met with HIPPI-800 interconnects, which provide high communication bandwidth with great flexibility, without imposing a restricting topology. Each of the 48 machines has 12 HIPPI ports, connected via a 3-dimensional toroidal interconnect using 36 HiPPI-800 16 port switches.

In 1999, the interconnect will be reconfigured with HiPPI-6400 32 port switches. HIPPI-6400 is a new ANSI standard for 6.4 gigabit/second data rates, with transport layer error control built into the hardware. Having this error control in the hardware permits the use of more lightweight protocols operating on each SMP node. The goal is to actually bypass the operating system that currently interacts with transfers between user space and the network. Another ANSI specification, Scheduled Transfer, provides the mechanism to remove this interaction and will be the technique used to increase interconnect performance between the 48 individual machines that make up the ASCI Mountain Blue system.

For the high performance needed by ASCI applications, the multiple machines must be used together as a single machine. This is primarily accomplished via the Message Passing Interface (MPI) software. MPI uses the OS bypass, a low-level protocol which achieves low latency. The objective is to write portable applications by using MPI but to optimize performance through the use of OS bypass. To further optimize performance, LANL is also writing a library which will use OS bypass without MPI. In performance comparisons, the MPI library has provided a bandwidth of 90 Mbytes/second sustained with 144 microseconds one-way latency, and the OS bypass library has given 140 Mbytes/second bandwidth with 104 microseconds one-way latency.

The Load Sharing Facility (LSF) software from Platform Computing Corporation is used for job scheduling and control on the system. LSF distributes jobs across the 48 machines using features such as queue or machine limits, queue priorities, processor reservation, and job backfilling to provide efficient utilization of the system. In addition to queueing batch jobs, the software allows interactive work spanning multiple machines of the system, a capability which facilitates the development and testing of applications.

Archival storage for the ASCI Blue Mountain system is provided by the High Performance Storage System, which is a new-generation storage system for extremely large amounts of data (petabytes) with the ability to access data at very high data rates (tens to hundreds of Mbytes/second). ASCI applications running on the system are expected to generate multigigabyte-sized files.

A team of approximately 45 people, involving both LANL employees and SGI personnel, has been assembled on-site for the installation and support of the Blue Mountain 3TOps system. Areas of support include networking, user consultation, documentation, problem tracking, platform integration and system management, distributed resource management, security, applications support, development of parallel tools, data storage, operations, and facilities management.

The ASCI Blue System requires extensive facilities support. It uses

- 10,000 square feet of floor space,
- 1.6 MWatts of power,
- 530 tons of cooling capability,
- 384 cabinets to house 6144 CPUs,
- 48 cabinets for the meta routers,
- 96 cabinets for the disks,
- 8 cabinets for the 36 HiPPI, switches, and
- ~476 miles of fiber cable.

The successful integration of the Blue Mountain 3TOps system represents one milestone on the road to scaling applications and supporting a fully operational simulation capability for stockpile stewardship. Building upon the experience and knowledge gained with the 3TOps system, LANL will procure and install a computational system that will achieve a peak performance level of 30 TeraOps by midyear 2001. It is expected that a 100 TeraOp capability is needed by the year 2004 in order to meet the goals of stockpile stewardship.

For more information about ASCI Blue Mountain, contact

John Morrison (jfm@lanl.gov or 505-667-1042),

Ray Miller (rdm@lanl.gov or 505-665-3222), or

Manuel Vigil (mbv@lanl.gov or 505-667-5243

The ASCI Blue Web site is http://www.lanl.gov/asci/bluemtn/.

The Accelerated Strategic Computing Initiative Program includes participants from Los Alamos National Laboratory, Lawrence Livermore National Laboratory, and Sandia National Laboratories.

LALP-98-127