NERSC NEWS

News from the National Energy Research Scientific Computing Center December 2005

NERSC, IBM Collaborate on New Software Strategy To Simplify Supercomputing

This month IBM announced an innovative software strategy in supercomputing which allows customers to leverage the General Parallel File System (GPFS) across mixed-vendor supercomputing systems for the first time. This strategy is the result of a direct partnership with NERSC.

GPFS is an advanced file system for high performance computing clusters that provides high speed file access to applications executing on multiple nodes of a Linux or AIX cluster. GPFS's scalability and performance are designed to meet the needs of data-intensive applications such as engineering design, digital media, data mining, financial analysis, seismic data processing and scientific research.

"Thank you for driving us in this direction," wrote IBM Federal Client Executive Mike Henesy to NERSC General Manager Bill Kramer as IBM announced the project. "It's quite clear we would never have reached this point without your leadership!"

Staff at NERSC used GPFS to create a scalable parallel file system that is capable of supporting hundreds of terabytes of storage within a single highly reliable file system. In November 2005, NERSC implemented a production version of the NERSC Global Filesystem (NGF) using GPFS. This was demonstrated at the SC|05 conference and used for several HPC Analytics and StorCloud Challenges at SC|05.

The production NGF grew out of the

(continued on page 2)

NERSC News

NERSC News, highlighting achievements by staff and users of DOE's National Energy Research Scientific Computing Center, is published every other month via email and may be freely distributed. NERSC News is edited by Jon Bashor, JBashor@lbl.gov or 510-486-5849.

New IBM Cluster to Go into Production in January

NERSC users who run jobs on up to 512 processors will benefit from "Bassi," the new IBM cluster which goes into production on Monday, Jan. 9, 2006. During the acceptance testing, users reported that codes ran up to 10 times faster on Bassi than on Seaborg.

The new machine is an IBM p575 POWER 5 system with 888 processors available for parallel scientific computation. Each Bassi processor has a theoretical peak performance of 7.6 Gflop/s. The processors are distributed among 111 compute nodes with eight processors per node. Processors on each node have a shared memory pool of 32 Gbytes. A Bassi node is an example of a shared memory processor, or SMP.

The compute nodes are connected to each other with a high-bandwidth, low-latency switching network. Each node runs its own full instance of the standard AIX operating system. The disk storage system is a distributed, parallel I/O system called GPFS. Additional nodes serve exclusively as GPFS servers. Bassi's network switch is the IBM "Federation" HPS switch which is connected to a two-link network adapter on each node.

Scientifically Productive

One of the early users during the acceptance testing was the combustion simulation team led by Jackie Chen of Sandia National Laboratories. Chen's group had been awarded 2 million hours on Seaborg under DOE's INCITE program, and were able to run their three-dimensional direct numerical simulation of turbulent non-premixed combustion code for an additional one millions hours on Bassi.

"We had a major success enabled by Bassi — the successful completion of our INCITE project to perform direct numerical simulation of a turbulent nonpremixed CO/H2 jet flame with detailed chemistry," recounted Evatt Hawkes, a member of Chen's group. "Our project required a very long stretch of using a large fraction of Bassi processors — 512 processors for essentially an entire month. During this period we experienced only a few minor problems, which is exceptional for a pre-production machine and enabled us to complete our project against a tight deadline. We were very impressed with the reliability of the machine."

Hawkes noted that their code also ported quickly to Bassi, starting with a code already ported to Seaborg's architecture.

"Bassi performs very well for our code — with Bassi's faster processors we were able to run on far fewer processors (512 as opposed to 4096), and still complete the simulations more rapidly," Hawkes wrote. "Based on scalar tests, it is approximately seven times faster than Seaborg, and one-and-a-half times faster than a 2.0 GHz Opteron processor. Also, the parallel efficiency is very good. In a weak scaling test we obtain approximately 78 percent parallel efficiency using 768 processors, compared with about 70 percent on Seaborg."

System Named in Honor of Italian Physicist

The machine is named in honor of Laura Bassi, a noted Newtonian physicist of the eighteenth century. Born on Oct. 31, 1711, in Bologna, she was educated privately. Bassi studied logic, metaphysics, philosophy, chemistry, hydraulics, mathematics, mechanics, algebra, geometry, and ancient and modern languages (Greek, Latin, French, and Italian).

Bassi was appointed professor of anatomy at the University of Bologna in 1731, and has been cited as the first woman to officially teach at a European university. She

(continued on page 2)

NERSC, IBM Partner to Develop Simplified Supercomputing Strategy (continued from page 1)

multi-year Global Unified Parallel File System (GUPFS) Project at NERSC to provide a scalable, high-performance, high-bandwidth, shared-disk file system for use by all of NERSC's high-performance production computational systems. NGF provides unified file namespace for these systems and is being integrated with the High Performance Storage System (HPSS), while performing at or very close to rates achieved by parallel filesystems within a cluster. It is also possible to distribute GPFS-based file systems to remote facilities as local file systems over the Internet.

"We spent a long time looking at all the possible configurations of storage hardware, fabric hardware and filesystem software during the GUPFS project. Eventually we realized the critical component was filesystem software," said Kramer. "It came down to a limited number of choices, and GPFS was superior in many aspects, but limited only to IBM hardware. NERSC helped convince IBM to make GPFS available on any vendor's hardware — a requirement from NERSC's point of view. We are happy IBM decided to take this step to make GPFS more open."

The typical state of many high performance computational environments is one in which each large computational and support system has its own large, independent disk store, with additional network attached storage (NAS), such as NFS or DFS, and an archival storage server such as HPSS. These approaches lead to wasteful replication of customer files on multiple systems and an increased, nonproductive workload on customers to move and manage these files. This, in turn, creates a burden on the infrastructure to support file transfers between the systems as well as to the storage server. In addition, the existing environment prevents the consolidation of storage between systems, thus limiting the amount of working storage available to each system's local disk capacity.

The environment envisioned by the NERSC GUPFS project is one in which the large high performance computational systems and support systems can access a consolidated disk store. NGF is the first step toward that vision and currently is supporting five major systems — IBM Power 3+ SP, IBM Power 5 SP, SGI Altix, Linux Networx Opteron/InfinBand cluster and the PDSF Intel/Ethernet cluster, totaling over 1,200 client nodes.

"In terms of vendors and number of nodes, this it the largest, most diverse implementation of a global filesystem we know of," said Greg Butler, a computer engineer at NERSC who has worked on GUPFS and NGF. "Not only is it leading edge technology but it is standing up to production requirements supporing a wide range of science. Our users are very pleased with the NGF because it makes their work simpler and easier."

A major use of the file system will be in support of parallel scientific applications performing high volume concurrent and simultaneous I/O. This environment will eliminate unnecessary data replication, simplify the customer environment, provide better distribution of storage resources, and permit the management of storage as a separate entity while minimizing impacts on the computational systems.

NGF directly accesses storage through multivendor shared-disk file systems with a unified file namespace. Storage servers, accessing the consolidated storage using the shared-disk file systems, would provide hierarchical storage management (HSM), backup, and archival services. The deployed file system will be integrated with the NERSC HPSS archival system in 2006.

A heterogeneous approach for NGF is a key component of "Science-Driven Computing," NERSC's five-year plan recently published at <http://www.nersc.gov/news/reports/>. This approach is important because NERSC typically procures a major, new computational system every three years, then operates it for five years to support DOE research. Consequently, NERSC operates in a heterogeneous environment with systems from multiple vendors, multiple platforms, different system architectures, and multiple operating systems. The deployed file system must operate in the same heterogeneous client environment throughout its life time.

GPFS/HPSS Development

NERSC's Mass Storage Group collaborated with IBM to develop a Hierarchical Storage Manager (HSM) that can be used with IBM's GPFS. The HSM capability with GPFS will provide a recoverable GPFS file system that is transparent to users and fully backed up and recoverable from NERSC's multi-petabyte archive on HPSS.

One of the key capabilities of the GPFS/HPSS HSM is that users' files will automatically be backed up on HPSS as they are created. Additionally, files on the GPFS which have not been accessed for a specified period of time will be automatically migrated from on-line resources as space is needed by users for files currently in use. Since the purged files were already migrated/backed up on HPSS, they can easily by automatically retrieved by users when needed.

"This gives the user the appearance of almost unlimited disk storage space without the cost," said NERSC's Mass Storage Group Leader Nancy Meyer.

New IBM Cluster Enters Production (continued from page 1)

was elected to the Academy of the Institute for Sciences in 1732, and was given the Chair of Philosophy at the University of Bologna in 1733. In 1738, she married her colleague Dr. Giuseppe Veratti. They had 12 children.

While raising her family, she successfully petitioned for wider responsibilities and a higher salary to cover the cost of equipment for physical and electrical experiments. She continued her lifelong interest in physics, lecturing from her home while her children were small then returning to the university at age 65 as a professor of experimental physics in 1776. She died on Feb. 20, 1778, at age 66.