NERSC NEWS

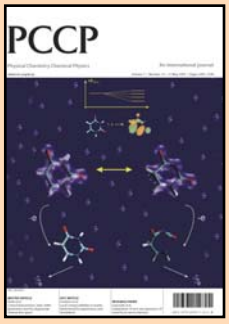
Berkeley Lab

Office of Science U.S. Department of Energy

News from the National Energy Research Scientific Computing Center February 2006

NERSC Users Report Over 1,400 Publications in 2005

As part of their 2006 allocations request, NERSC users were asked to submit lists of articles and conference papers resulting from their computations at NERSC over the preceding 12 months.



A total of 1,448 publications, which had either been published or accepted, were noted. The total does not include those papers listed as "submitted" or "in preparation."

"This is an inspiring list of publications, both in the number and the breadth of science covered," said NERSC Division Director Horst Simon. "By providing the systems and expertise which make it possible for our users to achieve such an impressive output, NERSC may well be the scientific computer center that enables the most science."

The complete list of publications can be found on the Web at <http://www.nersc.gov/news/reports/ ERCAPpubs05.php>.

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.

NERSC Global Filesystem Now Provides Seamless Data Access from All Systems

In February, NERSC deployed the NERSC Global Filesystem (NGF) into production, providing seamless data access from all of the Center's computational and analysis resources. With NGF, users can now run applications on Seaborg, for example, then use DaVinci to visualize the data without having to explicitly move a single data file.

NGF is intended to facilitate sharing of data between users and/or machines. For example, if a project has multiple users who must all access a common set of data files, NGF will provide a common area for those files. Alternatively, when sharing data between machines, NGF eliminates the need to copy large datasets from one machine to another. For example, because NGF has a single unified namespace, a user can run a highly parallel simulation on Seaborg, followed by a serial or modestly parallel post-processing step on Jacquard, and then perform a data analysis or visualization step on DaVinci.

"NGF stitches all of our systems together. When you go from system to system, your data is just there," said Greg Butler, leader of the NGF project. "Users don't have to manually move their data or keep track of it. They can now see their data simultaneously and access the data simultaneously."

NERSC staff began adding NGF to computing systems last October, starting with the DaVinci visualization cluster (an SGI Altix) and finishing with the Seaborg IBM SP system in December. To help test the system before it entered production, a number of NERSC users were given pre-production access to NGF.

Greg Butler

Early users helped identify problems with NGF so they could be addressed before the filesystem was made available to the general user community.

"I have been using the NGF for some time now, and it's made my work a lot easier on the NERSC systems," said Martin White, a physicist at Berkeley Lab. "I have at times accessed files on NGF from all three compute platforms (Seaborg, Jacquard and Bassi) semi-simultaneously."

NGF also makes it easier for members of collaborative groups to access data, as well

(continued on page 2)

NERSC Team Takes StorCloud Honors at SC05 Conference

One of the goals of providing comprehensive computing resources is to make the different components "transparent" to the end user. But if you are staff members demonstrating a groundbreaking approach for accessing distributed storage, such invisibility isn't so desirable.

Will Baird

That was the case for a NERSC/LBNL team competing in the StorCloud Challenge at the SC05 conference held Nov. 12-18, 2005, in Seattle.

The team, led by Will Baird of NERSC's Computational Systems Group, was given an award for "Best Deployment of a Prototype for a Scientific Application." Unfortunately, however, their award slipped through the cracks and the group was not recognized at the SC05 awards session, nor in the conference news

(continued on page 3)

NGF Goes Online in February *(continued from page 1)*

as ensure data consistency by eliminating multiple copies of critical data.

Christian Ott, a Ph.D. student and member of a team studying core-collapse supernovae, wrote that "the project directories make our collaboration much more efficient. We can now easily look at the output of the runs managed by other team members and monitor their progress etc. We are also sharing standard input data for our simulations."

NERSC General Manager Bill Kramer said that as far as he knows, NGF is the first production global file system spanning five platforms — Seaborg, Bassi, Jacquard, DaVinci and PDSF — three architectures and four different vendors. While other centers and distributed computing projects such as NSF's TeraGrid may also have shared file systems, Butler said he thinks NGF is unique in its heterogeneity.

NGF is the first production global file system spanning five platforms — Seaborg, Jacquard, PDSF, DaVinci and Bassi — three architectures and four different vendors.

A heterogeneous approach for NGF is a key component of "Science-Driven Computing," NERSC's five-year plan (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 lifetime.

Butler noted that the project, which is based on IBM's proven GPFS technology (General Parallel File System, in which NERSC was a research partner), started about five years ago. While the computing systems, storage and interconnects were mostly in place, deploying a shared file system among all the resources was a major step beyond a parallel file system. In addition to the different system architectures, there were also different operating systems to contend with. However, on Feb. 10, the last servers and storage were deployed. To keep everything running and ensure a "graceful shutdown" in the event of a power outage, a large uninterruptible power supply has been installed in the basement of the Oakland Scientific Facility.

While NGF is a significant change for NERSC users, it also "fundamentally changes the Center in terms of our perspective," Butler said. For example, when the staff needs to do maintenance on the file system, the various groups need to coordinate their efforts and take all the systems down at once.

Storage servers, accessing the consolidated storage using the shared-disk file systems, provide hierarchical storage management (HSM), backup, and archival services. The first phase of NGF is focused on function and not raw performance, but in order to be effective, NGF has to have performance comparable to native cluster file systems. The current capacity of NGF is approximately 70 TB of user-accessible storage, and 50 million inodes (the data structures for individual files). Default project quotas are 1 TB and 250,000 inodes. The system has a sustainable bandwidth of 3 GB/sec bandwidth for streaming I/O, although actual performance for user applications will depend on a variety of factors. Because NGF is a distributed network filesystem, performance will be only slightly less than that of filesystems that are local to NERSC compute platforms. This should only be an issue for applications whose performance is I/O bound.

NGF will grow in both capacity and bandwidth over the next several years, eventually replacing or dwarfing the amount of local storage on systems. NERSC is also working to seamlessly integrate NGF with the HPSS data archive to create much larger "virtual" data storage for projects. Once NGF is completely operational within the NERSC facility, Butler said, users at other centers such as NCAR and NASA Ames could be allowed to remotely access the NERSC filesystem, allowing users to read and visualize data without having to FTP the data. Eventually, the same capability could be extended to experimental research sites, such as accelerator labs.

"The availability of NGF has greatly facilitated my work, and I suspect it will be the way to go in the future," predicts LBNL's Martin White.

StorCloud Honors *(continued from page 1)*

release announcing the various prizes and awards given out at SC05.

The team, which also included Jonathan Carter and Tavia Stone of NERSC and Michael Wehner (not shown), Cristina Siegerist and Wes Bethel of Berkeley Lab's Computational Research Division, used the StorCloud infrastructure "to test the wide-area deployment of an unprecedented system in support of a groundbreaking climate modeling application," according to the award. The application was fvCAM — or Finite Volume Community Atmospheric Model — which is being used to predict hurricane formation.

Christina Siegerist

Tavia Stone

Jonathan Carter

Wes Bethel

StorCloud is a special initiative for building a high performance computing storage capability showcasing HPC storage technologies (topologies, devices, interconnects) and applications. The StorCloud Challenge invited applicants from science and engineering communities to use the unique StorCloud infrastructure to demonstrate emerging techniques or applications, many of which consume enormous amounts of network and storage resources.

Baird designed the TRI Data Storm prototype around the concept of using an integrated, multisystem file system to improve the analysis of results produced by the demanding HPC application — the Community Atmospheric Model (CAM).

"Our aim is to take the output of CAM from a high-resolution grid, filter out the data of interest, and visualize the formation of storms in the North Atlantic basin," according to Baird. "This tool will be used in a study comparing real hurricane data with simulations. While this is a fairly generic workflow that could hold true for virtually any HPC application, the unique aspect to our approach is that there is a single high-performance parallel file system serving all of the different systems and applications."

For TRI Data Storm, the team used an externalized GPFS (General Parallel File System), shared out by a dedicated cluster and mounted on all of the different computational resources used by their tool: the IBM Power5 cluster "Bassi" and the PDSF Linux cluster at NERSC, the GPFS servers and storage at the conference and an SGI Altix cluster in the LBNL booth at SC05.

"There are very few places that do this, especially with the variety of systems that we demonstrated," Baird said. "Additionally, all the communication between the systems was simply through the file system, not through anything else."

According to Baird, sharing data through multisystem file systems has been undergoing some radical changes in the last several years. Networked file storage first impacted the way data was shared among computers. Later on, the cluster file systems enabled whole clusters to be able to have access to the same data in a robust manner. In the last year, the cluster file systems have evolved to allow multiple systems to have access to that same data easily and reliably.

Such advancements in storage access have significant implications for HPC centers and users, who could access data from multiple systems in a way that is transparent. Using GPFS, a single file system can be presented to multiple computing systems to run extensive simulations without manual data transfers between systems.

201 Users Participate in Annual User Survey

The results from the 2005 user survey are now available and show generally high satisfaction with NERSC's systems and support.

Areas with the highest user satisfaction include account support services, the reliability and uptime of the HPSS mass storage system, and HPC consulting. The largest increases in satisfaction over last year's survey include the NERSC CVS server, the Seaborg batch queue structure, PDSF compilers, Seaborg uptime, available computing hardware and network connectivity.

Areas with the lowest user satisfaction include batch wait times on both Seaborg and Jacquard, Seaborg's queue structure, PDSF disk stability, and Jacquard performance and debugging tools. Only three areas were rated significantly lower this year: PDSF overall satisfaction and uptime, and the amount of time taken to resolve consulting issues. The introduction of three major systems in the last year combined with a reduction in consulting staff explain the latter.

As in the past, comments from the previous survey led to changes in 2005, including a restructuring of Seaborg's queuing polices, the addition of the new Jacquard and Bassi clusters, and upgrade of ESnet's connectivity to NERSC to 10 gigabits per second.

The complete survey results can be found at <https://www.nersc.gov/news/survey/2005/>.

DISCLAIMER
This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California. Ernest Orlando Lawrence Berkeley National Laboratory is an equal opportunity employer.