NERSC Global Filesystem Facilitates Data Sharing

Gregory F. Butler,* Rei Lee, Matthew Andrews, Cary Whitney, William Baird, James Craw, Howard Walter, NERSC, Lawrence Berkeley National Laboratory

Summary

The NERSC Global Filesystem (NGF) is a large, highly parallel, shared filesystem that provides seamless data access from all of NERSC's computational and analysis resources. This facilitates file sharing between platforms, as well as file sharing among NERSC users working on a common project.

In early 2006, the National Energy Research Scientific Computing Center (NERSC) deployed the NERSC Global Filesystem (NGF) into production. NGF facilitates sharing of data between users and/or machines and is a key piece of NERSC's Science Driven Analytics initiative. For example, if a project has multiple users who must all access a common set of data files, NGF provides a common area for those files. Alternatively, when sharing data between NERSC 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—all without having to explicitly move a single data file.

NGF's single unified namespace makes it easier for users to manage their data across multiple systems (Figure 1). Users no longer need to keep track of multiple copies of programs and data, and they no longer need to copy data between NERSC systems for pre- and post-processing. NGF provides several other benefits as well: storage utilization is more efficient because of decreased fragmentation; computational resource utilization is more efficient because users can more easily run jobs on an appropriate resource; and NGF provides improved methods of backing up user data.

Figure 1. NGF is the first production high-speed, global filesystem spanning five platforms (Seaborg, Jacquard, PDSF, DaVinci, and Bassi), three architectures, and four different vendors.

"NGF stitches all of our systems together," said Greg Butler, leader of the NGF project. "When you go from system to system, your data is just there. 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." Furthermore, NGF is parallel, so the speed of accessing the data is appropriate for NERSC's high performance systems.

* 510-486-8691, GFButler@lbl.gov

"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 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. We are also sharing standard input data for our simulations."

The heterogeneous approach of NGF is a key component of NERSC's five-year plan. 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 filesystem must operate in the same heterogeneous client environment throughout its lifetime.

The NGF project started about four years ago and is based on IBM's proven GPFS technology (Global Parallel File System, in which NERSC was a research partner). While the computing systems, storage, and interconnects were mostly in place, deploying a shared filesystem among all the resources was a major step beyond a parallel filesystem. In addition to the different system architectures, there were also different operating systems to contend with. However, the last servers and storage have now been 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.

NGF currently provides allocated project space on Seaborg, Bassi, Jacquard, DaVinci, and PDSF. It contains approximately 70 TB of user-accessible storage, and 50 million inodes. Default project quotas are 1 TB and 250,000 inodes. The system has a sustainable bandwidth of 1 GB/sec bandwidth for streaming I/O, although actual performance for user applications will depend on a variety of factors.

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 the National Center for Atmospheric Research and NASA Ames Research Center, could be allowed to remotely access the NERSC filesystem, allowing users to read and visualize data without having to execute file transfers. Eventually, the same capability could be extended to experimental research sites, such as accelerator labs.

For further information on this subject contact:
Barbara Helland, Program Manager
Mathematical, Information, and Computational Sciences Division
Office of Advanced Scientific Computing Research
Phone: 301-903-3127
Barbara.Helland@science.doe.gov