William R. Wiley
EMSL
Environmental Molecular Sciences Laboratory

EMSL's disk-based archive system allows researchers to quickly and efficiently manipulate data generated by the facility's 11.8-teraflop supercomputer and mass spectrometry systems. The archive allows the researcher to save to the system at up to 45 megabytes per second, and to read data at up to 70 megabytes per second.

Disk-Based Archive System—A First in a Supercomputing Environment

At the William R. Wiley Environmental Molecular Sciences Laboratory in Richland, Washington, researchers address some of the nation's most pressing scientific challenges in competitive research areas such as proteomics, computational simulations, nuclear magnetic resonance spectroscopy, and biology and bioinformatics. As part of their research, multitudes of data are generated using the facility's cutting-edge 11.8-teraflop supercomputer and mass spectroscopy systems.

NWfs—which replaces EMSL's significantly slower and outdated tape-based archive system—will ultimately provide EMSL's researchers and users with a massive 1 million gigabtyes of storage space.

Such data consume a great deal of space. Data generated by proteomics-related mass spectrometry experiments alone involve thousands of tests on countless biological samples—a feat that produces an accumulation of data files that currently consume approximately 22 terabytes of space and growing. Since analyses of these data files can take several hours to complete for each dataset, researchers must be able to easily store and access their data quickly and efficiently.

Staff from EMSL's Molecular Science Computing Facility have addressed the need for quick, efficient data storage and access capabilities by implementing NWfs, a disk-based archive system—the first used in a supercomputing environment. NWfs—which replaces EMSL's significantly slower and outdated tape-based archive system—will ultimately provide EMSL's researchers and users with a massive 1 million gigabtyes of storage space. NWfs provides the researcher with the ability to save their data at a rate of 30 to 45 megabytes per second, and to read data in excess of 25 to 70 megabytes per second.

NWfs is comprised of storage "bricks" built around off-the-shelf Advanced Technology Attachment drives and RAID controllers, each containing 1.7 terabytes or more of file system space. The "bricks" are clustered and integrated with system management and shared file system tools to provide a high-capacity file storage

Pacific Northwest National Laboratory
Operated by Battelle for the U.S. Department of Energy

At the William R. Wiley Environmental Molecular Sciences Laboratory, this new supercomputer operates 9,200 times faster than a PC.

space that seamlessly gives the user the impression of one single storage "pool." More bricks are added as additional storage capacity is required.

EMSL researchers and users can store and access NWfs following establishment of an account—and provided that they have a research proposal in place at EMSL. Users—who are generally not charged for using the archive system—can access NWfs via secure FTP, SFTP, SCP, or GridFTP services. A variety of browser platforms may be used to input or retrieve NWfs data, including the FileZilla Browser and Windows Explorer (for Microsoft Windows users); fetch (for MacIntosh users); and ftp, lftp, gftp, and uberftp (for Linux and Unix users).

The researcher may provide permission for any combination of read, write, or delete access to individuals or groups for file-sharing purposes.

Traditionally, tape-based systems for large-scale storage were preferred over disk-based systems because of the latter's prohibitive expense. However, costs associated with disk-based storage have dropped significantly faster than costs for tape, and it is now possible to build disk-based archives for approximately $2,500 per terabyte. This cost will continue to decline, providing EMSL with a high-end, effective system with dramatically better performance at a cost less than the annual maintenance fees for the tape library it replaces.

Contacts

Dave Cowley
Molecular Science Computing Facility
509-376-9181
de.cowley@pnl.gov

Scott Studham
Molecular Science Computing Facility
Operations Technical Lead
509-376-8430
Scott.Studham@pnl.gov

More information about NWfs is located at *http://mscf.emsl.pnl.gov/hardware/intro_nwfs.shtml.*

PNNL-SA-XXXXX 05/04