panasas

Customer Success Story

National Institutes of Health

National Institutes of Health

The National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH), serves as a national resource for molecular biology information serving research groups from around the world. Established in 1988, NCBI develops new information technologies to aid in the understanding of fundamental molecular and genetic processes that control health and disease. NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genomic data, and disseminates biomedical information. Some 450 people—ranging from NCBI researchers and staff scientists to programmers, curators, and indexers—generate, store, and access NCBI databases.

SUMMARY

Industry:
Life Sciences/Government

THE CHALLENGE

Meet demands of researchers from around the globe accessing the NCBI public database to conduct genome research. Eliminate I/O bottlenecks and maximize computing resources for public databases, including an estimated 1.5 PB of genetic information for the 1000 Genomes Project.

THE SOLUTION

Panasas ActiveStor running the PanFS™ parallel file system, 1800-core Dell PowerEdge Cluster, Cisco 6509 Network Switch

THE RESULT

- 5X application performance improvement
- Timelier database updates with faster time-to-results
- High performance irrespective of access patterns/dataset size
- Affordable scalability for fast-growing archives
- Administrative efficiencies across primary and secondary storage

The Challenge

Researchers at NCBI depend on high-performance compute clusters to run complex analyses of genotyping and sequencing data. The existing storage architecture did not effectively scale to support such efforts as the 1000 Genomes Project, an ambitious endeavor to sequence the genomes of at least 1,000 people from around the world. The project, creating the most detailed and medically useful picture to date of human genetic variation, is expected to generate more than 1.5 PB of genetic information. NCBI will be required to archive and provide timely investigator access to as much as 3 TB of new genome data arriving weekly from each of the six institutes participating in the 1000 Genomes Project. To accommodate the expected high demand for data access NCBI requires a storage solution that is reliable, manageable, and affordable.

The Solution

NCBI selected Panasas storage for the Center's Dell PowerEdge compute farm. The decision was based in part on testing results that indicated the Panasas ActiveStor solution delivered a significant performance improvement over existing installed storage. Designed specifically to accelerate the performance of applications deployed on Linux compute clusters, Panasas ActiveStor effectively eliminated the research-impacting I/O bottlenecks.

ActiveStor now provides scalable performance and capacity to multiple internal production systems (both Linux- and Windows-based platforms), including NCBI's 1800-core Dell PowerEdge cluster that provides computing resources to some 80 applications used by ten NCBI research groups. Panasas storage supports much of the daily computation that generates the data for such high-visibility services as NCBI's PubMed resource that brings together more than 18 million citations from MEDLINE and other life science journals for biomedical articles.

Most recently, NCBI implemented an ActiveStor system that provides economical second-tier storage for the high-density data requirements of the 1000 Genomes Project. The ActiveStor solution also provides storage resources

to projects such as the NCBI Short Read Archive (SRA), a central repository for short read sequencing data, and the dbGaP public repository of genotypes and phenotypes.

The Result

Flexibility and Efficiency Advances Discovery.
Technology advances that have brought down the cost of sequencing—from billions to millions per project and freefalling rapidly to the industry's goal of $10K or even as low as $1K for a single run—have also contributed to an explosion of data. Taking advantage of the ActiveStor solution for receipt and storage of genome and other project data helps NCBI keep pace with the volume and complexity of incoming information in a cost-effective manner.

Performance, Scalability for Fast-Growing Archives
Panasas solutions help address the research community's storage needs in spite of a very high unpredictability factor. Whether it's unexpected demand for particular research findings, storage requirements that mushroom from 150 TB to 1.5 PB almost overnight, or datasets that vary from 3 TB to 30 TB in size, the needs of the scientific community dictate storage flexibility and maximum uptime. In addition to the inherent administrative efficiencies of a common architecture, the Panasas unified storage platform for Tier 1 and secondary storage applications gives flexibility to support a scientific user community striving for discoveries that directly impact understanding of genetics and its role in health and disease analysis. NCBI's mission is to help researchers better leverage and build on the work of the larger biotechnology community, avoiding both the cost and the time penalities of reworking data.

"Technology advances that have brought down the cost of sequencing have contributed to an explosion of data...Panasas helps NCBI keep pace with the volume and complexity of incoming information in a cost-effective manner."

07282011 1029