Quantum

CASE STUDY

StorNext & Scalar Series

It's All About DNA: Genome Sequencing Center Relies on StorNext Data Management

There are few tasks more data intensive than sequencing the 3 billion chemical building blocks that make up the DNA of the 24 different human chromosomes. Sharing, managing and storing that data was a frustrating challenge until the Baylor College of Medicine's Human Genome Sequencing Center (HGSC) deployed Quantum's StorNext data management software.

LEGACY TECHNOLOGY INFRASTRUCTURE IMPACTS RESEARCH

As one of three U.S. federally funded centers driving the rapid growth of knowledge about genetic influences on human disease, HGSC has just under 200 employees, including approximately 40 research scientists who spend their time analyzing DNA sequencing data. With high volumes of data generated daily and the need for hundreds of terabytes of data to be accessible for analysis at any time, HGSC's piecemeal technology infrastructure that had been built over time was becoming a barrier to the important health research work under way.

In August 2008, Geraint Morgan was brought in as Director of Information Systems when it became apparent that the rate at which sequence data could be produced and analyzed was outpacing the ability of the technology on which it was residing to keep up.

HGSC has 32 genome sequencers, including 20 Applied Biosystems SOLiD Sequencing Instruments, 2 Illumina Genome Analyzers and 10 Roche/454Genome Sequencers. The most productive of these can generate up to approximately 1TB of raw data per day.

Morgan entered an environment where primary data from DNA sequencing was initially written locally to attached storage units from a variety of vendors and then moved to a centralized repository. There it was further processed by clusters of compute nodes and made accessible to the researchers through simple network file system mounts.

"The volume of data put a strain on the network infrastructure and limited accessibility to the data, which is critical in a research organization such as ours," says Morgan. "It also contributed to the huge management overhead needed to ensure that no issue could impact the tail-end of the researchers' sequencing pipeline operation."

To expand HGSC's storage capabilities, an additional data center was created. However, Morgan still faced the challenge of finding a way to centrally manage a complex heterogeneous environment of servers, networks and storage technology. Because funding for technology is often limited in such grant-based institutions, he needed to accommodate existing servers and storage arrays, rather than starting from scratch with an entirely new approach.

"I needed to find a solution that could not only utilize this existing hardware but also would be easily scalable to accommodate a predicted 20 petabytes of sequenced data over the next two years," Morgan says. "The solution had to require minimal changes to how the environment as a whole would be managed as the environment grew."

SOLD ON STORNEXT'S REPUTATION IN HIGH-PERFORMANCE, MULTI-PETABYTE ENVIRONMENTS

Morgan began his search with a number of vendors and consultants but found that many had difficulty understanding the complexity of the operation or could not provide adequate answers to how their solution would cope with the expected growth.

"Given the constraints of existing hardware, limited budget and the ability to accommodate significant data growth, there were surprisingly few solutions that I felt would be a good fit," says Morgan, reflecting on his three-month search for an answer. "However, one solution that kept coming up and impressed us was Quantum's StorNext data management software."

HGSC
HUMAN GENOME SEQUENCING CENTER

"By combining high-speed data sharing and cost-effective content retention in a single solution, StorNext has enabled our researchers to access the data they need quickly and easily and eliminated the significant management overhead we incurred with our legacy system."

Geraint Morgan
Director of Information Systems

SOLUTION OVERVIEW

- StorNext® File System
- StorNext Storage Manager
- Scalar® i2000 tape library system

KEY BENEFITS

- Enabled simultaneous access to huge volumes of data without impacting system users
- Provided cost-effective content creation through automated data management
- Allowed centralized management of heterogeneous environment
- Protected prior investments by integrating legacy resources
- Provided scalable foundation to meet anticipated storage growth of up to 20PB over next 2-3 years

Morgan liked the fact that there were a number of companies with similar high-performance data processing needs and large multi-petabyte storage requirements that were already successfully using StorNext. He also realized that the expectation that an open-source solution would be least expensive and least intrusive was wrong, once support contracts and other fees were factored in.

"StorNext offered the scalability we needed, support for existing storage hardware with no significant investment needed for additional hardware, and an easy to manage system," says Morgan.

HGSC purchased both the StorNext File System and Storage Manager to enable file sharing across multiple operating environments and automated data movement across storage tiers.

THE FOUNDATION OF A MASSIVELY SCALABLE SOLUTION

The StorNext implementation was straightforward, according to Morgan. HGSC currently has 2 metadata controllers and 8 StorNext File System SAN gateways. These gateways are connected to storage arrays via a 4 Gbp FC network, and the compute nodes access storage via StorNext Distributed LAN Clients over dual 4×10 Gbp Ethernet uplinks through the SAN gateways. The system started with 110TB of storage and an additional 560TB has been recently added.

Following the ingest of data on the local genome scanner devices and some initial pre-processing, the data is copied via NFS to a centralized StorNext File System. The pipeline analysis is then done by multiple genome analysis applications running on top of the StorNext Distributed LAN Client, which connects to the centralized storage to process the data in parallel.

In addition, HGSC uses StorNext Storage Manager for automatically moving data between different disk systems and a Quantum Scalar i2000 tape library, thereby protecting content at lower costs. Older genome projects can also be archived automatically on the Scalar i2000, freeing up the fast primary disk storage for newer sequencing workflows.

Since deploying StorNext, Morgan has been very pleased with the benefits it has provided from both a research and IT perspective.

"By combining high-speed data sharing and cost-effective content retention in a single solution, StorNext has enabled our researchers to access the data they need quickly and easily and also eliminated the significant management overhead we incurred with our legacy system," he says.

Looking toward the future, Morgan says that StorNext will serve as the foundation of a massively scalable solution that's needed to cope with expected storage requirements of two petabytes over the next 12 months and up to 20 petabytes over the coming two or three years.

"Because of the nature of genomics research—where data generated today might not have obvious value but could lead to important discoveries in the future—we preserve all the data generated at HGSC," says Morgan. "This creates an ever-growing archive, and StorNext will play a critical role in helping us to manage this growth. The exponential data growth is also one of the reasons we plan to leverage the data deduplication feature offered through StorNext—it will enable us to optimize the amount of storage capacity needed for archiving."

"The exponential data growth is also one of the reasons we plan to leverage the data deduplication feature offered through StorNext—it will enable us to optimize the amount of storage capacity needed for archiving."

Geraint Morgan
Director of Information Systems

ABOUT THE HUMAN GENOME SEQUENCING CENTER

Baylor College of Medicine's Human Genome Sequencing Center, founded in 1996, is a world leader in genomics. The fundamental interests of the HGSC are in advancing biology and genetics by improved genome technologies. One of three large-scale sequencing centers funded by the National Institutes of Health, the HGSC's location at the heart of the Texas Medical Center provides a unique opportunity to apply the cutting edge of genome technologies in science and medicine.

To contact your local sales office, please visit www.quantum.com

Quantum

CS00163A-v04