Complete Genomics Reduces Cost of Genome Data Management with SGI ArcFiniti

Key Facts

Organization:
Complete Genomics

Primary Location:
Mountain View, California

Industry:
Genome Sequencing

SGI has a long history and track record of success in the field of life sciences. Drawing on its superiority in graphics technology, the early Silicon Graphics offered scientists the ability to model complex molecular structures. Today, SGI is the trusted leader in technical computing, and is still meeting the complex needs of researchers, including those in genomics, by providing solutions for big data problems associated with next-generation sequencing (NGS) data storage and analysis.

Genomics: Partly a Data Management Problem

A key problem in life sciences, particularly with NGS data, is the sheer volume of data generated. Mountain View, California-based Complete Genomics provides whole human genome sequencing to researchers worldwide. The company was founded in 2005 with the goal of broadly enabling researchers to conduct large-scale whole human genome studies as a pathway toward more personalized approaches to medicine.

To meet this goal, Complete Genomics set out to offer affordable outsourced sequencing services to researchers, requiring higher throughput and a completely new set of data management and analytics capabilities. At the same time, the IT staff at Complete Genomics had to balance the requirement for a large and expandable storage and archiving solution – with the potential to grow from terabytes to tens and even hundreds of petabytes as new sequencers come into use – with the need to minimize total cost of ownership in order to remain competitive.

The scale of the data problem faced by Complete Genomics is typical of genomics environments everywhere: extreme amounts of data being produced in parallel by multiple sequencers. A single genome report, for example, weighs in at about 30GB, not including the over 500GB of reads and mapping data that goes along with it.

Although a single sequence run takes about 11 days, Complete Genomics' 24 production sequencers, running in parallel, produce 800 sequenced whole human genomes per month. This data feeds a post-processing pipeline across a cluster of servers. The result is that the Complete Genomics sequencing pipeline generates a relentless flood of about 30TB of new data each day. Not only does the storage infrastructure need to be sized to cost-effectively handle this volume, but it needs to do so in a way that absolutely ensures data integrity. An 11-day cycle is not something that can afford to be repeated.

"Complete Genomics is in the business of providing an accurate and reliable whole human genome sequencing service. SGI enables us to serve our customers, who are using our service to do critical research. Ensuring our customers' success ensures our success."

Tony Hansmann
Storage Infrastructure Manager
Complete Genomics

sgi

Another key issue with this type of workflow, which is seen in many big-data environments, is the fact that less than 10% of the data compiled is generally needed to be read again. The problem is that even though only a small amount of that data flood ever needs to be accessed again, when it is needed, it is absolutely critical to get it quickly.

This presents the challenge of how to justify the cost of storing inactive data on high-cost active disk arrays. Transactional disk arrays generate heat, burn power and have a high acquisition cost. And yet, when the data is needed, it is there. What was needed for this workflow was an active archive, something with a performance profile of transactional disk arrays, but with much lower acquisition and operational costs.

Finding a Solution

For several years, Complete Genomics relied upon a commercial cloud provider for its data storage needs. The cost, coupled with the fact that it was physically impossible to effectively move this volume of content into and out of the cloud in a timely fashion, caused them to look for a complementary solution to augment their data storage needs. They needed to find highly available, highly protected and highly reliable storage, with absolute protection for data loss or corruption. And they needed to find a way to do this in an active archive.

Complete Genomics considered tape, disk and hybrid solutions combining the two technologies. In the end, the company opted away from tape. While tape is valid in many workflows, the slow access time and lack of inherent data protection meant that cost and usability did not fit their requirements.

The SGI ArcFiniti active archive system rose to the top as the ideal solution that combines the speed, data protection and cost profiles needed for managing this type of workflow.

ArcFiniti provides high-performance network access with a disk-based archive solution that offers performance, accessibility and long-term data integrity advantages over tape. Leveraging patented SGI technology to significantly reduce power consumption and ensure data integrity, ArcFiniti can be scaled from an entry-level system of 156TB to 1.4PB of usable storage in a single rack before compression. This results in significant infrastructure savings over conventional archive systems, while also facilitating easy and immediate access to archived data to users.

ArcFiniti is designed not only to house archived data, but also to protect it for very long term retention. Employing patented software to proactively monitor the health of the hardware and the integrity of the data, ArcFiniti ensures that archive data is still valid and accessible many years later. If ArcFiniti detects any issues due to mechanical problems, it proactively migrates data and verifies data integrity while alerting administrators to replace the faulty part. This dramatically reduces the need to take the system offline for costly RAID rebuilds.

Virtualized Tiers for Better Performance and Control

As a fully integrated solution, ArcFiniti is designed to be easy to deploy and easy to use. Users can access ArcFiniti via high-performance network connections for high-volume throughput into its primary disk cache. This cache is virtualized to the archive tier and managed in the background by an automated archive policy engine. All files are always in an available online state to users, with ArcFiniti ensuring that archive content is protected for long-term retention in the most cost-effective storage tier.

For Complete Genomics, this translated into a total cost of ownership savings of over 40% over three years, with no compromise to the demands of their workflow.

The SGI ArcFiniti system was installed at Complete Genomics' data center after the system was pre-built and tested at the SGI manufacturing facility in Wisconsin. The final system, when configured, required only one day to install and integrate into Complete Genomics' work flow. Within the first 16 hours, around 26TB of data was migrated from their transactional arrays into ArcFiniti over the network, immediately freeing up online disk for more data. As of the end of 2011, the system has already been expanded to 1PB of storage capacity.

SGI Delivers Value and Performance to the Business

SGI and Complete Genomics continue to work together, beyond installation and maintenance, to provide key capabilities needed to deliver end-to-end human genome sequencing services to Complete Genomics' customers. As its business grows and data requirements increase, storage solutions from SGI will scale to meet these needs, providing continuity and reliability for end users.

Please visit sgi.com/go/genomics for more information.

Global Sales and Support: sgi.com/global