DDN® STORAGE

SUCCESS STORY

ACCELERATE: GENOMICS RESEARCH

The Wellcome Trust Sanger Institute Relies on Scalable, High-Performance Storage from DDN® to Reduce Global Health Burden

CHALLENGES

- Obtaining a robust IT infrastructure with large-scale, high-throughput performance is essential to supporting Sanger Institute's diverse research community, encompassing over 2,000 scientists worldwide
- Supporting major sequencing technology advancements, including powerful new machines, created a surge in data volume and computational analysis
- Resolving their classic Big Data problem was complicated by unpredictable data growth as the amount of data and computational analysis varied by workload and research project

SOLUTION

- DDN SFA® high-performance storage engine, EXAScaler™ Lustre® file system appliance & iRODS

RESULTS

- Solution delivers unprecedented levels of throughput and scalability to support tens of thousands of data sequences requiring up to 10,000 CPU hours of computational analysis
- Business goals of ensuring open data sharing by making it easy for the worldwide scientific community to collaborate and access the latest data and analytics are met

THE WELLCOME TRUST SANGER INSTITUTE, A CHARITABLY FUNDED GENOMIC RESEARCH CENTER located in the United Kingdom, is a world leader in studying the impact of genetics and genomics on global health. Since its inception in 1993, Sanger Institute has developed new understanding of genomes and their role in biology while delivering some of the most important advances in genomic research.

The organization played a pivotal role in the Human Genome Project (HGP), an international, collaborative research program whose goal was the complete mapping and understanding of all the genes of human beings. Sanger Institute also led the research on the first human chromosome and ultimately contributed one-third of the human genome to help the project reach its goals by April 2003.

Over the years, the Institute's work has led to other major contributions that have helped reduce the global health burden and lay the foundation for new diagnostics and therapeutics. For example, Sanger Institute has contributed 40 percent of the genome sequence of the malaria parasite, which kills around 600,000 people each year and causes debilitating illness in more than a half-billion people worldwide. In striving to better understand human health and disease, the Sanger Institute conducts significant research to determine the genetic basis of global diseases, such as diabetes, heart disease and cancer.

THE CHALLENGE

As one of the top five scientific institutions in the world specializing in DNA sequencing, Sanger Institute has remained in lockstep with major advancements in next-generation sequencing. For the organization, which has more than 900 employees including ten of whom are part of a team ensuring the well-being of back-end systems, the latest sequencing technologies created a surge in the volume of DNA sequencing that can be produced and analyzed.

To that end, having a robust IT infrastructure with large-scale, high-throughput performance is critical to the Sanger Institute's ability to meet the needs of its diverse research community, which encompasses more than 2,000 scientists from around the globe. Many of them access data through the organization's website, which results in 20 million hits and 12 million page impressions each week.

Equally important is scalable, reliable, high-performance storage that serves multiple purposes, ranging from a landing area for 30 DNA sequencers in the Institute's Illumina® Production Sequencing core facility, each of which pumps out about one terabyte of data daily, to sophisticated storage that supports a Lustre file system for complex computational analysis as well as the Integrated Rule-Oriented Data Management System (iRODS) for managing large data collections.

Over time, keeping pace with rapid data growth and around-the-clock access demands put extra pressure on the organization. With unpredictable data growth, it became difficult to scale storage

> *The sequencing machines that run today produce a million times more data than the machine used in the human genome project. Today, we produce more sequences in one hour than we did in our first 10 years. For instance, a single cancer genome project sequences data that requires up to 10,000 CPU hours for analysis and we're doing tens of thousands of these at once. The sheer scale is enormous and the computational effort required is huge.*
>
> **Phil Butcher**
> Director Of Information Communications Technology, Wellcome Trust Sanger Institute

> *If you need 10,000 cores to perform an extra layer of analysis in an hour, you have to scale a significant cluster to get answers quickly. You need a real solution that can address everything from very small to extremely large data sets.*
>
> **Tim Cutts**
> Acting Head of Scientific Computing, Wellcome Trust Sanger Institute

> *Our storage solution has to give us incredible scaling. If we need to add a new sequencer, we have to expand quickly and without disruption.*
>
> **Phil Butcher**
> Director Of Information Communications Technology, Wellcome Trust Sanger Institute

sufficiently without overburdening the Institute's existing 10-GigE network infrastructure or encroaching beyond its one petabyte per floor tile rule in the space-constrained data center.

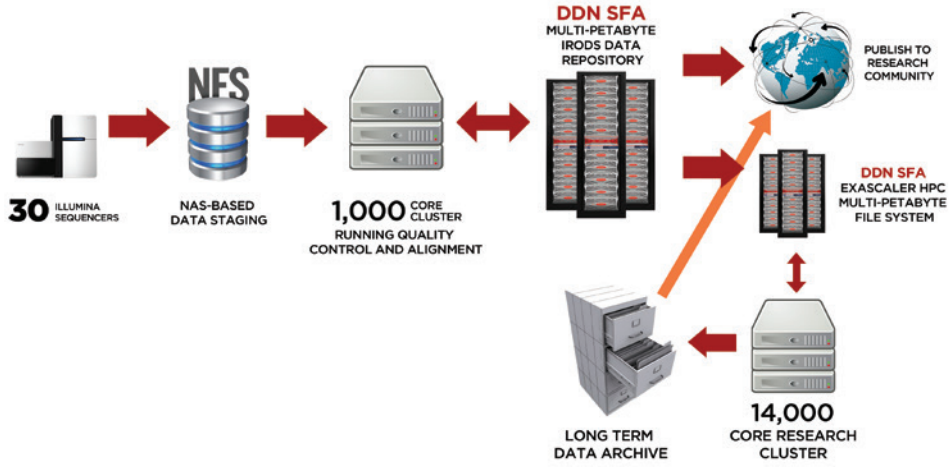
"Our goal is to have ample storage to accommodate different research workloads with varying levels of computational analysis and storage requirements as they all differ in the amount of data that's produced," says Butcher. "This makes it difficult to make data growth projections as it's more than just the sum of new data; the impact it has on the quality of data produced previously is equally important, so there's a lot of reuse to help teams accumulate new answers as they go along."

As a result, the Institute developed a classic "Big Data" problem that is further exacerbated whenever new advances in sequencer technology produce more sequencing data faster than ever before. "Almost overnight, we went from 100 machines to the equivalent of 700 machines over two months," notes Cutts. "As the scale got bigger, so did the need to support our Lustre file system with the fastest, most reliable storage available."

THE SOLUTION

Institute's new sequencing capabilities, the team needed a top performer that would work seamlessly with its existing Lustre file system. "We wanted to use technology we already had and understood," Cutts adds.

With more than two-thirds of the top 100 and one-third of the top 500 supercomputers in the world powered by DataDirect Networks (DDN), the team looked at the leader in HPC Storage. "We anticipated that our solution would come from a strong player in the top 500 and if you can serve some of the fastest computers on the planet, then you can help us," says Butcher.



DDN won the business by presenting a superior Lustre solution, which was reinforced by end-to-end technical expertise on all aspects of high-performance computing and scalable storage architectures. "One of our requirements was that our provider must understand what we are trying to achieve and recognize what would best meet our needs," says Cutts.

DDN also understood the Institute's demands for the highest throughput to push genome sequencing through the pipeline. For a cancer genome sequence, for instance, the Institute produces two-to-five times more data during analysis, which results in a massive amount of information that must be stored, managed, accessed and archived.

As part of its decision process, the Institute put DDN Storage Fusion Architecture® (SFA™) to the test to determine if the platform could meet its exploding capacity and performance requirements. With assistance from DDN, the Institute tested the hardware and servers, ran codes on the platform and got to know the DDN team that would potentially support them going forward.

BUSINESS BENEFITS

- DDN enabled Sanger Institute to further its exploration of groundbreaking scientific and medical research
- Sanger Institute is well positioned to keep pace with advancements in sequencing technology with storage that can scale seamlessly without replacement or forklift upgrades
- DDN enables Sanger Institute to achieve its business goal of ensuring open data sharing by making it easy for the worldwide scientific community to collaborate and access the latest data and analytics

TECHNICAL BENEFITS

- Flexible scaling ensures that Sanger Institute has sufficient storage performance to support downstream analysis, which is difficult to predict and varies by workload and project
- DDN's ability to engage at the software level ensures optimal Lustre and iRODS performance
- DDN's high-throughput storage enables Sanger Institute to upgrade its 10GbE network to 40GbE without replacing its storage infrastructure

After installing its initial SFA10K storage platforms powering EXAScaler™ parallel file systems running Lustre, the Institute could easily keep pace with ever-evolving computational and analysis demands. The team also took advantage of DDN's technology advancements to double initial performance from 3GB/s to 6GB/s, then grew to 10GB/s before a significant increase to 20GB/s.

The Institute also added an additional SFA10K to better leverage iRODS. "We needed to manage an avalanche of data and iRODS gave us a really useful way to deal with it as well as address our replication strategy and other good things rules can offer," notes Cutts. "Performance is critical to us as we have to manage this large pile of data most efficiently."

THE BENEFITS

With DDN's industry-leading performance and 22.5 petabytes of usable storage capacity, the Institute is well positioned to continue driving groundbreaking scientific research while accelerating new medical discoveries. The organization also is making major headway in taming Big Data growth while scaling its environment more cost-effectively.

Flexible scaling is also crucial to ensuring that scientists and researchers worldwide have sufficient storage performance to support downstream analysis once sequencing is completed. The Institute understands that rapid performance surges can occur quickly, which necessitates a rapid response. "If you need 10,000 cores to perform an extra layer of analysis in an hour, you have to scale a significant cluster to get answers quickly," says Cutts. "You need a real solution that can address everything from very small to extremely large data sets."

For the Institute, unprecedented levels of bandwidth and IOPS provide a storage foundation for its most demanding applications. "Our current I/O bottleneck is the network, which we're upgrading from 10GbE to 40GbE," says Butcher. "Then we'll be able to scale up our computing bandwidth easily so storage will never be a bottleneck."

Equally important is the ease with which DDN storage accommodates Sanger's Lustre and iRODS requirements. "Our storage solution must engage at the software layer to make sure the storage works together with Lustre and iRODS in the best and most effective way," Cutts adds.

To further improve data accessibility for its diverse and growing user community, the Institute is exploring DDN's WOS® distributed object storage platform, which is a turnkey appliance based on a cloud and object storage foundation. "We have to explore available emerging technologies that could play a significant role in our future architecture," says Cutts. "We need solutions that give us a much better way of providing storage with good access controls through iRODS. To that end, the Institute is examining solutions that enable increased collaboration and data sharing as part of a private cloud application."

As the Sanger Institute continues to push the boundaries of scientific research, DDN will be able to respond with stable storage solutions that can grow on-demand to meet a variety of computational demands without replacement or forklift upgrades. "When we make another step-change in sequencing performance, we expect to be able to anticipate it and then manage it," concludes Cutts.

ABOUT DDN®

DataDirect Networks (DDN) is the world's leading big data storage supplier to data-intensive, global organizations. For more than 15 years, DDN has designed, developed, deployed and optimized systems, software and solutions that enable enterprises, service providers, universities and government agencies to generate more value and to accelerate time to insight from their data and information, on premise and in the cloud. Organizations leverage the power of DDN technology and the deep technical expertise of its team to capture, store, process, analyze, collaborate and distribute data, information and content at largest scale in the most efficient, reliable and cost effective manner. DDN customers include many of the world's leading financial services firms and banks, healthcare and life science organizations, manufacturing and energy companies, government and research facilities, and web and cloud service providers. For more information, visit our website www.ddn.com or call 1-800-837-2298.

DDN® STORAGE

BOTTOM LINE RESULTS. ACCELERATED.

SALES@DDN.COM | +1.800.837.2298

v3 (4/15)