CASE STUDY

HIGH-PERFORMANCE STORAGE FOR GENOMICS RESEARCH

As a world leader in genomic research, the Wellcome Sanger Institute is tackling some of the most difficult challenges in healthcare and science. Their work spans five research fields — from understanding aging to demystifying infection genomics to closing in on a cure for malaria. Each area of study is urgent and each shares one constant: massive amounts of data. The Sanger Institute worked with Cray to set up an HPC storage solution that keeps their data working and their researchers discovering.

CHALLENGE

Data can accelerate the pace of discovery. Or it can bring it to a halt.

The latter isn't an option for the Wellcome Sanger Institute. As a world leader in genome research, the Institute provides critical insights into human and pathogen biology — insights that are answering questions of human and animal health and with them, changing approaches to science and medicine.

The Institute's Informatics Support Group (ISG) is one of the teams behind the discoveries. They provide the support, architecture design, and resources the scientists need.

A key challenge for the group is dealing with the wide range of applications and workloads up to very large scale generated by genomics research. The variety is compounded by codes that range from well-optimized to much less so. Then the jobs themselves are very I/O intensive.

"It's a huge scale-out I/O problem," says Peter Clapham, leader of ISG . That's because the high-value base files might be 100 GB in size.

But a researcher may need to look at only 2 MB slices. "That's a big I/O challenge. You have to see through to what you need, bring it into your workflow, load it into your memory. And it's happening thousands of times per second."

So, unlike a massively parallel compute scenario, the Institute's workloads are high throughput. Individual jobs tend to not use many processors but thousands run simultaneously — up to 2 million jobs per week.

CLUSTERSTOR DEPLOYED QUICKLY, DELIVERED INNOVATION, AND PERFORMED BETTER ON THE INSTITUTE'S WORKLOADS.

When it came time to replace an aging storage solution, the variety, volume and I/O-intensive nature of their workloads dictated the Institute's requirements. They needed a storage system with the performance and capacity to handle millions of jobs per week. And they needed one that wouldn't stall their research work through endless system modifications or user-facing complexities.

CRAY

SOLUTION

For an organization whose main product is insights, Clapham and the Institute's Head of Scientific Computing Tim Cutts knew they needed a turnkey solution that could be delivered and deployed quickly.

"We want our system administrators' time to be devoted to tackling the scientific questions our users have and not worrying about the details of how a particular piece of tin fits into the datacenter," says Cutts.

They chose the Cray® ClusterStor™ L300N storage system. This hybrid SSD/HDD solution had several selling points.

One, it features flash-accelerated NXD software that redirects I/O to the appropriate storage medium. It delivers cost-effective, consistent performance on mixed I/O workloads while shielding the application, file system and users from complexity through transparent flash acceleration.

When time spent retraining customers means lost insight, transparency jumps to the top of the priority list. "With Cray we felt we were getting some innovations with the NXD components," says Clapham. "It delivered better performance particularly with small files. We have a large number of small files that are indexes for those large files."

Two, the Institute doesn't use InfiniBand, so they were interested in the high-bandwidth Ethernet connectivity solutions ClusterStor offers.

Three, they wanted a short time to system deployment. "It's taken up to six months previously to get a system right for our environment," says Clapham. "This time around we got it done in under a month."

While the ClusterStor solution easily checked off all the technology specifications, what made another big impact on their decision was the expertise of the Cray team.

"We needed Cray to have a strong, knowledgeable Lustre team," says Cutts. "And we got that. It's been a high-value component."

CRAY CLUSTERSTOR L300N

Hybrid SSD/HDD storage

Flash-accelerated NXD software for mixed I/O patterns

Shields spplication, users and file system from complexity

Best value for mixed I/O workload performance

CUSTOMER PROFILE

The Wellcome Sanger Institute is one of the premier centers of genomic discovery and understanding in the world. It aims to deliver new insights into human and pathogen biology that change the course of biology and medicine. Using genome sequences, the Institute is increasing understanding of human and pathogen biology in order to improve human health.