Wellcome Trust Sanger Institute accelerates world-leading research

Using analytics to optimize high-performance computing capacity across 15,000 cores

Overview

The need

The Wellcome Trust Sanger Institute wanted to help its researchers stay at the leading edge of research by ensuring that their HPC workloads ran quickly—but legacy optimization tools made the task difficult.

The solution

The Institute implemented IBM® Platform™ Analytics software on top of IBM Platform LSF®, its policy-based workload scheduling solution, and designed dashboards for researchers and administrators.

The benefit

Ensures jobs run on the optimal compute nodes; dashboards help users identify ways to improve their jobs; utilization metrics strengthen the business case for future hardware investments.

Founded in 1993 to collaborate in the mapping, sequencing and decoding of the human genome, the Wellcome Trust Sanger Institute (Sanger Institute) is a non-profit research organization. From its campus in Hinxton, England, the institute's 1100 personnel collaborate in the global effort to understand the biology of genomes and their role in health and disease.

As a major participant in a number of international research projects—including the 1,000 Genomes Project and Cancer Genome Project—the Sanger Institute wanted to offer its researchers the tools they needed to collaborate effectively with other research teams around the globe.

Dr. Peter Clapham, Principal Systems Administrator, Informatics Systems Group at the Wellcome Trust Sanger Institute, explains: "Genetics and genomics research are fast-moving, competitive fields. Because many teams around the world are collaborating on individual projects, and others are working towards the same goal in parallel, securing funding and meeting collaboratively agreed deadlines are essential."

Need for optimal resource utilization

To meet agreed deadlines with on- and off-site collaborating partners, the Sanger Institute needed to ensure optimal utilization of its high-performance computing (HPC) resources.

"Our HPC environment was originally designed to help map the human genome, and has been growing ever since; today, we have 15,000 active cores spread between a mixture of high- and low-memory nodes across nine large clusters," says Dr. Clapham. "The nature of our research means that most of the distributed workloads we process are embarrassingly parallel. Although these types of workloads lend themselves to high throughput rates, they require careful management to ensure efficient resource utilization."

"Analytics is one of the most significant IT investments that we have made. We needed the IBM solution to deliver the HPC utilization required to help our research teams generate results rapidly, meet their publication deadlines and, ultimately, secure new funding."

—Dr. Peter Clapham, Principal Systems Administrator, Informatics Systems Group, Wellcome Trust Sanger Institute

In the past, the Sanger Institute found optimizing its HPC resources a considerable challenge. The volume and velocity of sequencing data were both steadily increasing, and a significant number of research projects depend on the clusters. Some of this processed data is subsequently presented to the world through various portals and websites, such as Ensembl and Cosmic. Similarly, sequencing data is made available via sequencing archives, allowing worldwide access to the data produced at the Sanger Institute.

"Our gene sequencing workloads are searching for the tiniest changes in massive data sets," says Dr. Clapham. "On top of our own sequencing work, we receive additional data from contributors at other institutions. In total, we now sequence approximately 520 terabases per year. We are now seeing an increasingly large proportion of directly clinically relevant research performed on our clusters, the end results of which are now being made available through public websites. With the demands on our finite compute resources increasing, tracking cluster resource usage was a key priority."

Using its legacy reporting tools, the Sanger Institute discovered that its low-memory nodes were dramatically underutilized—reducing the overall efficiency of the cluster. Dr. Clapham comments: "When submitting a job to the cluster, our researchers need to estimate the amount of memory required to complete the task. In the past, most of our users were creating jobs with high memory requirements, which were automatically allocated to the high-memory nodes."

"We believed a large proportion of these jobs were using far less memory than our users had estimated, and could run on the low-memory nodes. However, without deep insight into our HPC metadata, it was effectively impossible to determine which jobs these were."

Deploying IBM Platform Analytics

To deliver the insights it needed to optimize its HPC environment, the Sanger Institute used IBM Platform Analytics software. Platform Analytics provides rich, scalable, near real-time reporting capabilities straight out of the box through a set of generic dashboards. In addition, the framework itself can easily be extended to add external data, create business-specific dashboards and send scheduled reports via email.

"We have used IBM Platform LSF software to help schedule workloads on our clusters for a number of years," says Dr. Clapham. "Our positive experience with Platform LSF gave us confidence that the IBM Platform Analytics solution could help us to achieve our operational goals."

Working together with a team from IBM, the Sanger Institute implemented IBM Platform Analytics, which provides fine-grained insights into all jobs running on the cluster.

Solution components

- IBM® Platform™ LSF®
- IBM Platform Analytics

"When the time comes to make the case for new capacity, we will be able to use metrics from the IBM solution to demonstrate that our current compute usage is optimal, and show where additional capacity is required to boost performance."

—Dr. Peter Clapham, Principal Systems Administrator, Informatics Systems Group, Wellcome Trust Sanger Institute

"The deployment went very smoothly," says Dr. Clapham. "With the Platform Analytics solution in place, we designed a number of dashboards for our end users—enabling them to see the utilization of the cluster according to criteria such as job, job submitter and memory utilization. The dashboards are refreshed every ten minutes, which means we can offer near real-time insights to our users and systems administrators."

Optimizing HPC capacity usage

The IBM Platform Analytics solution provides the Sanger Institute with full visibility of the actual memory used by each job on the cluster. Using the results of this analysis, the organization has successfully modified its scheduling processes to utilize the full capacity of the HPC environment.

Dr. Clapham comments: "We can see a traffic-light dashboard of memory utilization across the cluster, with research groups that have consistently requested more memory than their jobs actually require highlighted in red. We can then drill down to see which individual jobs can run on our low-memory nodes, and make the adjustment manually—dramatically improving throughput."

In addition to correcting memory requirements, the IBM solution enables the Sanger Institute to help its users to design more efficient jobs. "Based on our insights from Platform Analytics, we've changed the fair-share mechanism to a hierarchical model—meaning that each research group now gets a fixed proportion of our overall compute capacity," says Dr. Clapham.

"The hierarchical fair-share policy, combined with full visibility of the jobs currently running on the cluster, encourages our research groups to self-police. As a result, we are now seeing more and more jobs submitted with accurate memory requirements—which means more efficient use of our compute resources overall. In fact, one group has now met all of its current research deadlines, and has actually started processing work from collaborating teams."

The Sanger Institute predicts that the ability to supply detailed metrics on cluster utilization will make it easier to determine the requirement for additional compute nodes in the future, and to make the business case for the investment.

"As a non-profit organization, we don't just have to optimize our compute utilization; we have to optimize our spending on compute capacity also," says Dr. Clapham. "When the time comes to make the case for new capacity, we will be able to use metrics from the IBM solution to demonstrate that our current compute usage is optimal, and show where additional capacity is required to boost performance."

Dr. Clapham concludes: “Analytics is one of the most significant IT investments that we have made. Thanks to the IBM solution, we can deliver the HPC utilization required to help our research teams generate results rapidly, meet their publication deadlines and, ultimately, secure new funding. If we turned Platform Analytics off tomorrow, capacity planning and user metrics would be severely impacted. Looking to the future, we consider the metrics provided by Platform Analytics an essential part of the LSF package.”

For more information

Contact your IBM sales representative or IBM Business Partner, or visit us at: **ibm.com**/platformcomputing

IBM Corporation
Systems and Technology Group
Route 100
Somers, NY 10589

Produced in the United States of America
November 2014

DCC03031-USEN-00