Life Sciences Institute Handles Data Deluge with Universal Storage and Access

See how Australian medical researchers overcame data access bottlenecks to speed complex analysis and modeling tasks

CREATED BY:

ActualTech MEDIA

SUCCESS STORY

Requirements

A simple and scalable all-flash solution that serves extreme data and access performance needs for medical researchers

Solution

Petabytes of VAST Data's Universal Storage architecture with high speed access to huge data volumes for analysis, animation, and complex 3D modeling

Results

Researchers can perform complex analysis and modeling runs without waiting overlong for results

CHALLENGE

The Complex Demands of Medical Research

As at other leading medical research facilities, WEHI is deeply engaged in the fields of genetics, structural biology, and life sciences. Working in each of these fields involves acquiring, managing, analyzing, and operating on huge amounts of data, where a single project may involve billions of files, both large and small.

WEHI applications related to cryo-electron microscopy (Cryo-EM), protein folding, and genomic analysis place unique and taxing demands on underlying computing technologies, particularly storage. Indeed, the majority of storage solutions—including both legacy systems and new technologies—struggle (and often fail) to provide affordable and acceptable performance across all the datasets involved in such work. Thus, fast, affordable flash storage is a critical element in providing capable and usable application performance for researchers.

What makes Cryo-EM datasets particularly taxing is the extraordinary variation between data sizes and access mechanisms used at each step in the processing pipeline. Though raw datasets are enormous, the process of cleanup and correction usually eliminates as much as half the original data in moving from one step to any next step. This changes the access behavior for datasets to progressively

smaller, random IO patterns as they move through the pipeline from one end to another. Random reads don't benefit from pre-fetching, so they put a much more taxing burden on storage retrieval systems. Indeed, this explains why all-flash storage is essential to handle such work, because device latency and access times are much faster than spinning disks (and "reading ahead" or "reading in line" confers no value).

Protein folding is another complex process with demanding performance requirements across multiple pipeline stages, which benefits greatly from AI-based systems and platforms. Thus also, WEHI has invested substantial time and effort into Deepmind's AI-based AlphaFold software and database. AlphaFold requires a massive amount of calculation across a huge number of individual data points.

The previous multi-tiered storage that WEHI used suffered from delays related to the highly random and variable I/O access patterns typical for Cryo-EM workloads. That's because they involve large amounts (1TB to 10TB) of different sized files and objects, where large objects can be up to hundreds of gigabytes in size, and small ones under 1KB. Multi-tiered storage has difficulties in staging and retrieving such data collections quickly and effectively, because I/O requirements for access to large objects involves vastly different access patterns, durations, and data rates than those for small ones.

Optimizing I/O latency and performance is crucial for moving data and images from intake all the way to final outputs.

Optimizing I/O latency and performance is crucial for moving data and images from intake all the way to final outputs. Only then can WEHI's researchers undertake the real work of interpreting data from Cryo-EM 3D models and AlphaFold protein structures, unlocking and decoding the information they contain.

SOLUTION

Universal Storage Accelerates Data Processing

The VAST Data Universal Storage solution has delivered substantial benefits to WEHI researchers working on Cryo-EM analyses and protein folding studies, among other data-intensive projects. These include:

Accelerating data collection from Cryo-EM devices to make it available quickly to applications at scale.

Researchers report that they need no longer wait for data to make its way into applications for modeling and analysis. VAST Data Universal Storage is fast enough to let work get underway on data coming through the front-end of the pipeline, even as the back-end is still filling up.

Exploiting GPU-based processing (such as NVIDIA's GATK Whole Genome Sequencing through CLARA Parabricks).

This means processing speeds for complex Cryo-EM and protein folding workloads run more quickly, and take better advantage of uniform and fast access speeds and enormous working sets for data in VAST Universal Storage to get more done, more quickly. This is a noticeable boost in an environment where big jobs are routinely scheduled months in advance, and where individual jobs can take a year or more to complete. WEHI estimates performance improvements from 10% to 20% (sometimes more) compared to its previous architecture. This means the same facilities can handle more and bigger jobs more quickly than they could using multi-tiered storage.

VAST speeds up the I/O-intensive Cryo-EM pipeline.

For the all-flash Vast Data Universal Storage, however, handling such collections of data objects works as quickly and efficiently for large objects as it does for small ones.

VAST provides fast and reliable data access to researchers working on protein folding.

VAST enables full pipelines for AlphaFold and other protein folding application pipelines, ensuring researchers never have to wait for data.

RESULTS

Overcoming Data Access Bottlenecks in Medical Research

WEHI researchers reported after deploying VAST Data's Universal storage on their networks, that IO delays were a thing of the past. In fact, none of the applications they use with VAST Data storage has been constrained by its ability to deliver data. This helps explain how VAST has delivered improvements in job completion times and total workload handling capacity.

In terms of relative metrics, that old system couldn't deliver more than 2 Gbps of data, or 20,000 IOPs. The VAST-based system has yet to be pushed beyond its limits. Says Martin: "We are now getting over 10 Gbps and more than 200,000 IOPs from VAST, and we're not even stretching the system's capabilities." This is adding impetus to migrate all of its users from any and all systems using legacy storage onto VAST, so that WEHI can realize further improvements to its data handling and job completion rates. This makes VAST Universal Data Storage a key component in WEHI's future systems design and development plans.

ABOUT WEHI

WEHI is a world-class medical research facility based in Australia. The Institute operates locations in Parkville and Bundora, and is affiliated with postgraduate programs at the University of Melbourne and the Royal Melbourne Hospital. As such, it has world-class needs in terms of data management.

CUSTOMER QUOTE

"VAST was much cleaner than everything else we looked at. Other options were too expensive and complex. Our previous system used to top out at 50 simultaneous jobs; now, we can run over 1,000 jobs at the same time. We spend less time troubleshooting (file systems, mostly), and can spend more time helping researchers."

Tim Martin

Director of IT Services Research Systems Group, WEHI

VAST

Explore the website, then request a demo. See how VAST Data can boost your organization's productivity.