Storage in life sciences

With data rates increasing and more complex challenges arising in life sciences, Panasas' Jim Donovan and Dale Brantly discuss the benefits of using HPC for these workloads

Innovations in genomics, predictive biology, drug discovery, bioinformatics, personalised medicine and other life science disciplines depend on the ability to rapidly store, retrieve, and analyse increasingly large volumes of data. For many institutions, legacy storage systems just aren't fast or scalable enough to meet the challenging requirements of modern life science applications, which is why researchers are increasingly turning to HPC computation and storage architectures.

The explosion in genomic and proteomic information, especially from next-generation sequencing, and the emergence of data-intensive molecular diagnostic techniques, demand computing infrastructure that can match the workload requirements. Parallel file systems used in HPC are one answer to this challenge, particularly for workloads in computational chemistry, bioinformatics, and other data-intensive applications. Parallel file systems allow scientists to manage large datasets, such as microarray or mass spectrometry images, in a single scalable namespace.

Praneetha Manthravadi, director of product management at Panasas, described the firm's Activestor system.

'It's a very flexible system. HPC workloads used to be large streaming files, but things have been changing in the past few years. Lots of different workloads and applications come with different requirements, and this system is exceptional at mixed workloads and delivering that consistent performance.

'Activestor is a parallel file system that has been completely engineered to deliver enhanced flexibility, due to the modular components. It delivers high performance of up to 35 Gb/s per rack and it is delivered on industry standard hardware, which helps to drive the cost efficiency,' added Manthravadi.

Portable storage

This latest system from Panasas is aimed at delivering a portable storage architecture which can help life science users, because it reduces the burden of designing and maintaining the latest storage technology for their organisation.

Jim Donovan, Panasas chief sales and marketing officer, said: 'We are delivering a high performance product with expectations of performance that goes through a qualification process on the hardware.

'We are using industry standard hardware from suppliers. We do the qualification, porting and certification that ensures the product delivers the performance manageability and reliability that the HPC customer expects.

'Users get the benefit of the latest hardware at every step along the way. They take advantage of the cost benefit of using standard hardware, so there is no custom design from us. What you get is the latest hardware, the highest level of performance and the highest level of manageability and reliability in a competitively priced product,' adds Donovan.

WhiteMocca/Shutterstock.com

"If you think about some of the popular applications such as genomics, you are dealing with a long string of data that is well suited to HPC"

Brantly said that many life science applications are similar to the traditional HPC workloads, such as CFD or weather forecasting, due to the nature of the large streaming I/O files which characterise the I/O patterns in these workloads. 'If you think about some of the popular applications, such as genomics, you are dealing with a long string of data that is well suited to HPC,' said Brantly.

'The other thing in the medical field is the categorisation and reading of medical images. Whether it be a brain scan or radiology, you are looking at images which are very large picture files. That is something that HPC is particularly good at,' added Brantly.

'We have these massive libraries of DNA sequences, The National Center for Biotechnology Information (NCBI) at the National Institute of Health of the Beijing genome institute. All of these organisations are collectors and disseminators of genome data,' said Donovan. 'The amount of data available to researchers is exploding. The problems are getting bigger, but there is also more data to analyse.

'Datasets are getting larger and larger and it is not just the sequencing, but the instruments for collecting images or reading other experimental data are getting finer and finer with larger capabilities so the output from a run is 10 or 100 times larger than it was a few years ago.'

As research programmes become more complex and the size of datasets gets larger, it becomes imperative to develop a computing infrastructure that can adapt to the growth in workloads and sustain the kind of data-intensive applications that are being used in life science industries, such as genomics and medical imaging.

'Not only is more detail available but the data is getting larger. You have several users running a number of different applications. Being able to put data out onto a file system that provides very good response to both super efficient streaming I/O and the less efficient random or smaller I/O for various applications, is key for Panasas customers ', concluded Donovan.