Case Study

Alamy

STORAGE > HIGH SPEED STORAGE

How Pure Storage FlashBlade Powered Infection Disease Research

The McArthur Lab at McMaster University ramped up storage power just in time for the COVID crisis. Learn how Pure Storage flash technology boosted the lab's research capabilities.

Karen D. Schwartz | Mar 09, 2023

Over the years, the lab has developed the world's largest database of data related to drug-resistant infections.

McArthur Lab's genomic datasets doubled every three months, creating a need for massive and responsive storage. For many years, the McArthur Lab got by with a farm of Hewlett Packard Enterprise (HPE) RAID-based spinning disks. However, as the lab grew from about five data scientists in 2014 to about two dozen by 2017, it frequently maxed out its storage capacity.

Related: Data Visualization Platform Boosts Geoscience Research Projects

"We started moving from serial-based, sequential analysis where you look at one case at a time, to situations like drug discovery where we needed to look at tens of millions of data points simultaneously," said Dr. Andrew McArthur, a genomics professor and researcher in charge of the lab. "That required more I/O and high availability. Even though the spinning disk was high performing and we had plenty of it, it couldn't keep up."

The Addition of Pure Storage FlashBlade and FlashArrays

McArthur prioritized performance, agility, and cost when evaluating products to increase the lab's storage power. He ultimately settled on a flash-based technology. He knew the compression capabilities would be invaluable as more of the lab's groups embraced imaging. However, the more important feature was strong I/O performance and flexibility.

"Scientists walk through the door with a different experiment every day, so we needed something that would be nimble enough to change direction," McArthur explained. "I needed to be able to shift people back and forth between spinning disk and flash so I could reprioritize for the experiment of the week."

The McArthur Lab added Pure Storage's all-flash NVMe-based FlashBlade and FlashArrays to the mix, favoring the ability to add more blades as needed and their high performance. That way, when research would require high-performance computing and high data throughput, teams could switch from spinning disk to flash. The original spinning disk farm is now used for projects without fast turnarounds and archival storage.

COVID Pandemic Creates New Demands

By the end of 2019, McArthur Lab had integrated the flash storage into its operations, just before the

datasets related to COVID. “Normally, we fight drug-resistant pathogens, which we call a ‘slow-moving pandemic,’ but if it takes us six weeks to build a dataset, that’s fine,” McArthur noted. “The [COVID-19] pandemic required much more speed.”

Because the lab already had a compute farm and a large DNA sequencing core, the team was asked to work with hospitals to perform sequencing and write the national platform for variants and mutations. The lab had flash storage installed along with an HPE Apollo 6500 high-performance computer and multiple standalone servers comprising 3TB of memory, but McArthur knew that it wouldn’t be enough.

To meet these demands, Pure Storage loaned the lab an additional FlashBlade. With the existing blades and arrays and the disk farm, the lab now had 90TB of storage across seven blades and a total of 1.2 petabytes of storage across all storage tiers. HPE loaned the team additional compute, and Cisco donated two 88-core blade servers. Funding eventually came through for the team to add an HPE Superdome Flex 2-node high CPU and high memory environment that could handle genomic computations at scale.

The extra storage and compute proved to be critical for the lab to keep up with rapid turnaround requirements during the pandemic. All pandemic-related raw data was stored on flash.

“As new variations or mutations showed up, the government would ask us if we had ever seen anything like it before. That meant that we’d have to recompute all the data, which grew exponentially over time, so having that capacity and performance was critical,” McArthur said.

What’s more, the time it took to analyze large COVID-related datasets dropped from up to 14 days to one to six hours.

McArthur Lab relied on the Superdome and Flashblade to process thousands of DNA sequences, which contributed to nearly 20% of Canada’s data on variants during the first and second waves of COVID. The lab processed up to 1,400 positive test results in a day, and FlashBlade’s ability to handle massive parallelism in I/O operations was critical to shortening the time from sequencing to reporting. “We never saw I/O limitations,” McArthur noted. “It almost always came down to compute limitations.”

A New Set of Research Capabilities

Thanks to its expanded capabilities, MacArthur Lab can now perform previously impossible experiments. For example, scientists can analyze gut flora to see how it affects depression using vast

amounts of patient and microbiome data. The lab also has robots that use AI and machine learning to test large amounts of data, which has led to the drug discovery team generating even more data than ever.

The McArthur Lab plans to use AI and ML increasingly for projects like predicting how drugs will work in a clinical setting. However, it requires significant compute power to produce these reference sets. The FlashBlade's massive parallelism is critical in ensuring that datasets aren't IOPS-limited.

In addition, the lab is exploring the use of AI in designing antibiotics, which will require the power of flash. Other projects are related to cancer and placenta development, which involves analyzing enormous datasets of RNA and DNA expression. While these datasets can start on spinning disks, the lab will move the data to flash during the analysis phase. The lab is also considering using flash storage for the growing stores of images created by some areas of the lab.

With all these projects underway, McArthur expects the lab to outgrow its storage every three months. MacArthur plans to add more flash storage once the lab receives additional funding.

"We used to sequence one or two genomes. Now we don't blink when we sequence 10,000. And instead of supporting five professors and scientists, we now support 30," McArthur said. "All of that means more data and more high-performance storage."

About the author

Karen D. Schwartz is a technology and business writer with more than 20 years of experience. She has written on a broad range of technology topics for publications including CIO, InformationWeek, GCN, FCW, FedTech, BizTech, eWeek and Government Executive.