Genomic Threat Surveillance

How McArthur Lab and Pure Storage Chase Superbugs to Discover New Treatments

> “Genomic sequencing and threat surveillance constantly push against the limit. From a storage perspective, we look to Pure Storage to lead the way.”
>
> **ANDREW MCARTHUR**, MCMASTER UNIVERSITY PROFESSOR AND HEAD OF THE MCARTHUR LAB

Overview

This interview is based on a real-life best practices story from the McArthur Lab, a leading research laboratory within McMaster University.

To capture this story, Pure Storage® interviewed Dr. Andrew McArthur, head of the McArthur Lab, to learn more about the fight against superbugs. Questions are followed by answers from Dr. McArthur.

Can you start by telling us a little bit about yourself and the McArthur Lab?

I am a professor at McMaster University and the David Braley Chair in Computational Biology. I am also a member of the Department of Biochemistry and Biomedical Sciences in the Michael G. DeGroote Institute for Infectious Disease Research.

The McArthur Lab focuses on genomic threat surveillance. We track infectious pathogens and build tools for public health and private industry. We were deeply involved in COVID-19 surveillance and mitigation.

What is involved in genomic threat surveillance and why is it important?

Pathogens like viruses and bacteria have genes that can mutate, creating variants that can potentially evade vaccines and drug therapies. By tracking their movement, we have a chance of identifying and mitigating risk. Our primary mission is to support discovery of new drug therapies and revive old ones to offset the loss of treatment options from drug resistance.

The stakes are high. In 2018, 26 percent of bacterial infections coming into hospitals in Canada were drug resistant. That could rise to 40 percent in only a few years because COVID-19 accelerated the use of antibiotics to fight secondary infections, leading to additional drug resistance. That equates to an estimated 400,000 lives, $1.2 billion in hospital costs, and $400 billion in GDP.

At McArthur Lab we run the globe’s largest database on drug resistant infections, pathogenic genes, and mutations. The database combines sequencing results shared by global databases and thousands of peer-reviewed scientific papers. Roughly every five minutes someone on the planet runs a superbug through this Comprehensive Antibiotic Resistance Database (CARD) to understand why their patient is in trouble or to track potential outbreaks.

What role does Pure Storage FlashBlade play in McArthur Lab?

We rely on FlashBlade® for our most critical workloads. No other storage can compete with its combination of performance and simplicity. When our research requires high performance computing, whether GPU or high-speed CPU driven, we also tend to need high data throughput. The only way we can get the IOPS we need is with FlashBlade. Researchers have the confidence to run extensive experiments because they know FlashBlade can handle the load. FlashBlade delivers 5-6 times the IOPS compared to our next closest storage tier.

A research grant for the Institute for Infectious Disease Research brought FlashBlade into the lab to support genomic screening at scale. As compute demands grow, FlashBlade continues to meet our IOPS requirements. With FlashBlade and HPC, a large COVID-19 analysis dropped from 7-14 days to 1-6 hours, which was unreal.

That’s tremendous. In addition to accelerated IOPS, what other benefits does FlashBlade give you?

When our CPUs and GPUs are waiting for storage operations to complete, they are stranded—unable to do any other work. FlashBlade enables us to use the full potential of our computing, which makes it possible to do more projects with the same resources. This increases the value of the compute farm, which is the most expensive part of the environment. These systems run far faster than before and can run many projects simultaneously.

FlashBlade also supports our agile framework, allowing us to support a broad range of applications and pivot quickly as needed. From an academic view, it’s a force multiplier. When we write a grant application for high performance computing, we get a more positive response when we indicate that it builds upon an existing Pure Storage FlashBlade environment.

We have almost zero administrative burden and upgrades are non-disruptive. The all-flash all-NVMe design cuts down on both power and cooling requirements. FlashBlade is extremely efficient, with compression that enables huge volumes of data with a small footprint. With these savings we can devote more power, space, and time to our computing resources. As a result, we lean on FlashBlade as much as we can. FlashBlade is a highly sustainable platform for us.

Why is CARD an important tool for genomic threat surveillance?

CARD is a biocuration or knowledge network that brings together thousands of research papers and predictions of drug resistance from hundreds of thousands of known isolates or variants. The 30-million-row SQL database is human curated, with algorithms in the background doing quality control.

Every month CARD ingests about 4,000 scientific papers using natural language processing (NLP) and algorithms to index the content according to a strict vocabulary. Artificial intelligence, machine learning and analytics correlate the data at massive scale.

We estimate that about 6,000 genes in mutation drive drug-resistant infections. When we combine clinical outcome data with genomic data and apply machine learning models, we can predict quite accurately how drugs will work in a clinical setting.

It takes massive computing power to produce this reference set that can then be used in surveillance. FlashBlade ensures that our high-performance computers are not IOPS limited by providing massive parallelism for high data throughput managed by many CPUs, cutting database recompute from six months to six weeks.

Drug resistance evolves every day. New genes appear seemingly out of nowhere. So basic knowledge changes all the time. If we analyze 500,000 genomes to identify prevalence and then our reference data adds a gene, we need to recompute all 500,000 genomes. Every time we learn something, we need to redo the database. The surveillance refresh process is endless.

Can you give some examples of analyses that can be done with your algorithmic framework?

We can run algorithms against any amount of data because of the high throughput capabilities of FlashBlade. We can run a sample against hundreds of thousands of genomes to see if it is already known. We can also see that, for example, it occurs three percent of the time in one geolocation and 80 percent of the time in another. We can see whether it is associated with high transmission. These are all examples of large-scale analysis and that's where FlashBlade really kicks in.

Another example uses artificial intelligence or machine learning to predict from raw sequencing data how an infection will behave, which drug will fail, and which will not. That requires a very large genomic dataset, sophisticated algorithms, high-performance computing, and FlashBlade. I fully expect to see more machine learning and AI driving significant advances in genomics, and in turn, the need for Pure technology.

How did you get involved with COVID-19 surveillance and what role did FlashBlade play in that effort?

A respiratory pandemic is amongst the worst to face. We were part of the group that isolated the live virus in Canada and started sequencing to figure out what variants existed and where they came from. The platform we designed for analyzing raw sequence data is 10 times faster than any platform previously built. At any given time, we were processing thousands of DNA sequences. We relied on an HPE Superdome Flex 2-node server and FlashBlade to keep up and sequence COVID-19 samples fast and efficiently.

We are programmers but we are also biologists. So, we were receiving DNA from the swabs of thousands of patients, doing the lab work to get the genomic data, and reporting it globally in an international database called GISAID. During the first and second waves in Canada, close to 20% of the data that Canada was sharing around variants came through our group and all of that was run on Pure FlashBlade.

We can usually archive results quickly in genomics, but because of the evolution of variants, there was a constant need to recompute. We stored massive amounts of raw COVID-19 sequencing data on FlashBlade so that when the call came, we had the fastest platform available to check if a newly found mutation was already recorded in any of the national data we housed. FlashBlade ended up being really, really valuable.

I never envisioned having to do sequencing and surveillance work on such a large scale, but it was a pandemic. It was massive. FlashBlade was the only option that could handle the load.

After about a year, health agencies started buying sequencers and we moved into a support role. We had identified about 18,000 variants at that point with the local health authority. We pivoted and built the backend compute for the national data framework VirusSeq in Canada, which ran on FlashBlade.

Your COVID-19 sequencing and analysis process needed to be closer to real time. How did FlashBlade help you achieve that?

It was not unusual to have 1,400 positive test results in a single day that required sequencing as fast as possible. Combining Pure with cloud computing and the ability to reconfigure on the fly was huge.

Traditional genomics work requires high memory. The methods that we invented for COVID-19 surveillance are not high memory, they are high throughput. So Pure was critical for that. FlashBlade can handle massive parallelism in I/O operations and move data through a multi-application workflow without re-copying the data set. This technical advantage resulted in a ten-fold decrease in time from sequencing to reporting.

What are the critical computing and storage technologies supporting McArthur Lab?

Think of our environment as a research as-a-service cluster that combines large-scale computing with hierarchical storage. When someone requests service, we match technology to need and start up containers with the required resources. We have a VMware hypervisor cluster and OpenVZ container-based virtualization.

For machine learning, NLP, and other compute-intensive research, we use an HPE Apollo 6500 system. Our Superdome Flex 2-node high CPU and high memory environment handles genomic computations at scale. We also have multiple standalone servers with about 164 cores and three terabytes of memory. We can deploy between five and 600 cores and have at least 300 available for on-demand use. We built this compute farm because of COVID-19.

FlashBlade is our preferred storage for high-performance computing workloads. When a researcher is using our Apollo or Superdome systems, we move all the data to FlashBlade because that's the only storage that can provide sufficient IOPS. We currently have 90 terabytes across seven blades, and a total of 1.2 petabytes of storage across all tiers.

How do you manage tiered storage and support massive parallelism?

We manage several tiers of storage to support lab operations. FlashBlade is at the pinnacle of our storage choices. We use it for all high-performance activities. Even as we've added GPUs and CPU cores, FlashBlade has not reached a performance limit.

During the peak of the pandemic, we learned that there is also great value in keeping data on FlashBlade for instant access, accelerating surveillance searches, regulatory and compliance requirements, and cybersecurity. We will soon add Pure FlashArray//X to the lab for fast structured databases and mission critical applications.

Because we've had great success and stellar performance with Pure Storage for multiple projects, I expect that when we have opportunities to add to our infrastructure through project-related grants, Pure technology will be top of mind.

What kinds of applications get priority for FlashBlade?

Our lab serves both research projects and the medical community, so we must be agile. A single discovery can change the whole direction of research, and consequently the kind of compute that's needed. One of the biggest challenges is trying to predict what researchers are going to require from a compute perspective.

For example, we have a new investigator who works on the development of the placenta in early pregnancy, using very complex advanced genomics. Her data sets are stuff I haven't seen before, and they're massive. Because we focus on simplicity, when we met with her, we said, okay, what's your volume? What's your compute? And because of FlashBlade's speed, simplicity, and scale we were able to easily redeploy and create an environment for her.

Anything that needs scale, like a massive feature set from chemical genetics for machine learning, goes on FlashBlade because we've learned without a doubt that this kind of application becomes I/O bound quickly. Projects involving machine learning and intelligent systems definitely have to go on FlashBlade. If we put that on anything else, we will be I/O bound almost immediately.

If we need to find an effective treatment for cardiovascular patients, we focus FlashBlade on that. It's an immediate need. And for projects with rapid turnaround, we need FlashBlade too.

What was your work process before FlashBlade?

It takes a highly efficient computing environment to do genomic surveillance and biocuration. Shared computing resources were available outside the lab, but they weren't highly performant or flexible in scheduling, and they raised patient data privacy concerns. So, we invested in local high-performance computing. This accelerated computations, but our storage systems could not keep up.

Before FlashBlade we simply couldn't do it.

How are you evolving technology to keep pace with genomic data?

Genomic data is growing exponentially, and we need to advance our computing environments at a similar rate just to keep up. Yet, the data pipeline is outpacing Moore's Law and Kryder's Law, which estimate how quickly computing and storage density doubles. It is even pushing up against Butter's Law, which estimates doubling of network speed.

We used to sequence one or two genomes; now we don't blink when we sequence 10,000. Instead of supporting three or four professors, now there are 30. They can sometimes generate ridiculously large data sets and need high performance storage to go with it. FlashBlade's speed and compression helps meet these demands.

Our goal is to evolve compute and analysis at the same rate that the genome sequencing itself is evolving. We started from zero knowledge and worked our way up. FlashBlade allowed us to close the gap and now stay closer to real time.

I'm more hopeful in the fight against superbugs. Surveillance has seen big changes in the last three years, as sequencing technology evolved to a point not expected for another 10 to 20 years. That tells me that great challenges can drive great advancements. Genomic sequencing and threat surveillance constantly push against the limit. From a storage perspective, we look to Pure to lead the way.

Learn more about McArthur Lab's journey and how Pure Storage FlashBlade can solve your unstructured data challenges.

purestorage.com

800.379.PURE

PURESTORAGE®
Uncomplicate Data Storage, Forever