University of California-Berkeley researchers employ the latest tools built on Apache Spark to accelerate DNA sequencing in pursuit of precision medicine. Adding Pure FlashBlade™ from Pure Storage® has significantly reduced the time needed to sequence data-intensive DNA samples and analyze results.

BUSINESS TRANSFORMATION

Researchers and clinicians can sequence a greater volume of DNA samples in far less time, generating valuable new insights and, in some cases, delivering faster answers to life-and-death questions.

GEO

North America

INDUSTRY

Healthcare

"Pure FlashBlade allows us to maintain high throughput while horizontally scaling bandwidth."

Frank Austin Nothaft, ***Graduate Student***

UNIVERSITY OF CALIFORNIA BERKELEY RESEARCHERS EXTEND THE BOUNDARIES OF GENOMICS USING APACHE SPARK WITH FLASHBLADE FROM PURE STORAGE

Seldom in history has a scientific breakthrough held such huge potential for humankind as the ability to sequence the human genome. Every day, researchers and clinicians are applying the results of genetic sequencing to treat, cure and even prevent thousands of diseases. And, they are doing so at a faster pace, and often at less cost, than traditional methods that do not employ genetic information.

Sequencing DNA would not be possible without advanced computational tools. But even as these tools are being improved, a major challenge remains: sequencing is a massive data-processing task. The term "big data" barely begins to describe genetic sequencing. A single sample of a person's DNA is about 300GB of raw data, and the latest DNA-sequencing research projects involve as many as 50,000 to 100,000 participants, each of whose DNA will be sampled several times over the course of a research study. The database for a single project can grow into the petabytes. And, there are multiple projects on this scale being conducted or planned worldwide.

A big reason so much genetic sequencing is being conducted is that the cost of doing so is dropping precipitously. Since completion of the Human Genome Project in 2003, the cost of sequencing a single human genome has plunged from about $100 million to less than $1,000.

But there's a more important reason behind the explosion of genetic sequencing — the potential impact on human lives.

"Medicine today is very much trial-and-error. By being able to analyze everyone's genetic makeup, we can deliver medicine that is more precise and tailored to each individual," said Anthony Joseph, Chancellor's Professor of Electrical Engineering and Computer Science at the University of California-Berkeley. He is also a core faculty member of the Center for Computational Biology at Berkeley.

SCALE-OUT STORAGE PLAYS AN ESSENTIAL ROLE IN RAPIDLY SEQUENCING DNA

At the RISELab on the Berkeley campus, graduate student Frank Austin Nothaft is part of the team developing ADAM, an open-source, high-performance, distributed library for genomic analysis that is the basis for Dr. Chiu's DNA-sequencing work.

Work on ADAM began in 2013, Nothaft said, "to address a significant shortcoming: many of the sequencing tools were not computationally efficient. The goal is to take a process that used to take thousands of lines of code and a month to run, and do it with 100 lines of code in a day.

COMPANY:

University of California-Berkeley
www.berkeley.edu

USE CASE:

- Data Analytics – Apache Spark®

CHALLENGES:

- Exponential increases in the volume of genomic data demanded new approaches to high-performance storage.
- Needed to rethink the genome sequencing process.
- Adding storage capacity required unacceptably large capital outlays and downtime.

IT TRANSFORMATION:

- Load times for a critical sequencing index accelerated by 3x.
- Decoupling of storage from compute engines simplifies the addition of storage capacity, delivering agility needed for production workflow while lowering cost.
- Multiple types of workflows supported concurrently with no impact on performance, giving researchers critical flexibility to explore a broader range of questions.

"Bit matching analysis is very important to genome processing. There was no other way we could do it without FlashBlade."

Frank Austin Nothaft, ***Graduate Student***

"We took a clean-slate approach to the tools used to conduct genomic data analysis and chose a cluster architecture that looks like a high-performance distributed database." ADAM is based on the Apache Spark open-source framework.

When it came to designing the IT infrastructure for ADAM, scale-out storage was a particular challenge. "Storage is about as critical as it gets," Nothaft observed. "That's because we're working with very data-heavy workloads."

Early implementations of the ADAM environment used spinning-disk storage technology and HDFS storage management. This approach eventually proved incapable of keeping up with the ever-increasing demands for high-performance storage.

"Our data-storage needs are much greater than our compute needs," Nothaft said. "In our HDFS system, our cluster typically ran at 10-30% utilization of compute capacity, but at 80-85% of storage capacity. So, we needed both high performance and scalability in our storage infrastructure."

Consider the case of Joshua Osborn, a lively 15-year-old from Cottage Grove, WI, who was diagnosed with a case of encephalitis (swelling of the brain) so severe that he had to be put into a medically induced coma. Extensive tests (the type of trial-and-error medicine Prof. Joseph cited) failed to identify a cause for Joshua's condition.

Finally, his case was brought to the lab of Dr. Charles Chiu at the University of California, San Francisco. Chiu has developed methods for rapidly reading out DNA sequences and analyzing them for matches with disease-causing pathogens. To achieve faster results, Chiu uses state-of-the-art DNA-sequencing machines that can generate millions of sequences at a time.[1]

Within 48 hours of receiving Josh's blood and spinal fluid, scientists at Dr. Chiu's lab had completed 3 million DNA sequences and discovered the culprit: a bacterial species native to Puerto Rico, where Joshua and his family had visited. Armed with this knowledge, Joshua's doctors in Wisconsin treated him with penicillin. He emerged from his coma, improved rapidly and after 76 days in the hospital was able to return home.

The computer system used to generate the rapid sequencing results achieved by Dr. Chiu and his team was developed across San Francisco Bay, at UC-Berkeley, by a team under the direction of Prof. Joseph. "Our focus is on building tools that allow people to analyze and process genomic data," he noted, "and the amount of data being collected is growing exponentially. We are sequencing more people every year; the cost of sequencing is dropping dramatically, and the depth of coverage is also growing, allowing us to collect more data."

FLASHBLADE MEETS ALL THE KEY CRITERIA

"With our current cluster configuration, trying to increase storage capacity built out around HDFS would have been impractical," Nothaft said. By 2015, "we were at the highest density of storage per compute node, and the only option for increasing capacity was a very large capital outlay."

In late 2015, the RISELab was approached by Pure Storage as a potential beta site for its new Pure FlashBlade storage product. The Pure FlashBlade storage platform is a new generation of all-flash storage architected from the ground up to support modern data analytics in a massively parallel environment.

1. For more detail about Joshua's case, go to https://www.ucsf.edu/news/2014/06/115116/teen's-recovery-shows-value-"next-generation"-dna-sequencing

FlashBlade met all of Berkeley's criteria for a storage solution: a parallel architecture capable of moving high-volume data stores at high bandwidth; high throughput and IOPS; infinitely scalable capacity; and greatly simplified management.

"FlashBlade allows us to maintain high throughput while horizontally scaling bandwidth simply by adding blades," Nothaft noted. "And, we can decouple the storage capacity from the number of compute nodes we have in our cluster. If I want to grow my storage capacity with FlashBlade, I can easily do so whenever I want, without complex planning and without any disruption to production operations. In genomics, this becomes a really big deal. You can't trade storage capacity for bandwidth, because your data sets pull such high bandwidth."

"We've moved some workloads that we couldn't run before to FlashBlade and run them with no problem. That's huge for our team."

Frank Austin Nothaft, ***Graduate Student***

ADDING FLASHBLADE IMPROVES PERFORMANCE, ADDS FLEXIBILITY

Since adding FlashBlade to the ADAM configuration, the Berkeley team has experienced several benefits. "Genomic sequencing can be accelerated using FlashBlade as the data platform providing improved performance, manageability and scalability," Nothaft reported. "Moving workloads over to FlashBlade, we have achieved end-to-end performance improvements of 3x. And, our throughput and latency numbers are far better."

One of the steps in genetic sequencing is variant calling. Using ADAM and the advanced infrastructure incorporating FlashBlade, performance of variant calling has improved 17x, while the cost has been cut in half.

Another key step in genome alignment is loading a large (5-40GB) index describing the genome sequence. Loading the index from FlashBlade instead of mounting it locally from HDFS enables alignment in 11minutes, down from 30 minutes, Nothaft reported.

In addition, "There's a process we sometimes employ called 'bit matching,' and it can involve dumping 100 million files to disk. When we tried it initially, it broke HDFS. But when we moved it to FlashBlade, it worked flawlessly. As a data platform, FlashBlade is calm under fire. And this is key, because bit matching analysis is very important to genome processing. There was no other way we could do it without FlashBlade."

Another benefit is flexibility. "We often have hybrid workflows — for example, large scans as well as point file lookups. FlashBlade allows us to do both concurrently with no loss of performance or recoding. It's given us new capabilities in ad-hoc analysis."

EASE OF MANAGEMENT MEANS MORE TIME FOR RESEARCH

With his focus on research and on helping develop ADAM, Nothaft doesn't have time for system-administration tasks. "FlashBlade has a very simple, easy-to-use reporting interface. I can quickly get the statistics I want," he noted. "The FlashBlade is architected for high reliability, and if something were ever to go wrong, it phones home and I get an e-mail from Pure Storage support saying they already have addressed the issue. That's a big benefit from a day-to-day management perspective."

The ADAM team is always focused on the future, looking for any improvement that will accelerate new discoveries from genomic sequencing. With the addition of FlashBlade, Nothaft sees a way to make even more contributions to success stories like that of Joshua.

"As a researcher, anything that opens up workloads that I can't run today is invaluable. We've moved some workloads that we couldn't run before to FlashBlade and run them with no problem. While that's huge for our team, it's even more important to the lives we might impact in the future."

info@purestorage.com
www.purestorage.com/customers