COMMENT

DAVID SALLAK
VICE PRESIDENT OF
PRODUCT MANAGEMENT
AND INDUSTRY MARKETING
PANASAS INC.

PANASAS IS THE PREMIER PROVIDER OF HIGH PERFORMANCE PARALLEL STORAGE FOR TECHNICAL APPLICATIONS AND BIG DATA WORKLOADS.

FIFTYFOLD PERFORMANCE INCREASE FOR GENOMICS SEQUENCING? GARVAN INSTITUTE OF MEDICAL RESEARCH DID IT

HOW ON-PREMISE PERFORMANCE STORAGE COMBINED WITH A STATE-OF-THE-ART SEQUENCING PLATFORM CAN DRIVE EXTRAORDINARY RESEARCH ACHIEVEMENTS.

The volume of scientific data is growing exponentially and nowhere more so than in life sciences and bioinformatics. Take genomic sequencing: Over the next 10 years, as many as 2 billion human genomes will be sequenced, generating as much as 40 exabytes (1 exabyte = 1018 bytes) of data—an almost incomprehensible amount of information.1 Some data scientists believe that genomics will soon overtake disciplines such as astronomy and particle physics as the top data-generating scientific activity.2 Looking ahead, proteomics and other fields promise to generate ever-expanding volumes of additional information in life sciences.

But data is only half the story. The tools that researchers use to perform their analyses (e.g., MapReduce, Schrodinger Glide, BLAST) are straining the computing resources of many institutions. Today, more than 25 percent of life scientists require high-performance computing capabilities, and that number is growing steadily.[3] Taken together, massive data sets and computing-intensive research applications put the spotlight on performance as a key enabler for discovery across a wide range of scientific disciplines.

GARVAN'S CHALLENGE

No one understands the need for performance better than Garvan Institute of Medical Research, one of Australia's leading biomedical research institutes, with a 50-year history of making significant breakthroughs in the understanding and treatment of diseases such as cancer, diabetes, Parkinson's disease, and osteoporosis. In recent years, Garvan has invested heavily in genomics, culminating in the creation of Kinghorn Centre for Clinical Genomics (KCCG) in 2012. KCCG conducts genomic research and provides state-of-the-art interpretation of genomic data directly to clinicians under an accreditation from National Association of Testing Authorities, Australia as a diagnostic pathology lab.

To drive its ambitious goals for excellence in genomics research, Garvan has fully embraced emerging genomics technologies. In 2014, Illumina, a San Diego–based manufacturer of sequencing equipment, asked Garvan to be a pilot site for its new HiSeq X Ten sequencing platform. Garvan jumped at the opportunity. HiSeq X Ten can sequence up to 50 complete human genomes per day, an unheard of level of throughput. Garvan saw the opportunity to leapfrog the industry by building a true genome sequencing factory. However, one problem loomed large: Handling such a torrent of data required a quantum leap in the performance of the institute's infrastructure.

PERFORMANCE AT CENTER STAGE

Enter Warren Kaplan, Ph.D., chief of informatics at Garvan. Kaplan joined the institute in 2002 and has managed the IT infrastructure for KCCG since its inception. Because the institute had greatly increased the number of researchers and was considering computing-intensive applications such as MapReduce, Kaplan was no stranger to performance issues. In the process of dealing with a range of performance-related complaints, Kaplan had learned that there was

WHAT IS PERFORMANCE NAS?

In the early days of the automobile, virtually every part of it was made in the factory from raw materials such as steel, rubber, and leather. It didn't make much difference how fast the materials were shipped to the factory or how long it took to ship the car to the consumer because most of the time was taken up just building the car. In much the same way, traditional IT architecture focuses on the computing operation (analogous to the manufacturing cycle) and considers storage performance to be less important. Put another way, the storage function was archival.

Modern automobile manufacturing is quite different. Most of the components are made by outside vendors, so the time required to ship parts to the factory is a major determinant of the length of the manufacturing cycle. Also, the level of global demand means that time to the dealer's lot matters, too. Managing the supply chain workflow is an essential part of automobile manufacturing and can even spell the difference between success and failure.

Similarly, modern IT infrastructure requires faster storage to achieve high levels of overall performance and reduced time to results. Garvan chose the Panasas® ActiveStor® system based on performance NAS technology that integrates a parallel scale-out file system with flash and disk storage in a single unit.

Why is that important? The answer is in the cost-performance equation. Flash storage is very fast but also expensive. On the other hand, disk storage is inexpensive but often not fast enough for computing-intensive applications. A modern file system that carefully manages which tasks are assigned to each type of storage ensures both the scalability and high performance that life sciences organizations require. In short, performance NAS is storage designed to support processing of large-scale compute workloads in place, without first copying data to compute.

more to solving performance issues than just adding computing power. Specifically, the system architecture and the storage technology are just as important as raw computing power for achieving performance goals.

To start with, Kaplan and the rest of the IT team reviewed Illumina's recommended system architecture. Their analysis quickly zeroed in on data movement. The data flowed from the sequencer to archival storage; then from storage to the computing node; and finally from the computing node back to archival storage (Figure 1). Kaplan wanted to streamline the workflow and eliminate moving data to and from the computing node. But was that approach possible?

Fortunately, the IT team had experience with a new kind of storage technology, that is, storage designed for high-performance applications. Replacing traditional network-attached storage (NAS) with performance NAS (see sidebar) avoided the need to copy data back and forth from storage to computing nodes. The result: a fifty-fold increase in performance. "Our sequencing data stays in the central repository throughout the analysis," Kaplan says. "This streamlined workflow saves time and bandwidth, enabling us to deliver results quickly to researchers around the world."

CLOUD CONSIDERATIONS

Given the industry buzz around the cloud, why wouldn't Garvan just move all its data—and data processing—into the cloud? After all, the cloud is touted as having significant advantages over on-premise storage, including flexible capacity, so-called pay-as-you-go pricing, and persistent access to data. Those assertions are all true, but there are also hidden pitfalls in using the cloud. It's important to understand some of the subtler considerations before plunging headlong into an approach that may not fit every situation. →

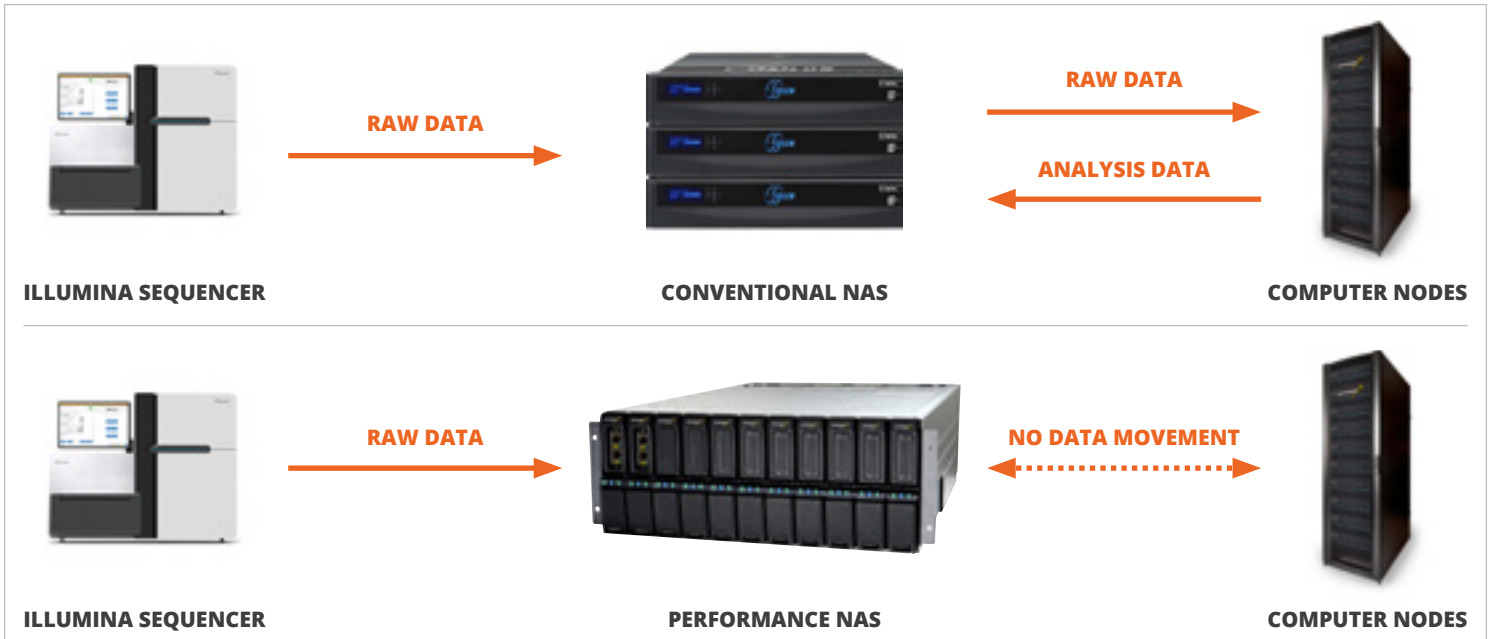

Figure 1. Data movement with conventional NAS and performance NAS

For starters, "the" cloud is a myth. Here's how one industry analyst sees it: "We talk about the cloud as if it were a singular, magically connected ether-verse, the same way we talk about the web, the Internet, or perhaps heaven. It's in the cloud, like that's one place, where your corporate accounting is hanging out with my Instagram pictures when no one's looking."[4]

There's more at stake than semantics. Offerings from cloud service providers differ substantially, even when it comes to the major players such as Google, Amazon, and Microsoft. When choosing a cloud service provider, it's a good idea to make sure that the cloud service offers the availability, accessibility, and security that your application requires. Many life sciences labs have found that a mixed workflow approach that takes advantage of the best of local compute/storage and cloud-based capabilities for burst or for exchanging results has proven to deliver the best time to results and investment value.

COMPLETE CONTROL

Cloud versus on-premise was never an either-or situation for Garvan. The institute keeps its proprietary research data on-premise and uses cloud services for sharing results with other research institutions (Figure 2). Keeping research data within the institute's walls ensures that Garvan protects its intellectual property rights and researchers can have unrestricted access to the information they need to do their work. By using cloud resources to distribute research results, the institute fulfills its mission to disseminate valuable information to researchers, clinicians, and others who can benefit from Garvan's work.

FULL SPEED AHEAD FOR INNOVATION

The pay-as-you-go model is one of the most enticing attributes of the cloud, but it can have an unexpected negative impact on innovation. The reason? Discovery requires imaginative thinking and a great deal of trial-and-error experimentation. When your analysis platform is on-premise, researchers can use storage and compute resources how they want; for example, there's no incremental cost to the organization to spend the entire weekend trying out a new analysis algorithm.

On the other hand, the knowledge that you are paying for every compute cycle can exert subtle pressure to economize ("Maybe my crazy idea isn't worth paying for an additional million computing cycles after all."). Thinking about a meter running can be a distraction—exactly what a researcher doesn't need in the quest for innovation and discovery. In addition, department managers and budget analysts are under significant pressure to control costs, which can discourage untested approaches and unproven algorithms that can run up the cloud service bill. By building out the high-performance infrastructure in-house, Garvan enables researchers to experiment—the key to innovation.

BENEFIT OF FIXED COSTS

"You only pay for what you use." That's a common phrase heard in discussions about cloud, but one that also bears investigation. To provide the flexibility you expect from the cloud, the service provider must build out extra capacity, which is unused by definition (how else could it provide it to you quickly?). That unused portion of infrastructure has associated costs; and because the provider makes a profit, those costs go into its pricing equation. In a way, you do pay for unused capacity; you just can't see it in your bill. If that's the case, why do it? The answer is that the cloud makes sense if your workload fluctuates wildly and you need the ability to address resources up and down rapidly. In these circumstances, using cloud resources to ensure flexibility can be worth the premium cost.

For Garvan, the predictability of its workload made an on-premise solution the better choice. In 2015, the institute employed 632 people, including 225 researchers and 120 visiting scientists.[5] With such a stable population of researchers, the overall workload falls within a narrow range. While the decision to migrate to performance storage was driven primarily by factors such as performance and ease of integration, Garvan benefits from fixed costs versus hundreds of meters running at once.

GARVAN TODAY

With its fifty-fold increase in performance, Garvan's researchers have realized important achievements in genomics research:

- In October 2016, Garvan announced that KCCG had sequenced 10,000 human genomes since it acquired the Illumina HiSeq X Ten system in 2014. It sourced most of the genomes from individuals who are part of research projects investigating the genetic components of human health and disease.
- In July 2016, a U.S. Food and Drug Administration truth challenge named Garvan a top performer when its KCCG team achieved 99.98 percent accuracy in detecting single-nucleotide polymorphisms in an unknown genome—the highest score of 36 entrants.
- Also in July 2016, Garvan launched Australia's first clinical whole-genome sequencing service at an event in Sydney. This new service is expected to triple the diagnosis rates for Australians living with rare and genetic conditions.

These and other recent successes are testimony to the hard work of Garvan's researchers but also reflect the importance of a high-performance infrastructure—including performance storage—for life sciences research. ■

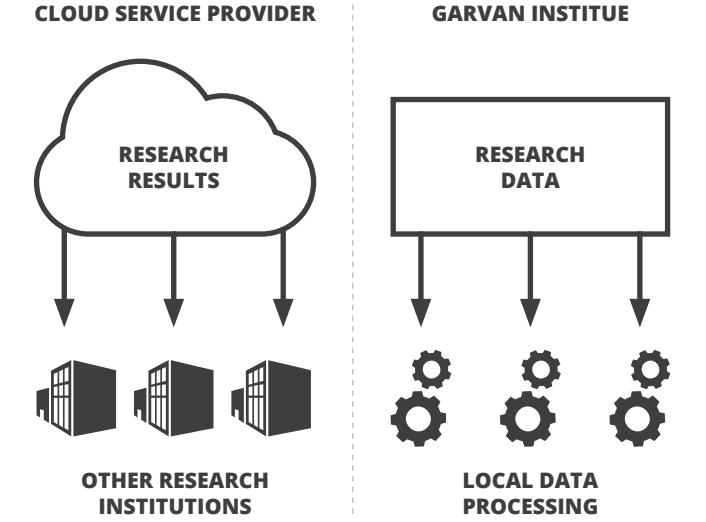

Figure 2: Partitioning research data and research results

REFERENCES

1. Z. D. Stephens, et al., "Big Data: Astronomical or Genomical?" PLOS Biology (2015), http://dx.doi.org/10.1371/journal.pbio.1002195.
2. Robert Gebelhoff, "Sequencing the Genome Creates so Much Data We Don't Know What to Do with It," The Washington Post (July 7, 2015), www.washingtonpost.com/news/speaking-of-science/wp/2015/07/07/sequencing-the-genome-creates-so-much-data-we-dont-know-what-to-do-with-it/.
3. John Russell, "25 Percent of Life Scientists Will Require HPC in 2015," HPCwire.com (May 18, 2015), www.hpcwire.com/2015/05/18/25-of-life-scientists-will-require-hpc-in-2015.
4. Addison Snell, "The Great Lies of Cloud Computing," The Next Platform (September 21, 2016), https://www.nextplatform.com/2016/09/21/three-great-lies-cloud-computing/.
5. "2015 Garvan's Annual Report" (April 2016), www.garvan.org.au/about-us/annual-report-files/garvan-annual-report-2015.pdf.