HPC DATA STORAGE ENABLES MICROBIOME RESEARCH AND UNDERSTANDING OF COVID-19 AND OTHER DISEASES

Rapid Microbiome Analysis at UC San Diego Center for Microbiome Innovation

Produced by Tabor Custom Publishing
in conjunction with:

HPC wire

PANASAS®

Introduction

Microbiome research has rapidly accelerated in the past few years. That is certainly the case at the UC San Diego Center for Microbiome Innovation (CMI), which leverages the university's strengths in clinical medicine, bioengineering, computer science, the biological and physical sciences, data sciences, and more to coordinate and accelerate microbiome research.

The great interest in the microbiome is because it is linked to a plethora of conditions and diseases, including COVID-19. Specifically, the microbes that exist in people's mouth and gut are considered key modulators of health and disease. Hence, the curiosity in learning more about the microbiome.

As a result of this growing awareness of the importance of understanding the microbiome, the United States and other countries have prioritized research efforts. Just as there was a national directive two decades ago to sequence the human genome, much of the new emphasis on microbiome research was spawned by the National Microbiome Initiative. When announced, the initiative had the goal of advancing the understanding of microbiome behavior and enabling protection and restoration of healthy microbiome function. The effort focused on three recommended areas, including:

1. **Supporting interdisciplinary research** to answer fundamental questions about microbiomes in diverse ecosystems.
2. **Developing platform technologies** that will generate insights and help share knowledge of microbiomes in diverse ecosystems and enhance access to microbiome data.
3. **Expanding the microbiome workforce** through citizen science, public engagement, and educational opportunities.

While life science labs around the world routinely have compute capacity to perform sequencing and analysis of the human genome, a microbiome has about 200 times the number of genes of the human genome. It thus requires much higher performance compute and storage systems to derive useful insights in a timely manner.

"Compared to more 'traditional' microbiologists, microbiome experts are not looking at one organism or a few at a time, but a multitude of them. Thus, we need a great amount of storage space to conduct our analysis properly," said Sandrine Miller-Montgomery, the executive director of the CMI.

Microbiome Research Challenges

Microbes are of great interest because they help people digest and process nutrients. They also constantly interact with — and help shape — people's immune systems. Several factors make microbiome research different and more challenging than typical life sciences research. Notably, the volume of data involved is much greater than traditional life sciences work based on genomic analysis.

Also, many researchers from different disciplines all need access to that data. The Center for Microbiome Innovation leverages UC San Diego's world-class experts across multiple disciplines and access to all the latest omics tools. These include genomics, metagenomics, metatranscriptomics, metabolomics, multiplex proteomics, artificial intelligence, and more to process hundreds of thousands of samples each year and analyze and collect data for some of the largest microbiome cohorts in the world. And they are using a variety of techniques that generate different workloads that place vastly different demands on storage performance. "We have more people that want to get work done, using data sets that are orders of magnitude larger than anything we've seen," said Jeff DeReus, systems administrator for CMI.

Researchers at UC San Diego and elsewhere are already finding that the makeup of gut microbiomes is associated with diseases and conditions you might expect, such as food allergies, obesity, inflammatory bowel disease, and colon cancer. They're also discovering that the gut microbiome plays a role in diseases you might not expect, such as rheumatoid arthritis, atherosclerosis, and asthma. Even more surprisingly, in mouse models, at least, the microbiome in the gut has even been linked to the brain. While human studies are still needed, this could mean that traits like how anxious you are, how outgoing you are, even how depressed you are, or whether you have autism, may depend on the microbes in your gut.

CMI's Data Storage Needs

"A lot of our work is collaborative," said Dr. Yoshiki Vázquez-Baeza, Associate Director of Bioinformatics Integration at CMI. "If we cannot share the data between users or with other partners, that creates a roadblock. Having a reliable storage resource facilitates a lot of the creative work that happens here."

Moreover, it is not just the amount of data or research that's done but the fact that the research is highly dynamic. New research directions are routinely undertaken as the science changes. Such changes have implications on the performance, capacity, and scalability requirements of the compute and storage systems used to conduct research.

Accelerating the impact of microbiome research means developing novel tools and methods for analyzing and manipulating microbiomes. Many of the computational analysis techniques used by the UCSD researchers are I/O-intensive. In contrast, much life sciences research is more compute-intensive. In such cases, simply adding more processing power helps, but that's not the case here. Using traditional storage technologies would result in degraded performance and hinder computational workflows, thus slowing research efforts. Having technology that takes the pain away from sluggish file systems or longer run times is required.

Performing genomic sequencing and analysis of microbiomes requires huge volumes of data. Within CMI, each of the nearly 300 users was assigned local storage across two clusters. However, those legacy resources became insufficient to process high I/O tasks such as DNA sequence processing or analysis of other multi-terabyte data sets. Attempts at processing current data volumes with traditional storage technologies resulted in degraded performance for the computational workflows, which compromised the pace of discovery.

Several years ago, one of the main research groups at the center began performing large-scale metagenomics sequencing, an advanced technology for microbial surveying. "That was the catalyst, said Vázquez-Baeza. "We knew then that we needed to make a change."

Any storage solution would need a certain set of characteristics to meet the group's research demands. Some of the top features required include:

- Highly scalable to accommodate the ever-growing volumes of data the UCSD researchers are working with
- High-performance, specifically a system that reduces I/O bottlenecks that can significantly degrade overall system performance and slow the execution of workflows
- Adaptable in that it can accommodate computational workloads with a wide range of I/O, throughput, Read/Write requirements
- Easy to use and support, minimizing issues that reduce productivity
- Cost-effectively scalability in capacity (thus, keeping a lid on CapEx costs)

CMI Partners with Panasas

Life sciences organizations have made significant infrastructure investments over the last ten years to meet their next-generation sequencing needs.

The common approach taken to keep pace with growing data volumes was to use traditional network-attached storage (NAS). NAS storage offered trade-offs. It is easy to install and manage as an individual file server. Due to limitations on file systems and the maximum number of files, it becomes highly complex to administer as the storage capacity scales. As more NAS systems are deployed, the system administration overhead increases faster than the number of NAS systems deployed. Administrators must constantly load balance each NAS system as well as migrate data between systems. Additionally, single file servers become a performance bottleneck as the number of clients accessing the server grows. This performance aspect is a particular bottleneck for high-performance clients running complex tasks such as those found in microbiome research.

An alternative was to use scale-out NAS solutions, which virtually aggregate file servers and present a global shared file system view to all clients. This provides a better way to scale up to a couple of petabytes than traditional NAS solutions. However, once you exceed this capacity, erradic and unpredicatable performance enters the environment. Scale-out NAS fails to address the performance requirements of intensive genomic analysis.

Such infrastructures cannot handle the current and future microbiome analysis requirements. Organizations that were used to genomics-only workloads will soon be challenged with analysis related to different microbiome-specific analysis technologies and the resulting increases in data sizes and volumes. The mixed workloads will stress storage infrastructure in unforeseen ways. As organizations plan for the changing landscape of research pipelines, they must ensure a data storage foundation based on scalability, performance, and adaptability.

It is also common for life sciences organizations to deploy a scale-out NAS solution in parts of the workflow and deploy a parallel storage solution for the more performance-demanding applications. However, having to access data from two separate systems slows collaboration and impacts productivity.

When CMI was looking to create an HPC storage infrastructure that could deliver consistently high performance, speed data exploration and discovery, and simplify storage management for administrators, they turned to Panasas, which has been their primary high-performance storage solution since 2015.

DeReus, the systems administrator, had worked with Panasas storage technology at previous academic institutes and was familiar with its functionality, manageability, and scalability. One additional factor: Panasas was easy to fully integrate into their current life sciences workflows.

Panasas delivers high-performance computing data storage solutions that support industry and research innovation around the world. The Panasas ActiveStor® data storage appliance gives CMI researchers access to a consistently fast, total-performance HPC storage solution that uses the PanFS® parallel file system to automatically adapt to the changing and evolving small file and mixed workloads that dominate today's HPC and AI landscape – all at the lowest total-cost-of-ownership (TCO) and without the need for tuning or manual intervention.

Using the ActiveStor system, CMI researchers can rapidly store, retrieve, and analyze unprecedented volumes of data. The fast, efficient PanFS parallel file system accelerates performance at every stage of the computational research process, and removes storage bottlenecks and system delays to deliver consistent high performance, regardless of the workloads being processed. For example, researchers generate many intermediate files to do their analysis. By staging data into the file system, they can alleviate the load on the storage cluster itself. That reduces the impact of these larger datasets on the other system users.

ActiveStor with PanFS also offers great flexibility and control. It's common for CMI researchers to need varying volumes of storage, depending on their projects and the current stage of the research. With ActiveStor, CMI can adapt to changing workload requirements without the need for labor- or skill-intensive tuning and administration efforts. Such capabilities are greatly appreciated since a lot of the work at the center is collaborative. The system enables data sharing between users or with other partners, without creating a roadblock. This facilitates a lot of the creative work that happens at the center and in the microbiome research field.

Looking ahead, CMI plans to continue using the latest generations of scientific technologies to understand more about microbiomes. With a focus on converting the research possibilities into solutions, CMI appreciates storage solutions that don't distract researchers from the scientific problems they aspire to solve.

Researchers working on the next big scientific discovery aren't typically aware of the storage infrastructure that supports their work, but they are aware when there are issues. By freeing researchers from worries about storage, CMI can accelerate the adoption of advanced scientific technologies.

For more information about Panasas high-performance storage solutions for investigating the microbiome and other demanding life sciences research, visit: www.panasas.com/industries/life-sciences or contact info@panasas.com.

HPC wire

PANASAS®