DDN® STORAGE

SUCCESS STORY

# ACCELERATE: LIFE SCIENCES

University of Miami's Center for Computational Science Correlates Viruses with Gastrointestinal Cancers for The Cancer Genome Atlas 400% Faster Using DDN Storage

## CHALLENGES

- Diverse, interdisciplinary research projects required massive compute and storage power as well as integrated data lifecycle movement and management
- Highly demanding I/O and heavy interactivity requirements from next-gen sequencing intensified data generation, analysis and management
- Powerful, flexible file system was required to handle large parallel jobs and smaller, shorter serial jobs
- Data surges during analysis created "data-in-flight" challenges

## SOLUTION

An end-to-end, high performance DDN GRIDScaler® solution featuring a GS12K™ scale-out appliance with an embedded IBM® GPFS™ parallel file system

## RESULTS

- Links between certain viruses and gastrointestinal cancers discovered with computation not possible before
- With DDN's high performance I/O, CCS has reduced genomics compute and analysis time from 72 to 17 hours

**UNIVERSITY OF MIAMI MAINTAINS ONE OF THE LARGEST CENTRALIZED, ACADEMIC, CYBER** infrastructures in the country, which is integral to addressing major scientific challenges and solving many of today's most challenging problems. At its Center for Computational Science (CCS), more than 2,000 researchers, faculty, staff and students across multiple disciplines collaborate on diverse and interdisciplinary projects requiring high performance computing (HPC) resources.

The Center provides hardware, software development and analytics expertise to support a variety of research areas, including genomics, computational biology, marine ecosystems, ocean modeling, climate and meteorology, computational economics, computational fluid dynamics as well as social systems informatics. According to Dr. Nicholas Tsinoremas, director of the Center for Computational Science at the University of Miami as well as a professor of medicine, computer science and health informatics, the center was founded on the premise that data drives discovery. Therefore, keeping pace with data growth is of paramount importance.

"Data-intensive discovery and multi-scale interdisciplinary approaches are becoming more prevalent in the way that sciences and engineering generate knowledge," he explains. "The speed at which scientific disciplines advance, depends in large part on how effectively researchers collaborate with one another and with experts in the areas of workflow management, data management, data mining, decision support, visualization and cloud computing."

Another guiding principle is the imperative to manage the entire data lifecycle as seamlessly as possible to streamline research workflow. "We have integrated the HPC environment with our data capture and analytics environments, so movement is transparent between different research steps," Tsinoremas adds. "This level of interactive processing speeds the delivery of data from sensors and instruments to the desktop of analysts and ultimately, into the hands of science-based decision makers."

## THE CHALLENGE

Unlike other high performance computing centers that originated as simulators, CCS has always put a lot of emphasis on data driving scientific results. Approximately 50 percent of the center's users come from University of Miami's Miller School of Medicine with ongoing projects at the Hussman Institute for Human Genomics, such as research into Alzheimer's disease and The Miami Project to cure paralysis. The remaining 50 percent of users cover marine and atmospheric sciences as well as engineering along with the university's schools of arts and sciences, architecture, music and business.

Other notable projects requiring massive compute and storage power include cancer biomarker research at the Sylvester Comprehensive Care Center, and the University of Miami's Grand Lagrangian Deployment project, sponsored by the Gulf of Mexico, which is the largest oceanic dispersion experiment of its kind to explore surface flows near the Deepwater Horizon site of the Gulf oil spill.

> “Translating research requirements into actionable technology is no small feat. Because of advances in computing, we now generate massive amounts of data from a multitude of models running multiple simulations simultaneously. All of this data must be stored, analyzed and distributed to decision makers.”

> “The process of validating data and detecting variants in DNA sequences through SNP calling requires both high throughput and extremely high levels of interactivity. Our goal was to reduce the time to perform the number crunching needed to map and merge hundreds of thousands of files so data could be analyzed faster.”

**Joel P. Zysman**
Director of High Performance Computing at the Center for Computational Science at the University of Miami

“Translating research requirements into actionable technology is no small feat,” says Joel P. Zysman, director of High performance Computing at the Center for Computational Science at the University of Miami. “Because of advances in computing, we now generate massive amounts of data from a multitude of models running multiple simulations simultaneously. All of this data must be stored, analyzed and distributed to decision makers.”

For example, the University of Miami’s Marine School regularly runs simulations of multiple climate models, which then are distributed to a variety of decision makers to determine the impact of climate change on water engineering, precipitation and local water management districts. In supporting research demands, CCS generates upwards of 50TB each year for this project alone.

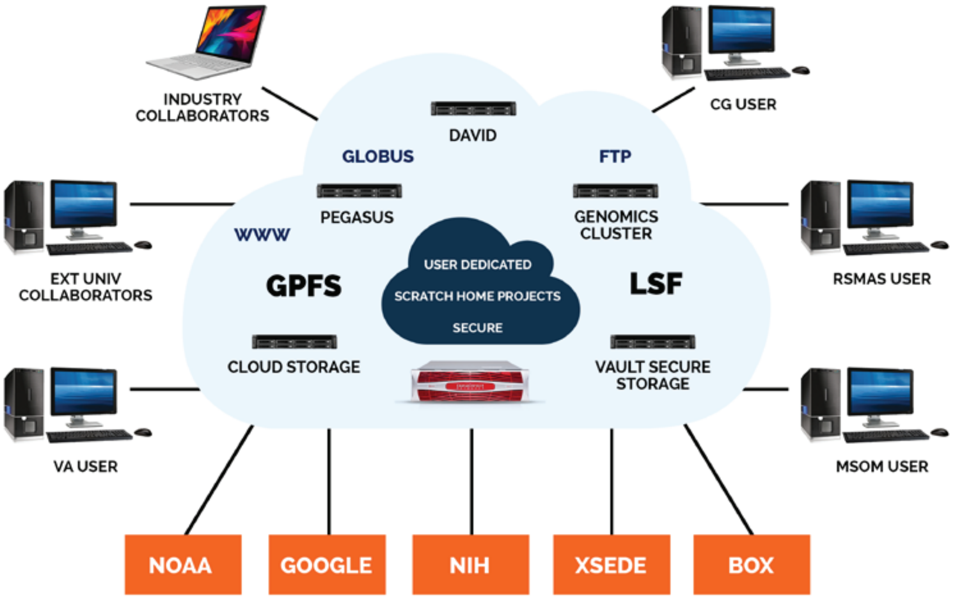
Additionally, the explosion of next-generation sequencing has had a major impact on compute and storage demands, as it’s now possible to produce more and larger datasets, which often create processing bottlenecks. At CCS, the heavy I/O required to create four billion reads from one genome in a couple of days only intensifies when the data from the reads needs to be managed and analyzed. “The process of validating data and detecting variants in DNA sequences through SNP calling requires both high throughput and extremely high levels of interactivity,” adds Zysman. “Our goal was to reduce the time to perform the number crunching needed to map and merge hundreds of thousands of files so data could be analyzed faster.”

Aside from providing sufficient storage power to meet both high I/O and interactive processing demands, CCS needed a powerful file system that was flexible enough to handle very large parallel jobs as well as smaller, shorter serial jobs. The key for CCS was the ability to take advantage of very high I/O as well as very low IOPS without having to move data around, which would have required an inordinate amount of time and administrative overhead.

Additionally, CCS had to address “data-in-flight” challenges, resulting from major data surges during analysis. As the creation of intermedia files often resulted in a 10X spike in storage, it was critical to scale and support petabytes of machine-generated data without adding a layer of complexity or creating inefficiencies.

## THE SOLUTION

The ideal storage solution for CCS would provide a single platform for both high-throughput genomics and highly interactive research collaboration. The center needed to accommodate its entire data lifecycle, so users didn’t have to deal directly with a lot of data movement. DDN Storage was superior to competing storage platforms with its ability to leverage one robust, easily managed platform for ensuring high performance, simplified collaboration and accelerated data analytics.



## BUSINESS BENEFITS

- The ability to meet varied research workflow demands enables CCS to accelerate data analysis and speed scientific discoveries
- Best-in-class performance for genomics assembly, alignment and mapping has proven invaluable in supporting major medical research into Alzheimer's, Parkinson's and gastrointestinal cancers
- High performance storage and transparent data movement lets CCS scale storage without adding complexity

> “*DDN enabled us to analyze thousands of samples for the Cancer Genome Atlas, which amounts to nearly a petabyte of data. Having a robust storage platform like DDN is essential to driving discoveries such as our recent study that revealed a link between certain viruses and gastrointestinal cancers. Previously, we couldn't have done that level of computation.*”

> “*With DDN's high performance I/O, we've been able to reduce both the time it takes to crunch numbers for genome mapping and the analysis of variants during the SNP calling process. Both workflow steps used to take three days. Now they can be completed in 17 hours. By accelerating these workflows, we can speed data analysis and scientific discovery.*”

**Dr. Nicholas Tsinoremas**
Director of the Center for Computational Science at the University of Miami

"Where DDN really stood out is in the ability to adapt to whatever we would need," says Zysman. "We have both IOPS-centric storage and the deep, slower I/O pool at full bandwidth. No one else could do that."

DDN's GS12K scale-out file storage appliance with one petabyte of storage was best suited for meeting the university's growing IOPS and bandwidth requirements while ensuring extremely fast application performance. Moreover, the embedded GPFS™ parallel file system eliminated the need to purchase and manage external servers, network adapters and switches. "DDN with GPFS was the best combination in terms of performance and integration," he adds. "Instead of having disparate file systems and different queues, we were able to centralize everything and then scale accordingly while managing it all through a single pane of glass."

Thanks to DDN's massively scalable GS12K clusters, CCS can meet its varied workflow demands, which are more file-set than file-system based. "With DDN, the ingest pool is the same as our processing area, so we don't need resources dedicated to pre-staging data," explains Zysman. "Once data is entered, it can be migrated automatically to a lower tier of storage, which allows researchers to easily interact and collaborate because data never goes offline." DDN's transparent data movement also is ideally suited for addressing the center's back-end genomics workflows and interactive jobs, as the team can leverage one platform to capture, download or move data.

With DDN, the CCS team has the confidence to process input from sequencing pipelines and handle the data computations generated by its 15 Illumina HiSeq next-generation sequencing instruments.  The team also can easily support all its application needs, including the use of BWA and Bowtie for initial mapping as well as SamTools and GATK for variant analysis.

In addition to supporting the center's internal sequencing requirements, DDN provides the necessary scalability and performance to support sequenced data from external databases, such as DBGap and The Cancer Genome Atlas (TCGA), along with data generated by a dozen collaborations with academic and commercial research partners, including companies in the biopharmaceutical industry.

## THE BENEFITS

DDN's best-in-class performance for genomics assembly, alignment and mapping has proven invaluable in helping CCS accelerate both research and collaboration efforts. For every genome that comes off a sequencer, CCS performs several intensive I/O steps, which generate tens of thousands of files just to prepare the data for further analysis.

"With DDN's high performance I/O, we've been able to reduce both the time it takes to crunch numbers for genome mapping and the analysis of variants during the SNP calling process," explains Tsinoremas. "Both workflow steps used to take three days. Now they can be completed in 17 hours. By accelerating these workflows, we can speed data analysis and scientific discovery."

Each year, the CCS team looks at thousands of genomes, with each containing up to hundreds of thousands of files. For example, CCS analyzed samples for TCGA as part of a project to study viruses in nine common cancers. "We analyzed more than 2,000 samples for TCGA amounted to nearly a petabyte of data," says Tsinoremas. "Having a robust storage platform like DDN is essential to driving discoveries, such as our recent study that demonstrated the link between certain viruses and gastrointestinal cancers. Quite honestly, we couldn't have done that level of computation previously."

DDN's high performance storage and transparent data movement also lets CCS scale its storage without adding complexity. "One of the biggest benefits with DDN comes from the fact that we can add storage no matter if it's needed in the high performance pool, high-transactional pool or slower, deeper storage layer," adds Zysman.

The ability to maintain an active archive with DDN Storage enables CCS to accommodate different types of analytics with varied I/O needs. For instance, the team recently embarked on a large scale, meta-analysis of genome-wide association data to identify six new risk locations for Parkinson's disease. "With DDN, our data is always live so researchers can access it easily and collaborate more freely," says Zysman. "And, moving it to different tiers of storage to meet different research needs is completely transparent. We now can serve out research data everywhere and it doesn't make any difference if it's being used by researchers on PCs, Macs or Linux systems."

Overall, the new DDN gateway will strengthen and streamline the center's collaboration with dispersed research teams. "Our arrangement is to share data or make it available to anyone asking, anywhere in the world," Tsinoremas concludes. "Now we have the storage versatility to attract researchers from both within and outside the HPC community, including investigators from the university's schools of business, music and communications. With DDN, we're well positioned to generate, analyze and integrate all types of research data to drive major scientific discoveries and breakthroughs."

## TECHNICAL BENEFITS

- Centralized storage with an embedded file system makes it easy to add storage where needed—in the high performance, high-transaction or slower storage pools—and then manage it all through a single pane of glass
- DDN's transparent data movement enables using one platform for data capture, download, analysis and retention
- The ability to maintain an active archive of storage lets the center accommodate different types of analytics with varied I/O needs

## ABOUT DDN®

DataDirect Networks (DDN) is the world's leading big data storage supplier to data-intensive, global organizations. For more than 15 years, DDN has designed, developed, deployed and optimized systems, software and solutions that enable enterprises, service providers, universities and government agencies to generate more value and to accelerate time to insight from their data and information, on premise and in the cloud. Organizations leverage the power of DDN technology and the deep technical expertise of its team to capture, store, process, analyze, collaborate and distribute data, information and content at largest scale in the most efficient, reliable and cost effective manner. DDN customers include many of the world's leading financial services firms and banks, healthcare and life science organizations, manufacturing and energy companies, government and research facilities, and web and cloud service providers. For more information, visit our website www.ddn.com or call 1-800-837-2298.