Success Story

Optimizes HPC Pipeline, Driving Research Discovery and New Drug Therapy

Van Andel Research Institute

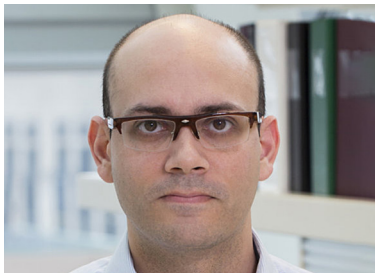
Infinitely scalable storage capacity available to keep pace with increased use of Cryo-EM and next-generation sequencing technologies.

Life Sciences
North America

- Simple, fast access to storage
- Research and instrument data
- Address active archive requirements
- Two sites for increased redundancy and protection

Solution

A flexible infrastructure with ability to implement private cloud for storing and crunching large amounts of data efficiently



Zachary Ramjan
Research Computing Architect

Van Andel Institute has a 20-year legacy of biomedical research and scientific education with a focus on improving health and enhancing the lives of current and future generations. Formed in 1996 by Jay and Betty Ann Van Andel, the Grand Rapids, Michigan-based organization has evolved into a premier center for research and education that supports more than 360 scientists, educators, and staff.

At Van Andel Research Institute (VARI), a dedicated team works tirelessly to determine the epigenetic, genetic, molecular and cellular origins of cancer and other diseases. Using state-of-the-art technologies and instrumentation, the Institute's scientists are working to translate discoveries into highly innovative and effective diagnostics and treatments. For example, scientists in VARI's Center of Epigenetics are shedding light on the mechanisms that control how genes are regulated to determine what happens when a cell transitions from a normal state into a rapidly dividing cancer cell. Epigenetics, is rapidly emerging and an important area of cancer research, according to Dr. Scott Rothbart, assistant professor in VARI's Center of Epigenetics.

"We're answering key questions about the molecular mechanisms of gene regulation," he explains. "These findings give us new vision for how to target cancers with innovative drug therapies. Of course, this research generates more data than ever before and requires more computing and storage than previously available."

The Challenge

- Fragmented storage pools were costly, cumbersome and lacked sufficient safeguards
- Addition of high-powered, cryo-electron microscopy was anticipated to quadruple existing storage
- Increasing requirement to ingest, process, store, archive, and share research

In the Rothbart Laboratory, scientists perform time- and data-intensive molecular dynamics simulations to model protein behavior in silico. "We employ molecular dynamics in part to simulate and predict protein-protein and protein-drug interactions," Rothbart says. "This technique is rapidly accelerated by high-performance computing."

Zachary Ramjan, the Institute's Research Computing Architect, joined the organization in 2014 to ensure there would be ample compute and storage power to continually push the research envelope. "I was given the freedom to build a high-performance compute and storage solution from the ground-up that would meet both current and future research needs," he explains. "Our goal was to create a progressive computing platform anchored by powerful, scalable storage."

In setting the stage, Ramjan sought to replace the organization's fragmented storage silos with primary shared storage harboring instrument and other research data. "Centralized storage produces major cost savings," he adds. "It also provides an extra measure of protection by moving irreplaceable research and instrumentation data from individual hard drives onto a single system."

Centralized storage and computing accommodate major growth, including the Institute's expanding structural biology research program, which is home to a suite of cryo-electron microscopes (cryoEM). This technology, which enables scientists to see the structure of molecules that are one-tenthousandth the width of a human hair, is transforming investigations of small yet vital components of human biology. "The installation of highly specialized electron microscopes quadrupled our current storage capacity," Ramjan says. "Performance and speed are crucial for accommodating these data-hungry instruments."

Approximately 30 researchers are getting trained on the Institute's FEI Tecnai Spirit G2 BioTWIN, FEI Talos Arctica and FEI Titan Krios microscopes. "The new electron microscopes will definitely push the boundaries of research while enabling us to see things at much higher resolutions," says Dr. Gongpu Zhao, Cryo-EM Core manager. "Having sufficient compute and storage capacity is critical, as these high-end microscopes generate about 20TB of data every three days."

Performance

Speed is crucial to accommodate data hungry instruments

Scale

Infinitely scalable storage capacity to keep pace with Cryo-EM technology

Flexibility

On-premises and cloud compute solution gives flexibility to grow and scale on demand

Experience

Unique ability to present a single, federated namespace across

"DDN makes it incredibly easy for us to put data where it best belongs, all within the context of a single system. The ability to store data in the most performance and cost-efficient place gives us flexibility to grow as research needs dictate."

Zarhary Ramjan

Research Computing Architect

The Solution

The organization began a thorough evaluation of next-gen HPC and storage solutions, including cluster and cloud computing, as well as parallel file and object storage. On one hand, VARI wanted to take advantage of the performance and scalability delivered by a traditional GPFS-based parallel file system. On the other hand, the organization also wanted the flexibility to implement a private cloud for storing and crunching large amounts of data efficiently and cost-effectively.

"I had a gut instinct that a hybrid on-premises and cloud computing solution was the best path to the future, which was proven out as we deployed our hybrid solution, says Ramjan. "This approach gave us the flexibility to grow and scale computing on demand."

The storage decision was driven by the need for a parallel file system and object storage to handle data ingest, processing, storage, collaboration, and archiving. The team reviewed several file-system solutions before deciding; including a closer look at its existing EMC Isilon storage before determining it would be too cost prohibitive to scale the platform in keeping pace with rapid storage growth.

In the decision process, Ramjan drew upon extensive experience with DataDirect Networks (DDN) from working on the next-generation sequencing team at University of Southern California. "We were really happy with the performance and scalability of our DDN parallel file system storage appliance," he says. "Because of that experience and its Spectrum Scale data management capabilities, we chose DDN's GRIDScaler® GS7K parallel file system appliance."

The Benefits

- Infinitely scalable storage capacity to keep pace with Cryo-EM and next-gen sequencing technologies.
- Simple, fast access to 2PB of storage for research and instrument data to address exponential storage growth and active archive requirements.
- Projected savings of hundreds of thousands of dollars by centralizing storage for state-of-the-art scientific instruments.

Betting on DDN has paid off several times for VARI, starting with the flawless implementation of two GS7Ks with 1.5PB and WOS with a half-petabyte of capacity. "Our DDN systems were very easy to deploy," says Ramjan. "In particular, DDN's OpenStack® driver support was a welcome surprise and a huge bonus because it significantly streamlined our OpenStack storage integration."

VARI took advantage of DDN's deployment flexibility to implement storage at two sites for increased redundancy and protection of the institute's most sensitive datasets. The addition of WOS further simplified data movement, tiering, protection and replication that are all handled automatically across file and object. "DDN makes it incredibly easy for us to put data where it best belongs, all within the context of a single system," Ramjan notes. "The ability to store data in the most performance- and cost-efficient place gives us flexibility to grow as research needs dictate."

Additionally, scientists can collaborate more effectively with other research centers as DDN's storage scales seamlessly. "I just had a request from a researcher who is pulling down a 100TB RNASeq dataset from The Cancer Genome Atlas project," says Ramjan. "Now we can accommodate this kind of request easily without causing storage constraints for anyone else."

In fact, DDN's scale-out capabilities can accommodate ongoing spikes in computation and data collection generated by a host of workflows, including GROMACS and AMBER for molecular dynamics simulations and GATK for next-gen genomics toolsets. "We now have the technology backbone to support novel research that translates into groundbreaking therapeutic interventions," says Rothbart. "The work we're doing in molecular dynamics, which is supported by DDN, is forming the future of chemistry."

Major cost savings will be realized with DDN as the primary storage for more than a dozen state-of-the-art scientific instruments, including Illumina NextSeq and iScan sequencers, Aperio slide scanners, and the new crop of electron microscopes. "We'll save hundreds of thousands of dollars by centralizing storage on DDN for data-intensive research and a dozen data-hungry scientific instruments," says Ramjan. "We can elevate the standard of protection, increase compliance, and push the boundaries of science on a single yet highly scalable storage platform. That's why DDN is core to our operation and a major asset for our scientists."

About DDN

DataDirect Networks (DDN) is the world's leading big data storage supplier to data-intensive, global organizations. DDN has designed, developed, deployed, and optimized systems, software, and solutions that enable enterprises, service providers, research facilities, and government agencies to generate more value and to accelerate time to insight from their data and information, on premise and in the cloud.

v2 (4/20)