Qumulo

Carnegie Institution for Science Case Study

Life Sciences

The Department of Embryology at the Carnegie Institution for Science Tackles Volume and Variety of Research Data with Qumulo File Fabric (QF2)

Wrestling with both terabyte-size data sets and millions of small sequencing files is a daily occurrence for one of the world's top research institutions. After considering many other storage systems, the Department of Embryology at the Carnegie Institution for Science turned to QF2, a modern, highly scalable file storage system, to deliver the performance, scalability and simplicity needed to keep pace with evolving research data requirements.

The Volume and Variety of Research Data

Founded by Andrew Carnegie in 1902, the Carnegie Institution for Science carries out a vast range of pure scientific research, from earth science and biology to magnetism and astronomy. The Institution's Department of Embryology, established in 1913, is globally recognized for its innovative experimental studies, using molecular biology, genetic techniques and animal models to investigate developmental processes from single-cell embryos to whole organisms – research that has led to numerous scientific insights and three Nobel Prizes.

Embryology's research data can be roughly divided into three categories: images collected from microscopes and other imaging systems, nucleotide sequencing data from next-generation sequencers, and the usual variety of common document files used to collect, report, and present results. While document storage is relatively straightforward, the imaging and sequencing data present substantial challenges. The Department's microscopy systems are often used to generate image data sets that can be as large as 1TB per experiment. Those images are stored and accessed for processing and analysis from client computers running all of the major operating systems (Windows, Mac OSX, Linux).

The sequencing data is a widely varied collection of files ranging in size from a few kilobytes – at last count the Department had approximately 10 million of these small files – up to 50 GB each. Processing and analysis, primarily run from Linux workstations, requires fast access to files across the entire range.

> ***"Our research organization falls between the cracks for most storage vendors, with giant imaging sets and millions of tiny genetic sequencing scraps. Finding a system that reasonably handled all our complex workflows was difficult, and in the end only QF2 was the right fit."***
>
> — **Bill Kupiec**
> IT Manager,
> Department of Embryology
> Carnegie Institution for Science

Embryology had relied on a legacy EMC/Isilon system as its primary storage. However, that system was approaching end-of-life, and a replacement needed to be found to deal with the Department's increasing demands for storage, performance and capacity.

"One of our major replacement criteria was finding a storage system that could bridge that file volume and variety," says Bill Kupiec, IT Manager for Carnegie's Department of Embryology. "It had to handle both the streaming needed for very large data sets and the fast processing required for millions of small files. That made locating a workable solution extremely challenging."

CARNEGIE SCIENCE

Solution Overview

- 4 Qumulo QC208 hybrid storage appliances, Mellanox 40GbE switch backbone
- SMB, NFS and REST protocols
- Qumulo Care enterprise support

Key Benefits for the Carnegie Institution for Science:

- Performs equally well with both terabyte-size data sets and millions of tiny sequencing files
- Scales throughput and capacity with each additional node to keep pace with rapidly growing data requirements
- Delivers immediate storage improvements in a familiar file-based architecture
- Evolves in capabilities and features through frequent and easy software updates
- Simplifies storage platform with convenient appliance-based design
- Ensures uninterrupted operation through proactive Qumulo Care support

Weighing Object Versus File Storage

"It's safe to say we conducted an exhaustive search," says Mahmud Siddiqi, microscopy facility manager for the Department of Embryology. "If you pulled a name out of a storage cluster hat, odds are excellent that over the years we spoke with them, often multiple times," he says, before naming more than a dozen vendors. Eventually, the Department settled on two very different finalists: Scality, with its object storage approach, and Qumulo, with its modern, highly scalable file storage system.

This wasn't the first time the team had considered Qumulo – eighteen months prior it had given the new storage system a close look. "Qumulo had just launched, and we came away thinking 'wow, that's really slick – but it doesn't yet have everything we want,'" notes Siddiqi. "While we looked at other vendors, Qumulo's feature set expanded at a really impressive pace. It had quickly grown into the type of system we needed."

QF2 is a modern, highly scalable file storage system that is fast, flexible and delivers the real-time analytics necessary for visibility into data usage and performance at scale. The combination provides the storage performance and scale the Department's team wanted, packaged in a simple and affordable appliance architecture that leverages commodity hardware.

Ultimately, that ability to combine performance, scalability and simplicity won the day. Although the Scality solution was extremely impressive, "We just didn't feel that we had the time or resources to properly implement and support a Scality solution," explains Kupiec. "Qumulo met our needs while still being familiar for the team and our users."

The Department selected Qumulo's QC208 hybrid storage appliances, deploying a four node, NFS and SMB-based cluster with almost a petabyte of raw capacity. Installation and operation of the system has been entirely straightforward.

"We started our search looking for an appliance: an all-in-one solution we can just turn on, that handles anything we throw at it," Kupiec says. "From the get-go, QF2 delivered with the right capacity and performance, really checking all the boxes with a single cluster that just works."

Eliminating Bottlenecks to the Next Breakthrough

With the QF2 cluster in place, the Department's challenge of maintaining system performance across file types and sizes is rapidly becoming a thing of the past.

> ***"We just didn't feel that we had the time or resources to properly implement and support a Scality solution."***
>
> — **Bill Kupiec**
> IT Manager,
> Department of Embryology
> Carnegie Institution for Science

"Most storage vendors tout aggregate bandwidth, which isn't relevant to us," explains Siddiqi. "We care about how quickly each client can get files back and forth from the storage system, or how it handles high volume from a metadata or directory standpoint. Virtually every storage system we looked at addressed our aggregate load, but all stumbled when pushed by a single client. Except for QF2."

The team finds the new QF2 cluster is able to quickly traverse large directories, feed high file volumes and easily deliver or ingest large streaming files.

When the team has needed help in configuring the system, the Qumulo Care support has been a quick call – or an easy Slack channel – away. "...our interaction with the Qumulo support team has been great," notes Kupiec. "It's so refreshing to have multiple people quickly, knowledgeably and pleasantly come together to help us sort issues."

The Department finds QF2's REST-based interface both highly informative and significantly extensible – an important consideration as its storage needs continue to evolve. The Qumulo team is helping to keep pace with this evolution through agile two-week software sprints that continually enhance the system, and help ensure it's always meeting the Department's needs.

"Part of the reason we had difficulty finding a storage solution is that our research organization falls between the cracks for most storage vendors," says Kupiec. "We have multiple simultaneous users and single clients that pound the system. We have giant imaging sets and millions of tiny genetic sequencing scraps. Finding a system that reasonably handled all our complex workflows was difficult, and in the end only QF2 was the right fit."

Considering the importance of this developmental science, it's critical that the QF2 scalable file storage system feeds Department of Embryology researchers the data they want when they need it. After all, it would be a shame to bottleneck the next breakthrough.

The Department of Embryology, founded in 1913 in affiliation with the Anatomy Department of Johns Hopkins University, is one of six departments within the Carnegie Institution for Science. During the succeeding decades a fundamental description of human development and path-breaking experimental studies emerged. Departmental staff have uncovered the role played by genes during embryogenesis, developed widely used experimental methodologies, trained several scientific generations of biologists, and were first appointed Investigators of the Howard Hughes Medical Institute.

Qumulo