

Storage Efficiency for Research Data

CAC: Deploying highly efficient, reliable storage for research data

The Cornell Center for Advanced Computing (CAC) is a leader in high-performance computing system, application, and data solutions that enable research success. As an early technology adopter and rapid prototyper, CAC helps researchers accelerate scientific discovery. Located on the Ithaca, New York campus of Cornell University, CAC serves faculty and industry researchers from dozens of disciplines, including biology, behavioral and social sciences, computer science, engineering, geosciences, mathematics, physical sciences, and business.

The center operates Linux, Windows, and Mac-based HPC clusters and the staff provides expertise in HPC systems and storage; application porting, tuning, and optimization; computer programming; database systems; data analysis and workflow management; Web portal design, and visualization. CAC network connectivity includes the national TeraGrid and New York State Grid.

The DataDirect Networks S2A9700 storage system is used as the central storage platform for a number of departments and applications. Initially deployed for backup and archival storage, CAC is increasingly using the S2A9700 as front-line storage for applications such as genome sequencing.

“We have been very impressed with the performance DDN’s S2A9700 delivers,” said David A. Lifka, CAC director. “For genomics research – Cornell uses Solexa Sequencers and the DDN storage system is directly connected to the compute cluster, while at the same time continuing to provide backup and archive storage for our other projects and departments.”

– David A. Lifka, CAC Director

Since CAC provides services to a wide range of Cornell departments and applications, implementing centralized storage platforms is critical in ensuring an efficient, reliable and cost-effective infrastructure.

Cornell researchers were considering buying commodity, off-the-shelf storage solutions to locally store their research data. While the cost of such technology appeared initially low – the lack of coordination, data protection and system reliability detracted from the long-term value of this approach. As research productivity and access to data are directly correlated – the primary focus of the storage solution had to be high reliability and scalability.

It was clear that an affordable, centrally managed, highly available research storage system was needed in order to control costs and also to ensure that researchers remained productive. Accommodating a variety of applications and departments would prove a challenge for ordinary storage systems, but the DDN S2A9700 proved capable even beyond the initial scope of the project.

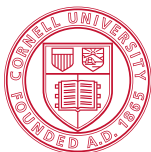


The Challenge:

Implementing a scalable, centralized storage infrastructure solution for a wide range of departments and applications. DDN provided an affordable, centrally managed, highly available research storage system accommodating a variety of research applications from genome sequencing to engineering and mathematics.

The Solution:

The center selected an S2A9700 storage system from DDN with 40TB unformatted capacity in RAID-6 configurations. DDN partnered with Ocarina Networks to provide transparent, content-aware storage optimization at CAC, reducing the overall capacity need by more than 50 percent. For some Microsoft SQL database applications, a compression rate of up to 82 percent was achieved.



Cornell University
Center for Advanced Computing

About CAC

The Cornell Center for Advanced Computing (CAC) is a leader in high-performance computing system, application, and data solutions that enable research success. As an early technology adopter and rapid prototyper, CAC helps researchers accelerate scientific discovery. Located on the Ithaca, New York campus of Cornell University, CAC serves faculty and industry researchers from dozens of disciplines, including biology, behavioral and social sciences, computer science, engineering, geosciences, mathematics, physical sciences, and business.

Solution

The center selected an S2A9700 storage system from DDN with 40TB unformatted capacity in RAID-6 configurations. DDN partnered with Ocarina Networks to provide transparent, content-aware storage optimization at CAC, reducing the overall capacity need by more than 50 percent. For some Microsoft SQL database applications, a compression rate of up to 82 percent was achieved. Ocarina's ECOsystem uses an innovative approach to data reduction.

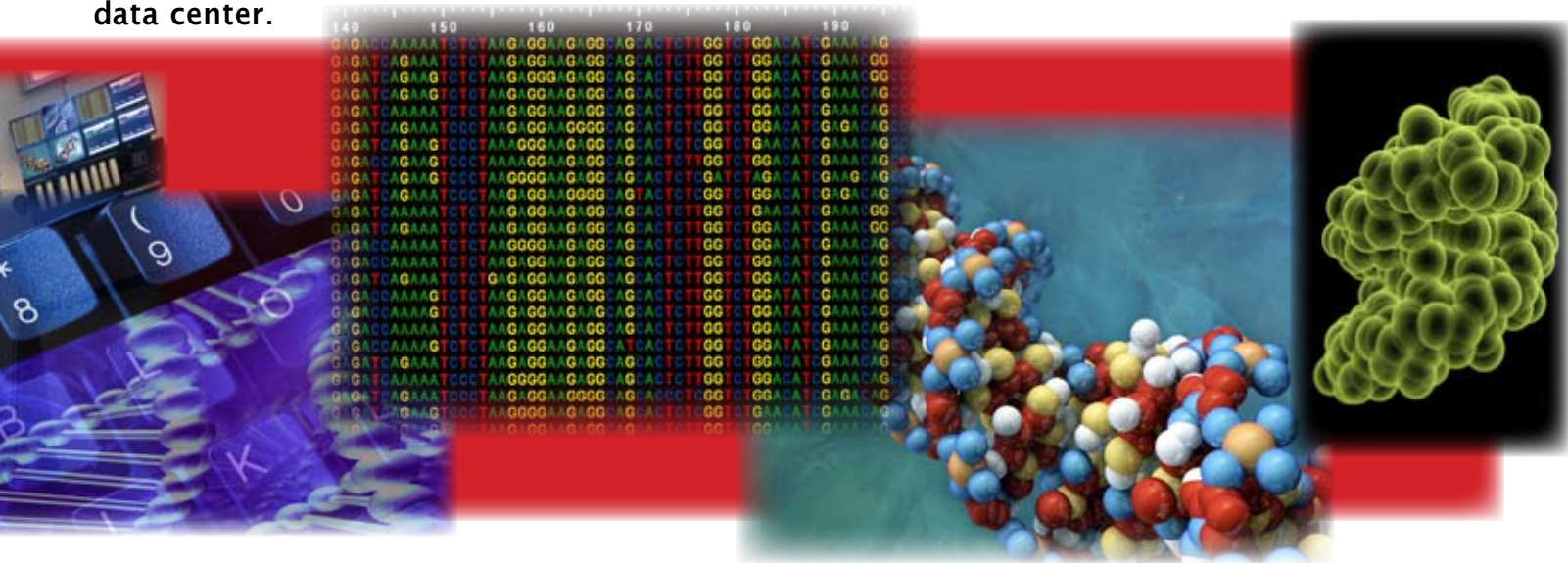
The ECOsystem first extracts files into raw binary data and applies object boundaries to the data. It then applies object dedupe and content-aware compression to the natural semantic objects found within. The object dedupe approach finds object duplicates in compressed, encoded data that would never be found using standard block dedupe. After processing object duplicates, the ECOsystem then applies content specific compression to the remaining unique object. This dual approach provides better space savings than either block dedupe or generic compression alone would. Ocarina's ECOsystem includes multiple data compressors for the types of files commonly found in research computing environments and includes over 100 algorithms that support 600 file types.



Benefits

The DDN S2A9700 combined with Ocarina Networks enables CAC to provide storage to its end-users

(researchers and other departments on campus and beyond) at very affordable prices – while also providing levels of uptime and management simplicity – which can't be found with commodity storage technology. Additional benefits include much higher availability and reliability due to DDN's DirectRAID and SATAssure features as well as high performance. Cornell Life Sciences, whose data center is located across campus, has been testing the DDN S2A9700 for central storage of life science data. They have found that the DataDirect Network system at CAC outperforms their local NFS servers by a factor of 2–3x for writes and is slightly faster for reads. These tests were performed under ordinary system and network conditions and utilized Cornell's WAN for connectivity between the Life Science data center and the CAC data center. The flexibility to implement SSD or mixed hard drive technology as well as Dynamic MAID for energy efficiency are options that CAC is investigating in order to reduce power consumption and cooling requirements in the data center.



Affordable Storage for Life Sciences

High-throughput sequencers parallelize the gene sequencing process and produce millions of sequences at once. As these machines become more powerful, individual gene sequencing and analysis enables affordable patient-specific diagnosis and treatment. Additionally, next-generation gene sequencing technologies are capable of delivering as much as 10 times the amount of sequence reads as compared to traditional sequencing systems. This new level of resolution has resulted in a sea change in processing and storage methodologies when serving and storing next-generation sequence data.

“In life sciences, next generation sequencing techniques are producing vast quantities of data that must be quickly processed and stored online for short periods of time,” said Dr. Jaroslaw Pillardy, a senior researcher at Cornell's Computational Biology Service Unit. “For example, one Solexa sequencing run produces .05 terabytes of raw data, and a single sequencer may be used multiple times per week.” Pillardy expects that new sequencing techniques will soon generate data at a rate of up to 2.5 terabytes per day.

DDN's S2A storage systems provide predictable performance and scalable capacity for the most demanding life sciences applications. Partnering with Ocarina Networks, the combined solution significantly reduced the required storage capacity, thus lowered cost – while at the same time provided the performance and reliability needed for front-line sequencing storage.

EXTREME STORAGE

DataDirect Networks, Inc. is the data infrastructure provider for the most extreme, content-intensive environments in the world—including the largest online gaming and music sites, social networking applications developers, photo and video sharing services, high performance computing environments, and more than 400 broadcast and post-production facilities around the globe. With more than 200 petabytes installed worldwide, the company's S2A™ (Silicon Storage Architecture™) technology delivers massive throughput, scalable capacity, consistency, efficiency and data integrity for today's extremely competitive and evolving markets. Founded in 1998, DataDirect Networks serves customers through its global partnerships with Dell, IBM, Sony and other industry leaders; and through its offices in Europe, India, Asia Pacific, Japan and throughout the U.S. For more information, go to www.ddn.com or call +1-800-TERABYTE (837-2298).

© 2009, DataDirect Networks, Inc. All Rights Reserved. DataDirect Networks, the DataDirect Networks logo, Silicon Storage Architecture and S2A are trademarks of DataDirect Networks. All other trademarks are the property of their respective owners.

DataDirect[™]
N E T W O R K S

ddn.com

1.800.TERABYTE

9351 Deering Ave.
Chatsworth, CA 91311
USA