University of Colorado Boulder

Supporting ground-breaking research with the pioneering PetaLibrary

Overview

The need

Successful computational research calls for participants to store and share huge amounts of data. How could CU-Boulder provide scalable, accessible storage resources to multiple teams?

The solution

CU-Boulder met demand for large-scale, low-cost data storage with the PetaLibrary—a hierarchical storage system—with IBM tape, disk systems, and Elastic Storage based on IBM® GPFS™ software.

The benefit

An estimated capacity of 1.5 Petabytes is expected to accommodate data growth over four years; researchers have access to enterprise-class technology and data management support.

Established in 1877, the University of Colorado Boulder (CU-Boulder) is the flagship university of the State of Colorado. Offering 32 different bachelor's and master's degree programs to nearly 32,000 students as of 2013, CU Boulder boasts an illustrious academic history that includes five Nobel laureates.

Recognizing the Big Data challenge

To maintain its reputation as a center for research excellence, CU-Boulder must ensure that its researchers are equipped with the tools they need to work effectively. This means tackling the challenge of escalating data faced by researchers every day.

By democratizing access to cutting-edge technology, CU-Boulder is succeeding in extending usage to non-traditional users. "With GPFS, we can present a familiar file-tree to users so that they need no special expertise to make use of sophisticated resources, meaning the solution has a wider appeal," says Peter Ruprecht, Senior HPC Analyst, University of Colorado Boulder.

Solution components

Hardware
- IBM® System Storage® DCS3700
- IBM System Storage TS3584 Tape Library

Software
- Elastic Storage based on IBM General Parallel File System (GPFS™)
- IBM Tivoli® Storage Manager

IBM Business Partner
- Re-Store, LLC

Peter Ruprecht, Senior HPC Analyst, University of Colorado Boulder, elaborates, "A variety of factors have led to research activities generating more data than ever—for example, running more sophisticated simulations and the growing popularity of digital archiving. Before, the university did not have a centralized data storage solution that could scale sufficiently to handle this kind of demand. As a result, some researchers found their own costly, unsecure and inconvenient solutions, such as USB drives, for example.

"We saw this as a missed opportunity: what if we could leverage economies of scale to give each of our researchers a slice of a much more sophisticated solution than they would otherwise have access to? By providing them with a specialized solution designed to take the headache out of data storage, we could free them up to focus more on research."

Moreover, the National Science Foundation—the principal source of scientific research funding for U.S. universities—changed its guidelines a few years ago so that every research proposal must include a robust data management plan before it can be approved.

Ruprecht comments: "For researchers who may have no previous experience with developing data management plans, it can be a time-consuming and even daunting prospect. This was another area where we believed that we could make a difference."

CU-Boulder PetaLibrary is born

CU-Boulder decided to create the PetaLibrary—a hierarchical storage solution combining disk and tape systems to provide an appealing cost-efficient model for research teams at the university. Partly funded by the National Science Foundation (NSF), the solution is designed to meet the demand for high-performance short-term storage, long-term archive storage and the ability to share data with collaborators both within CU-Boulder and across the United States.

"Without IBM GPFS and Tivoli Storage Manager, the PetaLibrary would simply not be possible."

—Peter Ruprecht, Senior HPC Analyst, University of Colorado Boulder

The disk storage layer of the solution is based on scalable high-density IBM System Storage DCS3700 and DDN SFA10k systems, clustered with IBM General Parallel File System (GPFS) software for performance and reliability. An IBM System Storage® TS3584 Tape Library system with four LTO-6 drives is the basis of the PetaLibrary's tape layer.

The university engaged IBM Premier Business Partner Re-Store to help with design and implementation of the PetaLibrary. "Without Re-Store's substantial expertise, developing the innovative Hierarchical Storage Management capabilities of the PetaLibrary would have been difficult if not impossible for us," comments Ruprecht.

He explains the cost model offered to users: "Our customers each buy a quota that includes 40 percent capacity on disk, the rest on tape. As soon as they reach their limit on disk, their oldest and largest files are automatically migrated to tape. They are able to access data stored on disk virtually instantly and over our high-capacity network. If they are accessing data that has been moved to tape, there is just a short delay while the tape cartridge is mounted and read. Since the tape is intended for archive data, this delay is generally not an issue."

Using the Hierarchical Storage Management (HSM) module in IBM Tivoli® Storage Manager and a number of custom scripts, CU-Boulder automates the migration of data between disk and tape, so that this process is totally invisible to users and requires little additional attention from administrators.

Facilitating pioneering research

At present, approximately 25 groups from multiple disciplines across the university utilize the PetaLibrary, and 400 TB of the estimated capacity of between 1.5 and 2 Petabytes is in use. CU-Boulder anticipates that the existing capacity will accommodate data growth over the next four years, and it can easily scale up the hardware by purchasing and installing additional disk drives and tape cartridges.

"By helping us more efficiently and effectively meet research needs at CU-Boulder, the PetaLibrary makes the university a more attractive place for research."

—Peter Ruprecht, Senior HPC Analyst, University of Colorado Boulder

The PetaLibrary now plays a crucial role in the university's research computing infrastructure, enabling groups to cost-effectively store and share data sets.

"Built on IBM technology, the PetaLibrary is designed to allow users to share data easily and securely," says Ruprecht. "For example, large-scale genetics studies generate huge data sets that need to be analyzed by various people. Using the PetaLibrary, researchers can make this data instantly available to collaborators rather than creating and sending copies between labs. As a result, they can work faster and their work has less impact on the university network."

Another key benefit is the associated data management support that CU-Boulder provides to PetaLibrary users. Ruprecht adds, "Research groups can utilize our data support services to help understand established data management policies, and to understand and implement best practices throughout the data life cycle. This helps expedite funding applications so researchers can get proposals done and submitted sooner."

Removing barriers to leading-edge technology

Comparing favorably with alternative options on cost—such as purchasing storage capacity in the cloud—the PetaLibrary brings enterprise-class technology within reach of users that may not otherwise have access.

"Developing a centralized solution enables us to take advantage of economies of scale to provide technology that would otherwise be too expensive, or too complex, for individual research groups," comments Ruprecht. "The alternatives to IBM GPFS and Tivoli Storage Manager are out of our price range, so without these tools the PetaLibrary would simply not be possible."

Moreover, combining disk and tape storage enables CU-Boulder to optimize operational costs with performance, as Ruprecht explains, "Our disk systems consume around 3.5 KW of power under normal operation while our tape system consumes much less, about 0.75 KW. By using much more energy-efficient tape storage for archive data, we can lower costs while retaining the high performance of disk storage for more frequently-accessed data."

"Built on IBM technology, the PetaLibrary is designed to allow users to share data easily and securely."

—Peter Ruprecht, Senior HPC Analyst, University of Colorado Boulder

Seeing the solution in action

The CU-Boulder Libraries group was an early adopter of the PetaLibrary, using the solution to digitize audio, video, images and text to nationally-accepted archival standards, and will soon add 3D objects to the list for digitization. The project calls for large-scale storage due to the huge size of the initial files—often as much as 120 GB per hour of digitized video. In 2014, the libraries anticipate that they could require more than 80 TB of data storage, so the PetaLibrary is a key tool in ensuring reliable archiving of the university's library collections.

Another major digitization project supported by the PetaLibrary is at the University of Colorado Museum of Natural History, which is creating digitized copies of its entire collection—numbering 4.5 million objects—to ensure they are saved for posterity. By storing images of each object alongside complex metadata, the museum can open up its collection in a way that was not possible before.

Elsewhere, CU-Boulder campus researchers rely on the PetaLibrary to store data generated by simulation studies and laboratory experiments. For example, researchers in the Psychology and Neuroscience department use the PetaLibrary to store large files from Magnetic Resonance Imaging scans and share the data with collaborators at other universities.

Ruprecht concludes, "We are very excited about the potential for the PetaLibrary in the future. By helping us more efficiently and effectively meet current and future research needs at CU-Boulder, it makes the university a more attractive place for research, enhancing our worldwide reputation as a place where innovation can flourish."

For more information

To learn more about IBM Platform Computing, please contact your IBM representative or IBM Business Partner, or visit the following website: **ibm.com**/systems/gpfs

IBM Systems
Route 100
Somers, NY 10589

Produced in the United States of America
January 2015

DCC03040-USEN-00