



PANASAS Case Study:

# Brookhaven National Laboratory Deploys Panasas Storage Cluster

June 2005 | CS 121 |

**Abstract:** Brookhaven National Laboratory (BNL), home to the Relativistic Heavy Ion Collider (RHIC), maintains the RHIC Computing Facility (RCF) for data reconstruction and analysis. The facility has experienced tremendous growth in the size of its Linux compute cluster, that now exceeds two thousand dual-CPU compute nodes. In 2004, administrators at BNL recognized that their NFS filesystem was failing with increasing frequency under the growing cluster workload. Since their plans called for ongoing cluster growth, a year-long search was launched for a more reliable, easy-to-manage, high-performance storage system. Their comprehensive evaluation led to the selection of the Panasas ActiveScale Storage Cluster. This case study provides details on the unique challenges BNL faced in moving from a traditional client/server model to that of a distributed network filesystem model, along with details of how the Panasas Storage Cluster met BNL's requirements for bandwidth, scalability, and reliability.

## Introduction

The Relativistic Heavy Ion Collider (RHIC) at Brookhaven National Laboratory (BNL) is a worldclass scientific research facility that began operations in 2000, following ten years of development and construction. Hundreds of physicists from around the world use RHIC to study what the universe may have looked like in the first few moments after its creation. RHIC drives two intersecting beams of gold ions in subatomic, head-on collisions. From these collisions, physicists hope to gain an improved understanding of why the physical world behaves the way it does, from the smallest subatomic particles to the largest stars.

The RHIC Computing Facility (RCF), constructed to warehouse and analyze the collider data, has relied on NFS servers and SAN storage to provide online access of the collider data to a Linux cluster that now exceeds two thousand dual-CPU Linux nodes. With the deployment of fast Linux cluster nodes, all having Gigabit Ethernet (GbE) network connections, it was apparent that the standard client/server model that was used for the previous six years would be insufficient to meet current and future computing demands.

## Architecture

Over the past few years, equipment from VA Linux, IBM, and Dell has been acquired through a competitive procurement process. These machines range from having dual Pentium III 450 MHz processors to dual 3.2 GHz Pentium IV processors. They have between 512 MB – 2 GB RAM, and 40 GB – 400 GB local disk storage. The older machines have 100 Mb network interfaces, but the newer, faster machines have GbE network interfaces, all of which are connected to a Cisco network with cat 5E or cat 6 copper cables. These machines were originally configured with the RedHat Linux distribution, but have migrated to Scientific Linux, which is RedHat Linux open source code, compiled and distributed by Fermi National Accelerator Laboratory.

There are 37 Sun SPARC systems running Solaris 9, which act as NFS servers. This group of servers is comprised of nine E450s, ten V480s, and 18 V240s. The older E450s have 4 x 400 MHz processors and 4 GB of RAM. The V480s have 2 x 900 MHz processors and 4 GB of RAM. The V240s have either 2x1 GHz or 2x1.5 GHz processors with 2 GB of RAM. The newer servers, V480s and V240s, have onboard GbE NICs, but the older E450s have GbE Sysconnect NICs installed. The servers utilize Veritas Foundation Suite, Veritas Volume Manager, and Veritas Filesystem to create, manage, and export filesystems. The servers use external storage connected through SAN switches.

The RCF has three SAN islands that connect NFS servers to their storage, which consist of 24 Brocade switches, models 2250 and 2800. The servers all have dual fibre channel host bus adapters, models Qlogic QLA2200 and QLA2300.

The SAN storage consists of fibre channel controllers and disks procured from various system integrators. The controllers run in active/active dual controller mode and the disks are either 73 GB SCSI disks or 73 GB or 146 GB fibre channel disks. The controllers export LUNS (logical volumes), which are typically configured with RAID 5 parity protection with not more than eight disks in a RAID set.

## System Characteristics

RHIC generates massive amounts of data that is buffered in storage by the detector, and eventually written to tape in the RCF. Over the last year alone, the collider generated over one Petabyte (PB) of data. When a reconstruction process is started, blocks of raw data are moved from tape to disk so that measurements gathered by different sensors can be sorted by time and packaged into individual events. The reconstruction program is individual to each experiment, but the process of going through all the files is locally developed and uses the Condor batch system to move and process large amounts of data. Production managers then make the reconstructed data available to scientists. They use either the LSF or CONDOR batch systems to submit their programs with event data for analysis. The batch system processes over one million jobs a month.

## Unique I/O Challenges

The RCF recognized that newer, faster Linux nodes needed greater access to data to be productive. The RCF team also recognized that failure to match growing cluster demand with new file sharing technology would result in NFS server crashes due to excessive load. This wasted valuable processor time and caused confusion among users when thousands of jobs had to be investigated to verify one successful job.

The following factors were considered when evaluating alternative file sharing technologies:

- Reliability needed to keep the cluster running
- Performance required to serve and write data for the cluster to keep it busy
- Parallel I/O for optimized performance for each node
- Scalability needed to support a large number of cluster nodes
- Cost, both in terms of acquisition and management
- The ability of the storage architecture to support the next generation of clusters.

## Storage Requirements

Recording and saving the significant amount of data generated by the RHC collider is accomplished using a hierarchical system of IBM servers, StorageTek tape silos, and HPSS (High Performance Storage System) software. Once archived, portions of the data are transferred to online storage for use in reconstruction and analysis. The workflow requires copying blocks of data from tape to disk, where it is then available for access by the Linux computer cluster. Capacity requirements are insatiable as the facility has petabytes of data. Users benefit to the extent that that data is on line and available. In addition to online capacity, key considerations include cost, performance, power, cooling, and floor space.

Performance requirements are continuously changing, depending on the quantity and speed of the Linux cluster nodes. Currently, the RCF manages 2,000 Linux nodes for the RHIC experiment and 400 Linux nodes for the USAtlas experiment. The primary factor requiring a change in online storage technology was the prevalence of GbE network interfaces on all of the Linux cluster nodes being purchased. At present, 800 of 2,000 RHIC nodes have GbE NICs (Network Interface Cards), and 300 of 400 USAtlas nodes have GbE NICs. Though the increasing processor speeds were cause for concern, the 100Mb NIC represented a bottleneck, that throttled access to the storage servers. Now that the fastest processors have fast network

adapters, the performance bottleneck has moved from the node's network interface to the network infrastructure, and finally to the storage server. The RCF had previously attempted to keep storage capacity of each NFS server under four Terabytes (TB), but under increased load from the Linux cluster, this would have to be greatly reduced.

Factors making expansion of the NFS server facility undesirable were:

- The cost of servers
- The cost of host bus adapters for Ethernet and storage networking
- The cost of SAN equipment (switches, GBICS, cables)
- The cost of Veritas licensing
- Space, power and cooling issues
- The effort involved in purchasing and coordinating all the above
- The management involved in building and maintaining hundreds of servers the attention required by users having so many filesystems available.

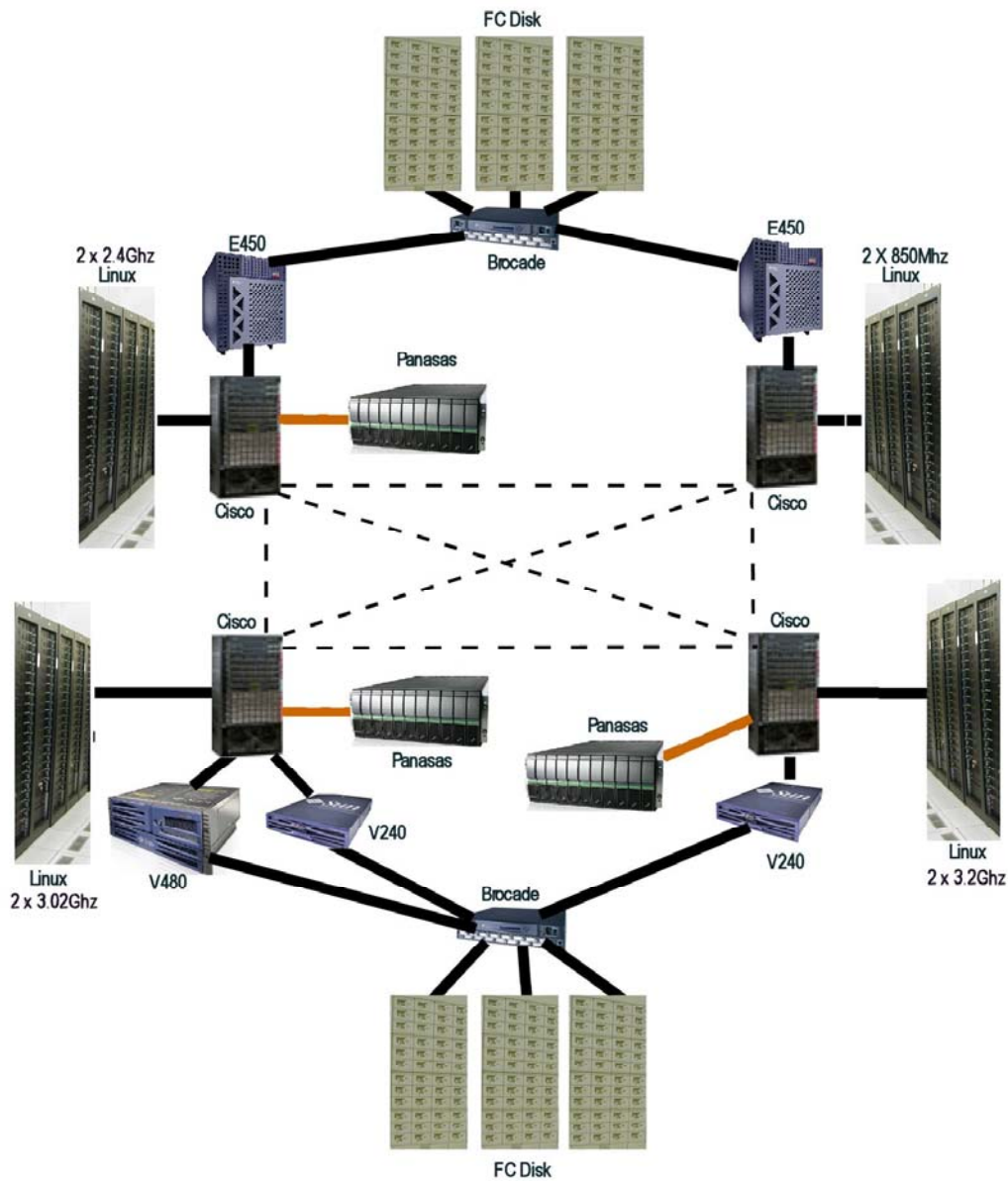
Once the possibility of deployment of more NFS-based storage systems was eliminated, the search for a storage system that could scale to support the large number of nodes in the RCF yielded only a few cutting edge technology companies. Performance and reliability, as well as the ability to easily grow the storage facility to meet future demands, were key factors in selecting the Panasas ActiveScale Storage Cluster.

## Deployment of Storage from Panasas

BNL already had several Cisco network switches, each populated with Linux nodes and NFS servers. The hardware components were purchased together in yearly procurements expanding the facility. The switches were connected by one, two, or four GbE links. Since each year additional nodes and storage were purchased along with a switch to bring them online, the newer, faster network switches also had the fastest nodes and servers.

The Panasas Storage Cluster offered the desired network bandwidth to access the growing storage pool. A rack of eight shelves provided 32 GbE network connections. RCF's previous model of buying a switch to connect storage to the network would be ineffective as all compute nodes would have a bottleneck at the switch-to-switch communication level. Therefore, the 20-shelf implementation was broken down into several smaller bladesets that were connected through several switches. Nodes on the same switch were configured to have unrestricted access to the storage, while nodes from other switches had less contention for switch interconnect bandwidth.

The 20 shelves were built into ten 2 shelf bladesets and deployed on the three network switches connected to the newest, fastest Linux nodes. The RCF believes this will get the greatest output from its processor farm. [Note: Each shelf was deployed using only two of the four GbE links available. Since switch ports are expensive, the RCF decided to postpone deployment of the other links until necessary based on load.] **Figure 1** below graphics depicts the BNL, RHIC Computing Facility architecture.



**Figure 1: RCF Configuration**

Meeting the RCF's storage needs required more than just deployment of a fast hardware architecture. There are several products available that merely speed up the NFS file serving performance. Going one step further, Panasas was able to maximize performance by allowing access to the data using the company's DirectFLOW client.

DirectFLOW empowers each BNL Linux cluster node, through the use of a small installable file system from Panasas, to access data directly from the Panasas StorageBlades without the delays normally associated with NFS filer servers. This unique, direct data transfer capability enables exceptional performance in the Panasas Storage Cluster.



A simple, three-step process is required to initiate direct data transfers:

1. Requests for I/O are made to a Panasas DirectorBlade, which controls access to data.
2. The DirectorBlade authenticates the requests, obtains the object maps of all applicable objects across the StorageBlades and sends the maps to the I/O nodes.
3. With authentication and virtual maps, I/O nodes access data on StorageBlades directly and in parallel.

This concurrency eliminates the bottleneck of traditional, monolithic storage systems, which manage data in small blocks, and delivers record-setting data throughput. The number of data streams is limited only by the number of StorageBlades and the number of compute nodes in the server cluster.

On the storage side, PanFS is a parallel file system that divides files into large virtual data objects. These objects can be stored on Panasas StorageBlades, or units of storage, enabling dynamic distribution of data activity throughout the storage system. Parallel data paths between compute clusters and the StorageBlades result in high performance data access to large files. The result is that the Panasas ActiveScale Storage Cluster delivers performance that scales almost linearly with capacity.

## Summary

BNL has successfully deployed a distributed network storage environment for higher reliability, higher performance, and unlimited scalability. The RCF processor farm uses the Panasas Storage Cluster as a centralized storage pool to reconstruct and analyze collider data to support 2,000 cluster nodes. The Panasas system enables the Linux cluster to stay busy processing jobs rather than waiting for I/O operations to complete. The combination of enhanced cluster utilization and faster completion of jobs is a direct result of the Panasas system acting as a Cluster Accelerator™. Specifically designed to support Linux clusters, the Panasas Storage Cluster scales performance in concert with capacity. As such, it is capable of meeting the needs of the world's leading high-performance computing clusters, both now and for future generations of cluster technology.



---

6520 Kaiser Drive Fremont, California 94555 Phone: 1-888-PANASAS Fax: 510-608-4798 [www.panasas.com](http://www.panasas.com)

© 2005 Panasas Incorporated. All rights reserved. Panasas, the Panasas logo, Cluster Accelerator, Panasas ActiveScale Storage Cluster, Panasas ActiveScale File System, Panasas ActiveRAID, Panasas ActiveSpares, Panasas DirectFLOW, Panasas StorageBlades, Panasas DirectorBlade, Panasas PanActive Manager, Panasas PanActive Support Program, Panasas PanActive Link and MyPanasas are trademarks of Panasas in the United States and other countries.

---