



Customer Success Story

Los Alamos National Laboratory

Panasas High Performance Storage Powers the First Petaflop Supercomputer at Los Alamos National Laboratory

June 2010

Highlights

First Petaflop Supercomputer

- #1 on the Top-500 list in 2009
- Over 3,250 Compute Nodes
- Over 156 I/O Nodes
- Over 12,000 Core Processors
- Hundreds of Thousands of Cell Processors

Panasas High Performance Storage Solutions

- 100 Panasas Storage Shelves
- 2 Petabytes Capacity
- 55 GB/s Throughput
- Throughput Scales Linearly with Capacity
- Non-Stop Availability & Simple to Deploy

Abstract

Scientists want faster, more powerful high-performance supercomputers to simulate complex physical, biological, and socioeconomic systems with greater realism and predictive power. In May 2009, Los Alamos scientists doubled the processing speed of the previously fastest computer. Roadrunner, a new hybrid supercomputer, uses specialized Cell coprocessors to propel performance to petaflop speeds capable of more than a thousand trillion calculations per second.

One of the keys to the project's success was the need for a highly reliable storage subsystem that could provide massively parallel I/O throughput with linear scalability that was simple to deploy and maintain. Los Alamos National Laboratory deployed Panasas High Performance storage to meet the stringent needs of the Roadrunner project. Panasas provides scalable performance with commodity parts providing excellent price/performance, scalable capacity and performance that scale symmetrically with processor, caching, and network bandwidth.

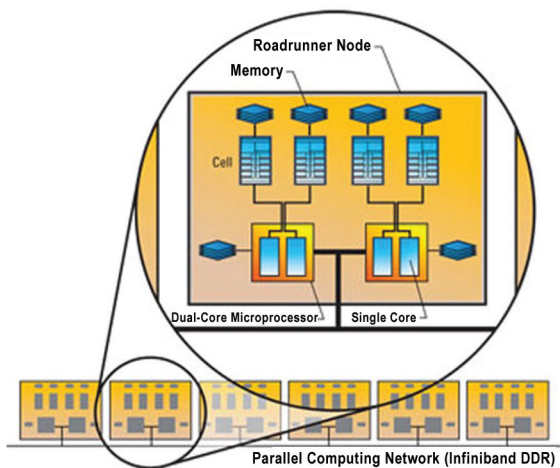
Introduction

From its origins as a secret Manhattan Project laboratory, Los Alamos National Laboratory (LANL) has attracted world-class scientists to solve the nation's most challenging problems. As one of the U.S. Department of Energy's multi-program, multi-disciplinary research laboratories, Los Alamos thrives on having the best people doing the best science to solve problems of global importance. The Laboratory's central focus is to foster science, encourage innovation, and recruit and retain top talent while balancing short-term and long-term needs with research driven by principal investigators and aligned with missions.

In 2002, when Los Alamos scientists were planning for their next-generation supercomputer, they looked at the commodity market for a way to make an end run around the speed and memory barriers looming in the future. What they found was a joint project by Sony Computer Entertainment, Toshiba, and IBM to develop a specialized microprocessor that could revolutionize computer games and consumer electronics, as well as scientific computing.

The major application areas addressed were radiation transport (how radiation deposits energy in and moves through matter), neutron transport (how neutrons move through matter), molecular dynamics (how matter responds at the molecular level to shock waves and other extreme conditions), fluid turbulence, and the behavior of plasmas (ionized gases) in relation to fusion experiments at the National Ignition Facility at Lawrence Livermore National Laboratory.





Source: Los Alamos National Laboratory

Roadrunner Architecture

Roadrunner is a cluster of approximately 3,250 compute nodes interconnected by an off-the-shelf parallel-computing network. Each compute node consists of two AMD Opteron dual-core microprocessors, with each of the Opteron cores internally attached to one of four enhanced Cell microprocessors. This enhanced Cell does double-precision arithmetic faster and can access more memory than can the original Cell in a PlayStation 3. The entire machine will have almost 13,000 Cells and half as many dual-core Opterons.

Unique I/O Challenges

LANL used a breakthrough architecture designed to achieve high performance. The cluster could run a large number of jobs. In order to make the best use of the cluster, there were some I/O challenges and considerations that had to be addressed, including achieving:

- 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

- Reliability needed to keep the cluster running
- A reasonable cost, both in terms of acquisition and management
- A storage architecture that could support future generations of clusters

Storage Requirements

LANL wanted a shared storage architecture where all their compute clusters on a network could have access to shared Panasas storage. This is in contrast to many other HPC sites that bind a storage cluster tightly to a compute cluster. LANL has been using Panasas for all their high performance storage needs for several years and so naturally, when they deployed RoadRunner, Panasas storage was the logical choice to satisfy their demanding performance, availability and scalability requirements.

In order to achieve the performance goals of RoadRunner, the network storage system would have to deliver superior bandwidth, lower latency and be superior in terms of the file creation rate per second and the aggregate throughput. Parallel I/O would be important, as this would enable parallel data streams to go to the 156 I/O nodes, which in turn provide I/O service to the compute nodes. In addition, the storage system would have to be able to scale to support storage capacities of 10 Petabytes (PB) and beyond, plus a growing number of nodes in the cluster. This eliminated the possibility of implementing NFS-based storage systems, as they would not be able to scale past a certain number of nodes.

Availability was another key consideration. Because RoadRunner is such a high demand resource, downtime for maintenance, expansion and repair is extremely scarce, so the storage system would need to support automatic provisioning for easy growth

LANL designed the cluster architecture to be simple, easy to manage, and cost effective. One aspect of simplicity and cost was to use I/O nodes that interface with network attached storage, lowering cost by reducing the number of GbE connections from a few

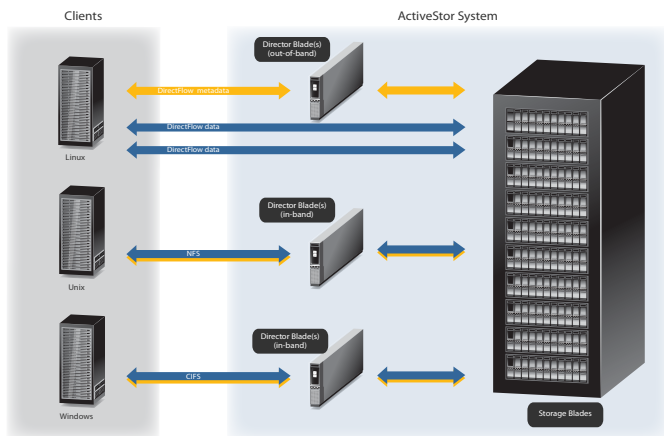
thousand to a few hundred. The storage system used, likewise, would have to be easy to manage, provide a reasonable total cost of ownership, and fit into the established architecture.

Last but not least, the storage system architecture needed to have “headroom” to provide I/O to larger future cluster configurations. Instead of being able to support just a single large cluster, the storage architecture needs be able to scale to support multiple clusters from a single, central storage pool.

Deployment of Storage from Panasas

Panasas ActiveStor™ high performance storage met all the criteria dictated by LANL’s computer cluster architecture. Panasas utilizes a parallel file system and provides GbE or InfiniBand connectivity between some

PANASAS Multi-Protocol Access



of the cluster nodes and storage. In fact, the Panasas storage system itself is a cluster. It uses an object-based file system that is an integral part of the PanFS™ Storage Operating System.

PanFS divides files into large virtual data objects. These objects can be stored on Panasas storage blades or units of storage, enabling dynamic distribution of data activity throughout the storage system.

Parallel data paths between compute clusters and the storage blade modules result in high performance data access to large files. The result is that Panasas Scale-out NAS delivers performance that scales almost linearly with capacity. In fact, the current implementation supports more than 2 PB capacity while delivering a massive 55 GB per second throughput.

Parallel access is made possible by empowering each of the LANL cluster I/O nodes with a small installable file system from Panasas—the DirectFlow® client access software. This enables direct communication and data transfer between the I/O nodes and the storage blades.

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

1. Requests for I/O are made to a Panasas director blade, which controls access to data.
2. The director blade authenticates the requests, obtains the object maps of all applicable objects across the storage blade and sends the maps to the I/O nodes.
3. With authentication and virtual maps, I/O nodes access data on storage blade modules 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 storage blades and the number of I/O nodes in the server cluster.

Performance is also a key factor in evaluating the cost effectiveness of storage for large, expensive clusters. It is important to keep a powerful cluster busy doing computations and processing jobs rather than waiting for I/O operations to complete. If a cluster costs \$3.5M and is amortized over 3 years, the cost is approximately \$3200 per day. As such, it makes sense to keep the cluster utilized and completing jobs as fast as possible. In order to do this, outages have to be minimized and the cluster must be kept up and running. Therefore, system availability is another key factor.

“
 Rather than having a cluster node failure at least once a week, as a comparable system with local disks would experience, the time between node failures was increased to once every 7 weeks.
 ”

*Ron Minnich,
 Leader of cluster
 research team, LANL*

improves reliability by providing even more paths to serve data to the compute cluster. Furthermore, by having a centralized pool of high-performance storage, there is no need to copy data for different kinds of jobs. After the computation jobs, visualization tasks can take place with a “compute in place” approach rather than copying the data to another storage system.

Summary

The Roadrunner project has proven to be a tremendous asset to the Laboratory’s nuclear weapons program simulations as well as for other scientific endeavors like cosmology, antibiotic drug design, HIV vaccine development, astrophysics, ocean or climate modeling, turbulence, and many others.

The Panasas architecture is designed specifically to support Linux clusters, scaling performance in concert with capacity. Panasas ActiveStor scale-out NAS is capable of meeting the needs of the world’s leading high performance computing clusters, both now and for future generations of cluster technology.

The Panasas ActiveStor system provided the availability that LANL was looking for. In terms of simplicity of administration, the Panasas architecture allows management of all data within a single seamless namespace. There is no NFS root as NFS is replaced by a global file system that is scaleable. Data objects can be dynamically rebalanced across storage blades for continual ongoing performance optimization. Furthermore, the object-based architecture enables faster data reconstruction in the event of a drive failure because storage blade modules have the intelligence to reconstruct data objects only, not unused sectors on a drive.

Finally, the Panasas storage architecture is capable of supporting future generations of more complex cluster configurations, including the scalability to support multiple clusters from one central storage pool. Instead of using one big, expensive GbE switch through one subnet, Panasas storage can be configured across many subnets through smaller, less expensive network switches that connect to the I/O nodes. This

