



LOS ALAMOS NATIONAL LAB “DISKLESS CLUSTER” DEPLOYS PANASAS PARALLEL STORAGE

CASE STUDY | APRIL 2005

ABSTRACT

In 2002, Los Alamos National Laboratory (LANL) deployed a Linux cluster supercomputer, one of the five fastest in the world at that time. In order to increase reliability, LANL chose to go with a “diskless” architecture. By storing data outside the cluster rather than on internal disks within the cluster nodes, LANL was able to achieve an order of magnitude improvement in reliability. This case study provides details on the unique I/O challenges in implementing a “diskless” approach and on how the Panasas storage architecture met the requirements in terms of bandwidth, scalability, reliability, and cost.



INTRODUCTION

In late 2002, the Los Alamos National Laboratory (LANL) deployed a Linux cluster supercomputer, then one of the five fastest in the world. Capable of handling 10 trillion operations per second, this machine has since been used for the compute-intensive work of the Laboratory's unclassified projects. In addition, it serves as a prototype for future Linux supercomputer clusters. These clusters will deliver the massive computational capacity required for a wide scope of scientific study, including molecular dynamics analysis, weather pattern simulation, research into cures for viruses, and simulations of scenarios involving nuclear materials.

To effectively deploy the cluster, LANL used a unique diskless architecture. Without local disks in the compute nodes, LANL reduced the number of components in the cluster that generate heat or are vulnerable to failure. This removed a huge reliability issue. Instead of having to replace nodes whose disks presented points of failure, a network storage system would centralize the data with RAID protection.

ARCHITECTURE

Based on its computational needs, LANL used a 1,024-node Linux Network cluster with 2,048 processors. Rather than using proprietary software, LANL collaborated with the Linux community to customize Linux for easier installation and management of the sizable compute cluster.

First, LANL replaced the BIOS bootstrap mechanism with LinuxBIOS, which allows for fast booting and simplifies nodes. No local file system or local disk is required, and power supply fans are the only moving parts on the node. Nodes come up in under 8 seconds, and the operating system then contacts a master node to obtain and boot a production-level system kernel comprised of the Beowulf Distributed Process Space (BProc) and Linux. At that point, initialization is completed, taking about 2.5 minutes for the entire cluster of 1024 nodes.

BProc, a small set of kernel modifications, utilities and libraries, provides a single system image of the entire cluster, and allows for booting in either the master or slave mode. A master node will go to local disk to boot, but the other nodes will go to the network and contact a master node. The master node then sends the other nodes a kernel for booting and registers the nodes as part of the cluster.

Users submit programs to the master nodes, but jobs can be migrated and assigned to the other nodes in the cluster. The BProc software has a component called BPsh that is used to start tasks for parallel processing. The set of software that LANL runs to facilitate cluster setup and management is called Clustermatic, and LANL distributes it at the www.clustermatic.org web site.

All binaries are available only on the master node. Every program is launched on the master and migrated to one of the slave nodes automatically. Even so, LANL achieved Message Passing Interface (MPI) startup times for 16 MB jobs of 3 seconds on 1000 nodes. The slave nodes have no local storage. The input for a program or the results generated by a program can not be stored on the master node as it will quickly become the bottleneck. This cluster architecture requires a scalable parallel file system to store and retrieve data.

LANL configured the cluster nodes into 64 groups of 16 nodes. LANL designed the nodes in the cluster to be diskless, that is, without storage for data, except for two master nodes, which had disks for booting the system.

This innovative architecture employs 1 node per group to act as an I/O node, freeing the rest of the nodes to concentrate on computation. Thus, each group of 15 compute nodes has its own, associated I/O node; however, each group of compute nodes can also “borrow” I/O nodes from 3 other groups if necessary. The I/O nodes are connected to the cluster compute nodes with Myrinet as the network infrastructure. To interface with external network attached storage, the I/O nodes utilized gigabit Ethernet (GbE) connections.

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 the next generation of clusters.

Storage Requirements

The diskless cluster was already up and running before storage was implemented. The cluster was built with the idea of having a high-performance, high-reliability central storage system instead of local disks. The performance of the network storage system would have to be superior in terms of the file creation rate per second and the aggregate throughput. The storage system would have to deliver superior bandwidth and lower latency as compared to local disks.

Parallel I/O would be important, as this would enable parallel data streams to go to the 64 I/O nodes, which in provide I/O service to the compute nodes. In addition, the storage system would have to be able to scale to support the large 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.

Reliability was a key challenge and consideration. The rationale for a diskless cluster configuration was very clear. If a clustered system of over 1000 nodes were to utilize internal disks, say 5 disks per node, mean time between failures (MTBF) for the cluster’s compute nodes would be much less than one week. A comparable cluster with just one internal disk per node would already have an MTBF of 1 week. The diskless cluster architecture would have to improve the node failure rate significantly.

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,

an ideal storage architecture would be able to scale and support multiple clusters from a single, central storage pool.

Deployment of Storage from Panasas

LANL deployed Panasas storage as it fit the criteria dictated by the computer cluster architecture. Panasas has a parallel file system and provides for GbE connectivity between some of the cluster nodes and storage. In fact, the Panasas storage system itself is a cluster. It uses an object-based file system called the Panasas PanFS™.

The PanFS file system is a parallel file system that divides files into large virtual data objects. These objects can be stored on Panasas StorageBlade® modules or units of storage, enabling dynamic distribution of data activity throughout the storage system.

Parallel data paths between compute clusters and the StorageBlade modules result in high performance data access to large files. The result is that the Panasas ActiveStor Parallel Storage Cluster delivers performance that scales almost linearly with capacity.

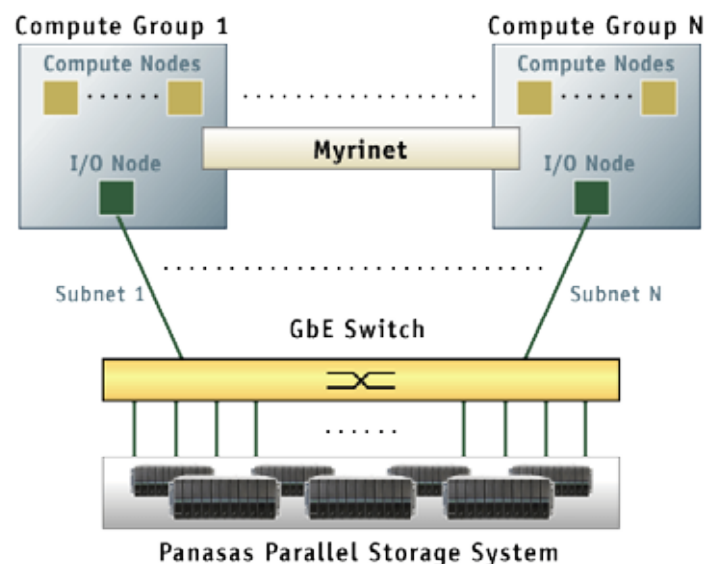
Parallel access is made possible by empowering each of the LANL cluster I/O nodes with a small installable file system from Panasas -- the DirectFLOW® client access software. This enables direct communication and data transfer between the I/O nodes and the StorageBlade modules.

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

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

Performance is a key factor in evaluating 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, reliability is another key factor.



In addition to significant performance improvement, parallel I/O, and scalability, the Panasas ActiveStor parallel storage system provided the reliability that LANL was looking for. Ron Minnich, leader of the cluster research team at LANL, said, “Rather than having a cluster node failure at least once a week, as a comparable system with local disks would experienced, the time between node failures was increased to once every 7 weeks.”

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 the StorageBlade module for continual ongoing performance optimization. Furthermore, the object-based architecture enables faster data reconstruction in the event of a drive failure because StorageBlade 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 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

In conclusion, LANL has successfully deployed a diskless cluster for higher reliability, high performance, scalability, and lower cost. The LANL supercomputer cluster uses the Panasas Parallel Storage Cluster as a centralized storage pool to support the 1024 nodes in the cluster. The ActiveScale architecture is designed specifically to support Linux clusters, scaling performance in concert with capacity. The Panasas ActiveStor Storage Cluster is capable of meeting the needs of the world’s leading high performance computing clusters, both now and for future generations of cluster technology.



Accelerating Time to Results™

6520 Kaiser Drive Fremont, California 94555 Phone: 1-888-PANASAS Fax: 510-608-4798 www.panasas.com
1-888-PANASAS (US & Canada) 00 (800) PANASAS2 (UK & France) 00 (800) 787-702 (Italy) +001 (510) 608-7790 (All Other Countries)

©2007 Panasas, Inc. All rights reserved. Panasas, the Panasas logo, and DirectFlow are trademarks or registered trademarks of Panasas, Inc. in the United State and other countries. All other trademarks are the property of their respective owners.

Information supplied by Panasas, Inc. is believed to be accurate and reliable at the time of publication, but Panasas, Inc. assumes no responsibility for any errors that may appear in this document. Panasas, Inc. reserves the right, without notice, to make changes in product design, specifications and prices. Information is subject to change without notice.