

Serial Attached SCSI Storage

for High-Performance Computing

Serial Attached SCSI (SAS) is the successor of SCSI technology, and is designed to meet the performance requirements of new-generation servers by taking advantage of serial connection to help improve performance and scaling. This article discusses deployment of SAS storage in high-performance computing clusters.

BY AZIZ GULBEDEN, AMINA SAIFY, ANDREW BACHLER, AND RAMESH RADHAKRISHNAN, PH.D.

Related Categories:

Characterization

Cluster management

Dell PowerEdge RAID
Controller (PERC)

Dell PowerEdge servers

Dell PowerVault storage

High-performance
computing (HPC)

Performance

RAID controllers

Serial Attached SCSI (SAS)

Storage

Visit www.dell.com/powersolutions
for the complete category index.

Serial protocols are becoming widespread as performance requirements advance beyond what traditional bus-based systems can provide. Traditionally, the SCSI disks are connected to a shared parallel bus. However, the accuracy of parallel connections decreases at high speeds, and performance requirements limit the bus length.

The Serial Attached SCSI (SAS) protocol is designed to help avoid these shortcomings. Serial connections between devices can provide much higher throughput than parallel bus-based connection schemes. The current SAS technology uses point-to-point connections with a maximum 300 MB/sec capacity per disk—in contrast, SCSI disks share a bus that has a maximum 320 MB/sec capacity. The SCSI Trade Association expects the performance of SAS links to double with SAS 600 (expected to be available in 2007) and double again with SAS 1200 (expected to be available

in 2010).¹ Additionally, SAS overcomes the SCSI limit of 16 devices per channel. As a result of these improvements, even though SAS uses the same set of SCSI commands, it is not backward compatible with SCSI.

Storage needs in high-performance computing clusters

In a high-performance computing (HPC) environment, storage is a key component. HPC applications require high performance and highly available access to large storage. For example, a fluid dynamics simulation starts parallel processes on several compute nodes that produce a large amount of data. The data is shared among different processes through a distributed file system on the shared storage, and the stored data is reused and updated as the simulation proceeds.

¹ "Serial Attached SCSI – Roadmap," by the SCSI Trade Association, January 2004, www.scsita.org/aboutscsi/sas/SAS_roadmap2004.html.

In addition to high performance and high availability, the storage systems must also be manageable and expandable. Manageability allows setting the appropriate configuration for the storage and enables hardware-related problems to be fixed easily. Furthermore, the storage can be tuned for various applications so that they benefit from custom cache- or block-size settings. Extensibility is necessary to allow growth as requirements change. These factors make using external disk storage attractive for clusters because the external disk arrays are optimized to address these challenges.

Test environment for comparing SAS and SCSI

To compare the performance of SAS storage with SCSI storage, a team of Dell engineers ran benchmarks on two Dell™ PowerVault™ storage enclosures in April 2006. The test environment used one Dell PowerVault MD1000 SAS storage enclosure and one Dell PowerVault 220 SCSI enclosure; both were connected to a Dell PowerEdge™ 1950 server with dual 3.2 GHz processors (see Figure 1).

PowerVault MD1000 configuration: External SAS storage

The PowerVault MD1000 is an external disk storage enclosure utilizing SAS disks; it is classified as a JBOD (Just a Bunch of Disks). RAID configuration on the storage is managed by a Dell PowerEdge Expandable RAID Controller (PERC) 5/E adapter that resides inside a PowerEdge server. The PERC 5/E is required on the server to communicate with the PowerVault MD1000. The PERC 5/E is a PCI Express card designed to deliver four times the bandwidth of PCI Extended (PCI-X); it supports RAID-0, RAID-1, RAID-5, RAID-10, and RAID-50. One x4 SAS cable, which can provide bandwidth of up to 1.2 GB/sec, connects the PERC 5/E to the PowerVault MD1000. The PERC 5/E settings can be changed with the Dell OpenManage™ Server Administrator application or from its BIOS, which is accessed by pressing Ctrl + R during system startup.

The setup used one enclosure with 14 disks; however, for achieving higher performance than what was demonstrated in this environment, up to three enclosures can be daisy-chained, which allows a server to connect to a maximum of 45 disks through one port when all enclosures are fully populated. For maximum storage capacity, both of the ports on the PERC 5/E can be used to connect the server to two sets of daisy-chained PowerVault MD1000

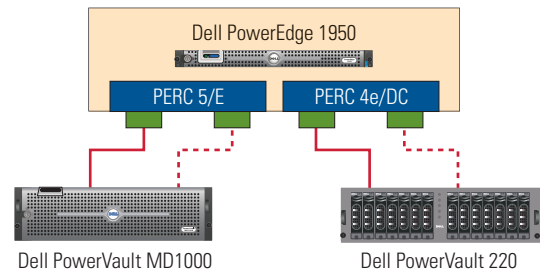


Figure 1. Test environment for comparing SAS and SCSI storage

enclosures containing 45 disks each, which allows a maximum of 90 SAS disks connected to one server.

PowerVault 220 configuration: External SCSI storage

Similar to the PowerVault MD1000, the PowerVault 220 is an external disk storage enclosure that uses SCSI disks; it is managed by a PERC 4e/DC adapter that supports the same RAID levels as the PERC 5/E. The PowerVault 220 can host a maximum of 14 disks.

In the setup, both storage enclosures contained fourteen 73 GB, 15,000 rpm hard drives. The PowerVault MD1000 was deployed with fourteen 73 GB Fujitsu SAS disks, and the PowerVault 220 with fourteen 73 GB Seagate Cheetah disks.

Both the PERC 4e/DC and PERC 5/E are dual-port controllers. The Dell HPC team ran I/O benchmarks on two configurations: single channel (one port on the controller connected to the storage) and dual channel (both ports on the controller connected to the storage). For the PERC 5/E, the results include only a single channel because no significant performance difference was observed between single-channel and dual-channel configurations for the PERC 5/E using either the IOzone benchmark or OOCORE application. Other applications with different data-access patterns may benefit from the two-channel configuration with the PowerVault MD1000.

Performance tuning and benchmark study

The Dell HPC team used the IOzone benchmark and OOCORE application to compare SAS and SCSI performance. Before doing so, however, the team conducted initial performance tuning to determine the appropriate settings for the test environment.

Performance study with the Linux® OS read-ahead cache showed that this setting affects the disk storage performance significantly, especially for streaming reads on fast storage devices such as SCSI, SAS, and external disk storage. When a block is accessed, the read-ahead algorithm prefetches and caches the subsequent blocks for enhanced performance. The algorithm turns itself off when it finds that the data is being accessed randomly.

The highest performance was achieved with a 4 MB read-ahead cache; therefore, for all the benchmark tests, the cache was set to 4 MB (8,192 blocks of 512-byte sectors) using the command

`blockdev -setra 8192 /dev/device` or, alternatively, `hdparm -a 8192 /dev/device`. In the Linux 2.6 kernel, this value is stored under `/sys/block/device/queue/read_ahead_kb` in kilobytes.

IOzone benchmark study

The IOzone benchmark² was used to measure the performance of both storage enclosures for sequential write and sequential read operations. The test file size was set to 12 GB to eliminate the cache effect. Figure 2 shows the performance of different configurations relative to the PERC 4e/DC single-channel RAID-0 write performance.

For sequential write and read operations, SAS storage provided much higher performance than SCSI storage with the same number of similar hard drives. Read operations in particular performed two to three times faster on SAS storage than on SCSI storage.

OOCORE application study

The OOCORE application,³ an out-of-core matrix solver, was used for application study. *Out-of-core matrix solving* refers to solving the matrices that do not fit into the CPU memory. The OOCORE application uses the disk to store the matrix and temporary files generated during the process of solving the matrix, which makes the benchmark I/O-intensive. The configuration parameters were adjusted to increase the number of disk I/O operations, and the amount of RAM was reduced to help ensure that the benchmark performance reflected the storage performance.

Figure 3 shows the performance results relative to the PERC 4e/DC single-channel RAID-0 configuration performance. The results show that SAS storage outperformed SCSI storage in all test

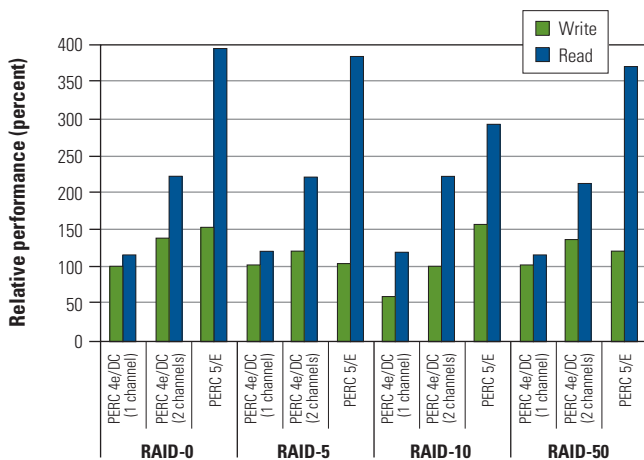


Figure 2. IOzone sequential-access performance: SAS versus SCSI storage

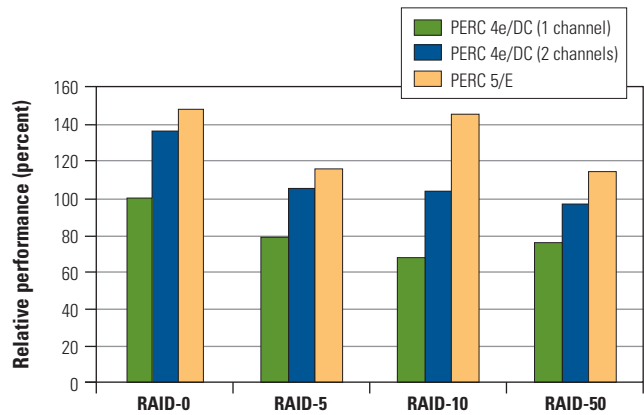


Figure 3. OOCORE performance: SAS versus SCSI storage

scenarios. For example, in a RAID-5 configuration, the SAS storage performed 47 percent better than the one-channel PERC 4e/DC SCSI storage and 14 percent better than the two-channel PERC 4e/DC SCSI storage.

Enhanced storage performance

SAS technology with serial architecture can help provide high performance for HPC applications that use storage extensively. The IOzone benchmark and OOCORE application performance results described in this article show that SAS storage consistently provided better throughput than SCSI storage in similar configurations. Performance can also be enhanced by taking advantage of the large number of disks supported with SAS by connecting additional enclosures. [↪](#)

Aziz Gulbeden is a systems engineer in the Scalable Systems Group at Dell. His current areas of focus include scalable high-performance file and storage systems. He has a B.S. in Computer Engineering from Bilkent University in Turkey and an M.S. in Computer Science from the University of California at Santa Barbara.

Amina Saify is a member of the Scalable Systems Group at Dell. Amina has a bachelor's degree in Computer Science from Devi Ahilya University in India and a master's degree in Computer and Information Science from the Ohio State University.

Andrew Bachler is a systems engineer in the Scalable Systems Group at Dell. He has an associate's degree in Electronic Engineering and 12 years of UNIX® and Linux OS experience.

Ramesh Radhakrishnan, Ph.D., is a member of the Scalable Systems Group at Dell. His interests include performance analysis and characterization of enterprise-level benchmarks. Ramesh has a Ph.D. in Computer Engineering from the University of Texas at Austin.

² The IOzone benchmark can be downloaded from www.iozone.org.

³ The OOCORE application can be downloaded from www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf0605.