

Designing High-Performance Clusters

with the Dell PowerEdge SC1425 Server

Hardware selection for high-performance computing clusters is often driven by the characteristics of the parallel applications that will be deployed. This article discusses different classes of parallel applications and presents the Dell™ PowerEdge™ SC1425 server as a viable, low-cost platform on which to build clusters for different classes of applications.

BY RON PEPPER AND RINKU GUPTA

An important consideration when designing a high-performance computing (HPC) cluster is the characteristics of the parallel application that will run on the cluster. Application characteristics go a long way in determining the components needed for the cluster. For example, a long-running parallel application that exchanges many small messages between nodes in the cluster may require a special network communications infrastructure. In this case, the cluster design should specify that the compute servers be connected to each other by a fast network interconnect that can send and receive many small messages very quickly.

Defining the types of HPC applications

A parallel application runs on a distributed collection of nodes. Hence, all applications consist of communication steps (to communicate between themselves) and computation steps (to perform independent computation). Based on the degree of communication and computation, parallel applications in the HPC field fall into three broad categories.

Coarse-grained parallel applications. Beowulf clusters—that is, parallel-processing HPC clusters comprising industry-standard components—have traditionally been built to exploit coarse-grained parallel applications. Parallel applications for which the overall time spent in computation is much higher than the time spent in communication fall into this category. A coarse-grained application is an ideal candidate for running Beowulf

clusters, because there is a great probability of obtaining higher performance at a lower price when additional servers or CPUs are added to the cluster. Because of their highly parallel nature, these applications are also called embarrassingly parallel applications. A classic example is the Monte Carlo simulation problem.

Medium-grained parallel applications. Applications for which the computation time is greater (but not much greater) than the communication time are considered medium-grained parallel applications. As compared to coarse-grained parallel applications, medium-grained applications have a much lower computation-to-communication ratio. The communication overhead is apparent for medium-grained parallel applications.

Fine-grained parallel applications. In a fine-grained application, the time intervals spent on communication tasks are greater than or equal to the time spent on computation tasks. Such an application is not considered the best candidate for Beowulf clusters. Even if the application is parallelized and run on a cluster of nodes, the scalability of the application may be limited.

Choosing the right cluster components

Components in a cluster are generally selected based on the generic applications that will be run on them. Every part of the cluster plays an important role in the cluster environment, depending upon the specific application's ability to exploit it. Two of the most important components

in a cluster are server type and network interconnect. The servers in a cluster provide the computational power and are connected through the interconnect fabric.

For coarse-grained parallel applications, communication is generally less of a concern. Hence, organizations running a high-performing, coarse-grained parallel application can select high-performance servers, which offer good computation power. Gigabit networking can meet the interconnect requirements of coarse-grained parallel applications while maintaining a low price point. Using a traditional low-cost interconnect with a high-performance server can provide a low-cost HPC solution for these applications.

For medium-grained parallel applications, the communication overhead can be high. For such applications, organizations should consider a high-performance interconnect like Myricom Myrinet¹ or InfiniBand. Both interconnects offer low latency and high bandwidth. *Latency* is the amount of time taken to transmit a message from the source server to the destination server. *Bandwidth* is the rate at which information can be sent over the interconnect. A slow interconnect with a fast processing system will cause the interconnect to become a bottleneck for the application.

For fine-grained parallel applications running on Beowulf clusters, a fast interconnect like Myrinet or InfiniBand is recommended. Using a slower interconnect could cause the communication time to overshadow the computation time, rendering the parallelization of the application unsuccessful.

Server types and interconnects are two high-level components of a cluster. Choosing the most appropriate server is in itself a broad topic with many components such as memory subsystem, processor speed, and cache sizes to be considered. Dell supports a wide range of rack-mount servers—including the PowerEdge 1850, PowerEdge 1855, PowerEdge 2850, and PowerEdge 3250 servers—that are suitable for HPC clusters and offer varied architectures to satisfy computational needs.

For coarse-grained applications, choosing the appropriate industry-standard components can enable organizations to create a low-cost cluster that will still meet their application needs. A recently released server from Dell, the PowerEdge SC1425, is one such component that can help provide a viable low-cost alternative for HPC needs.

Testing the Dell PowerEdge SC1425 server as a cluster node

In October 2004, a team of Dell engineers tested cluster performance using a PowerEdge SC1425 server as the building block for a low-cost cluster. For this study, each PowerEdge SC1425 server was configured with two Intel® Xeon™ processors. However, the PowerEdge SC1425 can be configured with either single or dual CPUs and can run in either 32-bit mode or Intel Extended Memory 64 Technology (EM64T) mode

for 64-bit applications. For memory subsystem needs, the PowerEdge SC1425 supports dual-banked double data rate 2 (DDR2) memory on an 800 MHz frontside bus (FSB). It can be equipped with either serial ATA (SATA) or SCSI hard drives. Because most of the data in this cluster performance test was shared from the master node, a local SCSI hard drive on each compute node was not necessary. Two embedded Gigabit² Ethernet controllers were included in the base system, which eliminated the need for any additional network hardware.

The following sections demonstrate the performance of the PowerEdge SC1425 cluster for different communication patterns. Using various well-known benchmarks—Pallas PingPong, NAS Parallel Benchmarks, IOzone, and STREAM—the team conducted tests on a single node, two nodes, and the entire cluster to profile the PowerEdge SC1425 for cluster performance. The lab setup consisted of four Dell PowerEdge SC1425 compute server nodes, each with 2 GB of main memory and dual symmetric multiprocessing (SMP) processors at 3.6 GHz. Two types of interconnect were examined: Gigabit Ethernet and Myrinet. The Dell PowerConnect™ 5324 switch was used to form the Gigabit Ethernet fabric, while Myricom Myrinet switches formed the Myrinet fabric.

The impact of interconnects on cluster performance

Using the Pallas³ PingPong benchmark, the Dell test team examined latency at various message sizes for two PowerEdge SC1425 cluster nodes. The two nodes—one using Gigabit Ethernet and the other using Myrinet—sent data to each other via the Message Passing Interface (MPI). As shown in Figure 1, the latency for Myrinet with small messages of 4 bytes was about 6.5 microseconds (μ s) compared to the latency of 35 μ s for 4-byte messages using Gigabit Ethernet.

Figure 2 shows the MPI message throughput between these two cluster nodes at various message sizes. In this study, Gigabit

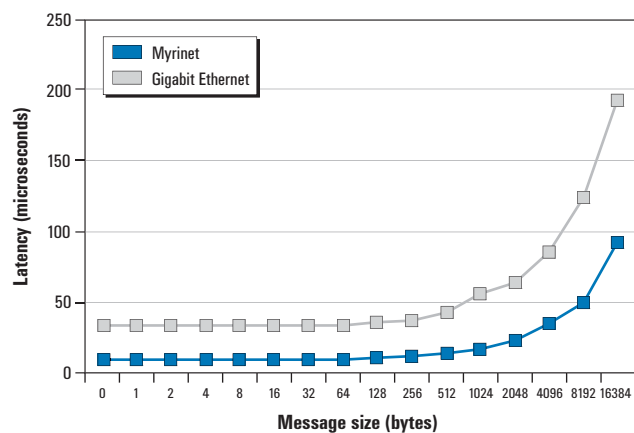


Figure 1. Latency between two PowerEdge SC1425 cluster nodes using different interconnects

¹For more information about Myrinet, see www.myri.com.

²This term does not connote an actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

³For more information about Pallas MPI benchmarks, see www.pallas.com/e/products/pmb/index.htm.

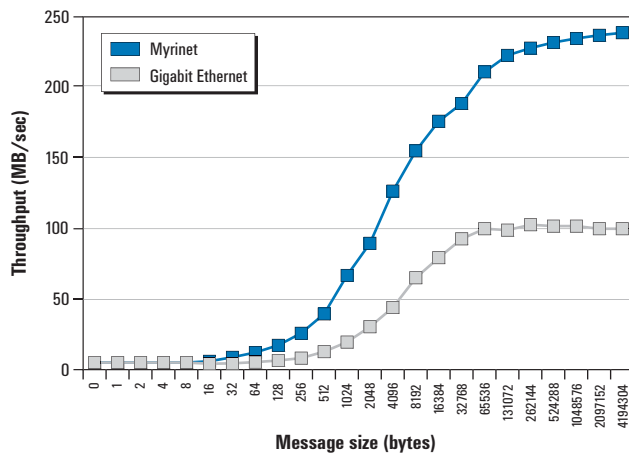


Figure 2. Throughput between two PowerEdge SC1425 cluster nodes using different interconnects

Ethernet scaled up to 95 percent of the theoretical bandwidth, offering close to 100 MB/sec. Myrinet, in contrast, demonstrated bandwidth of nearly 235 MB/sec (close to its theoretical peak of 250 MB/sec) for message sizes of up to 4 GB.

For coarse-grained applications, it is usually sufficient to have a Gigabit Ethernet communication fabric. The on-board Gigabit Ethernet network interface cards (NICs) built into each PowerEdge SC1425 server can be used for connecting the systems together. Gigabit Ethernet provides a good price/performance ratio for coarse-grained applications.

For medium- and fine-grained applications, high-speed, low-latency interconnects like Myrinet can be instrumental in improving the overall performance of an application. An example of this can be seen in Figure 3, which shows relative performance results from the NAS Parallel Benchmarks (NPB) suite.⁴ The NPB suite is a set of eight programs derived from computational fluid dynamics code, and results are measured in millions of operations per second (MOPS). Each of the eight programs represents a particular function of highly parallel computation for aerophysics applications. NPB measures overall cluster performance, so the Dell team conducted the NPB tests on the entire four-node PowerEdge SC1425 cluster. Figure 3 shows the results of three NPB benchmarks on the four-node cluster with one instance on each node. The Embarrassingly Parallel (EP) benchmark from the NPB suite falls into the category of extreme coarse-grained application. The EP test generates pairs of Gaussian random deviates according to a specific scheme. Because EP does not perform any interprocessor communication, the results obtained in this study using different interconnects show the same performance characteristics—thereby supporting the assertion that clusters running applications similar to the EP benchmark suffice with a Gigabit Ethernet interconnect.

The Integer Sort (IS) benchmark from the NPB suite tests both integer computation speed and communication performance. It is a parallel integer sort program that is used in particle method codes. The IS benchmark involves no floating-point arithmetic but does have intense data communication. As shown in Figure 3, the type of interconnect used can significantly affect the performance of Integer Sort. In this study, the IS benchmark performed 1.75 times better on the Myrinet interconnect when compared to Gigabit Ethernet for a four-node cluster (with one instance). Hence, for such medium- and coarse-grained applications, low-cost clusters can benefit from Myrinet or InfiniBand interconnects. The same trend can be seen in the results of other NPB benchmarks like the Conjugate Gradient (CG) benchmark used in this study, whose performance increased by almost 50 percent (on a four-node cluster) when a Myrinet interconnect was used.

Disk I/O performance

Figures 4 and 5 show the relative performance of SCSI and SATA drives supported by the PowerEdge SC1425 server. The Dell test team generated this data using the IOzone benchmark⁵ with a file size of 6 GB. IOzone measures the I/O performance of a single server, so the test was conducted on only one PowerEdge SC1425 cluster node.

Figure 4 shows the read performance with varying record sizes for a 6 GB file. In this study, the SCSI reads showed about 12 to 15 percent better performance as compared to the SATA reads. Figure 5 shows the write performance for varying record sizes. The SCSI writes showed more than 40 percent improvement as compared to the SATA writes.

The compute nodes in a cluster typically mount and use shared storage or storage located on the master node. When compute nodes do not use the local hard drives, it may be sufficient to use SATA drives, depending on specific application needs, and thus achieve a good price/performance ratio. However, if the PowerEdge SC1425

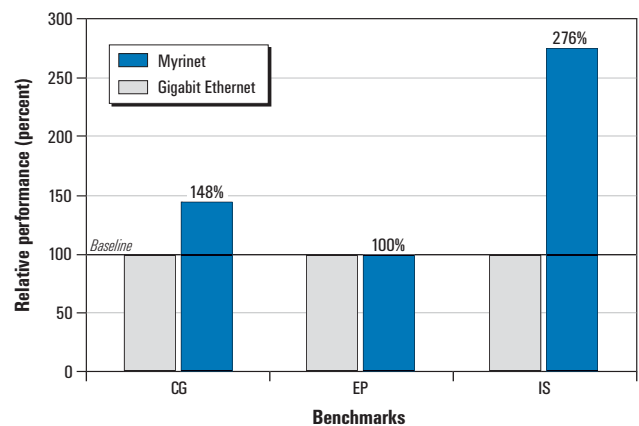


Figure 3. Results of the CG, EP, and IS benchmarks from the NPB suite for four nodes with one instance each

⁴For more information about the NAS Parallel Benchmark suite, see www.nas.nasa.gov/Software/NPB.

⁵For more information about IOzone, see www.iozone.org.

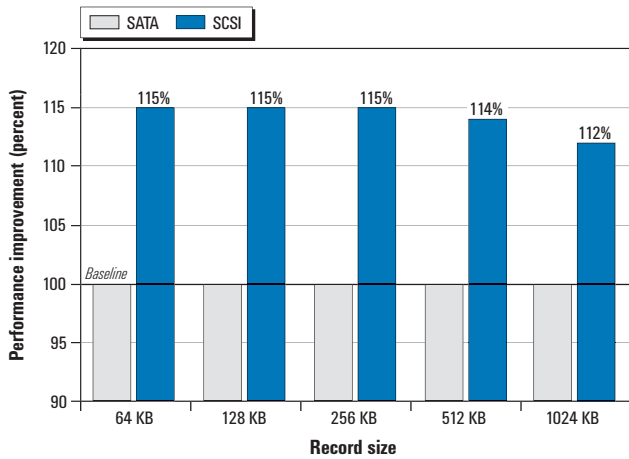


Figure 4. Comparison between SCSI and SATA drives: reads with IOzone benchmark

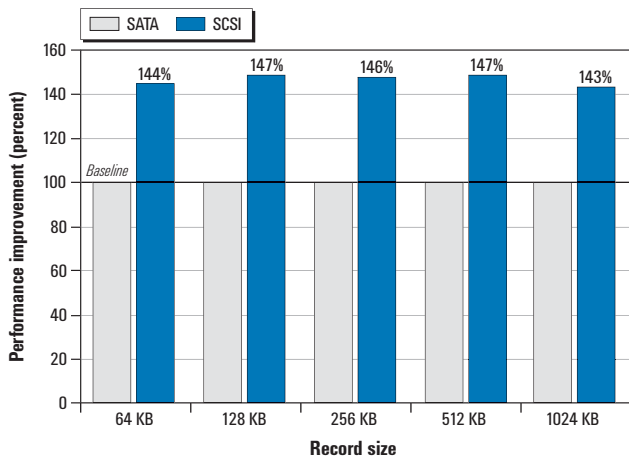


Figure 5. Comparison between SCSI and SATA drives: writes with IOzone benchmark

server is part of an I/O cluster in which each node uses its hard drives for local storage, then using SCSI drives can provide better performance than SATA drives.

Memory subsystem performance

Using the STREAM benchmark,⁶ the Dell test team compared the memory subsystem performance of the PowerEdge SC1425 server with that of a PowerEdge 1750 server. STREAM is a synthetic benchmark that measures memory bandwidth in megabytes per second (MB/sec) and can be useful in understanding the speed at which the compute nodes can process large data sets that are too big to remain in the CPU cache. Memory bandwidth is defined as the amount of memory traffic that a computer system can move from memory to CPU.

For this memory subsystem performance study, the PowerEdge SC1425 was configured with an 800 MHz FSB and dual Intel Xeon processors at 3.6 GHz with 1 MB of level 2 (L2) cache; the PowerEdge 1750 was configured with a 533 MHz FSB and dual Intel Xeon processors at 3.2 GHz with 512 KB of L2 cache and 2 MB of level 3 (L3)

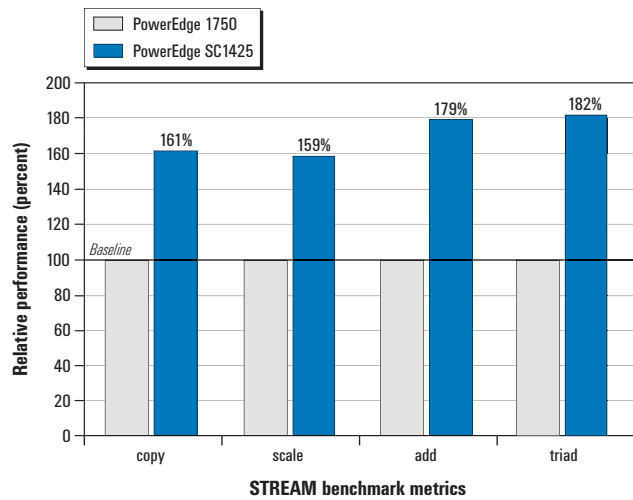


Figure 6. Comparison of memory subsystem performance in a PowerEdge 1750 server and a PowerEdge SC1425 server

cache. The PowerEdge SC1425 used 2 GB of 400 MHz error-correcting code (ECC) DDR2 memory; the PowerEdge 1750 used 4 GB of 266 MHz ECC DDR memory.

Figure 6 shows that the PowerEdge SC1425 server demonstrated high memory bandwidth—achieving up to 82 percent more bandwidth than the PowerEdge 1750 server in this study. The 82 percent increase can be attributed mainly to the high-speed 800 MHz FSB and the 400 MHz DDR2 memory of the PowerEdge SC1425 as compared to the 533 MHz FSB and 266 MHz DDR memory of the PowerEdge 1750.

Deploying HPC clusters based on the PowerEdge SC1425 server

The PowerEdge SC1425 server can be an excellent choice for organizations that plan to deploy coarse-grained parallel applications in high-performance cluster environments. Dell supports standard configuration bundles with the PowerEdge SC1425 server, with node counts ranging from 8 to 256 nodes—and offers custom solution consulting services for larger node counts. The standard bundles support Gigabit Ethernet, Myrinet, and InfiniBand interconnects. PowerEdge SC1425 Gigabit Ethernet clusters can be optimal for coarse-grained applications that do not require a high-speed interconnect. In addition, the PowerEdge SC1425 server can be combined with high-speed interconnects to create a cluster suitable for a range of HPC application categories. [↗](#)

Ron Pepper is a systems engineer and advisor in the Scalable Systems Group at Dell. He works on the Dell HPC Cluster team developing grid environments. Ron attended the University of Madison at Wisconsin, where he worked on a degree in computer science; he is continuing his degree at Saint Edwards University.

Rinku Gupta is a systems engineer and advisor in the Scalable Systems Group at Dell. Her current research interests are middleware libraries, parallel processing, performance, and interconnect benchmarking. Rinku has a B.E. in Computer Engineering from Mumbai University in India and an M.S. in Computer Information Science from The Ohio State University.

⁶For more information about STREAM, see www.cs.virginia.edu/stream.