

# Using PCI Express Technology

## in High-Performance Computing Clusters

Peripheral Component Interconnect (PCI) Express is a scalable, standards-based, high-bandwidth I/O interconnect technology. Dell™ HPC clusters use PCI Express–based products including InfiniBand interconnects and Dell PowerEdge™ Expandable RAID Controller (PERC) storage. To demonstrate the benefits of PCI Express, a team of Dell engineers compared the performance of a Dell HPC cluster using PCI Express–based InfiniBand and PERC components with a cluster using previous-generation PCI Extended (PCI-X) technology.

BY RINKU GUPTA; SAEED IQBAL, PH.D.; AND ANDREW BACHLER

### Related Categories:

Benchmarks

Clustering

Dell PowerEdge servers

High-performance computing (HPC)

InfiniBand

Interconnects

PCI Express

Performance

Visit [www.dell.com/powersolutions](http://www.dell.com/powersolutions) for the complete category index.

High-performance computing (HPC) clusters comprising industry-standard servers, storage, and interconnect components are designed to provide high performance at a relatively low cost compared to traditional monolithic supercomputers. As technology advances, IT organizations have more options regarding the choice of the most suitable cluster components. For example, fast interconnects such as InfiniBand and Myricom Myrinet are now available for HPC clusters. Server technology is transitioning to dual-core processors, and storage technology is moving toward Serial ATA (SATA) and Serial Attached SCSI (SAS). SAS is designed to replace SCSI parallel cable with a much smaller form factor; it uses point-to-point connectivity that promises many advantages over its parallel predecessor (SCSI). Along with these developments, Peripheral Component Interconnect (PCI) Express technology has been created to provide higher I/O bus throughput in servers than previous-generation PCI technologies.

The use of PCI Express technology can affect the choice of other components within an HPC cluster. In particular, using PCI Express with a cluster interconnect

such as InfiniBand and with storage-based RAID cards such as the Dell PowerEdge Expandable RAID Controller (PERC) can offer certain advantages.

### The history of PCI, PCI-X, and PCI Express

In the 1980s, a typical PC used Instruction Set Architecture (ISA), which was a 16-bit wide I/O bus that operated at 4.77 MHz and had a bandwidth of 9 MB/sec. Subsequently, ISA was enhanced and new implementation technology was developed—for example, the Extended ISA (EISA) was 32 bits wide and operated at 8 MHz, and the VESA local bus (VL-bus) was 32 bits wide. A major issue that limited scalability in these I/O buses was the possibility of interference among communicating devices when more than two devices were connected simultaneously.

Developed by Intel in 1992, the PCI bus standard combined and improved upon the features of ISA and VL-bus. PCI introduced a bridge between the I/O devices and the CPU via the frontside bus (FSB), enabling higher scalability than had been possible using previous I/O technology. With these innovations, up to five devices could be connected to the PCI bus. The PCI bus was

32 bits wide, operated at 33 MHz, and had a bandwidth of 132 MB/sec. Since its introduction, PCI has been universally accepted as the standard I/O interconnect in servers. In addition, several design and technology improvements have increased the performance of PCI to 64 bits at 133 MHz in the PCI Extended (PCI-X) specification, which has a maximum bandwidth of 1 GB/sec.<sup>1</sup>

Because the upper limit of PCI-X bandwidth is 1 GB/sec, further improvement in performance is not economical. Several computer manufacturers have proposed new bus technologies in response to the numerous requirements of I/O devices in use today. In general, the trend is to move away from shared-bus technology toward point-to-point (direct connection) technology among communicating devices.

### PCI Express architecture

Today, microprocessors and I/O devices—such as 10 Gigabit Ethernet, 10 Gbps Fibre Channel, InfiniBand, SATA, and SAS—demand more bandwidth than PCI-X can provide. PCI-X is an extension of PCI, and the basic architectures of PCI and PCI-X are very similar (parallel shared buses). In the PCI and PCI-X architectures, I/O devices are connected to the memory controller through an I/O bridge. The I/O bridge limits scalability because the bus is shared among all connected devices. In addition, only one I/O device can be connected in a point-to-point configuration to the PCI-X 2.0 bus.

The PCI Express architecture is very different from the PCI-X architecture. PCI Express is based on high-speed point-to-point technology, which uses serial interfaces to connect devices. The point-to-point architecture improves scalability by allowing multiple lanes of data at each PCI Express slot. The PCI Express architecture has a host bridge with several end points. Devices are connected to these end points, and traffic is routed through the host bridge. To add more devices, IT administrators can add a switch to the end points in the host bridge. Several devices can then be connected to the switch. Traffic can be routed through the switch from one device to another in a peer-to-peer configuration without going through the host bridge.

The PCI Express architecture is designed to improve performance substantially by directly connecting I/O devices to a memory controller via PCI Express links. Each PCI Express link can have multiple lanes, with each lane capable of 250 MB/sec of bidirectional bandwidth. Thus, an 8x PCI Express channel (8 lanes per link) can achieve 2 GB/sec in each direction. The basic PCI Express slot is 1x, which means one lane carries data. Different slots may be 2x, 4x, 8x, or 16x, depending on how many lanes are designed in them.

PCI Express has a layered architecture, which enhances scalability and eases backward compatibility. For example, PCI Express is compatible with PCI, and components based on previous versions of PCI can be used in PCI Express slots. Similarly, silicon technology can be improved without altering other layers, or a technology such as Fibre Channel can be used to implement the physical layer.

The PCI Express architecture consists of five layers. The lowest layer is the physical layer, which offers pairs of low-voltage differential, high-bandwidth dedicated serial channels. Each channel is dual-simplex—that is, it can transmit and receive signals simultaneously.<sup>2</sup> The link layer is above the physical layer and adds packet sequencing and error detection to help ensure data integrity during transmission. Above the link layer is the transaction layer, which receives read and write requests from the layers above it and creates packets according to these requests; the transaction layer supports 32-bit and 64-bit addressing. The transaction layer also offers several standard packet-switching features. On top of the transaction layer is the software layer, which is primarily used for the driver software. The software layer generates read and write requests to I/O devices. The topmost layer is the configuration/OS layer, which uses the PCI plug-and-play specification to initialize, enumerate, and configure any connected device. It does this with cooperation from the software layer below it; the result is robust device initialization and system configuration. System architects can choose from various mechanical form factors currently available for PCI Express connectors.<sup>3</sup>

PCI Express offers several advanced features, but the most important are advanced power management, support for real-time traffic, hot swapping and hot plugging, data integrity, and error handling. PCI Express can adjust power consumption when the bus is not in use to save power. Some multimedia devices require guaranteed processing in real time, which is supported under PCI Express by implementing virtual channels. PCI Express has native support for hot plugging and hot swapping of I/O devices, which can help minimize required server downtime.

PCI Express has a layered architecture, which enhances scalability and eases backward compatibility. For example, PCI Express is compatible with PCI, and components based on previous versions of PCI can be used in PCI Express slots.

<sup>1</sup> For more information about PCI Express, visit [www.intel.com/technology/pciexpress/devnet](http://www.intel.com/technology/pciexpress/devnet).

<sup>2</sup> For more information, visit [www.express-lane.org](http://www.express-lane.org).

<sup>3</sup> For more information, visit [www.PCI-sig.org](http://www.PCI-sig.org).

## PCI Express and cluster interconnects in Dell HPC clusters

Interconnects play an important role in clusters because they connect industry-standard computing components. To help communication-intensive applications achieve high performance on clusters, various high-speed interconnects have been developed over the last few years. While many of these interconnects are proprietary (such as Myrinet and Quadrics QsNet), there has been an initiative to develop a low-latency, high-bandwidth interconnect based on industry standards. InfiniBand is the result of such an initiative led by the InfiniBand Trade Association (IBTA),<sup>4</sup> a consortium of hardware and software vendors.

The InfiniBand architecture is a point-to-point, switched fabric architecture that connects various end points, where each end point can be a storage controller, a network interface card (NIC), or an interface to a host system. A host channel adapter (HCA) provides the host interface. This HCA has traditionally been connected to the host processor through a standard PCI or PCI-X bus. Because most Dell servers now support the PCI Express interface, this bus has become the primary choice for connecting HCAs to the host processor.

InfiniBand architecture defines different link speeds: 1X, 4X, and 12X. These link speeds are designed to yield data rates of 2.5 Gbps, 10 Gbps, and 30 Gbps, respectively. At the physical layer, InfiniBand uses 8B/10B encoding. In September 2004, the IBTA completed the InfiniBand 1.2 specification, which specifies double data rate (DDR) and quad data rate (QDR) modes of operation. These modes define increased signaling rates over existing 1X, 4X, and 12X InfiniBand links and are designed to effectively double or quadruple bandwidth. While current 4X InfiniBand HCAs have specified speeds of 10 Gbps, these speeds cannot be achieved via the current PCI-X bus. PCI-X buses, which have a maximum bidirectional throughput of 1 GB/sec, can act as a major bottleneck and limit the performance achievable by the InfiniBand 4X cards. The PCI Express bus, in contrast, is designed to eliminate this bottleneck, helping achieve the full bandwidth potential of InfiniBand cards. Dell partners with Topspin Communications, Inc., to incorporate InfiniBand hardware and software into Dell HPC clusters.

Dell HPC clusters can range from 8 to 256 nodes and include PCI Express-based components. Servers such as the Dell PowerEdge 1850 and PowerEdge 1855 support PCI Express. The PowerEdge 1855 can accommodate up to 10 server blades and includes an InfiniBand daughtercard; the PowerEdge 1850 uses a PCI Express slot-based HCA. For further information, refer to [www.dell.com/hpcc](http://www.dell.com/hpcc).

## InfiniBand performance analysis

In November 2004, a team of engineers from the Scalable Systems Group at Dell tested Dell HPC clusters to demonstrate the

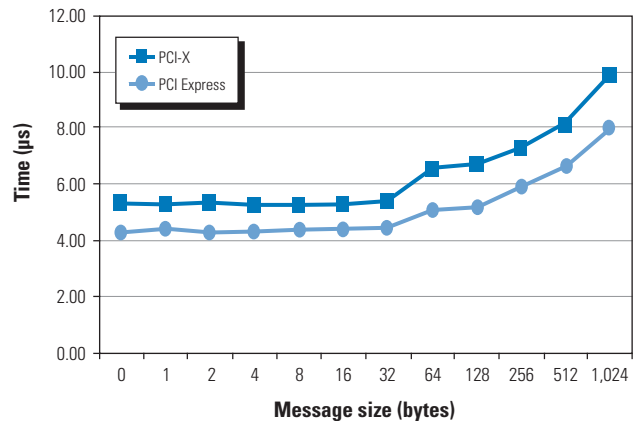


Figure 1. Pallas Ping-Pong latency test results for small message sizes

advantages of PCI Express-based InfiniBand HCAs as compared to PCI-X-based HCAs. The test environment comprised a cluster of 16 identically configured Dell PowerEdge 1850 servers running the Red Hat® Enterprise Linux® 3 OS, interconnected with InfiniBand HCAs and switches. Each PowerEdge 1850 server had two Intel® Xeon™ processors running at 3.2 GHz and 4 GB of RAM operating on an 800 MHz FSB.

For the PCI-X systems, each server was equipped with a PCI-X riser consisting of two PCI-X slots operating at 100 MHz and 133 MHz. The InfiniBand card was inserted in the 133 MHz slot. For the PCI Express systems, each server used a PCI Express riser, which had 4x and 8x PCI Express slots. The InfiniBand card was inserted into the 8x slot. The PCI-X and PCI Express InfiniBand components used in the study were obtained from Dell partner Topspin Communications. Each PCI-X and PCI Express HCA had dual 4x ports.

The Dell test team used two popular benchmarks: the Pallas Message Passing Interface (MPI) Benchmarks (PMB) 2.2.1 and the NASA Advanced Supercomputing (NAS) Parallel Benchmarks (NPB) 2.4.

## Testing performance with Pallas benchmarks

To determine the advantages of PCI Express-based InfiniBand over PCI-X-based InfiniBand, the test team used the Pallas Ping-Pong test, which provides point-to-point bandwidth and latency measurements between two nodes; and the Pallas Send-Receive test, which provides bidirectional bandwidth between two nodes. Figures 1 and 2 show the point-to-point link performance results from these tests. Figure 1 shows that the latency for small messages for the PCI Express-based InfiniBand was approximately 4.3 microseconds ( $\mu$ s) as compared to the 5.32  $\mu$ s obtained for PCI-X-based InfiniBand. Figure 2 shows that the bandwidth for

<sup>4</sup>For more information about InfiniBand and the IBTA, visit [www.infinibandta.org](http://www.infinibandta.org).

PCI-X-based InfiniBand peaked at approximately 700 MB/sec, while PCI Express-based InfiniBand achieved up to 915 MB/sec for large message sizes.

In the Pallas Send-Receive benchmark, the nodes in a group first send a message to the node on the right and receive a message from the node on the left. The total number of messages with respect to each node is two: one sending and one receiving. The benchmark is based on the MPI\_SendRecv primitive implementation in MPI; the Pallas Send-Receive test between two nodes serves as a bandwidth test and consists of two nodes sending and receiving messages from each other simultaneously. As shown in Figure 3, for a PCI-X-based InfiniBand HCA, peak bidirectional bandwidth was limited to approximately 813 MB/sec; for PCI Express, this bandwidth scaled up to approximately 1,763 MB/sec. These test results indicate that PCI Express-based InfiniBand can leverage the additional bandwidth made available by PCI Express buses and can help provide a major performance boost to applications.

### Testing performance with the NPB suite

The NPB suite consists of eight programs derived from computational fluid dynamics (CFD) code. These programs measure overall cluster performance, and results are measured in millions of operations per second (MOPS). Different classes of these programs represent different problem sizes. For the Dell tests, the team used Class C, which represents the largest problem size. Figure 4 shows the percentage improvement shown by PCI Express-based HCAs as compared to the PCI-X-based HCAs (which served as a baseline) for six of the programs in the NPB suite. The test team used six of the eight NPB programs: the kernel programs FT (Fast Fourier Transform), MG (Multigrid), CG (Conjugate Gradient), and IS (Integer Sort) emulate the computational core of different numerical methods used by CFD applications; BT (Block Tridiagonal) and SP (Scalar Pentadiagonal) are simulated CFD applications.

The IS benchmark from the NPB suite tests both integer computation speed and communication performance. It is a parallel program that is used in particle method codes. The IS benchmark involves no floating-point arithmetic but does have intense data communication. When run on PCI Express-based InfiniBand, the IS program achieved nearly a 12 percent performance improvement compared to PCI-X. As indicated in Figure 4, other benchmark programs show improvement depending upon the amount of communication that takes place within them. Applications with less communication can suffice with PCI-X-based InfiniBand cards or possibly a slower interconnect such as Gigabit Ethernet.<sup>5</sup>

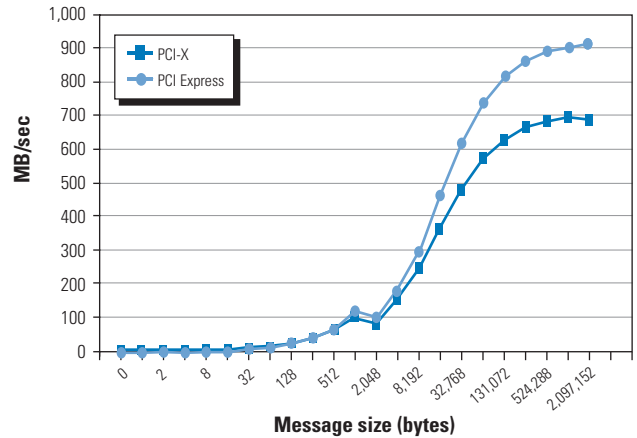


Figure 2. Pallas Ping-Pong bandwidth test results

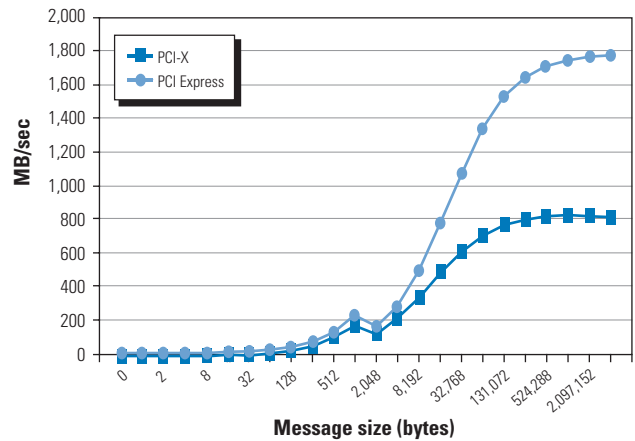


Figure 3. Pallas Send-Receive test results between two nodes

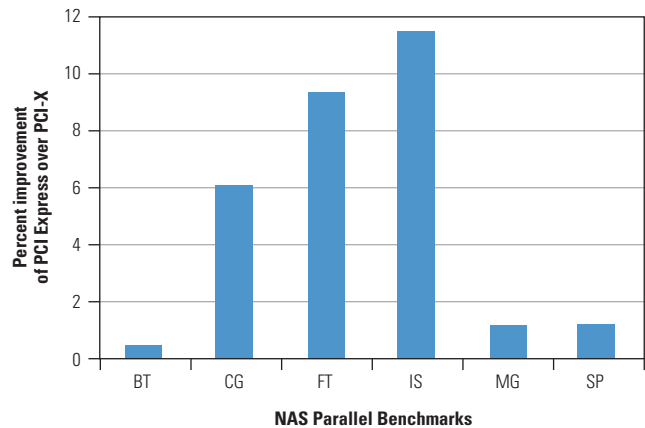


Figure 4. Relative percentage improvement in NAS parallel benchmarks of PCI Express versus baseline PCI-X-based InfiniBand HCAs

<sup>5</sup> This 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.

## PCI Express and storage in Dell HPC clusters

HPC clusters are commonly built with low-cost SCSI RAID storage. In the Dell tests, the team also compared a PCI-X PERC 4, Dual Channel (PERC 4/DC) SCSI adapter with a new-generation PCI Express PERC 4, Extended Dual Channel (PERC 4 e/DC) SCSI adapter. These two RAID adapters were connected to an external Dell PowerVault™ 220S storage system fully populated with fourteen 73 GB 10,000 rpm drives in RAID-0 and RAID-5 configurations.

The PERC 4/DC and the PERC 4 e/DC each has a 128 MB read/write cache that supports up to 14 drives per channel with a maximum of 28 drives. Both adapters can also handle up to 40 logical drives in RAID-0, RAID-1, RAID-5, RAID-10, and RAID-50 configurations.

To evaluate the performance comparisons, the Dell test team used a public domain benchmark utility called IOzone, which is widely used in testing file system I/O performance for sequential and random read/write processes. The testing was performed on a Dell PowerEdge 2850 server configured with 8 GB of RAM and connected to a Dell PowerVault 220S storage system using a PERC 4/DC. The OS was Red Hat Enterprise Linux AS 4 using Intel Extended Memory

64 Technology (EM64T). The first test compared the PCI-X PERC 4/DC with the PCI Express PERC 4 e/DC utilizing both external channels and all 14 drives in a RAID-0 configuration for a direct baseline comparison. Running IOzone on the server, the test team used a test file size of 12 GB—

one and a half times the available memory, which saturated the 8 GB of total system memory to simulate the storage load of an HPC environment. The PCI Express PERC 4 e/DC achieved on average a performance increase of more than 150 percent for writes and more than 100 percent for reads when compared with the PCI-X PERC 4/DC (see Figure 5).

The second test used the same hardware, but instead of a RAID-0 configuration, a seven-drive RAID-5 configuration was implemented to help ensure redundancy in the event of a drive failure. Again, the team used a 12 GB test file, and the PCI Express system performed significantly better—write performance increased more than 200 percent and read performance increased more than 50 percent—when compared with the PCI-X PERC 4/DC adapter (see Figure 5).

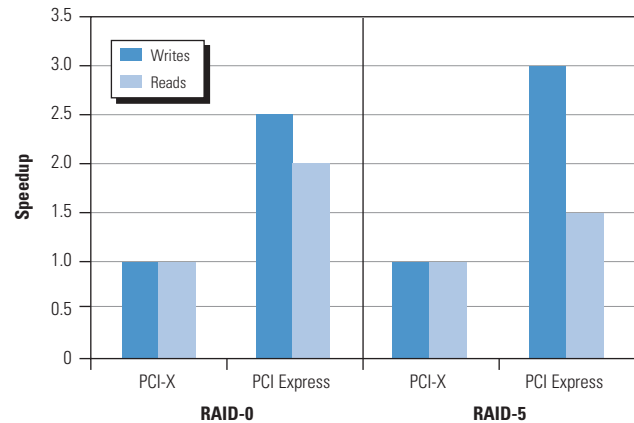



Figure 5. IOzone benchmark results comparing read/write performance for PCI-X and PCI Express in RAID-0 and RAID-5 configurations

The ability to aggregate bandwidth through the use of wide ports and expanders will help provide the performance scalability required by next-generation enterprise servers and storage systems.

## The role of PCI Express in HPC clusters

PCI Express technology has helped achieve the improved performance that has been a feature of interconnects such as InfiniBand and has served as an impetus for the development of communication devices that can take advantage of this improved performance. On the InfiniBand side, the industry is moving toward memory-free PCI Express HCAs; on the storage side, SAS technology is the next-generation storage interface for SCSI. The ability to aggregate bandwidth through the use of wide ports and expanders will help provide the performance scalability required by next-generation enterprise servers and storage systems. 

**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.

**Saeed Iqbal, Ph.D.**, is a systems engineer and advisor in the Scalable Systems Group at Dell. His current work involves evaluation of resource managers and job schedulers used for standards-based clusters. He is also involved in performance analysis and system design of clusters. Saeed has a B.S. in Electrical Engineering and an M.S. in Computer Engineering from the University of Engineering and Technology in Lahore, Pakistan. He has a Ph.D. in Computer Engineering from The University of Texas at Austin.

**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 experience with UNIX® and Linux platforms.