In increases in the number of processors and cores within system architectures have created a constant challenge to balance computational and communications data in environments using powerful multi-core processors. An efficient communications subsystem is critical to help sustain high application performance. InfiniBand—a low-latency, high-bandwidth interconnect widely used in high-performance computing (HPC) cluster environments for Interprocess Communication (IPC)—can provide a key component in these environments.

ConnectX is the fourth-generation InfiniBand architecture from Mellanox Technologies, following the third-generation InfiniHost III architecture, and includes multiple enhancements designed to increase performance in HPC clusters. To help demonstrate these enhancements, Dell engineers ran a variety of benchmarks on an HPC cluster of Dell™ PowerEdge ™ servers using Mellanox ConnectX double data rate (DDR) InfiniBand, InfiniHost III Lx DDR InfiniBand, and Gigabit Ethernet interconnects. The results illustrate the advantages ConnectX can provide in reduced latency and increased bandwidth in HPC cluster environments.

UNDERSTANDING THE INFINIBAND ARCHITECTURE

The designs and specifications for the InfiniBand switched-fabric I/O architecture are created by the InfiniBand Trade Association. The nodes connect to the switch through a host channel adapter (HCA) that plugs into a PCI Express (PCIe) slot. HCAs have programmable direct-memory-access engines that process packets transferred to and from the host.1 The packets are transferred through Remote Direct Memory Access (RDMA), which can help reduce processor overhead. Data transfers directly from host memory to sender memory without using host processor cycles.

Many software stacks and libraries can support InfiniBand hardware; one commonly used unified software stack is the OpenFabrics Enterprise Distribution (OFED), available from the OpenFabrics Alliance. Major InfiniBand vendors and developers, including research and national labs, have collaborated and contributed to the OFED stack. The most recent release, OFED 1.3, includes a comprehensive InfiniBand stack, including drivers, libraries, diagnostic tools and utilities, and message-passing libraries. OFED 1.3 also includes

other advanced features supported by InfiniBand in the kernel and user space, such as Internet SCSI (iSCSI) Extensions for RDMA (iSER), IP over InfiniBand (iPoIB), SCSI RDMA Protocol (SRP), and Sockets Direct Protocol (SDP).²

ConnectX HCAs are 4X dual-port adapters designed to support PCIe 1.1 and 2.0. ConnectX includes enhanced packet processing and offload capabilities compared with the InfiniHost III chipset, which can help utilize additional bandwidth out of the PCIe slot. The advanced packet scheduler for ConnectX is implemented at the hardware level, and does not involve software or firmware components for its RDMA and send and receive operations.

The enhanced ConnectX architecture is designed to support a messaging rate of up to 25 million messages per second, which helps reduce latency and increase bandwidth compared with InfiniHost III Lx DDR InfiniBand. The theoretical peak bidirectional throughput for a DDR InfiniBand HCA is 4 GB/sec. The ConnectX architecture is designed for enhanced performance and bandwidth, providing a greater messaging rate compared with the previous-generation InfiniBand chipset.

**COMPARING HPC INTERCONNECTS**

In July 2008, engineers from the Dell HPC team tested the performance of ConnectX DDR InfiniBand, InfiniHost III Lx DDR InfiniBand, and Gigabit Ethernet interconnects on an HPC cluster. Figure 1 shows the cluster configuration used in the test environment.

![HPC cluster configuration in the test environment](image)

### Server
Four Dell PowerEdge SC1435 servers

### Processors
Two dual-core AMD Opteron™ 2216 processors at 2.4 GHz per server

### Memory
Eight 1 GB DDR2 dual in-line memory modules (DIMMs) at 667 MHz per server

### OS
Red Hat® Enterprise Linux® 5.1 OS running kernel version 2.6.18-53.el5

### InfiniBand HCA
- Mellanox InfiniHost III Lx DDR HCA (MHGS18-XTC) in an x8 PCIe slot
- Mellanox ConnectX DDR HCA (MHGH28-XTC) in an x8 PCIe slot

### InfiniBand switch
24-port Cisco SFS 7000D switch

### Gigabit Ethernet network interface card (NIC)
On-board Broadcom NetXtreme II BCM5708 NIC

### Gigabit Ethernet switch
Dell PowerConnect™ 6248 switch

![Latency for different interconnects with varying message sizes measured using the Ohio State University MPI-level latency test](image)

- Gigabit Ethernet
- InfiniHost III Lx DDR InfiniBand
- ConnectX DDR InfiniBand

The test team used three Ohio State University Message Passing Interface (MPI)-level benchmarks—latency, unidirectional bandwidth, and bidirectional bandwidth—to compare the point-to-point performance of ConnectX DDR InfiniBand with that of InfiniHost III Lx DDR InfiniBand and Gigabit Ethernet.² The team performed these tests with two nodes connected through a single switch. In the latency test, the sender node sends a message of a certain data size to the receiver node. The receiver node receives the message and sends back a reply of the same data size, with latency then calculated as half the round-trip time. As shown in Figure 2, ConnectX DDR InfiniBand had a lower latency than InfiniHost III Lx DDR InfiniBand: the small-message latency for ConnectX was approximately 1.6 microseconds, while the small-message latency for

"ConnectX is a fourth-generation InfiniBand architecture and includes multiple enhancements designed to increase performance in HPC clusters."

²For more information, see “Configuring InfiniBand on HPC Clusters Using the OpenFabrics Enterprise Distribution,” by Munira Hussain, Arun Rajan, Sreeram Vedantham, and Jacob Liberman, in Dell Power Solutions, November 2007, DELL.COM/Downloads/Global/Power/pkkg07-20070551-Hussain.pdf.

²For more information on these benchmarks, visit mvapich.cse.ohio-state.edu/benchmarks.
InfiniHost III Lx was approximately 3.75 microseconds.

The test team next compared the unidirectional and bidirectional bandwidth for ConnectX DDR InfiniBand with those of InfiniHost III Lx DDR InfiniBand and Gigabit Ethernet. In the unidirectional bandwidth test, the sender node sends messages of varying sizes ranging from 1 MB to 4 MB. These messages saturate the interconnect to the receiver node by sending back-to-back messages. As shown in Figure 3, the unidirectional bandwidth for ConnectX was considerably higher than that of InfiniHost III Lx: the maximum unidirectional bandwidth for ConnectX was approximately 1,600 MB/sec, while the maximum unidirectional bandwidth for InfiniHost III Lx was approximately 1,470 MB/sec.

In the bidirectional bandwidth test, both nodes send back-to-back messages to each other simultaneously. As shown in Figure 4, the bidirectional bandwidth was also much higher for ConnectX than for InfiniHost III Lx: the maximum bidirectional bandwidth for ConnectX was approximately 3,050 MB/sec, while the maximum bidirectional bandwidth for InfiniHost III Lx was approximately 2,400 MB/sec. Bidirectional bandwidth results can be tuned and optimized by, for example, changing the message size when moving from the Eager protocol to the Rendezvous protocol from 8 KB to 16 KB, which helps increase the throughput for ConnectX.

The test team also used four other benchmarks—Conjugate Gradient (CG), Fast Fourier Transform (FT), Integer Sort (IS), and Lower-Upper Symmetric Gauss-Seidel (LU)—from the NASA Advanced Supercomputing (NAS) Parallel
Benchmarks (NPB) suite to simulate applications with different message-passing characteristics and to analyze the effect of ConnectX DDR InfiniBand on these applications. The CG, FT, and IS benchmarks use a mix of large and small messages for IPC; the LU benchmark primarily uses small message sizes. The test team compiled the benchmarks for 16 processors and used the Class C problem size, then executed them on the four nodes of the cluster using ConnectX DDR InfiniBand, InfiniHost III Lx DDR InfiniBand, and Gigabit Ethernet. Figure 5 shows the average performance speedup results for InfiniHost III Lx and ConnectX compared against a Gigabit Ethernet baseline. ConnectX provided consistently higher performance than InfiniHost III Lx at various scales, depending on the message-passing characteristics of the specific benchmark.

This performance increase was achieved with four nodes, and will vary with different numbers of nodes, processors, and cores. Actual speedup for a particular application depends on the application’s exact message-passing characteristics. Because ConnectX DDR InfiniBand has extremely low latency and enhanced bandwidth, latency- and bandwidth-intensive applications in particular can benefit from use of this interconnect.

OPTIMIZING INTERCONNECT PERFORMANCE IN HPC CLUSTERS

Because different HPC applications have unique characteristics, patterns, and requirements, generalizing optimal HPC configurations is difficult. For communication-intensive applications, however, low latency and high bandwidth have become two key factors that contribute to overall cluster performance. As the number of processors and cores in HPC systems continues to increase, the next-generation Mellanox ConnectX DDR InfiniBand architecture can help meet the I/O requirements for these increasingly powerful clusters.

Munira Hussain is a systems engineer and adviser in the Dell High-Performance Computing Group specializing in HPC systems and architecture design. She has a bachelor’s degree in Electrical Engineering from the University of Illinois at Urbana-Champaign.

Toby Sebastian is a senior engineering analyst in the Enterprise Solutions Group at the Dell Bangalore Development Center. His current interests include high-end interconnects, HPC clustering packages, and performance analysis of parallel applications. Toby has a B.Tech. in Computer Science and Engineering from the University of Calicut.

Shivaraj Nidoni is a senior engineering analyst in the Dell High-Performance Computing Group. His current interests include high-end interconnects, parallel file systems, and HPC clustering packages. Shivaraj has an M.Tech. in Computer Network Engineering from Visvesvaraya Technological University.

Glad Shainer is a director of HPC and technical marketing at Mellanox Technologies focusing on HPC, high-speed interconnects, and performance characterization. He has B.S. and M.S. degrees in Electrical Engineering from the Technion - Israel Institute of Technology, and is a member of the PCI-SIG PCI Extended and PCI Express Working Groups, cochair of the OpenFabrics Legal Working Group, and chairman of the HPC Advisory Council.