



An Introduction to DDR InfiniBand

BY RINKU GUPTA
VISHVESH SAHASRABUDHE
TOBY SEBASTIAN
RIZWAN ALI

The increasingly popular InfiniBand high-speed interconnect now supports both single data rate (SDR) and double data rate (DDR) transmission, with DDR providing significantly increased speeds. This article discusses how InfiniBand has evolved and presents test results comparing the performance of Gigabit Ethernet, SDR InfiniBand, and DDR InfiniBand.



The rapid adoption of high-performance computing (HPC) clusters has motivated significant advances in the technology of cluster components, including interconnects such as InfiniBand, 10 Gigabit Ethernet, and Myricom Myrinet. InfiniBand, in particular, has seen tremendous advances during the last few years: the TOP500 Supercomputer Sites lists, which track the top supercomputers and HPC clusters in the world, show that InfiniBand deployment in these clusters grew from 0.2 percent in June 2003 to 15.6 percent in November 2006.¹ Defined by the InfiniBand Trade Association in 2001, InfiniBand was designed to replace the traditional shared-bus PCI architecture with a switched-fabric architecture that could help increase scalability.

InfiniBand was initially designed to support single data rate (SDR) unidirectional signaling rates of up to 2.5 Gbps using PCI Extended (PCI-X) technology, but has since matured to use PCI Express (PCIe) technology and now supports double data rate (DDR) unidirectional signaling rates of up to 20 Gbps. This article outlines the InfiniBand architecture, discusses how it has evolved since its introduction, and presents test results comparing the performance of Gigabit Ethernet, SDR InfiniBand, and DDR InfiniBand.

InfiniBand architecture

The InfiniBand architecture (see Figure 1) defines a serial, point-to-point, switched-fabric I/O architecture providing a dedicated

link between any two endpoints, where each endpoint can be a direct interface to a host system or channel adapter. Channel adapters can be either server host channel adapters (HCAs) or storage system target channel adapters (TCAs).

The physical layer of InfiniBand can consist of bidirectional links capable of operating at signaling rates of 1X (the base link rate, defined at a theoretical 2.5 Gbps), 4X (10 Gbps), and 12X (30 Gbps). InfiniBand uses 8B/10B encoding at the lowest level, thereby providing theoretical data rates of 2 Gbps, 8 Gbps, and 24 Gbps for 1X, 4X, and 12X, respectively. The physical lanes are partitioned into virtual lanes at the transport level, including a dedicated virtual lane for management, with different priority levels to help ensure quality of service. Endpoints communicate with each other using *queue pairs*, which consist of a send queue and a receive queue located on source and destination adapters; applications use these queue pairs to communicate information to the adapters about data to be sent or received.

InfiniBand has a hardware-offloaded protocol stack. It also provides capabilities such as a zero-copy mechanism and Remote Direct Memory Access (RDMA). The zero-copy mechanism helps avoid extra memory copies that may be generated when a message is sent from an application to an adapter. RDMA allows data to be moved from local memory to remote memory by the InfiniBand adapter without involving the receiver host processor. Both the zero-copy mechanism

Related Categories:

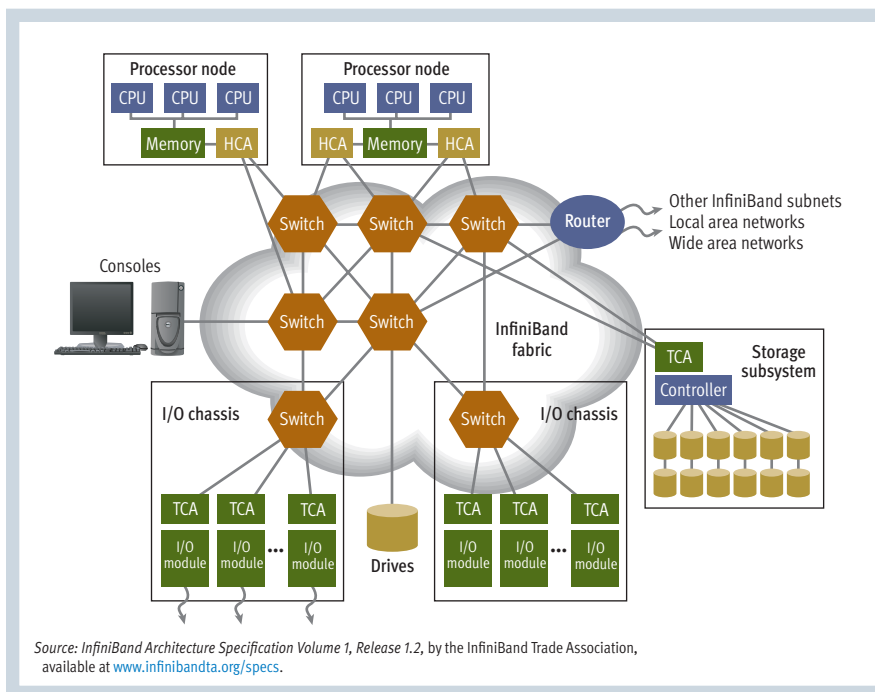
High-performance computing (HPC)

InfiniBand

Interconnects

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

¹TOP500 Supercomputer Sites, Interconnect Family statistics, June 2003 list, www.top500.org/stats/21; and Interconnect Family statistics, November 2006 list, www.top500.org/stats/28.



Source: InfiniBand Architecture Specification Volume 1, Release 1.2, by the InfiniBand Trade Association, available at www.infinibanda.org/specs.

Figure 1. Example deployment using the InfiniBand architecture

and RDMA are designed to optimize message transfer times and reduce processor utilization. All these features help make InfiniBand well suited for high-performance applications.

Evolution of InfiniBand

SDR InfiniBand adapters using 4X signaling rates can achieve a theoretical maximum bidirectional throughput of 2 GB/sec. However, the underlying PCI-X technology of these adapters has limited their bidirectional throughput to 1 GB/sec (for 64-bit PCI-X buses at 133 MHz), and thus severely limited their performance. The introduction of PCIe helped reduce this bottleneck, complementing InfiniBand by upgrading PCI slot performance and helping increase performance on servers. Unlike standard PCI and PCI-X, PCIe uses a serial, point-to-point interface composed of lanes, with each lane operating at a unidirectional signaling rate of 2.5 Gbps. PCIe helps reduce latency by allowing I/O devices to connect directly to memory controllers, and can deliver scalable bandwidth by using multiple lanes in each dedicated point-to-point link.

The InfiniBand specification also supports DDR and quad data rate (QDR) transmission on InfiniBand links. DDR InfiniBand uses the same

number of physical transmission and receiving signal wires and lanes as SDR InfiniBand, but each 2.5 Gbps InfiniBand lane is clocked at twice the original speed, thereby doubling the signaling rate. Thus, 4X DDR InfiniBand is capable of a unidirectional signaling rate of 20 Gbps and a bidirectional signaling rate of 40 Gbps, making it well suited for applications requiring high throughput. Figure 2 summarizes the unidirectional speeds of different SDR and DDR InfiniBand links.

Although an x8 PCIe slot has a backplane capacity of 16 Gbps, theoretically making it a

good fit for 4X DDR InfiniBand adapters, server architecture implementation details and signal or interface changes may cause it to become a slight bottleneck, in which case an x16 PCIe slot may help increase performance. The performance of DDR InfiniBand adapters can also depend on adapter hardware, fabric topology, InfiniBand protocol, and middleware implementation as well as application characteristics.

InfiniBand performance test configuration

In March 2007, engineers from the Dell high-performance computing team tested an HPC cluster to compare the performance of Gigabit Ethernet, SDR InfiniBand, and DDR InfiniBand interconnects. The cluster consisted of eight Dell™ PowerEdge™ 1950 servers—each with two quad-core Intel® Xeon® X5355 processors at 2.66 GHz with two 4 MB level 2 (L2) caches and a 1,333 MHz frontside bus, and four 1 GB PC2-5300 fully buffered dual in-line memory modules (DIMMs)—running the Red Hat® Enterprise Linux® 4 Update 4 OS with kernel version 2.6.9-42.EL.smp.

In the Gigabit Ethernet tests, the servers used on-board Broadcom NetXtreme II BCM5708 Gigabit² Ethernet network interface cards (NICs) along with the OS-native bnx2 driver. In addition, these tests employed a 24-port Dell PowerConnect™ 5324 switch.

For the InfiniBand tests, each server was equipped with a PCIe riser consisting of two x8 PCIe slots, with the SDR and DDR InfiniBand cards inserted in each of these slots. These tests used

		SDR InfiniBand	DDR InfiniBand	Recommended bus
1X	Signaling rate	2.5 Gbps	5 Gbps	<ul style="list-style-type: none"> SDR: 64-bit PCI-X at 133 MHz DDR: N/A (DDR InfiniBand adapters do not include a PCI-X interface)
	Data rate	2 Gbps	4 Gbps	
4X	Signaling rate	10 Gbps	20 Gbps	<ul style="list-style-type: none"> SDR: x8 PCIe DDR: x8 or x16 PCIe
	Data rate	8 Gbps	16 Gbps	
12X	Signaling rate	30 Gbps	60 Gbps	N/A (12X InfiniBand chips are typically used internally in InfiniBand switches)
	Data rate	24 Gbps	48 Gbps	

Figure 2. Unidirectional speeds of different SDR and DDR InfiniBand links

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.

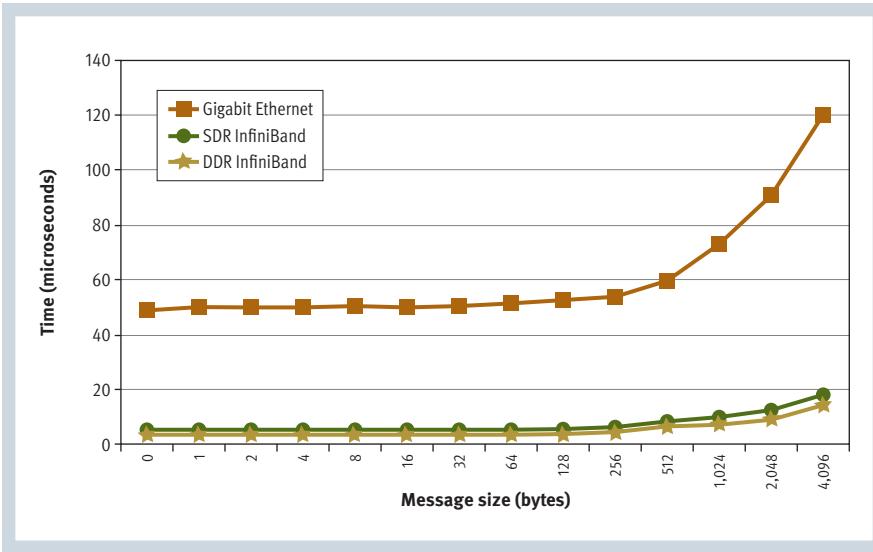


Figure 3. Latency times for different interconnects and message sizes measured using the Intel MPI Benchmarks latency test

a single-port, memory-free, x8 PCIe-based Cisco SDR InfiniBand HCA (SFS-HCA-312-A1) and a single-port, memory-free, x8 PCIe-based Cisco DDR InfiniBand HCA (SFS-HCA-320-A1). The InfiniBand fabric consisted of a 24-port Cisco SFS 7000P SDR switch and a 24-port Cisco SFS 7000D DDR switch, both including an embedded subnet manager for topology management.

slightly lower latency than SDR InfiniBand—when sending small messages of 4 bytes, for example, SDR InfiniBand had a latency of approximately 4 microseconds, while DDR InfiniBand had a latency of approximately 3.75 microseconds.

The Dell team next measured throughput using the Ohio State University MPI-level unidirectional and bidirectional bandwidth tests

between two nodes connected by a single switch. In the unidirectional bandwidth test, the sender node sends back-to-back messages to the receiver node in an attempt to saturate the interconnect pipe. The bidirectional bandwidth test is similar, except that both nodes send and receive messages at the same time. As shown in Figure 4, both SDR and DDR InfiniBand provided much higher unidirectional throughput than Gigabit Ethernet, with maximums of approximately 800 MB/sec for SDR InfiniBand and 1,200 MB/sec for DDR InfiniBand. As shown in Figure 5, both SDR and DDR InfiniBand also provided much higher bidirectional throughput than Gigabit Ethernet, with maximums of approximately 1,200 MB/sec for SDR InfiniBand and 2,000 MB/sec for DDR InfiniBand. Optimizing the InfiniBand protocol and middleware stack and using x16 PCIe slots may have helped increase DDR InfiniBand throughput even further.

In addition to performing basic latency and throughput tests, the Dell team also examined the impact of these interconnects on the synthetic benchmarks in the NPB suite. The MPI programming version of the NPB suite contains a set of eight programs, four of which represent computational cores of different numerical methods used

InfiniBand performance test results

The Dell team used three benchmarks to evaluate the performance of the different interconnects: the Intel Message Passing Interface (MPI) Benchmarks latency test, the Ohio State University MPI-level unidirectional and bidirectional bandwidth tests, and four NASA Advanced Supercomputing (NAS) Parallel Benchmarks (NPB) tests.³ The team first measured latency between two nodes using the ping-pong latency test from the Intel MPI Benchmarks suite, in which the sender node sends a message to the receiver node, and the receiver node then sends a message of the same size back to the sender node. Latency is typically measured as half the round-trip time. As shown in Figure 3, both SDR and DDR InfiniBand had much lower latency than Gigabit Ethernet, and DDR InfiniBand had a

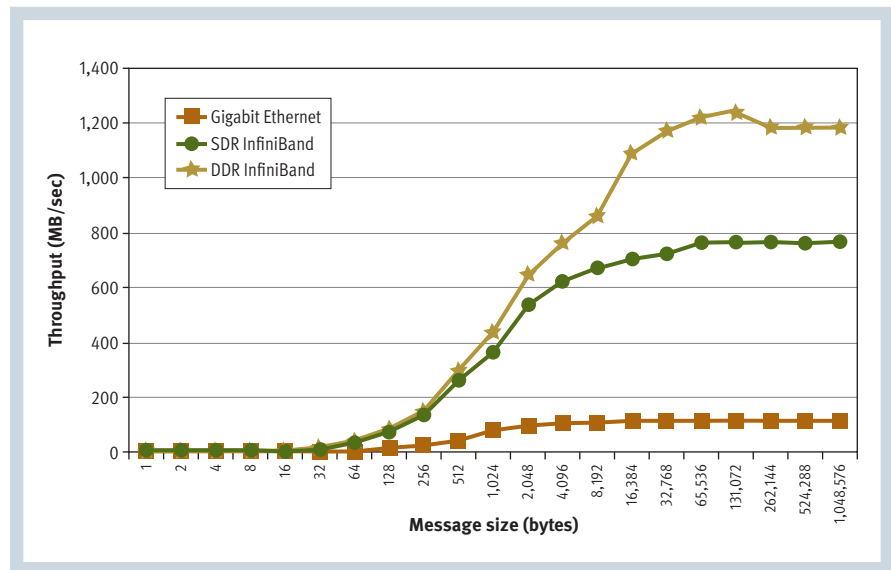


Figure 4. Unidirectional throughput for different interconnects and message sizes measured using the Ohio State University MPI-level unidirectional bandwidth test

³For more information about these benchmarks, visit www.intel.com/cd/software/products/asm-na/eng/307696.htm#mpibenchmarks, mvapich.cse.ohio-state.edu/benchmarks, and www.nas.nasa.gov/Software/NPB.

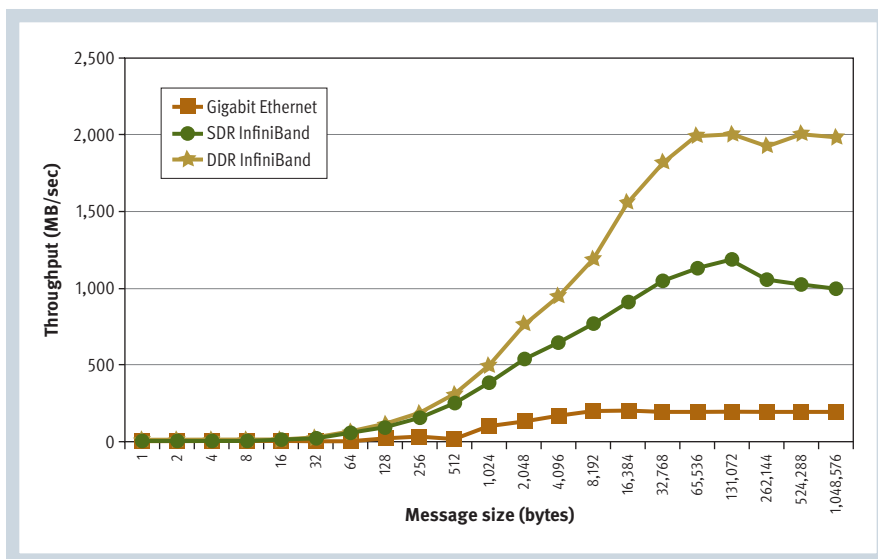


Figure 5. Bidirectional throughput for different interconnects and message sizes measured using the Ohio State University MPI-level bidirectional bandwidth test

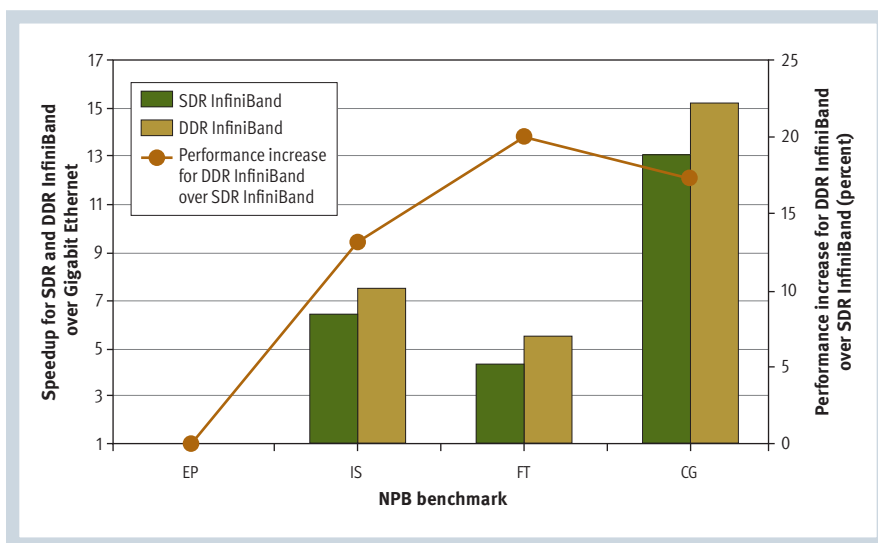


Figure 6. Speedup for SDR and DDR InfiniBand over Gigabit Ethernet and performance increases for DDR InfiniBand over SDR InfiniBand measured using four NAS Parallel Benchmarks tests

by computational fluid dynamics applications. The Embarrassingly Parallel (EP) benchmark involves no communication. The Integer Sort (IS) benchmark, based on a parallel sorting algorithm, tests both integer computation speed and communication performance, which is important in particle method codes. The Fast Fourier Transform (FT) and Conjugate Gradient (CG) benchmarks, based on solving matrix vector multiplication and partial differential equations, represent high-computation and long-distance communication applications. Each benchmark

includes multiple problem-size classes, denoted as S, W, A, B, C, and D. The Dell tests used the Class C problem size for each benchmark. The NPB benchmarks were run on 64 processor cores, with eight processes on each of the eight PowerEdge 1950 nodes.

Figure 6 shows the average performance speedup results for SDR and DDR InfiniBand compared against a Gigabit Ethernet baseline as well as the performance increases for DDR InfiniBand over SDR InfiniBand. The EP test, as expected, showed no difference in performance

between the different interconnects. The IS test, which involves a large number of small messages, resulted in significant speedup for both SDR and DDR InfiniBand over Gigabit Ethernet, with DDR InfiniBand performing approximately 13 percent better than SDR InfiniBand because of its lower latency. The FT and CG tests, which involve a mix of small and large messages, again resulted in significant speedup for both SDR and DDR InfiniBand over Gigabit Ethernet, with DDR InfiniBand performing approximately 20 percent better than SDR InfiniBand in the FT test and 17 percent better on the CG test.

The characteristics of specific applications can have a significant effect on interconnect performance. In particular, applications that communicate frequently using large message sizes may benefit more from DDR InfiniBand than SDR InfiniBand.


InfiniBand cluster deployment

Dell has partnered with Cisco Systems to deliver DDR InfiniBand technology as part of its HPC clusters. Dell validates and verifies preconfigured homogenous InfiniBand-based clusters ranging from 8 to 256 nodes and designs custom clusters using Intel Xeon processor-based Dell PowerEdge 1950 and PowerEdge 2950 servers as well as AMD Opteron™ processor-based PowerEdge SC1435 servers. Dell designs its InfiniBand-based clusters using single-port, memory-free x8 PCIe-based adapters rated at 20 Gbps and 24-port, 144-port, and 288-port DDR InfiniBand switches. In contrast to copper-based SDR cables, which are available in lengths of up to 15 meters, signal and data integrity requirements limit the length of supported copper-based DDR cables to 8 meters.

Administrators can transparently install Cisco InfiniBand software on Dell clusters using the Platform Open Cluster Stack (OCS) deployment package. Platform OCS is an enterprise-level semi-automated software stack developed by Platform Computing that enables the deployment, maintenance, and management of Dell clusters. Platform OCS 4.4.1 packages a single set of Cisco Linux host drivers that can support

both SDR and DDR InfiniBand hardware. In addition to low-level drivers, kernel modules, and diagnostic utilities, the Cisco Linux host driver stack also contains fully supported MPI software, which enterprises can use to run parallel distributed applications on their clusters.

High-throughput interconnect for HPC clusters

InfiniBand has made rapid advances since its introduction, and now includes support for both SDR and DDR transmission. Taking advantage of the reduced latency and increased throughput of DDR InfiniBand can help enterprises increase the performance of throughput-sensitive applications in HPC cluster environments. 

Rinku Gupta is a senior software developer in the Mathematics and Computer Science Division at Argonne National Laboratory. Her research interests include middleware parallel libraries,

fault tolerance in HPC, and performance and interconnect benchmarking. Rinku has a B.E. in Computer Engineering from Mumbai University and an M.S. in Computer Science from the Ohio State University.

Vishvesh Sahasrabudhe is a member of the Scalable Systems Group at Dell. His current research interests include high-speed interconnects and performance benchmarks. He has a B.Tech. in Electrical Engineering from the Indian Institute of Technology and an M.S. in Computer Science and Engineering from the Ohio State University.

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

in Computer Science and Engineering from the University of Calicut.

Rizwan Ali is a systems engineer and a senior member of the Scalable Systems Group at Dell. His research interests include performance benchmarking, cluster architecture, parallel applications, and high-speed interconnects. He has a B.S. in Electrical Engineering from the University of Minnesota.

more
Online
www.dell.com/powersolutions

QUICK LINK

InfiniBand Trade Association:
www.infinibandta.org