



InfiniBand SDR, DDR, and QDR Technology Guide

The InfiniBand standard supports single, double, and quadruple data rate that enables an InfiniBand link to transmit more data. This paper discusses the characteristics of single data rate (SDR), double data rate (DDR), and quad data rate (QDR) InfiniBand transmission with respect to performance, transmission distances, and cabling considerations.

INFINIBAND DDR

InfiniBand supports DDR and QDR transmission to increase link bandwidth. In the context of InfiniBand, DDR and QDR differ with respect to computer DDR and QDR transmission as the InfiniBand 2.5-Gbps lane is clocked two times (DDR) or four times (QDR) faster, instead of transferring two bits (DDR) or four bits (QDR) per clock cycle. By increasing the clock-rate by a factor of two or four times, DDR and QDR transmission provide a simple method of increasing bandwidth capacity and reducing serialization delay.

Note: Commercial InfiniBand QDR products are not available at this time.

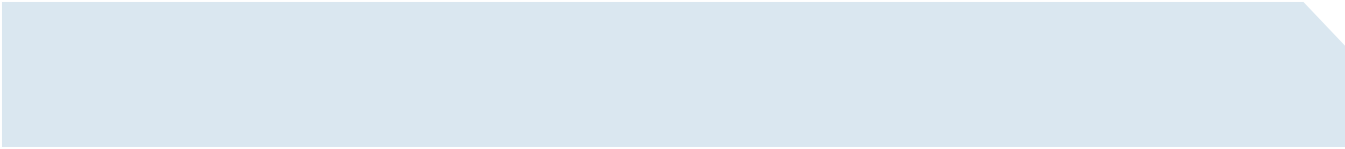
Because the link data rate is twice as fast (5 Gbps) or four times faster (10 Gbps) than InfiniBand's base 2.5 Gbps, bandwidth is increased by a factor of two or four and latency is reduced because packets are serialized faster. This benefits applications that are either bandwidth-intensive (also referred to I/O-intensive) or are particularly sensitive to interprocess latency. See Table 1.

Table 1. InfiniBand SDR and DDR Link Characteristics

InfiniBand Link	Signal Pairs	Signaling Rate	Data Rate (Full Duplex)
1X-SDR	2	2.5 Gbps	2.0 Gbps
4X-SDR	8	10 Gbps (4 x 2.5 Gbps)	8 Gbps (4 x 2 Gbps)
12X-SDR	24	30 Gbps (12 x 2.5 Gbps)	24 Gbps (12 x 2 Gbps)
1X-DDR	2	5 Gbps	4.0 Gbps
4X-DDR	8	20 Gbps (4 x 5 Gbps)	16 Gbps (4 x 4 Gbps)
12X-DDR	24	60 Gbps (12 x 5 Gbps)	48 Gbps (12 x 4 Gbps)
1X-QDR	2	10 Gbps	8.0 Gbps
4X-QDR	8	40 Gbps (4 x 10 Gbps)	32 Gbps (4 x 8 Gbps)
12XQDDR	24	120 Gbps (12 x 10 Gbps)	96 Gbps (12 x 8 Gbps)

Note: Although the signaling rate is 2.5 Gbps, the effective data rate is limited to 2 Gbps because of the 8B/10B encoding scheme: $(2.5 \times 8) \div 10 = 2$ Gbps

For applications that move large data files, such as distributed databases and data-mining applications, InfiniBand 4X DDR provides significant performance benefits. If an application transmits many small messages, there may be limited performance improvement realized by deploying InfiniBand DDR depending upon the size of the high-performance computing (HPC) cluster and application. However, because DDR serializes packets to line twice as fast as SDR, end-to-end latency can be significantly reduced in large, multi-stage clusters. Although the difference in latency is quite small—in the order of 120 to 600 nanoseconds (ns) for back-to-back configurations—



the *cumulative* gains can be significant when large multi-stage switch and multi-tier network architectures are considered, or where the application run-time is measured in days or weeks.

However, the actual gain in performance realized may be limited by the efficiency of the server BUS architecture—PCI-X, PCI-Express, etc.—which must be taken into account to ensure the most effective utilization of the available bandwidth. For PCI and PCI-X 1.0 servers, InfiniBand 4X SDR provides balanced transmission because they support 4 and 8 Gbps of bandwidth capacity (Table 2).

Table 2. Server PCI and InfiniBand Link Matrix

Bus Technology	Backplane Capacity	InfiniBand			
		4X-SDR (8 Gbps)	4X-DDR (16 Gbps)	12X-SDR (24 Gbps)	12X-DDR (48 Gbps)
PCI-X 1.0 (64 bit at 100 MHz)	6.4 Gbps	X			
PCI-X 1.0 (64 bit at 133 MHz)	8.5 Gbps	X			
PCI-X 2.0 (64 bit at 266 MHz)	17 Gbps	X	–		
PCI-X 2.0 (64 bit at 533 MHz)	34 Gbps	X	–	–	–
PCI-Express (X8)	16 Gbps	X	X	X	
PCI-Express (X12)	24 Gbps	X	X	X	

Note: InfiniBand data rates are quoted that are 20 percent less than the signal rate.

Note: Table denotes ideal configurations for PCI-X based servers. Additional cards attached to the PCI-X bus will reduce available bandwidth.

It should be noted that although DDR and QDR may imply that bandwidth is doubled or quadrupled, several factors such as CPU, memory speed, PCI architecture (PCI-X, PCI-Express 8X or 12X), application characteristics, drivers, and physical cable plant may not deliver the full potential of InfiniBand DDR transmission, and it is important to consider all components when looking to increase overall system performance. For example, although a PCI-Express 8X (16-Gbps) server can theoretically saturate a 4X DDR link, the actual performance is less. Additionally, the application may also cause performance to drop if large volumes of data are transferred that require being paged to hard disk if the physical RAM cannot accommodate the volume of data.

Note: PCI-Express currently utilizes a maximum packet size of 88 bytes that has a 24-byte header and 64-byte data payload. This limits the efficiency and the effective throughput of an 8X PCI-Express server to approximately 72% of 16 Gbps, or 11.6 Gbps. If the throughput of current 64-byte PCI-Express 8X InfiniBand SDR and DDR implementations are compared, the performance difference between SDR and DDR is approximately 30%. However, new PCI-Express chip-sets that increase the PCI-Express packet size to 128 bytes will increase efficiency to 85% of 16 Gbps or approximately 13.5 Gbps. If the throughput of current PCI-Express 8X InfiniBand SDR and DDR implementations with 128-byte packets are compared, the performance difference between SDR and DDR is expected to be approximately 60%. It should also be noted that for smaller packet sizes, such as those less than 64 bytes, throughput drops due to the relationship between payload and 24-byte header.

In common with InfiniBand SDR, DDR and QDR transmission also use cut-through switching, although it should be noted that mixing transmission modes within the same network can be problematic because of the difference in how the packet is physically transmitted. If different transmission rates are used, either the InfiniBand subnet manager must be topology-aware and only switch SDR packets to SDR links, and DDR packets to DDR links, or the switch fabric must be able to store and forward the packets to provide rate matching.

Note: When switching between SDR and DDR links, the additional store-and-forward delay is one half of the packets serialization delay. As an example, if a 2KB packet is serialized to an SDR link, the serialization of half the packet—1,024 bytes—is approximately 1 microsecond. The reasons behind a “half-packet” delay when switching between SDR and DDR hosts are two-fold. If the switching decision is delayed until half the packet is received at the SDR interface, the egress port is not “blocked” while waiting for the remainder of the SDR transmitted packet to be received, and it enables other DDR-to-DDR transmissions to the same egress port to complete while the SDR packet is being received, and also enables the remainder of the SDR packet to be streamed through to the egress port once the switching decision has been made. For switching between DDR and SDR enabled hosts, the DDR host rate-limits transmission depending upon queue-pair (QP) parameters that are exchanged during connection set up. This behavior reduces the amount of buffers consumed to provide rate matching between DDR and SDR interfaces.

Although switching packets between DDR and SDR links incurs a small store-and-forward delay to provide rate matching, for applications that can tolerate the small increase in latency, using DDR links can reduce the number of uplinks to core switches and also better utilize the switch ports for server connections. As an example, a nonblocking design for a 24-port 4X switch would require 12 server connections and 12 uplinks. By using 4X-DDR connections for the uplinks, it is possible to connect sixteen 4X SDR attached servers and eight 4X DDR uplinks to maintain the desired oversubscription ratio.

The Cisco® InfiniBand portfolio includes a topology-aware High-Performance Subnet Manager that provides optimized forwarding for SDR and DDR traffic, and Cisco SFS 7000D Series InfiniBand Server Switches that support DDR-to-SDR switching. The ability to support DDR-to-SDR switching enables investment in dual-speed DDR/SDR technology to support PCI-X architectures today, with a clear migration to higher-capacity architectures later.

INFINIBAND PHYSICAL-LAYER CHARACTERISTICS

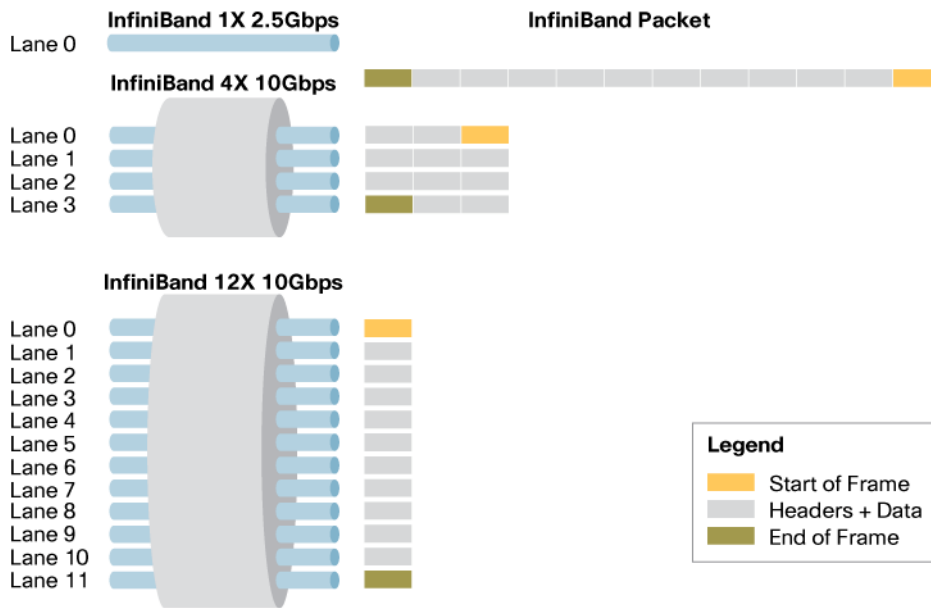
The InfiniBand physical-layer specification supports three data rates, designated 1X, 4X, and 12X, over both copper and fiber optic media. The base data rate, 1X single data rate (SDR), is clocked at 2.5 Gbps and is transmitted over two pairs of wires—transmit and receive—and yields an effective data rate of 2 Gbps full duplex (2 Gbps transmit, 2 Gbps receive). The 25 percent difference between data rate and clock rate is due to 8B/10B line encoding that dictates that for every 8 bits of data transmitted, an additional 2 bits of transmission overhead is incurred.

InfiniBand 4X and 12X SDR interfaces use the same base clock rate that uses multiple pairs, referred to as lanes, to increase bandwidth capacity. Therefore, an InfiniBand 4X interface realizes a data rate of 8 Gbps using 8 pairs of wires and a 12X interface realizes 24 Gbps using 24 pairs. Although it is possible for InfiniBand switches to forward frames between an InfiniBand 1X and 4X, or 4X and 12X link, there is a performance penalty because of the way that InfiniBand transmits packets.

InfiniBand switches use cut-through switching to provide low-latency transport between devices, which does not require the entire packet to be buffered before being forwarded. This enables InfiniBand to provide low-latency network transmission, but at the same time requires that the ingress and egress link speeds are the same and that the packet is received and transmitted in the same format.

When different link speeds are mixed in InfiniBand networks, the latter statement does not hold true as InfiniBand transmits portions of the packet *in parallel*. Therefore, for an InfiniBand 1X link, a packet is transmitted serially; for an InfiniBand 4X link, the packet is divided into four chunks and transmitted in parallel across the four 2.5-Gbps lanes. For a 12X link, a packet is divided into 12 chunks and transmitted in parallel across twelve 2.5-Gbps lanes (Figure 1).

Figure 1. InfiniBand 1X, 4X, and 12X Packet Transmission Scheme



Because the packet is transmitted in parallel, if a packet is switched between a 12X and a 4X interface, a store-and-forward penalty is incurred to change transmission formats. To maintain the performance that cut-through switching delivers, most InfiniBand fabrics consist of a single link speed; for example, all links are 4X SDR with multi-path forwarding used to provide additional bandwidth within the InfiniBand network. However, if an application can tolerate the small store-and-forward delay, mixing 12X and 4X in the same InfiniBand network may be an acceptable solution.

INFINIBAND CABLING

InfiniBand uses a base 2.5-Gbps clock rate that necessitates that InfiniBand ports and cables are thicker and bulkier than Universal Twisted Pair or UTP, commonly found in Ethernet applications. Generally, the faster a link is clocked, the more crosstalk is generated because of the rapid rise and fall of the carrier voltage as bits are transmitted. Additionally, because of the signal attenuation, the receiver must be able to interpret the signal correctly when it is received. At higher InfiniBand speeds, the effects of crosstalk and data skew limit cable lengths and consequently limit the number of nodes that can be supported within a cluster. See Table 3.

Table 3. InfiniBand DDR and SDR Cabling Lengths

Cable Type	Link Rate	Distance	Notes
CX-4 Copper	<ul style="list-style-type: none"> • 1X-SDR • 4X-SDR • 12X-SDR 	<ul style="list-style-type: none"> • 0–20m • 0–15m • 0–8–10m 	Cisco supports up to 15m because of BER degradation.
	<ul style="list-style-type: none"> • 4X DDR • 12X DDR 	<ul style="list-style-type: none"> • 0–8–10m • 0–5–7m 	4X DDR = 5-Gbps signal, 4-Gbps data rate 12X DDR = 20-Gbps signal, 16-Gbps data rate At this time, Cisco only supports 4X DDR cables for lengths up to 8 meters.
Optical Fiber: <ul style="list-style-type: none"> • 62.5 micron multimode • 50 micron at 500 MHz per Km • 50 micron at 2000 MHz per Km • 50 micron at 500 MHz per Km • 50 micron at 2000 MHz per Km 	<ul style="list-style-type: none"> • 4X-SDR • 4X-SDR • 4X-SDR • 12X-SDR • 12X-SDR 	<ul style="list-style-type: none"> • 2–75m • 2–125m • 2–200m • 2–125m • 2–200m 	InfiniBand Specification: 2–200m <ul style="list-style-type: none"> • 12 core ribbon • 12 core ribbon. Cisco supported. • 12 core ribbon • 24 core ribbon • 24 core ribbon
Optical Fiber: <ul style="list-style-type: none"> • 62.5 micron multimode • 50 micron at 500 MHz per Km • 50 micron at 2000 MHz per Km 	<ul style="list-style-type: none"> • 4X-DDR • 4X-DDR • 4X-DDR 	<ul style="list-style-type: none"> • 2–75m • 2–125m • 2–200m 	InfiniBand Specification: 2–200m <ul style="list-style-type: none"> • 12 core ribbon • 12 core ribbon • 12 core ribbon

Note: The 12 core ribbon cable uses the first four fibers as the transmit path, the center four fibers are unused and the last four fibers are used for the receive path. Each fiber strand supports transmission at 2.5 Gbps.

For InfiniBand 4X installations, although the copper cabling lengths are relatively limited when compared to Gigabit Ethernet, by carefully planning cabling runs and equipment placement, very large clusters can be constructed. As an example, the world’s largest InfiniBand-based HPC cluster, built using Cisco SFS 7000 Series InfiniBand Server Switches, consists of more than 4,500 dual-processor servers. By contrast, using copper cabling only, the largest InfiniBand 12X HPC cluster that can be built is 256 nodes using standard “side-by-side” rack configuration. By arranging the racks in a “U” shape, it is possible to build a 512-node cluster, although this complicates the cabling and maintenance aspects of the cluster.

Fiber optic cabling mitigates the copper cable distance limitations due to crosstalk and also provides a solution to some of the problems related to cable management, although there are considerations that need to be made with respect to cost. If a nonblocking InfiniBand 12X configuration is considered using a 24-port InfiniBand edge switch, this equates to 12 fiber optic uplinks and 24 fiber optic transceivers. When the cost of provisioning multiple 12X optical transceivers is considered, this can significantly increase the cost of the HPC deployment when compared to an all-copper cabling solution.

12X InfiniBand switching provides a good solution for high-performance mid-range or high-end systems that can drive the bandwidth delivered using 12X connections. Additionally, these systems tend to be less densely clustered, and the cable length limitations of high-speed transmission do not adversely affect the scalability of the system. By contrast, for larger clusters that are comprised of multiple low-end systems, InfiniBand 4X provides excellent price-to-performance ratio and scalability characteristics.

The effect that using DDR has is similar to the issues encountered when considering 4X or 12X InfiniBand; namely the reduction in cabling distance limits the size of the cluster that can be built (refer to Table 3).

Cisco Systems currently offers 4X copper wire SKUs for both SDR and DDR signal rates. Table 4 describes these SKUs.

Table 4. Cisco InfiniBand SDR and DDR Cabling SKUs

SKU	Wire Gauge	Description	Distance	SDR Support	DDR Support
CAB-04XS-01=	30AWG	Cisco 4XIB SuperFlex Cable, 1m, DDR Ready	1 meter	✓	✓
CAB-04XS-03=	30 AWG	Cisco 4XIB SuperFlex Cable, 3m, DDR Ready	3 meters	✓	✓
CAB-04XS-05=	30 AWG	Cisco 4XIB SuperFlex Cable, 5m	5 meters	✓	
CAB-04XD-05=	26 AWG	Cisco 4XIB Cable, 5m, DDR Ready	5 meters	✓	✓
CAB-04XD-08=	24 AWG	Cisco 4XIB Cable, 8m, DDR Ready	8 meters	✓	✓
CAB-04X-10=	24 AWG	Cisco 4XIB Cable, 10m	10 meters	✓	
CAB-04X-15=	24 AWG	Cisco 4XIB Cable, 15m	10 meters	✓	

INFINIBAND CABLE MANAGEMENT

Most data centers are designed and implemented to accommodate Category 5 or 6 copper and fiber optic cabling. Because these cables have a relatively small diameter and are reasonably flexible, maintaining a minimum bend radius is relatively simple and equipment cabinets can be positioned side-by-side with little or no space between them. As is typically used within the data center, Category 5 or 6 copper inter-rack cabling may be routed across overhead cable management systems up to 100 meters.

InfiniBand cable presents a challenge within this environment because the cables are considerably thicker, heavier, and shorter in length to mitigate the effects of cross-talk and signal attenuation and achieve low bit error rates (BERs). To assure the operational integrity and performance of the HPC cluster, it is critically important to maintain the correct bend radius, or the integrity of the cable can be compromised such that the effects of cross-talk introduce unacceptable BERs.

To address these issues, it is essential to thoroughly plan the InfiniBand implementation and provide a good cable management solution that enables easy expansion and replacement of failed cables and hardware. This is especially important when InfiniBand 12X or DDR technologies are being deployed because the high transmission rates are less tolerant to poor installation practices.

SUMMARY

InfiniBand is the technology of choice for building high-performance compute cluster MPI networks because of the low latency and high bandwidth offered by the technology. The development of InfiniBand DDR and future QDR technology, which increase available bandwidth by a factor of two and four respectively, further increase performance for applications that are I/O-intensive. However, when deploying these technologies, consideration of the overall system—CPU, memory, PCI architecture, etc.—and the technical and deployment considerations regarding reduced cable lengths and cable management must also be taken into account.

Many InfiniBand vendors are actively promoting DDR capability for their respective InfiniBand switch and host channel adapters (HCAs). DDR technology will be commercially available in mid-2006 and promises to deliver increased link bandwidth capacity. QDR-capable switches and HCAs that increase the link capacity by a factor of four are under development.



Corporate Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 526-4100

European Headquarters
Cisco Systems International BV
Haarlerbergpark
Haarlerbergweg 13-19
1101 CH Amsterdam
The Netherlands
www-europe.cisco.com
Tel: 31 0 20 357 1000
Fax: 31 0 20 357 1100

Americas Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
www.cisco.com
Tel: 408 526-7660
Fax: 408 527-0883

Asia Pacific Headquarters
Cisco Systems, Inc.
168 Robinson Road
#28-01 Capital Tower
Singapore 068912
www.cisco.com
Tel: +65 6317 7777
Fax: +65 6317 7799

Cisco Systems has more than 200 offices in the following countries and regions. Addresses, phone numbers, and fax numbers are listed on the **Cisco.com Website at www.cisco.com/go/offices.**

Argentina • Australia • Austria • Belgium • Brazil • Bulgaria • Canada • Chile • China PRC • Colombia • Costa Rica • Croatia • Cyprus • Czech Republic
Denmark • Dubai, UAE • Finland • France • Germany • Greece • Hong Kong SAR • Hungary • India • Indonesia • Ireland • Israel • Italy
Japan • Korea • Luxembourg • Malaysia • Mexico • The Netherlands • New Zealand • Norway • Peru • Philippines • Poland • Portugal
Puerto Rico • Romania • Russia • Saudi Arabia • Scotland • Singapore • Slovakia • Slovenia • South Africa • Spain • Sweden
Switzerland • Taiwan • Thailand • Turkey • Ukraine • United Kingdom • United States • Venezuela • Vietnam • Zimbabwe

Copyright © 2006 Cisco Systems, Inc. All rights reserved. CCSP, CCVP, the Cisco Square Bridge logo, Follow Me Browsing, and StackWise are trademarks of Cisco Systems, Inc.; Changing the Way We Work, Live, Play, and Learn, and iQuick Study are service marks of Cisco Systems, Inc.; and Access Registrar, Aironet, BPX, Catalyst, CCDA, CCDP, CCIE, CCIP, CCNA, CCNP, Cisco, the Cisco Certified Internetwork Expert logo, Cisco IOS, Cisco Press, Cisco Systems, Cisco Systems Capital, the Cisco Systems logo, Cisco Unity, Enterprise/Solver, EtherChannel, EtherFast, EtherSwitch, Fast Step, FormShare, GigaStack, HomeLink, Internet Quotient, IOS, IP/TV, iQ Expertise, the iQ logo, iQ Net Readiness Scorecard, LightStream, Linksys, MeetingPlace, MGX, the Networkers logo, Networking Academy, Network Registrar, Packet, PIX, Post-Routing, Pre-Routing, ProConnect, RateMUX, ScriptShare, SlideCast, SMARTnet, The Fastest Way to Increase Your Internet Quotient, and TransPath are registered trademarks of Cisco Systems, Inc. and/or its affiliates in the United States and certain other countries.

All other trademarks mentioned in this document or Website are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (0601R)

Printed in USA

C11-352004-01 09/06