Mellanox Technologies

White Paper

Introduction to InfiniBand™

Executive Summary

InfiniBand is a powerful new architecture designed to support I/O connectivity for the Internet infrastructure. InfiniBand is supported by all the major OEM server vendors as a means to expand beyond and create the next generation I/O interconnect standard in servers. For the first time, a high volume, industry standard I/O interconnect extends the role of traditional "in the box" busses. InfiniBand is unique in providing both, an "in the box" backplane solution, an external interconnect, and "Bandwidth Out of the box", thus it provides connectivity in a way previously reserved only for traditional networking interconnects. This unification of I/O and system area networking requires a new architecture that supports the needs of these two previously separate domains.

Underlying this major I/O transition is InfiniBand's ability to support the Internet's requirement for RAS: reliability, availability, and serviceability. This white paper discusses the features and capabilities which demonstrate InfiniBand's superior abilities to support RAS relative to the legacy PCI bus and other proprietary switch fabric and I/O solutions. Further, it provides an overview of how the InfiniBand architecture supports a comprehensive silicon, software, and system solution. The comprehensive nature of the architecture is illustrated by providing an overview of the major sections of the InfiniBand 1.1 specification. The scope of the 1.1 specification ranges from industry standard electrical interfaces and mechanical connectors to well defined software and management interfaces.

The paper is divided into four sections.

The introduction sets the stage for InfiniBand and illustrates why all the major server vendors have made the decision to embrace this new standard. The next section reviews the effect InfiniBand will have on various markets that are currently being addressed by legacy technologies. The third section provides a comparison between switch fabrics and bus architecture in general and then delves into details comparing InfiniBand to PCI and other proprietary solutions. The final section goes into details about the architecture, reviewing at a high level the most important features of InfiniBand.

1.0 Introduction

Amdahl's Law is one of the fundamental principles of computer science and basically states that efficient systems must provide a balance between CPU performance, memory bandwidth, and I/O performance. At odds with this, is *Moore's Law* which has accurately predicted that semiconductors double their performance roughly every 18 months. Since I/O interconnects are governed by mechanical and electrical limitations more severe than the scaling capabilities of semiconductors, these two laws lead to an eventual imbalance and limit system performance. This would suggest that I/O interconnects need to radically change every few years in order to maintain system performance. In fact, there is another practical law which prevents I/O interconnects from changing frequently - *if it ain't broke don't fix it*.

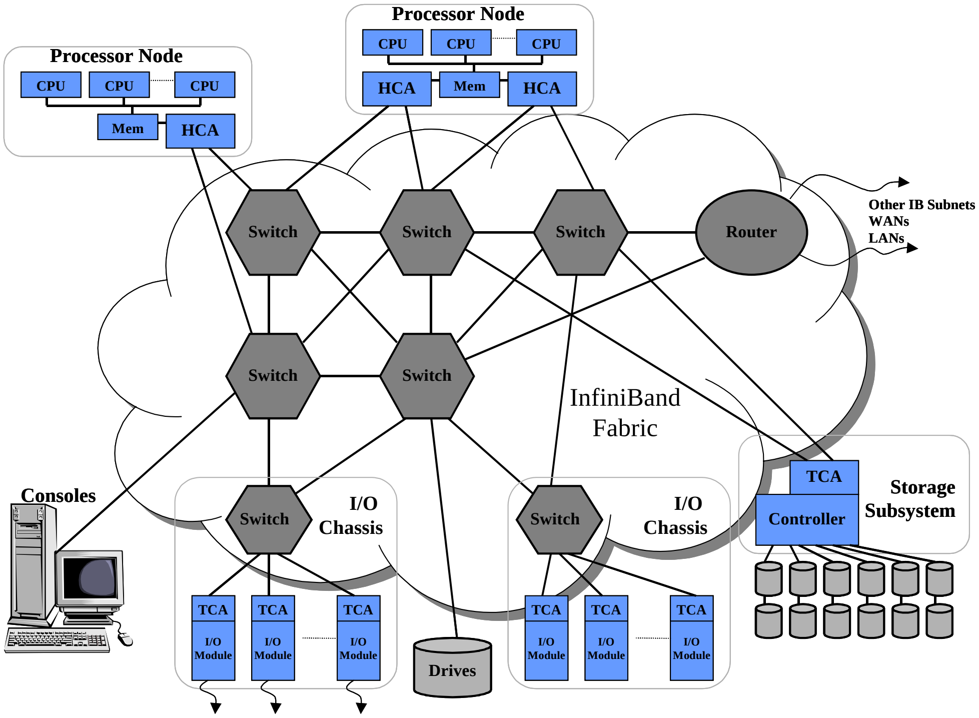
Bus architectures have a tremendous amount of inertia because they dictate the mechanical connections of computer systems and network interface cards as well as the bus interface architecture of semiconductor devices. For this reason, successful bus architectures typically enjoy a dominant position for ten years or more. The PCI bus was introduced to the standard PC architecture in the early 90's and has maintained its dominance with only one major upgrade during that period: from 32 bit/33 MHz to 64bit/66Mhz. The PCI-X initiative takes this one step further to 133MHz and seemingly should provide the PCI architecture with a few more years of life. But there is a divergence between what personal computers and servers require.

Personal Computers or PCs are not pushing the bandwidth capabilities of PCI 64/66. PCI slots offer a great way for home or business users to purchase networking, video decode, advanced sounds, or other cards and upgrade the capabilities of their PC. On the other hand, servers today often include clustering, networking (Gigabit Ethernet) and storage (Fibre Channel) cards in a single system and these push the 1GB bandwidth limit of PCI-X. With the deployment of the InfiniBand architecture, the bandwidth limitation of PCI-X becomes even more acute. The InfiniBand Architecture has defined 4X links, which are being deployed as PCI HCAs (Host Channel Adapters) in the market today. Even though these HCAs offer greater bandwidth that has ever been achieved in the past, PCI-X is a bottleneck as the total aggregate bandwidth of a single InfiniBand 4X link is 20 Gb/s or 2.5 GB/s. This is where new "local" I/O technologies like HyperTransport and 3GIO will play a key complementary role to InfiniBand.

The popularity of the Internet and the demand for 24/7 uptime is driving system performance and reliability requirements to levels that today's PCI interconnect architectures can no longer support. Data storage elements; web, application and database servers; and enterprise computing is driving the need for fail-safe, always available systems, offering ever higher performance. The trend in the industry is to move storage out of the server to isolated storage networks and distribute data across fault tolerant storage systems. These demands go beyond a simple requirement for more bandwidth, and PCI based systems have reached the limits of shared bus architectures. With CPU frequencies passing the gigahertz (Ghz) threshold and network bandwidth exceeding one gigabit per second (Gb/s), there is a need for a new I/O interconnect offering higher bandwidth to support and scale with today's devices.

Introduce InfiniBand, a switch-based serial I/O interconnect architecture operating at a base speed of 2.5 Gb/s or 10 Gb/s in each direction (per port). Unlike shared bus architectures, InfiniBand is a low pin count serial architecture that connects devices on the PCB and enables "Bandwidth Out of the Box", spanning distances up to 17m over ordinary twisted pair copper wires. Over common fiber cable, it can span distances of several kilometers or more. Furthermore, InfiniBand provides both QoS (Quality of Service) and RAS. These RAS capabilities have been designed into the InfiniBand architecture from the beginning and are critical to its ability to serve as the common I/O infrastructure for the next generation of compute server and storage systems at the heart of the Internet. As a result, InfiniBand will radically alter the systems and interconnects of the Internet infrastructure[1]. This paper discusses the features inherent to InfiniBand that enable this transformation.

Figure 1. InfiniBand System Fabric



InfiniBand is backed by top companies in the industry, including the steering committee members: Compaq, Dell, Hewlett Packard, IBM, Intel, Microsoft, and Sun. In total, there are more than 220 members of the InfiniBand Trade Association. The InfiniBand architecture offers all the benefits mentioned, but, to realize the full performance bandwidth of the current 10Gb/s links the PCI limitation must be removed, and this is where currently developing interconnect technologies will assist InfiniBand. This paper illustrates how to realize the full potential of InfiniBand, including full bandwidth, even up to the 12X link specification (or 30 Gb/s in each direction), in a later section.

2.0 Markets

Important markets such as Application Clustering, Storage Area Networks, Inter-Tier communication and Inter-Processor Communication (IPC) require high bandwidth, QoS, and RAS fea-

[1] The white paper InfiniBand and the Internet Data Center covers this transformation

tures. Also, many embedded systems (including routers, storage systems, and intelligent switches) utilize the PCI bus, often in the Compact PCI format, for their internal I/O architecture. Such systems are unable to keep up with high-speed networking interconnects such as Gigabit Ethernet and ATM, and therefore many companies are developing proprietary I/O interconnect architectures. Building on the experience of developing Ethernet Local Area Networks (LAN), Fibre Channel Storage Area Networks, and numerous Wide Area Network (WAN) interconnects, InfiniBand has been networked to go beyond the needs of today's markets and provide a cohesive interconnect for a wide range of systems. This is accomplished with direct support for highly important items such as RAS, QoS, and scalability.

2.1 Application Clustering

The Internet today has evolved into a global infrastructure supporting applications such as streaming media, business to business solutions, E-commerce, and interactive portal sites. Each of these applications must support an ever increasing volume of data and demand for reliability. Service providers are in turn experiencing tremendous pressure to support these applications. They must route traffic efficiently through increasingly congested communication lines, while offering the opportunity to charge for differing QoS and security levels. Application Service Providers (ASP) have arisen to support the outsourcing of e-commerce, e-marketing, and other e-business activities to companies specializing in web-based applications. These ASPs must be able to offer highly reliable services that offer the ability to dramatically scale in a short period of time to accommodate the explosive growth of the Internet. The cluster has evolved as the preferred mechanism to support these requirements. A cluster is simply a group of servers connected by load balancing switches working in parallel to serve a particular application.

High priority transactions between devices can be processed ahead of the lower priority items through quality of service mechanisms built into InfiniBand.

InfiniBand simplifies application cluster connections by unifying the network interconnect with a feature-rich managed architecture. InfiniBand's switched architecture provides native cluster connectivity, thus supporting scalability and reliability inside and "out of the box". Devices can be added and multiple paths can be utilized with the addition of switches to the fabric. High priority transactions between devices can be processed ahead of the lower priority items through QoS mechanisms built into InfiniBand.

2.2 Inter-Processor Communication (IPC)

Inter-Processor Communication allows multiple servers to work together on a single application. A high bandwidth, low-latency reliable connection is required between servers to ensure reliable processing. Scalability is critical as applications require more processor bandwidth. The switched nature of InfiniBand provides connection reliability for IPC systems by allowing multiple paths between systems. Scalability is supported with fully hot swappable connections managed by a single unit (Subnet Manager). With multicast support, single transactions can be made to multiple destinations. This includes sending to all systems on the subnet, or to only a subset of these systems. The higher bandwidth connections (4X, 12X) defined by InfiniBand provide backbone capabilities for IPC clusters without the need of a secondary I/O interconnect.

2.3 Storage Area Networks

Storage Area Networks are groups of complex storage systems connected together through managed switches to allow very large amounts of data to be accessed from multiple servers. Today, Storage Area Networks are built using Fibre Channel switches, hubs, and servers which are attached through Fibre Channel host bus adapters (HBA). Storage Area Networks are used to provide reliable connections to large databases of information that the Internet Data Center requires. A storage area network can restrict the data that individual servers can access, thereby providing an important "partitioning" mechanism (sometimes called zoning or fencing).

The fabric topology of InfiniBand allows communication to be simplified between storage and server. Removal of the Fibre Channel network allows servers to directly connect to a storage area network without a costly HBA. With features such as Remote DMA (RDMA) support, simultaneous peer to peer communication and end to end flow control, InfiniBand overcomes the deficiencies of Fibre Channel without the need of an expensive, complex HBA. A bandwidth comparison can be seen later in this white paper.

3.0 I/O Architectures - Fabric vs. Bus

The shared bus architecture is the most common I/O interconnect today although there are numerous drawbacks. Clusters and networks require systems with high speed fault tolerant interconnects that cannot be properly supported with a bus architecture. Thus all bus architectures require network interface modules to enable scalable, network topologies. To keep pace with systems, an I/O architecture must provide a high speed connection with the ability to scale. Table 1 provides a simple feature comparison between a switched fabric architecture and a shared bus architecture.

To keep pace with systems, an I/O architecture must provide a high speed connection with the ability to scale.

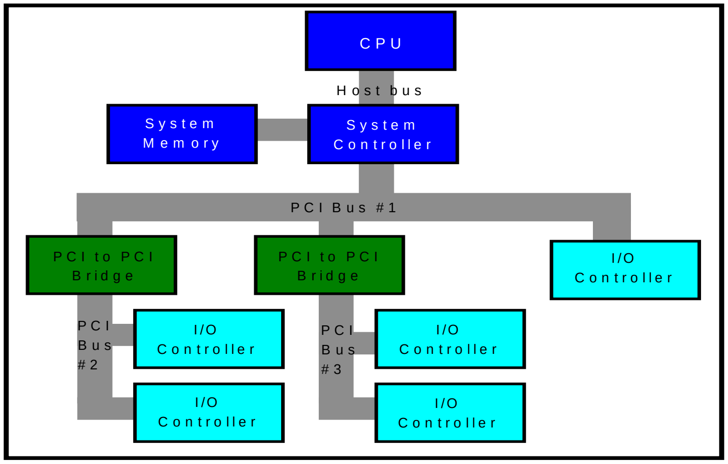
Table 1. Fabric vs. Shared Bus Architecture

Feature	**Fabric**	**Bus**
Topology	Switched	Shared Bus
Pin Count	Low	High
Number of End Points	Many	Few
Max Signal Length	KMs	Inches
Reliability	Yes	No
Scalable	Yes	No
Fault Tolerant	Yes	No

3.1 Shared Bus Architecture

In a bussed architecture, all communication shares the same bandwidth. The more ports added to the bus, the less bandwidth available to each peripheral. They also have severe electrical, mechanical, and power issues. On a parallel bus, many pins are necessary for each connection (64 bit PCI requires 90 pins), making the layout of a board very tricky and consuming precious printed circuit board (PCB) space. At high bus frequencies, the distance of each signal is limited to short traces on the PCB board. In a slot-based system with multiple card slots, termination is uncontrolled and can cause problems if not designed properly.

Figure 2. Conventional Shared Bus Architecture

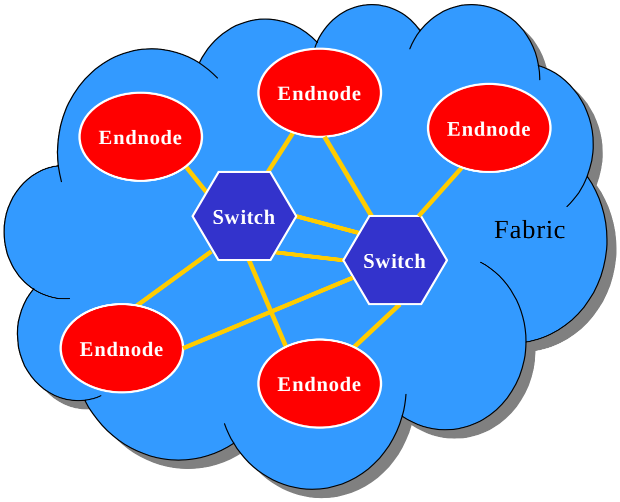


There is a load limit to a bus design that only allows a few devices per bus. Adding a bridge device to provide another bus with a new load limit behind the bridge overcomes this limitation. Although this allows for more devices to be connected to the system, data still flows through the central bus when accessing devices on other parts of the system. Latency and congestion increases with each bridge added to the system. A bus must be designed to operate under fully loaded conditions assuming the worst case number of devices allowed by the specification, which fundamentally limits the bus frequency. One of the major issues with a bus is that it can’t support “out of the box” system interconnects. To get systems to talk together, a separate interconnect is required, such as Ethernet (server-to-server communication) or Fibre Channel (storage networking).

3.2 Switched Fabric Architecture

A switched fabric is a point-to-point switch-based interconnect designed for fault tolerance and scalability. A point-to-point switch fabric means that every link has exactly one device connected at each end of the link. Thus the loading and termination characteristics are well controlled and (unlike the bus architecture), with only one device allowed, the worst case is the same as the typical case and therefore I/O performance can be much greater with a fabric.

Figure 3. Switched Fabric



The switched fabric architecture provides scalability which can be accomplished by adding switches to the fabric and connecting more endnodes through the switches. Unlike a shared bus architecture, the aggregate bandwidth of a system increases as additional switches are added to the network. Multiple paths between devices keep the aggregate bandwidth high and provide fail-safe, redundant connections.

3.3 I/O Interconnect Comparison

Numerous standards are today vying for supremacy in the interconnect market. Along with InfiniBand, these include PCI-X, Fibre Channel, 3GIO, Gigabit Ethernet, and RapidIO.

With the introduction of PCI-X, the reigning leader of the I/O interconnect market makes an attempt at supporting next generation high speed systems. Although PCI will not disappear completely, its deficiencies will make growth very limited. To make the comparison with InfiniBand is trivial. PCI-X operating at 133Mhz is a 90 pin bus with a fanout of one. This severely limits its reach in the market and without the notion of switches, PCI-X does not provide scalability. In contrast, InfiniBand supports 64k nodes per subnet, utilizing high port density switches requiring only four pins for each connection.

An additional advantage is the low power requirements of InfiniBand physical layer devices (PHYs) relative to other serial interconnect technologies. An InfiniBand copper PHY requires only about 0.25 watts per port. In contrast, a Gigabit Ethernet PHY requires roughly two watts per port. The order of magnitude difference is explained by realizing that Gigabit Ethernet PHY's are designed to support local area networks (LAN's) requiring connections spanning at least 100 meters. InfiniBand addresses only server and storage connections within the Internet Data Center and thus does not need to span such great lengths, and can therefore operate at a reduced power level.

InfiniBand is the only architecture that can be used on the PCB while also providing "out of the box" system interconnects (via fiber or copper cabling).

The much lower PHY power results in both integration and RAS cost advantages for InfiniBand. In this case, semiconductor integration is limited by the maximum power level of a chip (this time Moore's Law runs into the first law of thermodynamics). Thus for Gigabit Ethernet it may not be feasible to integrate PHYs into switches with 8, 16, 24, or more ports. In contrast, with InfiniBand's reduced PHY power requirements, these higher port count devices are entirely within reach. Reducing a multi-chip system to a single chip solution provides substantial cost as well as area savings. Whether or not the PHYs are integrated, InfiniBand's reduced power consumption results in cost savings for highly available applications. High availability requires uninterruptable power supplies in case of power failures. In this era of increased rolling brown-outs, this is a very real issue that facility managers must account for when designing their Internet Infrastructure. With even a moderately sized network, the use of InfiniBand can result in power savings of hundreds of watts with a corresponding reduction in cost savings.

Table 2, "Feature Comparison" examines features supported in hardware by InfiniBand and by other interconnect hardware. Each of the other technologies has serious drawbacks that point to InfiniBand as the clear choice for a system level interconnect. Most notably, InfiniBand is the only architecture that is designed to be used on the PCB while also providing "out of the box" system interconnects (via fiber or copper cabling). InfiniBand is designed as an I/O fabric and delivers transport level connections in hardware and thus is the only technology designed to support all of the data center on a single unified fabric. Each of these technologies offer benefits to the specific problems they were designed to solve, but only InfiniBand offers a single unified wire interconnect for clustering, communication and storage.

Following is a feature comparison among a number of compute I/O technologies.

Table 2. Feature Comparison

Feature	**InfiniBand**^{SM}	**PCI-X**	**Fibre Channel**	**1Gb & 10Gb Ethernet**	**HyperTransport™**	**Rapid I/O**	**3GIO**
Bus/Link Bandwidth	2.5, 10, 30Gb/s[a]	8.51 Gb/s	1, 2.1Gb/s[b]	1 Gb, 10Gb	12.8, 25.6, 51.2 Gb/s[g]	16, 32Gb/s[c]	2.5, 5, 10, 20,... Gb/s[d]
Bus/Link Bandwidth (Full Duplex)	5, 20, 60Gb/s[a]	Half-Duplex	2.1, 4.2Gb/s[b]	2 Gb, 20Gb	25.6, 51.2, 102 Gb/s[g]	32, 64Gb/s[c]	5, 10, 20, 40,... Gb/s[d]
Pin Count	4, 16, 48[e]	90	4	4 (GbE), 8 (10GbE-XAUI)	55,103,197[g]	40/76[c]	4, 8,16, 32...

Table 2. Feature Comparison

Feature	InfiniBand^{SM}	PCI-X	Fibre Channel	1Gb & 10Gb Ethernet	Hyper-Transport™	Rapid I/O	3GIO
Transport Media	PCB, Copper & Fiber	PCB only	Copper and Fiber Cable	PCB, Copper & Fiber	PCB only	PCB only	PCB & connectors[h]
Max Signal Length PCB/ Copper Cable	30in, 17m	inches	NA, 13M	20in, 100m	inches	inches	30in, NA
Maximum Signal Length Fiber	Km		Km	Km			
Simultaneous Peer to Peer communication	15 VLs + Mngt Lane			X		3 Transaction Flows	
Native HW Transport Support with Memory Protection	X						
In-Band Management	X		Out-of-band mngt	Not native, can use IP			
RDMA Support	X						
Native Virtual Interface Support	X						
End-to-End Flow Control	X			X	X	X	X
Partitioning/ Zoning[f]	X		X	X			
Quality of Service	X		X	Limited		X	limited
Reliable	X		X		X[g]	X	X
Scalable Link Widths	X				X	X	X
Backwards Compatible	n/a	X (PCI 3.3v only)	X	No, 10 GbE new physical signaling	n/a	n/a	n/a
Maximum Packet Payload	4 KB	Not Packet Based	2 KB	1.5KB (10GbE no jumbo support)	64 bytes	256 bytes	256 bytes

a. The raw bandwidth of an InfiniBand 1x link is 2.5Gb/s (per twisted pair). Data bandwidth is reduced by 8b/10b encoding to 2.0Gb/s for 1X, 8 Gb/s for 4X and 24Gb/s for 12x. In comparison to a half duplex bus the full duplex serial connections yields twice the data rate: 4/16/48 Gb/s.

b. The bandwidth of 2Gb Fibre Channel is 2.1Gb/s but the actual raw bandwidth (due to 8b/10b encoding) is 20% lower or around 1.7Gb/s (twice that for full duplex).

c. Values are for 8 bit/16 bit data paths peak @ 1GHz operation. Speeds of 125, 250 & 500 MHz are supported.

d. The raw bandwidth of an 3GIO link is 2.5Gb/s (per twisted pair). Data bandwidth is reduced by 8b/10b encoding to 2.0Gb/s, 4.0Gb/s, 8.0Gb/s, etc. In comparison to a half duplex bus the full duplex serial connections yields twice the data rate: 4,8,16, Gb/s etc.

e. The pin count for a 1x link is 4 pins, a 4X links uses 16 pins, and a 12X link uses 48 pins.

f. Memory partitioning enables multiple hosts to access storage endpoints in a controlled manner based on a key. Access to a particular endpoint is controlled by this key, so different hosts can have access to different elements in the network.

g. Based upon 8, 16, 32 bit HyperTransport (it can support 2 & 4 bit modes) with up to 800 Million transfers per second operation (modes from 400 MHz DDR can be supported). Error management features will be refined in future revisions of the specification.

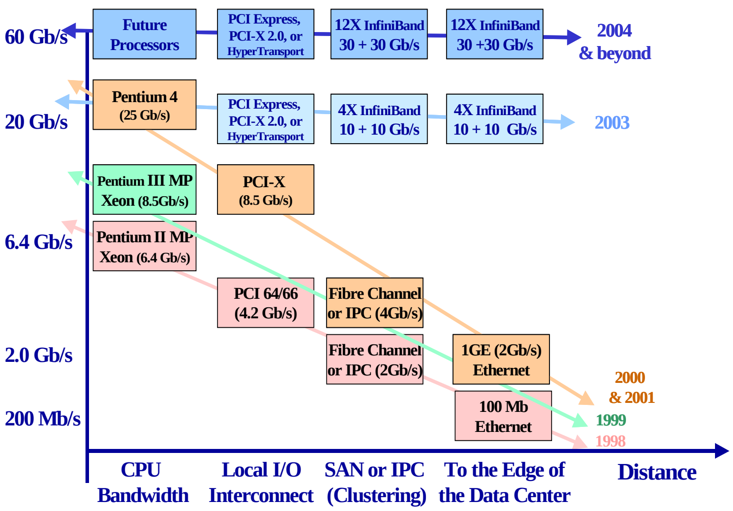
h: 3GIO has 1X, 2X, 4X, 8X, 16X and 32X lane widths. Copper, optical and emerging physical signaling media.

3.4 Interconnects Complement InfiniBand

Several of these new interconnects are actually key enablers for InfiniBand as they provide access to new levels of processor bandwidth and allow InfiniBand to extend this bandwidth outside the box. Technologies such as 3GIO, HyperTransport, and Rapid I/O are being developed and this will provide InfiniBand with a point of attachment to system logic that can support the 20 Gb/s required by 4X InfiniBand links and even the 60 Gb/sec needed by12X InfiniBand links. These technologies complement InfiniBand nicely.

3.5 Bandwidth Out of the Box

Figure 4. Bandwidth Out of the Box


60 Gb/s
20 Gb/s
6.4 Gb/s
2.0 Gb/s
200 Mb/s
Future Processors
PCI Express, PCI-X 2.0, or HyperTransport
12X InfiniBand 30 + 30 Gb/s
12X InfiniBand 30 +30 Gb/s
2004 & beyond
Pentium 4 (25 Gb/s)
PCI Express, PCI-X 2.0, or HyperTransport
4X InfiniBand 10 + 10 Gb/s
4X InfiniBand 10 + 10 Gb/s
2003
Pentium III MP Xeon (8.5Gb/s)
PCI-X (8.5 Gb/s)
Pentium II MP Xeon (6.4 Gb/s)
PCI 64/66 (4.2 Gb/s)
Fibre Channel or IPC (4Gb/s)
Fibre Channel or IPC (2Gb/s)
1GE (2Gb/s) Ethernet
100 Mb Ethernet
2000 & 2001
1999
1998
CPU Bandwidth
Local I/O Interconnect
SAN or IPC (Clustering)
To the Edge of the Data Center
Distance

A fundamental aspect of the InfiniBand Architecture is the concept of *“Bandwidth Out of the Box”*. InfiniBand has the ability to take bandwidth, which has historically been trapped inside the server, and extend this across the fabric. InfiniBand enables 10Gb/s performance to be effectively utilized, by delivering the data precisely where it is needed anywhere in the fabric. Historically bandwidth goes down the farther from the CPU data travels. Figure 4, “Bandwidth Out of the Box” illustrates this phenomena and the historical trends. Outside of the box means bandwidth from the processor to I/O, between servers for clustering or inter-processor communication (IPC), to storage, and all the way to the edge of the data center. Current state of the art processors have front side busses able to communicate with other processors and memory at 25 Gb/sec, but the PCI-X systems available today constrain the bandwidth available “outside the box” to only 8 Gb/s. Actual bandwidth within the data center is even further limited, with IPC bandwidth constrained to 1 or 2 Gb/s, Fibre Channel or storage communication is at best 2 Gb/sec and communication between systems, typically over Ethernet is limited to 1Gb/s. This illustrates that from the processor to the edge of the data center an order of magnitude of bandwidth is lost.

As discussed, new interconnects, namely 3GIO, HyperTransport or Rapid I/O can be used to increase I/O bandwidth well past 30 or even 60 Gb/s. As new processors and/or system chip sets incorporate these interconnects, the current PCI limitations are overcome. From there the bandwidth of InfiniBand can be unleashed as the HCA will connect into these interconnects and this allows clustering, communication and storage all to be connected at native InfiniBand speeds. 1X (2.5 Gb/s) and 4X (10 Gb/s) links are being deployed since 2001, and 2003 marks the deployment of 12X or 30 Gb/s links.

The figure below illustrates how InfiniBand releases *Bandwidth Out of the Box* by giving a historical look at bandwidth in the data center. Circa 1998: Intel’s Pentium II offered world class performance but the overall design of the compute server architecture limited the processors bandwidth to "inside the box." The further data travels from the processor the lower the bandwidth, until more than an order of magnitude of bandwidth is lost at the edge 100 Mb/sec. Circa 1999: The Pentium III improves processor performance but the equation stays the same. Bandwidth is lost over distance as data centers still communicate at only 100 Mb/sec at the edge. In 2000 & 2001 the Pentium 4 and all other data center sub systems improve bandwidth, but the equation still remains the same: there is more than an order of magnitude loss of bandwidth from the Processor to the edge of the data center. The InfiniBand Architecture changes the equation. The InfiniBand architecture provides 20Gb/s bandwidth (aggregate baud rate) from the processor to the edge of the data center including LAN/WAN and storage connection. InfiniBand enables *Bandwidth Out of the Box* allowing processor level bandwidth to move all the way to the edge of the data center. Additionally, the InfiniBand Architecture provides the headroom at 12X to scale to 60 Gb/s in 2003.

It is important to note that InfiniBand delivers not only bandwidth, but also delivers the data right where it is needed; through RDMA transfers from system memory to system memory. InfiniBand implements a reliable in-order transport connection in hardware and thus data is delivered extremely efficiently, with low latencies and without host CPU assistance. This is a huge benefit compared to Ethernet which has much longer latencies and consumes significant CPU cycles to run the TCP stack.

4.0 InfiniBand Technical Overview

InfiniBand is a switch-based point-to-point interconnect architecture developed for today’s systems with the ability to scale for next generation system requirements. It operates both on the PCB as a component-to-component interconnect as well as an “out of the box” chassis-to-chassis interconnect. Each individual link is based on a four-wire 2.5 Gb/s bidirectional connection. The architecture defines a layered hardware protocol (Physical, Link, Network, Transport Layers) as well as a software layer to manage initialization and the communication between devices. Each link can support multiple transport services for reliability and multiple prioritized virtual communication channels.

To manage the communication within a subnet, the architecture defines a communication management scheme that is responsible for configuring and maintaining each of the InfiniBand elements. Management schemes are defined for error reporting, link failover, chassis management and other services, to ensure a solid connection fabric.

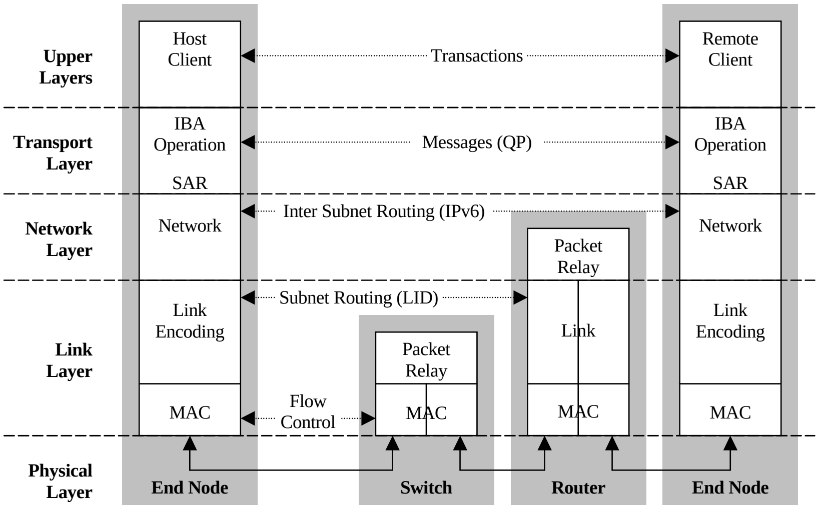
InfiniBand Feature Set

- Layered Protocol - Physical, Link, Network, Transport, Upper Layers
- Packet Based Communication
- Quality of Service
- Three Link Speeds
  1X - 2.5 Gb/s, 4 wire
  4X - 10 Gb/s, 16 wire
  12X - 30 Gb/s, 48 wire
- PCB, Copper and Fiber Cable Interconnect
- Subnet Management Protocol
- Remote DMA Support
- Multicast and Unicast Support
- Reliable Transport Methods - Message Queuing
- Communication Flow Control - Link Level and End to End

4.1 InfiniBand Layers

The InfiniBand architecture is divided into multiple layers where each layer operates independently of one another. As shown in Figure 5, “InfiniBand Layers” InfiniBand is broken into the following layers: Physical, Link, Network, Transport, and Upper Layers.

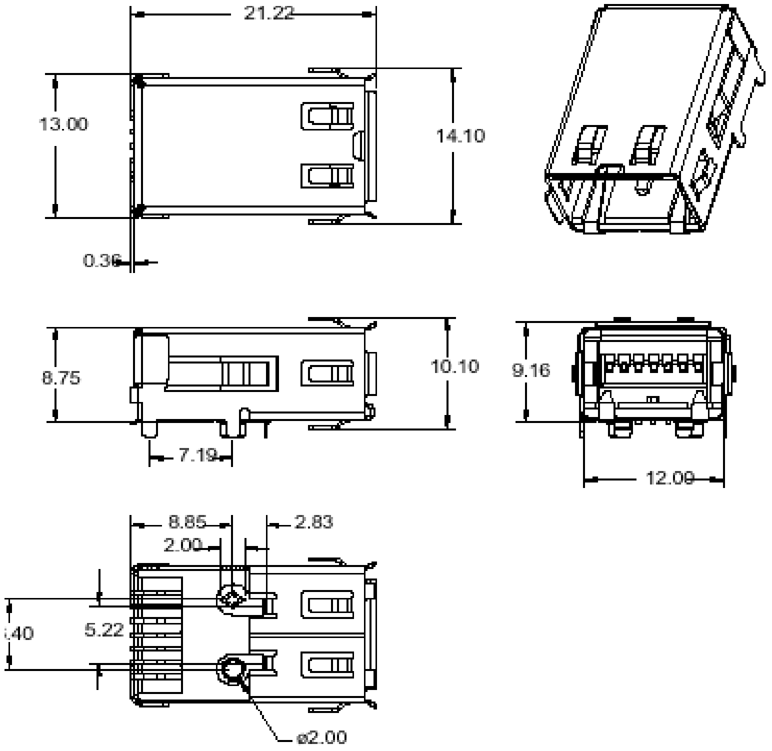
Figure 5. InfiniBand Layers



4.1.1 Physical Layer

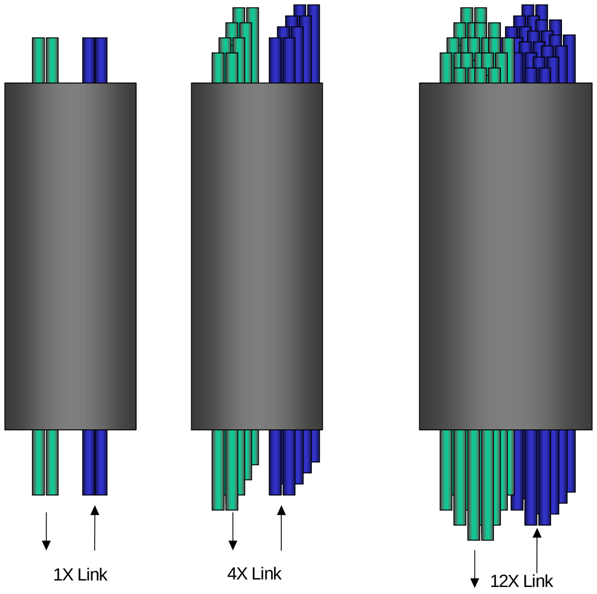
InfiniBand is a comprehensive architecture that defines both electrical and mechanical characteristics for the system. These include cables and receptacles for fiber and copper media, backplane connectors, and hot swap characteristics.

Figure 6. InfiniBand Architecture Specification v1.0, Sample Connector - Mechanical Characteristics



InfiniBand defines three link speeds at the physical layer, 1X, 4X, 12X. Each individual link is a four wire serial differential connection (two wires in each direction) that provide a full duplex connection at 2.5 Gb/s. These links are illustrated in Figure 7, "InfiniBand Physical Link".

Figure 7. InfiniBand Physical Link



The data rates and pin counts for these links are shown in Table 3, “InfiniBand Link Rates”.

Table 3. InfiniBand Link Rates

InfiniBand Link	Signal Count	Signalling Rate	Data Rate	Fully Duplexed Data Rate
1X	4	2.5 Gb/s	2.0 Gb/s	4.0 Gb/s
4X	16	10 Gb/s	8 Gb/s	16.0 Gb/s
12X	48	30 Gb/s	24 Gb/s	48.0 Gb/s

Note: The bandwidth of an InfiniBand 1X link is 2.5 Gb/s. The actual raw data bandwidth is 2.0 Gb/s (data is 8b/10b encoded). Due to the link being bi-directional, the aggregate bandwidth with respect to a bus is 4 Gb/s. Most products are multi-port designs where the aggregate system I/O bandwidth will be additive.

InfiniBand defines multiple connectors for “out of the box” communications. Both, fiber and copper cable connectors as well as a backplane connector for rack-mounted systems, are defined.

4.1.2 Link Layer

The link layer (along with the transport layer) is the heart of the InfiniBand Architecture. The link layer encompasses packet layout, point-to-point link operations, and switching within a local subnet.

- Packets

  There are two types of packets within the link layer, management and data packets. Management packets are used for link configuration and maintenance. Device information, such as Virtual Lane support is determined with management packets. Data packets carry up to 4k bytes of a transaction payload.

- Switching
  Within a subnet, packet forwarding and switching is handled at the link layer. All devices within a subnet have a 16 bit Local ID (LID) assigned by the Subnet Manager. All packets sent within a subnet use the LID for addressing. Link Level switching forwards packets to the device specified by a Destination LID within a Local Route Header (LRH) in the packet. The LRH is present in all packets.

- QoS

  QoS is supported by InfiniBand through Virtual Lanes (VL). These VLs are separate logical communication links which share a single physical link. Each link can support up to 15 standard VLs and one management lane (VL 15). VL15 is the highest priority and VL0 is the lowest. Management packets use VL15 exclusively. Each device must support a minimum of VL0 and VL15 while other VLs are optional.

QoS is supported by InfiniBand through Virtual Lanes (VL). These VLs are separate logical communication links which share a single physical link.

  As a packet traverses the subnet, a Service Level (SL) is defined to ensure its QoS level. Each link along a path can have a different VL, and the SL provides each link a desired priority of communication. Each switch/router has a SL to VL mapping table that is set by the subnet manager to keep the proper priority with the number of VLs supported on each link. Therefore, the InfiniBand Architecture can ensure end-to-end QoS through switches, routers and over the long haul.

- Credit Based Flow Control

  Flow control is used to manage data flow between two point-to-point links. Flow control is handled on a per VL basis allowing separate virtual fabrics to maintain communication utilizing the same physical media. Each receiving end of a link supplies credits to the sending device on the link to specify the amount of data that can be received without loss of data. Credit passing between each device is managed by a dedicated link packet to update the number of data packets the receiver can accept. Data is not transmitted unless the receiver advertises credits indicating receive buffer space is available.

- Data integrity

  At the link level there are two CRCs per packet, Variant CRC (VCRC) and Invariant CRC (ICRC) that ensure data integrity. The 16 bit VCRC includes all fields in the packet and is recalculated at each hop. The 32 bit ICRC covers only the fields that do not change from hop to hop. The VCRC provides link level data integrity between two hops and the ICRC provides end-to-end data integrity. In a protocol like ethernet which defines only a single CRC, an error can be introduced within a device which then recalculates the CRC. The check at the next hop would reveal a valid CRC even though the data has been corrupted. InfiniBand includes the ICRC so that when a bit error is introduced, the error will always be detected.

4.1.3 Network Layer

The network layer handles routing of packets from one subnet to another (within a subnet, the network layer is not required). Packets that are sent between subnets contain a Global Route Header (GRH). The GRH contains the 128 bit IPv6 address for the source and destination of the packet. The packets are forwarded between subnets through a router based on each device's 64 bit globally unique ID (GUID). The router modifies the LRH with the proper local address within each subnet. Therefore the last router in the path replaces the LID in the LRH with the LID of the destination port. Within the network layer, InfiniBand packets do not require the network layer information and header overhead when used within a single subnet (which is a likely scenario for Infiniband system area networks).

4.1.4 Transport Layer

The transport layer is responsible for in-order packet delivery, partitioning, channel multiplexing and transport services (reliable connection, reliable datagram, unreliable connection, unreliable datagram, raw datagram). The transport layer also handles transaction data segmentation when sending, and reassembly when receiving. Based on the Maximum Transfer Unit (MTU) of the path, the transport layer divides the data into packets of the proper size. The receiver reassembles the packets based on a Base Transport Header (BTH) that contains the destination queue pair and packet sequence number. The receiver acknowledges the packets and the sender receives these acknowledgements as well as updates the completion queue with the status of the operation. Infiniband Architecture offers a significant improvement for the transport layer: all functions are implemented in hardware.

InfiniBand specifies multiple transport services for data reliability. Table 4, “Support Services” describes each of the supported services. For a given queue pair, one transport level is used.

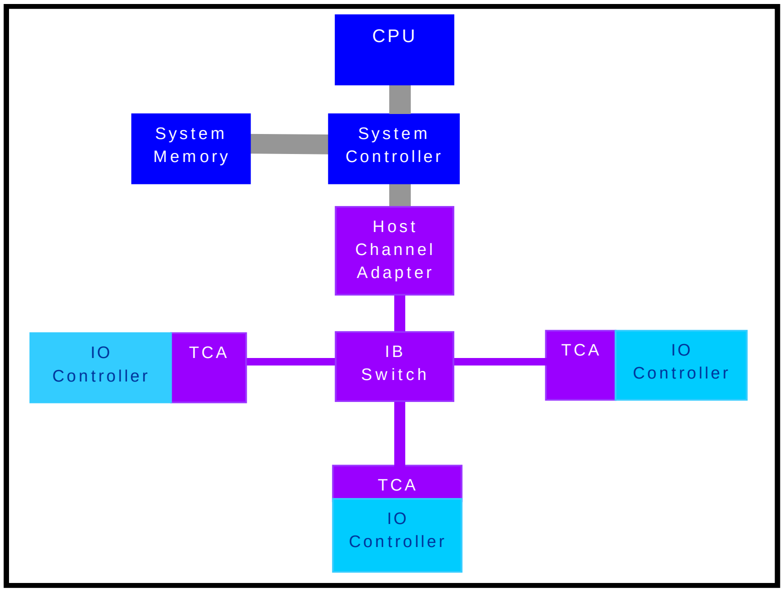
Table 4. Support Services

Class of Service	Description
Reliable Connection	acknowledged - connection oriented
Reliable Datagram	acknowledged - multiplexed
Unreliable Connection	unacknowledged - connection oriented
Unreliable Datagram	unacknowledged - connectionless
Raw Datagram	unacknowledged - connectionless

4.2 InfiniBand Elements

The InfiniBand architecture defines multiple devices for system communication: a channel adapter, switch, router, and a subnet manager. Within a subnet, there must be at least one channel adapter for each endnode and a subnet manager to set up and maintain the link. All channel adapters and switches must contain a Subnet Management Agent (SMA) required for handling communication with the subnet manager.

Figure 8. InfiniBand Architecture



4.2.1 Channel Adapters

A channel adapter connects InfiniBand to other devices. There are two types of channel adapters, a Host Channel Adapter (HCA) and a Target Channel Adapter (TCA).

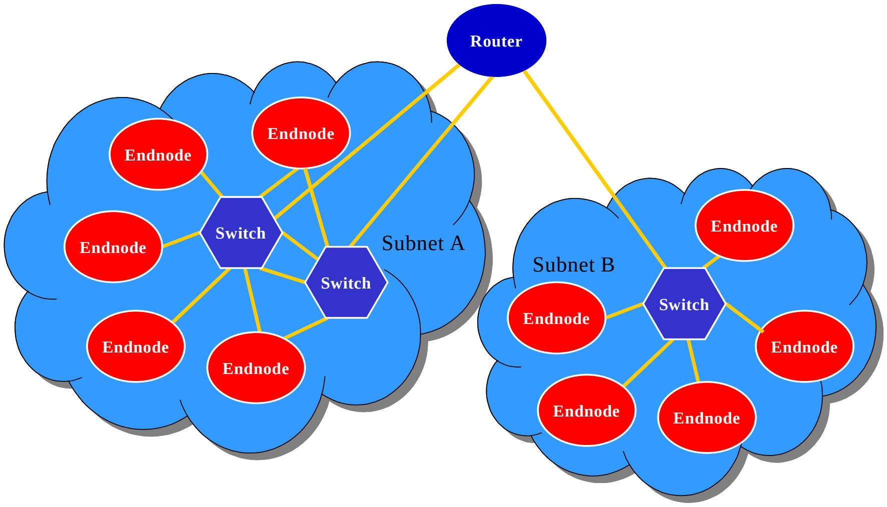
An HCA provides an interface to a host device and supports all software Verbs defined by InfiniBand. Verbs are an abstract representation which defines the required interface between the client software and the functions of the HCA. Verbs do not specify the application programming interface (API) for the operating system, but define the operation for OS vendors to develop a usable API.

A TCA provides the connection to an I/O device from InfiniBand with a subset of HCA features necessary for specific operations of each device.

4.2.2 Switch

Switches are the fundamental components of an InfiniBand fabric. A switch contains more than one InfiniBand port and forwards packets from one of its port to another based on the LID contained within the layer two Local Route Header. Other than management packets, a switch does not consume or generate packets. Like a channel adapter, switches are required to implement a SMA to respond to Subnet Management Packets. Switches can be configured to forward either unicast packets (to a single location) or multicast packets (addressed to multiple devices).

Figure 9. InfiniBand Components



4.2.3 Router

InfiniBand routers forward packets from one subnet to another without consuming or generating packets. Unlike a switch, a router reads the Global Route Header to forward the packet based on its IPv6 network layer address. The router rebuilds each packet with the proper LID on the next subnet.

4.2.4 Subnet Manager

The subnet manager configures the local subnet and ensures its continued operation. There must be at least one subnet manager present in the subnet to manage all switch and router setups and for subnet reconfiguration when a link goes down or a new link comes up. The subnet manager can be within any of the devices on the subnet. The Subnet Manager communicates to devices on the subnet through each dedicated SMA (required by each InfiniBand component).

There can be multiple subnet managers residing in a subnet as long as only one is active. Non-active subnet managers (Standby Subnet Managers) keep copies of the active subnet manager's forwarding information and verify that the active subnet manager is operational. If an active subnet manager goes down, a standby subnet manager will take over responsibilities to ensure the fabric does not go down with it.

4.3 Management Infrastructure

The InfiniBand architecture defines two methods of system management for handling all subnet bringup, maintenance, and general service functions associated with the devices in the subnet. Each method has a dedicated queue pair (QP) that is supported by all devices on the subnet to distinguish management traffic from all other traffic.

4.3.1 Subnet Management

The first method, subnet management, is handled by a Subnet Manager (SM). There must be at least one SM within a subnet to handle configuration and maintenance. These responsibilities include LID assignment, SL to VL mapping, Link bringup and teardown, and Link Failover.

All subnet management uses QP0 and is handled exclusively on a high priority virtual lane (VL15) to ensure the highest priority within the subnet.

All subnet management uses QP0 and is handled exclusively on a high priority virtual lane (VL15) to ensure the highest priority within the subnet. Subnet Management Packets (SMPs - pronounced "sumps") are the only packets allowed on QP0 and VL15. This VL uses the Unreliable Datagram transport service and does not follow the same flow control restriction as other VLs on the links. Subnet management information is passed through the subnet ahead of all other traffic on a link.

The subnet manager simplifies the responsibilities of client software by taking all configuration requirements and handling them in the background.

4.3.2 General Services

The second method defined by InfiniBand is the General Services Interface (GSI). The GSI handles functions such as chassis management, out-of-band I/O operations, and other functions not associated with the subnet manager. These functions do not have the same high priority needs as subnet management, therefore the GSI management packets (GMPs - pronounced "gumps") do not use the high priority virtual lane, VL15. All GSI commands use QP1 and must follow the flow control requirements of other data links.

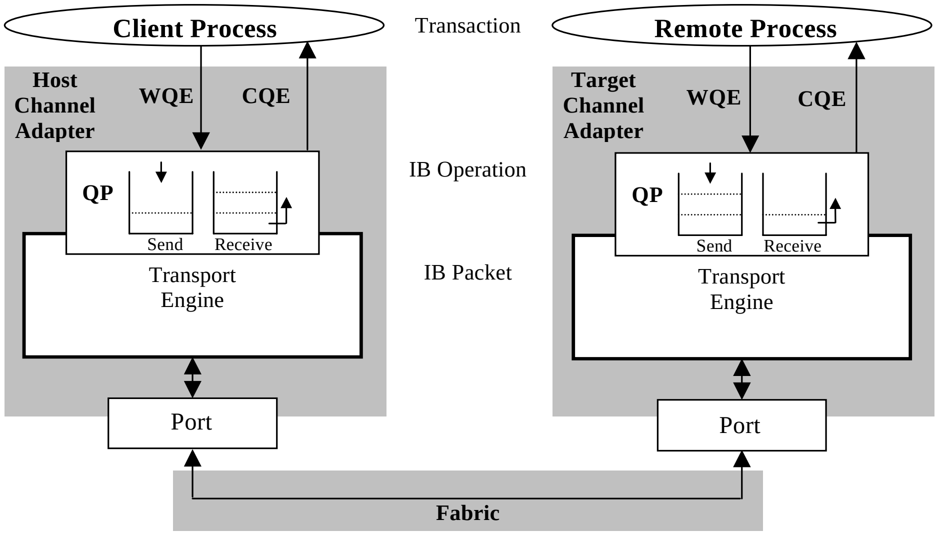
4.4 InfiniBand Support for the Virtual Interface Architecture (VIA)

The Virtual Interface Architecture is a distributed messaging technology that is both hardware independent and compatible with current network interconnects. The architecture provides an API that can be utilized to provide high-speed and low-latency communications among peers in clustered applications.

InfiniBand was developed with the VIA architecture in mind. InfiniBand off-loads traffic control from the software client through the use of execution queues. These queues, called work queues, are initiated by the client, and then left for InfiniBand to manage. For each communication channel between devices, a Work Queue Pair (WQP - send and receive queue) is assigned at each end. The client places a transaction into the work queue (Work Queue Entry - WQE, pronounced "wookie"), which is then processed by the channel adapter from the send queue and sent out to the remote device. When the remote device responds, the channel adapter returns status to the client through a completion queue or event.

The client can post multiple WQEs, and the channel adapter's hardware will handle each of the communication requests. The channel adapter then generates a Completion Queue Entry (CQE) to provide status for each WQE in the proper prioritized order. This allows the client to continue with other activities while the transactions are being processed.

Figure 10. InfiniBand Communication Stack



4.5 Realizing the Full Potential of Blade Computing

To fully realize the TCO (total cost of ownership) benefits of blade based server computing the blade technology must deliver at a minimum the following core functions: scalability, fault tolerance, hot swap, QoS, clustering, support for I/O connectivity (both memory and message semantics), reliability, redundancy, active stand-bye for failover, interconnect manageability, and error detection. It is fairly straight forward to understand why the IT Manager would require these attributes in every new server platform that is deployed. As outlined in this paper, all of these attributes are delivered natively within the InfiniBand Architecture and they will truly unleash the full potential that blade computing promises. The white paper, Realizing the Full Potential of Server, Switch & I/O Blades with InfiniBand Architecture, document # 2009WP, explores the attributes and TCO benefits in detail.

5.0 Summary

The collective effort of industry leaders has successfully transitioned InfiniBand from technology demonstrations to the first real product deployments. The IBTA currently has over 220 members, the specification is mature, multiple vendors have publicly shown working silicon and systems, and interoperatbility between InfiniBand silicon vendors has been demonstrated. The benefits of the InfiniBand architecture are clear and include: support for RAS (Reliability, Availability, Serviceability), a fabric that works both in-the-box and enables *Bandwidth Out of the Box*, and scalability well into the future.

The IBTA has a vision to improve and simplify the data center through InfiniBand technology and the fabric it creates as an interconnect for servers, communications and storage. Imagine a data center made of servers that are all closely clustered together. These servers have only processors and memory that connect to storage and communications via InfiniBand ports. This allows for much greater processor performance via Virtual Interface clustering, greater processor and memory density (as most of the peripheral devices have moved out of the server racks), and much greater (InfiniBand) I/O bandwidth. Best of all, all these improvements are based upon an architecture designed for RAS. Now imagine all this wrapped around the flexibility of upgrading your existing servers, though PCI (upgraded by 3GIO) and through InfiniBand Server Blades. The rapid adoption of Infiniband continues and its being welcomed as more businesses and consumers utilize the Internet more often and at higher bandwidths.

6.0 About Mellanox

Mellanox is the leading supplier of InfiniBand semiconductors, providing complete solutions including switches, host channel adapters, and target channel adapters to the server, communications, data storage, and embedded markets. Mellanox Technologies has delivered more than 100,000 InfiniBand ports over two generations of 10 Gb/sec InfiniBand devices including the InfiniBridge, InfiniScale and InfiniHost devices. Mellanox InfiniBand interconnect solutions today provide over eight times the performance of Ethernet, and over three times the performance of proprietary interconnects. The company has strong backing from corporate investors including Dell, IBM, Intel Capital, Quanta Computers, Sun Microsystems, and Vitesse as well as, strong venture backing from Bessemer Venture Partners, Raza Venture Management, Sequoia Capital, US Venture Partners, and others. The company has major offices located in Santa Clara, CA, Yokneam and Tel Aviv Israel. For more information on Mellanox, visit www.mellanox.com.