



# **An Introduction to InfiniBand**

**Bringing I/O up to speed**

Copyright 2001, JNI Corporation

InfiniBand™ – a play on the words “infinite bandwidth” – is an advanced input/output (I/O) architecture designed to increase the communication speed between CPUs, devices within servers and subsystems located throughout a network. Unlike any previous I/O architecture, InfiniBand extends its feature set outside the server to devices on the network. This paper will introduce InfiniBand and describe its unique benefits and components.

*CPU advancement has outpaced the I/O bus, leading to serious performance bottlenecks*

### **The I/O Bottleneck Problem: The Need for Speed**

In 1967, Gene Amdahl argued that the sequential nature of computers limits the speed at which they can solve problems, regardless of how many processors are working on a solution. Known as *Amdahl's law*, the principle demonstrates that a balance must exist between CPU speed, memory bandwidth and I/O performance to achieve effective computing results. Over the last decade, the speed of CPUs has grown much faster than the capabilities of the I/O bus, presenting a serious performance mismatch and a bottleneck for servers.

### The Limits of the PCI Bus

The leading I/O bus architecture, Peripheral Component Interconnect (PCI), was introduced in 1992 and has become the standard bus architecture for servers. Today's exploding need for greater I/O performance demanded by the Internet, e-commerce, symmetric multiprocessing, server clustering and remote storage has outpaced the ability of the PCI bus. With only one major upgrade widely implemented in the last decade, the PCI bus grew from 32-bit and 33 megahertz (MHz) to 64-bit and 66 MHz. The PCI-X protocol, a 64-bit solution that can perform at speeds up to 133 MHz, was developed in the late 1990s to improve the PCI standard; however, many experts feel the shelf life of PCI-X will be limited to a few years.

With CPU performance surpassing 1 gigahertz (GHz) and network bandwidth exceeding 1 gigabit per second (Gb/s), there is a critical need for an I/O architecture that meets and exceeds the performance capabilities of processors and networks. The PCI bus severely diminishes a processor's ability to push and retrieve data to external devices quickly. In addition, a slow I/O bus contributes to bottlenecks inside the server.

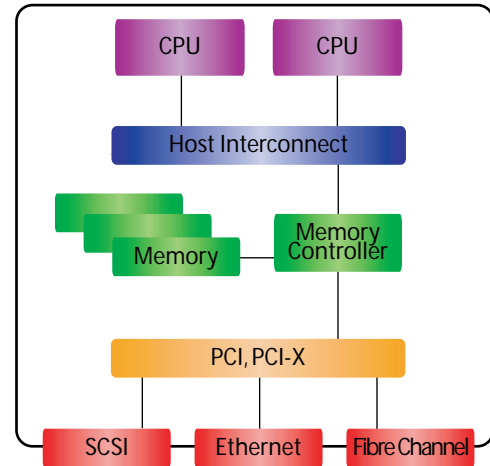
Other PCI concerns include:

**Bus Sharing:** The PCI bus is like a single-lane highway with a multiple-lane onramp. The bus requires all devices plugged into it to share a single communication path to memory and the CPU. Only one device can communicate at a time, competing for I/O resources, limiting overall system performance, reducing I/O bandwidth and decreasing CPU utilization.

**Bus Speed:** The PCI and PCI-X bus architectures have reached a performance limit, beyond which the technologies can go no further. The theoretical performance maximum for PCI is 528 megabytes per second (MB/s) and 1064 MB/s for PCI-X. Signal integrity and timing issues have stretched PCI-X as far as it can go. Even 1064 MB/s is not sufficient to meet the needs of tomorrow's enterprises.

**Scalability:** Scaling the PCI bus is expensive and limited, requiring bridge chips on the system board. With PCI-X, as the bus frequency increases, the number of cards supported decreases. For example, at 66 MHz, PCI-X supports up to four cards, but at 133 MHz, it supports only a single card.

**Fault Tolerance:** Generally, when a PCI card fails, a server must be taken offline. Because the PCI bus is shared, it cannot isolate faults, thus introducing a single point of failure.



**Example configuration of a PCI or PCI-X shared bus**

### The Solution: InfiniBand

*InfiniBand was developed from scratch by leading technology corporations to bring advanced performance to the enterprise*

To meet the I/O needs of the enterprise, many leading computer and technology manufacturers came together to form an open standard for moving high volumes of data between processors and I/O devices, known as the *InfiniBand Architecture (IBA)*. Released in the year 2000, major technology leaders including Compaq, Dell, HP, IBM, Intel, Microsoft and Sun co-developed the IBA. IDC estimates that 80% of all servers will be InfiniBand-capable by the year 2005 (May 2001).



Designed from the ground up as a new universal I/O standard, the IBA was developed to connect servers with remote storage, networking devices and other servers, as well as for use inside servers for interprocessor communications. The IBA standard was also designed to eventually replace the PCI bus, yet provide a smooth migration path from PCI shared bus configurations.

The IBA offers many unique benefits, including:

- Bandwidths that are an order of magnitude greater than existing I/O media capabilities.
- An open and industry-inclusive standard designed to benefit component vendors, systems suppliers (from storage to networking) and end-users.
- Improved connection flexibility and scalability because storage and I/O are separated from the processor and the memory.
- Improved reliability via redundant links and hot-plug components.
- Reduced cost of ownership via multiple redundant paths between components, thus reducing hardware expenses. This model is “pay as you grow,” allowing enterprises to add capacity without impacting operations.
- Offloaded communications processing from the OS and CPU, eliminating traditional communications overhead.
- Wide access to a variety of storage systems.
- Simultaneous device communication, rather than waiting for other devices to complete their communication (as seen in shared bus technologies).
- Built-in security, quality of service and improved usability.
- Support for Internet Protocol version 6 (IPv6) for effective communications between IBA fabrics and the Internet or intranets.
- Fewer and better-managed system interrupts.
- Support for up to 64,000 addressable devices.
- Support for copper cable, optical fiber and printed circuit backplanes.

The following table demonstrates some of the IBA advantages:

<b>Feature</b>	<b>InfiniBand Fabric</b>	<b>PCI Bus</b>
Raw Bandwidth	2.5 to 30 billion bits per second	1 billion bits per second
Simultaneous Data Flow (Full-Duplex)	Yes; Talk <i>and</i> listen	No; Talk <i>or</i> listen
Topology	Switched fabric, no device sharing	Shared bus, devices must be shared up to a max. of approx. 4 per bus
Distance Between Points	Up to 1000 meters or more	Up to 1 meter per bus
Pin Count	Low	High
Number of End Points	Many	Few

## InfiniBand Basics

*InfiniBand is a point-to-point switched I/O fabric architecture that increases its bandwidth as switches are added*

InfiniBand is a point-to-point, switched I/O fabric architecture. Each end point, or *node*, can vary from an inexpensive single SCSI chip or Ethernet adapter to complex host systems. *Point-to-point* means that each communication link extends between only two devices. Both devices at each end of a link have full and exclusive access to the communication path. To go beyond a point and traverse the network, switches come into play. By adding switches, multiple points can be interconnected to create a fabric. As more switches are added to a network, aggregated bandwidth of the fabric increases. By adding multiple paths between devices, switches also provide a greater level of redundancy.

There are five primary components that make up an InfiniBand fabric:

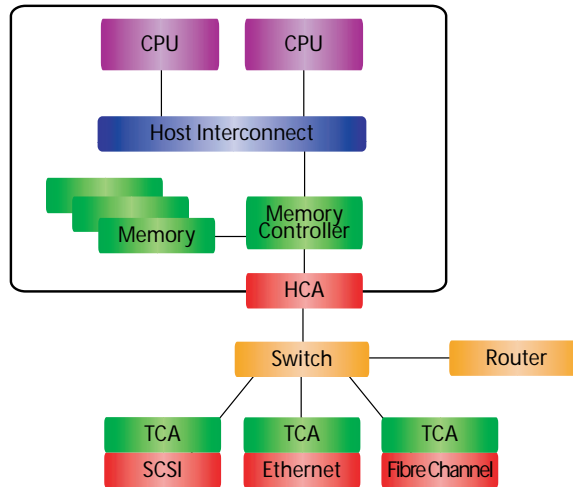
**Host Channel Adapter (HCA):** An HCA is an interface that resides within a server and communicates directly with the server's memory and processor as well as the IBA fabric. The HCA guarantees delivery of data, performs advanced memory access and can recover from transmission errors. HCAs can communicate with a target channel adapter or a switch. An HCA can be a PCI to InfiniBand card or it can be integrated on a system motherboard.

**Target Channel Adapter (TCA):** A TCA enables I/O devices, such as disk or tape storage, to be located within the network independent of a host computer. The TCA includes an I/O controller that is specific to its particular device's protocol (i.e.; SCSI, Fibre Channel or Ethernet). TCAs can communicate with an HCA or a switch.

**Switch:** An IBA switch is the virtual equivalent of a major intersection with a traffic cop. The switch allows many HCAs and TCAs to connect to it and handles network traffic. The switch looks at the "local route header" on each packet of data that passes through it and forwards it to the appropriate location. The switch is a critical component of the IBA that offers higher availability, higher aggregate bandwidth, load balancing, data mirroring and much more. A group of switches is referred to as a *fabric*. If a host computer is down, the switch still continues to operate. The switch also frees up servers and other devices by handling network traffic.

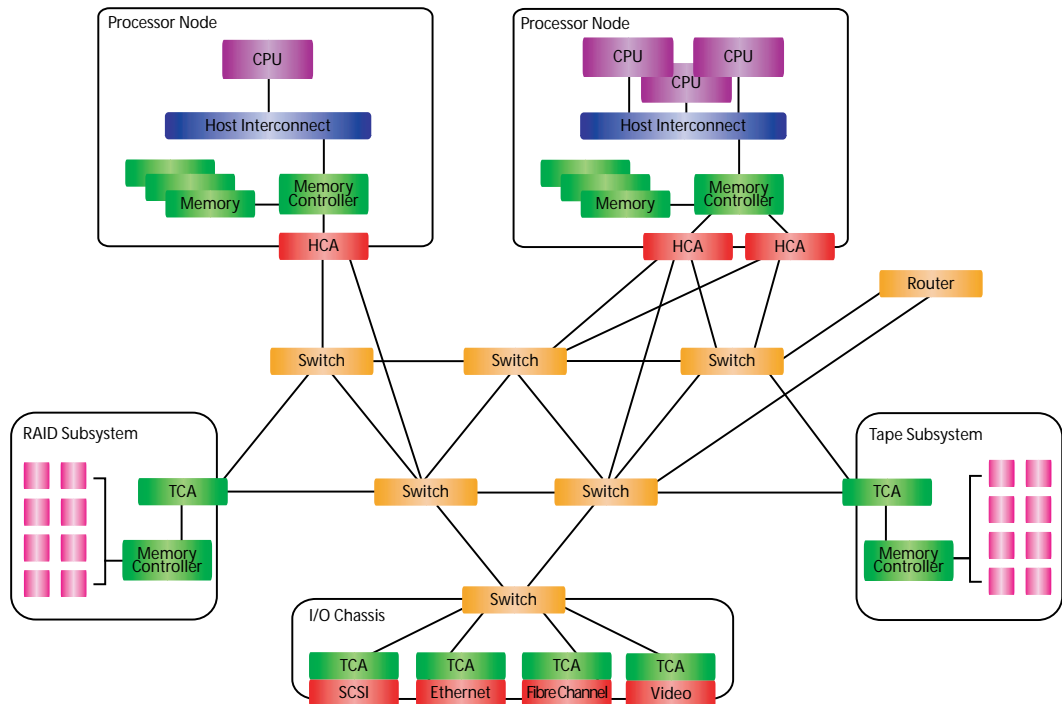
**Router:** A router forwards data packets from a local network (called a *subnet*) to other external subnets. The router reads the "global route header" and forwards packets based on the IPv6 network layer address. The router rebuilds each packet with the proper local address header as it passes it to the new subnet.

**Subnet Manager:** The subnet manager is an application responsible for configuring the local subnet and ensuring its continued operation. Configuration responsibilities include managing switch and router setups and reconfiguring the subnet if a link goes down or a new one is added.



**With InfiniBand, I/O devices that traditionally reside within the server are moved to the network**

InfiniBand essentially separates and distributes I/O devices to locations outside the server. This enables servers to shrink to 1U (each U, or standard unit of measurement, is 1.75 inches high) because servers only require a single InfiniBand HCA card (see above illustration). In the future, server “blades,” servers with only a CPU, system controller, memory and InfiniBand, will be as small as 4 by 6 inches.



**The above illustration is a possible configuration of an InfiniBand fabric**

## InfiniBand Layers

The IBA is comprised of four primary layers that describe communication devices and methodology. The layers are the physical layer, the link layer, the network layer and the transport layer.

### Physical Layer

The InfiniBand physical layer defines the electrical and mechanical characteristics of the IBA, including the cables, connectors and hot-swap characteristics. IBA connectors include fiber, copper and backplane connectors. There are three link speeds specified as 1X, 4X and 12X. The speeds are a function of the pin counts or wires within each cable. With a 1X link cable, there are four wires, two for each direction of communication (read and write). The 4X speed has four times as many pins and wires and the 12X has twelve times as many pins and wires as a 1X link cable.

InfiniBand Link	Pin Count	Signaling Rate	Data Rate	Full-Duplex Data Rate
1X	4	2.5 Gb/s	2 Gb/s	4 Gb/s
4X	16	10 Gb/s	8 Gb/s	16 Gb/s
12X	48	30 Gb/s	24 Gb/s	48 Gb/s

The bandwidth for a 1X InfiniBand link is 2.5 Gb/s, which can achieve an actual raw data bandwidth of 2 Gb/s because 8b/10b data encryption is used on all transmissions, resulting in a 20% performance overhead. Because all links are bidirectional, the aggregate bandwidth can be doubled. Many InfiniBand products have multiple ports, further increasing I/O bandwidth.

### Link Layer

*The link layer defines packet types, switching instructions and data integrity*

The link layer is central to the IBA and includes packet layout, point-to-point link instructions, switching within a local subnet and data integrity. There are two types of packets, management and data. Management packets handle link configurations and maintenance. Data packets carry up to 4 kilobytes of transaction payload. Packet forwarding and switching within a local subnet is also part of the link layer's responsibilities. Every device in a local subnet has a local ID (LID). Packets of data are forwarded to the appropriate LID by reading the local route header found in each packet of data.

Virtual lanes are also part of the link layer. A *virtual lane* is a unique logical communication link that shares a single physical link. Each link can have up to 15 virtual lanes and a management lane. As a packet travels through the subnet, it can be assigned a priority or service level. Higher-priority packets are sent down special virtual lanes ahead of other packets.

The link layer also handles data integrity by including variant and invariant cyclic redundancy checking (CRC) (CRCs are methods that check for errors in data transmitted on a communications link). The variant CRC checks fields that change from point-to-point and the invariant CRC provides end-to-end data integrity.

### Network Layer

The network layer is responsible for routing packets from one subnet to another. The global route header located within a packet includes an IPv6 address for the source and destination of each packet. Using a router, packets are forwarded through different subnets. For single-subnet environments, the network layer information is not used.

### Transport Layer

The transport layer handles the order of packet delivery as well as partitioning, multiplexing and transport services that determine reliable connections.

## **InfiniBand Reliability, Availability and Serviceability**

*InfiniBand offers advanced RAS features not present in the PCI bus*

A key requirement for enterprises is high reliability, availability and serviceability (RAS). InfiniBand represents a significant RAS improvement over the PCI bus.

### Reliability

The basic InfiniBand link connection is comprised of only four signal wires. When compared to the more than 100 signals on a PCI bus, IBA results in a smaller link failure multiplier. InfiniBand connectors also provide a positive seating mechanism that prevents wear common on PCI edge connectors. In addition, the IBA accommodates multiple ports for each I/O unit, enhancing reliability by providing multiple routes to a physical device. Furthermore, multiple CRCs provide greater error detection capabilities.

### Availability

An IBA fabric is inherently redundant, with multiple paths to sources, assuring data delivery. IBA also includes a failover mechanism that allows the network to heal itself if a link fails or is reporting errors. In addition, IBA breaks the one-to-one relationship between server and I/O elements by removing the I/O from the server. With PCI, if a network interface card fails, the entire server-I/O unit is unavailable. IBA has a many-to-many server-to-I/O relationship. Thus, if an I/O device fails, communication simply fails over to another redundant I/O device, saving time, resources and keeping the server online.

### Serviceability

With IBA, all I/O devices are designed to be hot-pluggable, increasing serviceability. If a stand-alone I/O device must be serviced, replacing the device is as simple as unplugging one device and plugging in a new device. In situations where multiple cards reside in a server or I/O chassis, an administrator can simply swap cards while the unit is online.

### **InfiniBand Roadmap**

The first release of InfiniBand products was in mid-2001. This first phase will include InfiniBand cards that plug into existing PCI or PCI-X buses for use in clusters or server-to-I/O interconnects to increase server interconnection performance. The next phase, expected in 2002, will include hybrid servers that have PCI/PCI-X buses and built-in IBA HCAs. The IBA port will bypass the PCI bus and have direct access to the memory controller. This phase will allow enterprises to easily migrate from PCI to IBA. The third phase is expected to begin in 2003 and will include IBA server designs (i.e., server blades), optimized for fabrics.

### **About JNI**

JNI is a leading manufacturer of enterprise-level storage network products and the leading provider of Fibre Channel-based HBAs for Solaris servers. JNI offers a broad line of FibreStar HBAs, Emerald ASICs, as well as DriverSuite™ and EZ Fibre software for storage networks. JNI's PCI products operate on Solaris, Windows 2000, Windows NT, HP-UX, AIX, Novell, Linux and Mac OS systems. JNI's SBus products run on Solaris. Customers include Amdahl, Avid, Chaparral, Compaq StorageWorks, Consan, EMC, Eurologic, Hewlett-Packard, Hitachi Data Systems, IBM, LSI Logic, McData, StorageTek and Sun Microsystems. JNI is headquartered in San Diego, California, with offices throughout the U.S. and in Munich (München), Germany. For more information, visit JNI on the Internet at [www.jni.com](http://www.jni.com).