White Paper

Understanding PCI Bus, PCI-Express and InfiniBand Architecture

1.0 Overview

There is some confusion in the market place concerning the replacement of the PCI Bus (Peripheral Components Interface) by either the InfiniBand[SM] Architecture (IBA), PCI-Express (formerly 3GIO - Third Generation I/O), or upgrades to the PCI bus itself. Initially the InfiniBand architecture was viewed as the replacement for the PCI bus, yet the PCISIG (PCI Special Interest Group) recently announced that PCI-Express would replace the PCI bus. To make matters more confusing the PCISIG also recently introduced a draft specification for PCIX 266 (also called PCI-X DDR - Double Data Rate) as the next upgrade to the PCI bus. So which is correct? Actually none of these technologies "replace" the PCI bus but rather provide upgrade paths with different levels of backwards compatibility and new capabilities. It is important to make a distinction between these technologies to help clear the confusion. Much of this confusion stems from viewing InfiniBand as strictly a chip-to-chip or local bus connection such as PCI and PCI-Express. In reality InfiniBand is much more than a local bus and thus it does not make sense to compare PCI or PCI-Express to InfiniBand. This paper will discuss how these technologies interact and complement each other.

2.0 Contrasting The Architectures

The InfiniBand Architecture is substantially different than local bus architectures such as PCI and PCI-Express. Because both PCI-Express and PCI are strictly local interconnect technologies, it is much more natural to consider PCI-Express as a "replacement" for the PCI bus. In reality even this is an over-simplification as PCI-Express offers an *upgrade* path rather than replacement. A separate PCI-Express connector has been defined to augment the PCI connector. Over time PCI-Express may effectively replace the PCI bus in PCs however history has proven that this may take a decade or more. In some sense PCI-Express may finally force the legacy ISA slots out of PCs (ten years after the addition of PCI slots) and thus it can be viewed as the replacement for ISA

slots. The following table highlights the key features that distinguish InfiniBand from simple local bus architectures such as PCI and PCI-Express:

Feature	InfiniBand	PCI/PCI-Express
I/O Sharing	Yes	No
System Hierarchy	Channel Based Fabric	Memory Mapped Tree
Kernel Bypass	Memory Protection & Address Translation	No
System Software Interface	Transport level connections	Low Level load/store

As can be seen from the table above, the InfiniBand architecture supports features that make it substantially more powerful than a simple local bus point of attachment. Thus it does not make sense to classify InfiniBand as a local bus like these other technologies. InfiniBand delivers reliable, transport level connections in hardware and supports both message passing and memory semantics on the wire. In addition InfiniBand's I/O sharing capabilities make it possible to package computers in fundamentally new ways. Utilizing InfiniBand as an I/O connection to external RAID arrays or JBODs (just a bunch of disks) enables storage to be shared by many servers connected via the InfiniBand fabric. The PCI model requires that each server include a PCI card to connect to clustering, communications, and storage networks constraining the system configuration to be a hierarchy of many I/O nodes with a single host. InfiniBand supports a system configuration of many-to-many where all I/O resources are shared by all the servers.

3.0 The PCI Bus

The PCI bus was developed in the early 1990's by a group of companies with the goal to advance the interface allowing OEM's or users to upgrade the I/O (Input-Output) of personal computers. The PCI bus has proven a huge success and has been adopted in almost every PC and Server since.

The latest advancement of the PCI bus is PCI-X. PCI-X is a 64-bit parallel interface that runs at 133 MHz enabling 1GB/s (8Gb/s) of bandwidth. Though other advancements are in the works, including DDR, for the PCI bus, they are perceived as falling short. They are too expensive (too many pins in the 64-bit versions) for the PC industry to implement in the mass volumes of PCs and that they don't offer sufficient bandwidth and advanced feature set required for the servers of the future.

Many would argue that there is no need to advance the bandwidth of PCI on PCs since few I/O cards are taxing the 250 to 500 MB/s bandwidth that is currently available. This is not the case for Servers. High performance servers are frequently equipped with clustering, communication and storage I/O cards, that together tax the bandwidth of the PCI-X bus. Another key limitation of PCI-X is that it can support only one slot per controller, which means that multiple controllers (and their expense) are needed on each server.

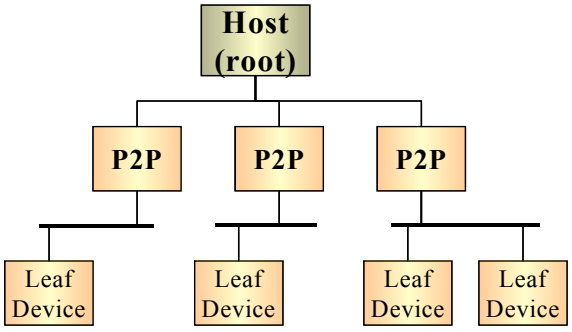
Memory Mapped Versus Channel Based Architectures:

The PCI bus essentially defines a low level interface between a host CPU and peripheral devices. The PCI architecture utilizes PCI to PCI (P2P) bridges to extend the number of devices that can be supported on the bus. By definition a system built from P2P bridges forms a hierarchical tree with a primary bus extending to multiple secondary PCI bus segments. Across all of these PCI bus segments there is a single physical memory map and a given memory address uniquely specifies an individual bus segment and device on this segment.

This architecture fundamentally defines a single global view of resources, which works well to build systems based on a single master (host CPU) sitting at the top of the hierarchy controlling multiple slaves on peripheral bus segments. In this case "master" and "slave" refer to the initialization and control of the devices in the system rather than to whether an individual slave device is actually capable of initiating a bus transaction (i.e. acting as a PCI bus master). The PCI architecture does not lend itself well to applications with multiple host CPUs each of which operates as a master.

InfiniBand is fundamentally different as devices are designed to operate as peers with channels (queue pairs or QPs) connecting them. These channels may each have their own independent Virtual and Physical Address spaces. This allows any node to be an initiator to any other node throughout the fabric. The InfiniBand architecture provides a large QP space to support up to 16 million channels, and every channel is capable of delivering a reliable, low latency, transport level connection. Such a channel based architecture functions ideally as a multi-protocol I/O fabric by providing fine grained quality of service to each channel which allows application level optimizations.

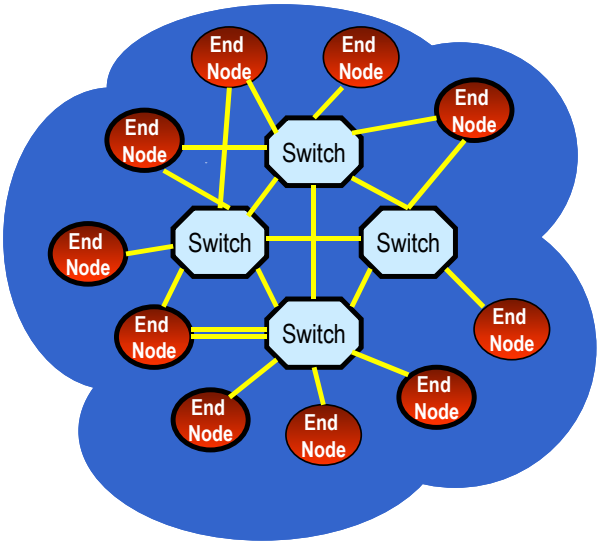
3.1 Hierarchical PCI Tree Structure vs Peer-to-Peer Fabric



As mentioned the PCI architecture utilizes a flat unified memory architecture and use of PCI to PCI bridges for scaling means that systems must be constructed using a hierarchical tree structure. The PCI architecture requires this as a PCI to PCI bridge has a primary (or upstream) bus and a secondary (or downstream) bus. At the top of a PCI system there is a single master (or root) host PCI segments and at leaf PCI segments at endpoints. Standard PCI boot software understands this hierarchy and discovers and numbers PCI bus segments accordingly. This type of hierarchy is good for systems with a single host at the root communicating with leaf I/O nodes. The "single" host may indeed consist of multiple CPU's configured as an SMP (symmetrical multi processor), but still discovery and communication to the leaf devices is from a single root PCI bus segment.

By contrast, the InfiniBand architecture supports a true peer-to-peer fabric where devices on the fabric are identical and there is no single root of the datapath. InfiniBand architecture does indeed distinguish between Host and Target Channel Adapters however essentially this distinction is a practical one between host CPU and I/O connections. In any case an InfiniBand fabric can have many of both types of channel adapters (or end nodes) each communicating with all others as peers. This type of symmetrical fabric with no notion of upstream or downstream is ideal for building peer to peer applications and clustering, I/O sharing, fault tolerance, etc.

InfiniBand enables peer-to-peer communications using a channel based fabric hierarchy rather than the flat unified memory space that PCI and PCI-Express use. InfiniBand enables a much broader class of scalable clustering and distributed computing applications than can be supported by systems built on the single flat memory space architecture of PCI and PCI-Express. Furthermore InfiniBand offers significant performance benefits by implementing in hardware core functions normally delivered by the kernel of the operating system. InfiniBand also delivers transport level connections and memory protection (across multiple CPUs) enabling a broader class of applications than a load/store interface can support.



3.2 PCI-X 266

This is touted as a straight forward technology. The specification uses the same 133MHz clock but simply clocks the data on both rising and falling edge of the clock to double the effective bandwidth to 266 MHz. PCI-X 266 uses much the same technology as memory and system ven-

dors used to implement DDR SDRAM. Proponents estimate that this technology can be implemented in 2002. While, others raise questions about how robust the backwards compatibility will be.

3.3 Overcoming the Limitations of PCI:

Thus there are three PCI limitations or issues that need to be resolved:

1. Advance the bandwidth of PCI at equal or lower cost than 32-bit versions of PCI on PC's (although some question whether more bandwidth is required for PCs)
2. Find a mechanism that scales bandwidth beyond PCI-X's 8Gb/s (greater b/w for servers)
3. Find a mechanism that advances server I/O capabilities in respect to I/O sharing and relationship hierarchies (overcome the one slot per PCI-X limit and allow many-to many relations)

If a way could be found to advance the bandwidth of PCI, while keeping the costs low and maintaining software compatibility, then the industry would have the means to solve the first two issues. Intel believes that this solution is PCI-Express. Others believe that PCI-X 266 and 532 (QDR or Quad Data Rate) may be a more straightforward route to enhancing the PCI bus. Both PCI-Express and the PCI-X extensions have pros and cons and it is likely that both technologies will be deployed to extend the PCI local bus. For issues three expanding the I/O capabilities is no simple matter and InfiniBand goes well beyond the just expanding I/O capabilities. Sharing I/O is only one of the reasons why Dell, IBM, HP, Intel, Microsoft, Sun and many others created InfiniBand.

4.0 PCI-Express

This past August at the Intel Developers Forum it was announced that PCI-Express would be developed as a new local bus (chip-to-chip interface) and used as a way to UPGRADE the PCI bus. The PCI-Express architecture is being defined by the Arapahoe Working Group and eventually when the specification is finalized it will be turned over to the general PCISIG organization for administration.

PCI-Express is defined as serial I/O point-to-point interconnect that uses dual simplex differential pairs to communicate. The intent of this serial interconnect is to establish very high bandwidth communication over a few pins, versus low bandwidth communication over many pins (aka: 64-bit PCI). It also leverages the PCI programming model to preserve customer investments and to facilitate industry migration. This way PCI bandwidth can be economically upgraded without consuming a great number of pins while preserving software backwards compatibility.

PCI Bridging over Switches

It is possible to implement a system that looks like PCI using serial links and switches. Each serial point-to-point link can act like a PCI to PCI bridge. However in order to make this transparent to system software requires that the switches preserve the PCI notion of a hierarchical tree structure and the ideas of upstream (towards the root) and downstream (towards leaf segments). The advantage of such a system is that it can be made completely transparent to system software. The disadvantage is that it breaks the normal symmetry of a switched fabric because a switch port needs to be defined as “upstream” while others become “downstream”. This break in the symmetry allows the switches to behave like a tree hierarchy, but also requires different behavior and features for an upstream port then for a downstream port. Furthermore where human interaction is involved it is necessary to physically key the upstream ports differently than downstream ports (like USB connections), otherwise inevitably the connections will be made incorrectly. Though constructed from switches the logical architecture is still a tree and is best suited for a single host CPU controlling I/O devices. Therefore, it is not well suited for peer-to-peer communications.

The stated goal of PCI-Express is to provide:

- A local bus for chip-to-chip interconnects
- A method to upgrade PCI slot performance at lower costs

PCI-Express does not meet the needs, nor is it intended to be an external wire protocol. As previously stated, PCI-Express, since it’s a local bus, does not implement the I/O sharing, transport level communication, kernel bypass support, memory protection and other higher level functions to support a large external fabric. Simply stated PCI-Express supports layers 1 & 2 of the OSI model while InfiniBand supports, in hardware, all of layers 1 through 4. Indeed the backers of the PCI-Express specification have stated it is not intended to create a cabled coherent interconnect for memory or clustering (aka: InfiniBand).

(It is important to note that there are two other emerging and open chip-to-chip interconnects standards in the market: namely HyperTransport and Rapid I/O. Both of these standards are designed as chip-to-chip interconnects only and both have defined specifications and products that are

available in the market today. The PCI-Express specification is not expected to be released till Q2 of 2002 with first products available in late 2003 or early 2004.

5.0 The InfiniBand Architecture (IBA)

In August of 1999 all of the major server vendors and Microsoft combined to develop a new I/O "fabric" for Servers within a data center and developed the InfiniBand Architecture. InfiniBand was developed to create a common fabric for Data Centers or the first 50 feet of the Internet. Many touted InfiniBand as a replacement for the PCI bus. Why was this?

As previously stated, the PCI bus was being taxed for bandwidth due to the use of InterProcess Communication (IPC) or Clustering cards, Ethernet Cards, Fibre Channel cards and others cards all in a single server. It was very easy to tie the functionality and the limitations of PCI to InfiniBand. So many members of the IBTA and the IBTA itself promoted InfiniBand as a replacement for PCI. Why is that? Because PCI was the means by which servers enabled I/O, and InfiniBand with its greater bandwidth is intended to be the predominate I/O of servers for the future. But, InfiniBand is not the actual physical replacement of a PCI slot. Initially, InfiniBand will be enabled through a PCI slots and much later though a direct connection to system logic.

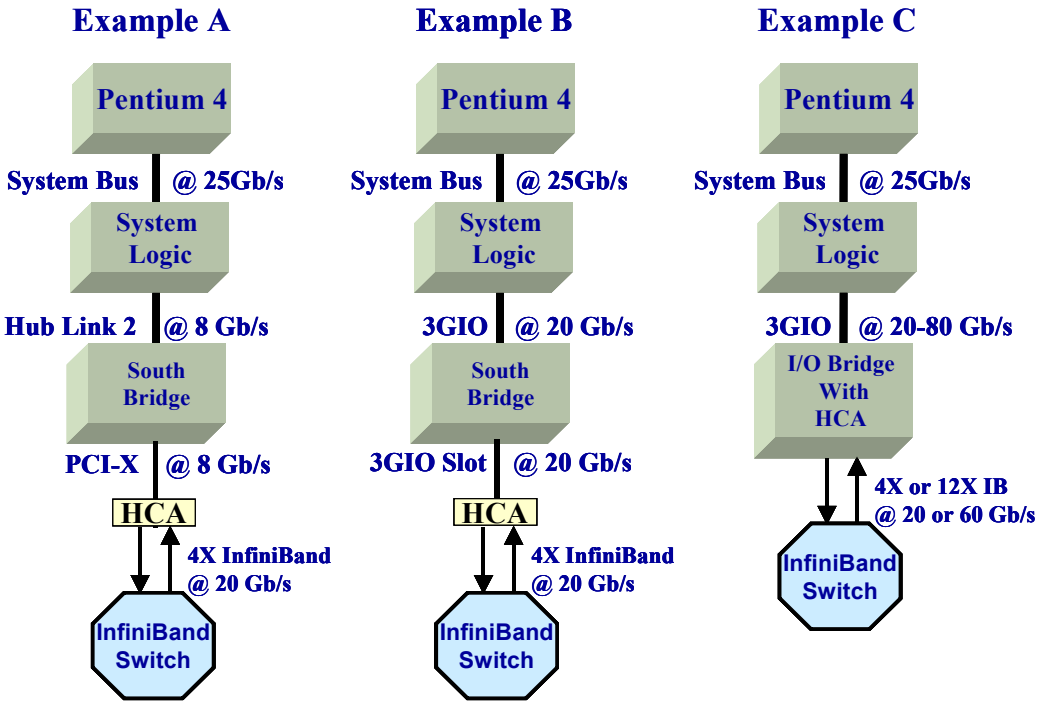
6.0 System Design Impacts

Why isn't InfiniBand a replacement for PCI slots? Foremost, InfiniBand addresses Server technology and not Personal Computers. The fabric that InfiniBand creates is designed to enhance the data center for the first 50 feet of the Internet and not to support consumer installation of graphics and networking cards in the office or home. InfiniBand does not have the goal to become a slot in a PC that maintains PCI software compatibility and lets the tens of millions of new PC buyers use it as a way to upgrade their systems. InfiniBand does have the goal of enabling processor level bandwidth all the way from the processor through system I/O and throughout the data center. (See Mellanox's White Paper: Introduction to InfiniBand for more on this goal.)

InfiniBand does however require an industry standard point of attachment to the CPU and memory subsystem that offers bandwidth to match the 20 Gb/sec that an InfiniBand 4X link offers. Here, PCI-Express complements InfiniBand by upgrading PCI slot performance and offering InfiniBand a way to high performance on servers without requiring the HCAs to be soldered down on the server logic board.

As mentioned, InfiniBand HCAs (host channel adapters) can be added either on the board or as an add-in card through either PCI or PCI upgraded by PCI-Express. A 4X or 10 + 10 Gb/s InfiniBand implementation needs ~ 20 Gb/sec of bandwidth to system logic. Example A shows a server block diagram where the 4X InfiniBand HCA bandwidth is limited by PCI-X and the system logic

to south bridge interconnect (shown is Intel's Hub Link 2). Certainly PCI-X bandwidth is more than sufficient to meet the needs of 1X InfiniBand links.



Example B shows how PCI-Express can be used to upgrade the bandwidth of both the interconnect between system logic and the south bridge, as well as, the PCI slot. In the implementation shown PCI-Express can be scaled to support the 20 Gb/sec InfiniBand requires in both locations. This implementation enables InfiniBand 4X bandwidth (20Gb/s) from a Pentium 4 processor, across the I/O subsystem and onto the InfiniBand fabric.

Example C shows how the SouthBridge and the PCI bus could be eliminated if the InfiniBand logic is integrated into the chip set. PCI-Express can be implemented at higher bandwidths (multiple sets of wires), so it's very likely that PCI-Express will be used to scale the bandwidth of the system logic as the needs for greater InfiniBand bandwidth grows to 12X or 60 Gb/s. It is expected that this level of system logic integration require several years before it can be achieved. Nonetheless when InfiniBand logic is integrated into the chip set and the PCI bus is no longer needed for the path to the processor, it is still expected that PCI slots will continue to be included on all server logic boards, just as ISA slots continued on boards for years after PCI was introduced. For the emerging market of InfiniBand server blades it is expected that no PCI functionality will be required since all the communication will take place over the fabric.

These examples clearly illustrate how PCI-Express complements InfiniBand by enabling the bandwidth that InfiniBand needs to communicate with the server's processor.

7.0 Summary

PCI, PCI-Express, and InfiniBand technologies all share the same general goal of improving I/O bandwidth, but each technology attacks a different problem. Two of these technologies, namely PCI and PCI-Express are local buses with PCI-Express positioning itself both as a chip to chip interconnect and as an upgrade for today's PCI slots.

InfiniBand enables a fabric that offers substantially more features and capabilities than a local bus interconnect. InfiniBand delivers reliable, transport level connections in hardware while still providing both message passing and memory semantics on the wire. InfiniBand creates a single fabric that reduces the complexity of the data center and enables flexible servers that are able to carry clustering, communication and storage traffic all on a single fabric. InfiniBand does NOT have the goal of upgrading PCI slots in PC's, but to realize its full potential it does require access to greater bandwidth to system logic than PCI-X can offer. Thus PCI-Express, both as a chip-to-chip interconnect between system logic and the south bridge and as an upgrade to PCI slot bandwidth, plays a key role in realizing the full potential of 4X and 12X InfiniBand.

Circling back to the three PCI issues that this paper mentioned needed solving. There is not a single solution for all three, they need to be resolved by different technologies.

- Intel contends that PCI-Express is the solution to the first problem as it is intended as a lower cost replacement for today's 32-bit PCI, and by using multiple links it will advance PCI bandwidth so it addresses the second issue, as well. Also it provides another benefit in that it can act as a local interconnect (for example: connecting the north bridge and south bridge together).
- PCI-X 266 & 532 targets the second issue as it advances the bandwidth of PCI slots, but it does nothing to solve PCI's flat memory architecture so PCI's I/O interconnect capabilities remain limited. PCI-X 266 proponents have stated that it can be implemented in the next year. So there is a time to market factor between PCI-Express and PCI-X 266.
- InfiniBand is clearly the solution to the third issue as it enables a much richer many-to-many I/O and system interconnect at much greater bandwidths.

In summary, this paper should have made it clear that InfiniBand does not compete with local bus architectures such as PCI-Express and PCI. Instead PCI-Express (and other chip-to-chip interconnects) provide improvements to local I/O bandwidth and/or PCI slots that can complement the InfiniBand fabric. When these technologies are combined they provide powerful solutions that enable processor level bandwidth all the way to the edge of the data center.

8.0 About Mellanox

Mellanox is the leading supplier of InfiniBand semiconductors, providing Switches, Host Channel Adapters, and Target Channel Adapters to the server, communications, and data storage markets. In January 2001, Mellanox Technologies delivered the InfiniBridge™ MT21108, the first 1X/4X InfiniBand device to market, and is now shipping second generation InfiniScale silicon. The company has raised more than $33 million to date and has strong corporate and venture backing from Intel Capital, Raza Venture Management, Sequoia Capital, and US Venture Partners.

In May 2001, Mellanox was selected by the Red Herring Magazine as one of the 50 most important private companies in the world and to Computerworld Magazine Top 100 Emerging Companies for 2002. Mellanox currently has more than 200 employees in multiple sites worldwide. The company's business operations, sales, marketing, and customer support are headquartered in Santa Clara, CA; with the design, engineering, software, system validation, and quality and reliability operations based in Israel. For more information on Mellanox, visit www.mellanox.com.