Creating a Third Generation I/O Interconnect
Ajay V. Bhatt
Desktop Architecture Labs
Intel Corporation

Summary

This paper looks at the success of the widely adopted PCI bus and describes a higher performance next generation of I/O interconnect, called PCI Express* Architecture, that will serve as a standard local I/O bus for a wide variety of future computing platforms. Key PCI attributes, such as its usage model and software interfaces are maintained whereas its bandwidth-limiting, parallel bus implementation is replaced by a long-life, fully-serial interface. A split-transaction protocol is implemented with attributed packets that are prioritized and optimally delivered to their target. The new PCI Express Architecture comprehends a variety of form factors to support smooth integration with PCI and to enable new system form factors. PCI Express Architecture will provide industry leading performance and price/performance.

Introduction

The PCI bus has served us well for the last 10 years and it will play a major role in the next few years. However, today's and tomorrow's processors and I/O devices are demanding much higher I/O bandwidth than PCI 2.2 or PCI-X can deliver and it is time to engineer a new generation of PCI to serve as a standard I/O bus for future generation platforms. There have been several efforts to create higher bandwidth buses and this has resulted in the PC platform supporting a variety of application-specific buses alongside the PCI I/O expansion bus as shown in Figure 1.
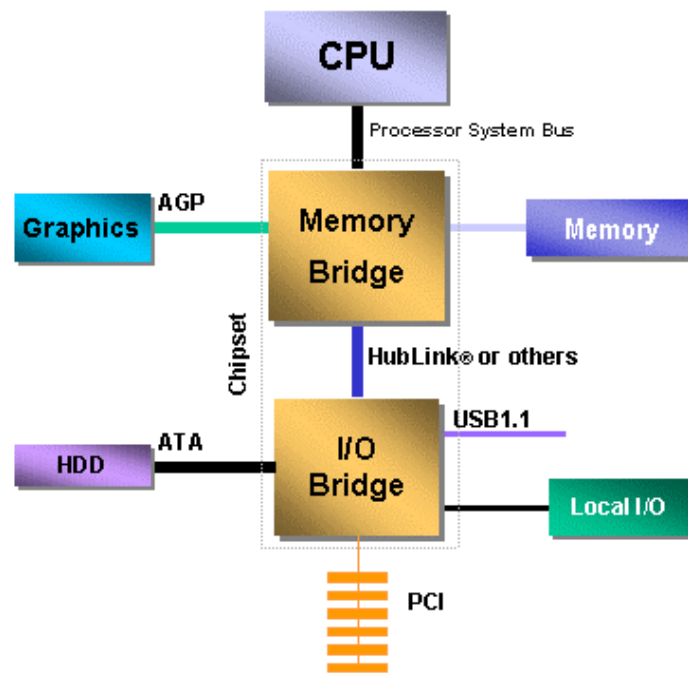

Figure 1. Today's PC has multiple local buses with different requirements

The processor system bus continues to scale in both frequency and voltage at a rate that will continue for the foreseeable future. Memory bandwidths have increased to keep pace with the processor. Indeed, as shown in Figure 1, the chipset is typically partitioned as a memory hub and an I/O hub since the memory bus often changes with each processor generation. One of the major functions of the chipset is to isolate these ever-changing buses from the stable I/O bus.

Close investigation of the 1990's PCI signaling technology reveals a multi-drop, parallel bus implementation that is close to its practical limits of performance: it cannot be easily scaled up in frequency or down in voltage; its synchronously clocked data transfer is signal skew limited and the signal routing rules are at the limit for cost-effective FR4 technology. All approaches to pushing these limits to create a higher bandwidth, general purpose I/O bus result in large cost increases for little performance gain. The desktop solution of Figure 1 is only part of the problem of diverging local I/O bus standards. To PCI's credit it has been used in applications not envisaged by the original specification writers and variants and extensions of PCI can be found in desktop, mobile, server and embedded communications market segments.

Today's software applications are more demanding of the platform hardware, particularly the I/O subsystems. Streaming data from various video and audio sources are now commonplace on the desktop and mobile machines and there is no baseline support for this time–dependant data within the PCI 2.2 or PCI-X specifications. Applications such as video-on-demand and audio re-distribution are putting real-time constraints on servers too. Many communications applications and embedded-PC control systems also process data in real-time. Today's platforms, an example desktop PC is shown in Figure 2, must also deal with multiple concurrent transfers at ever-increasing data rates. It is no longer acceptable to treat all data as equal – it is more important, for example, to process streaming data first since late real-time data is as useless as no data. Data needs to be "tagged" so that an I/O system can prioritize its flow throughout the platform.

Applications, such as Gigabit Ethernet* and InfiniBand*, require higher bandwidth I/O. A third generation I/O bus must include additional features alongside increased bandwidth.
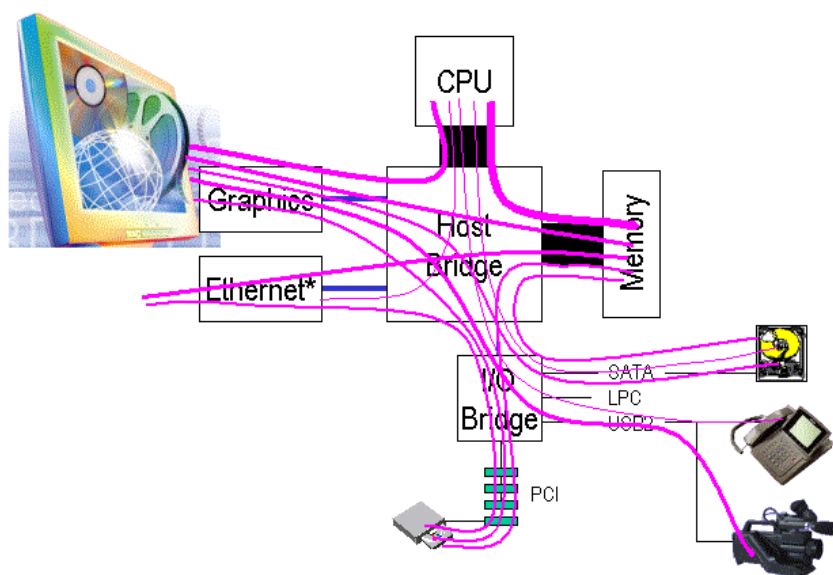

Figure 2. Multiple concurrent data transfers.

The following list summarizes the requirements for a third generation local I/O bus.

- Supports multiple market segments and emerging applications:

  - Unifying I/O Architecture for Desktop, Mobile, Server, Communications Platforms, Workstations and Embedded Devices.

- Low cost and high volume:

  - Cost at or below PCI cost structure at the system level

- PCI Compatible software model

  - Boot existing operating systems without any change. PCI compatible configuration and device driver interfaces.

- Performance:

  - Scalable performance via frequency and additional lanes. High Bandwidth per Pin. Low overhead. Low latency.

- Support multiple platform connection types:

  - Chip-to-chip, board-to-board via connector, docking station and enable new form factors.

- Advanced features:

  - Comprehend different data types. Power Management. Quality Of Service. Hot Plug and Hot Swap support. Data Integrity and Error Handling. Extensible. Base mechanisms to enable Embedded and Communications applications.

- Non-Goals:

  - Coherent interconnect for processors, memory interconnect, cable interconnect for cluster solutions.

PCI Express* Architecture Overview

Recent advances in high-speed, low-pin-count, point-to-point technologies offer an attractive alternative for major bandwidth improvements. A PCI Express multi-drop, parallel bus topology contains a Host Bridge and several endpoints (the I/O devices) as shown in Figure 3. Multiple point-to-point connections introduce a new element, the switch, into the I/O system topology also shown in Figure 3. The switch replaces the multi-drop bus and is used to provide fan-out for the I/O bus. A switch may provide peer-to-peer communication between different endpoints and this traffic, if it does not involve cache-coherent memory transfers, need not be forwarded to the host bridge. The switch is shown as a separate logical element but it could be integrated into a host bridge component.

The low signal-count, point-to-point connections may be constructed with connectors and cables. The PCI Express mechanicals will enable new classes of system partitioning (the "boring biege box" is no longer required!).
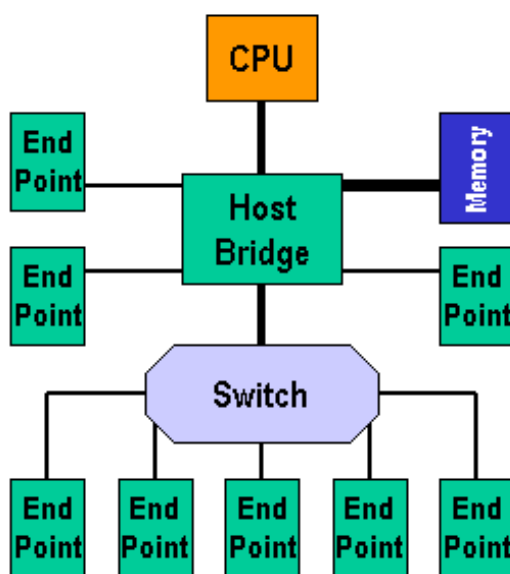

Figure 3. A switch is added to the system topology

Figures 4 through 6 show typical 2004 platforms using the PCI Express Architecture. The multiple, similar parallel buses of today's platform are replaced with PCI Express **links** with one or more lanes. Each link is individually scalable by adding more lanes so that additional bandwidth may be applied to those links where it is required – such as graphics in the desktop platform and bus bridges (e.g. PCI Express-to-PCI-X) in the server platform.

A PCI Express switch provides fanout capability and enables a series of connectors for add-in, high performance I/O. The switch is a logical element that may be implemented within a component that also contains a host bridge, or it may be implemented as a separate component.

It is expected that PCI will coexist in many platforms to support today's lower bandwidth applications until a compelling need, such as a new form factor, causes a full migration to a fully PCI Express based platform.
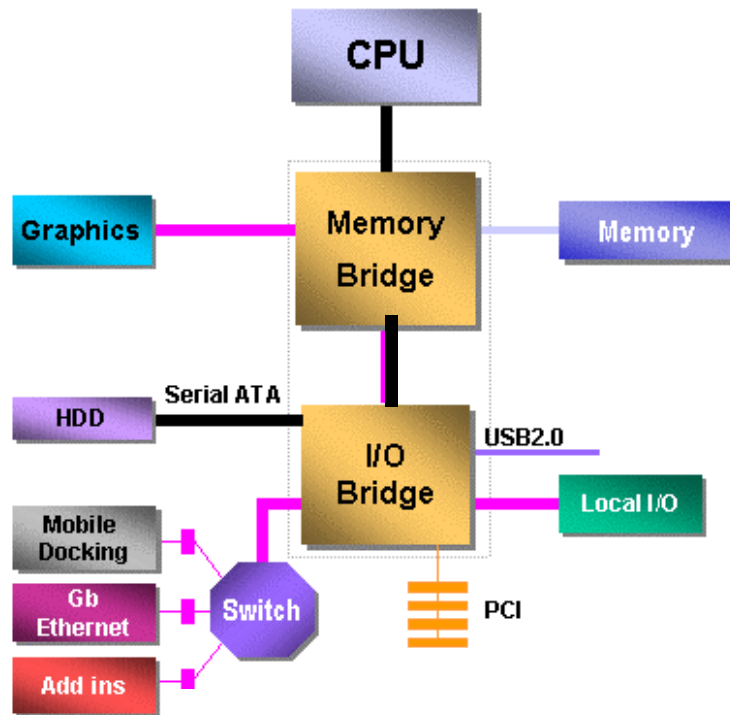
Figure 4. General purpose desktop/mobile I/O interconnect for 2003 and beyond

The server platform requires more I/O performance and connectivity including high bandwidth PCI Express links to PCI-X slots, Gigabit Ethernet* and an InfiniBand* fabric. Figure 5 shows how PCI Express provides many of the same advantages for servers, as it does for desktop systems. The combination of PCI Express for "inside the box" I/O, and InfiniBand fabrics for "outside the box" I/O and cluster interconnect, allows servers to transition from "parallel shared buses" to high speed serial interconnects..
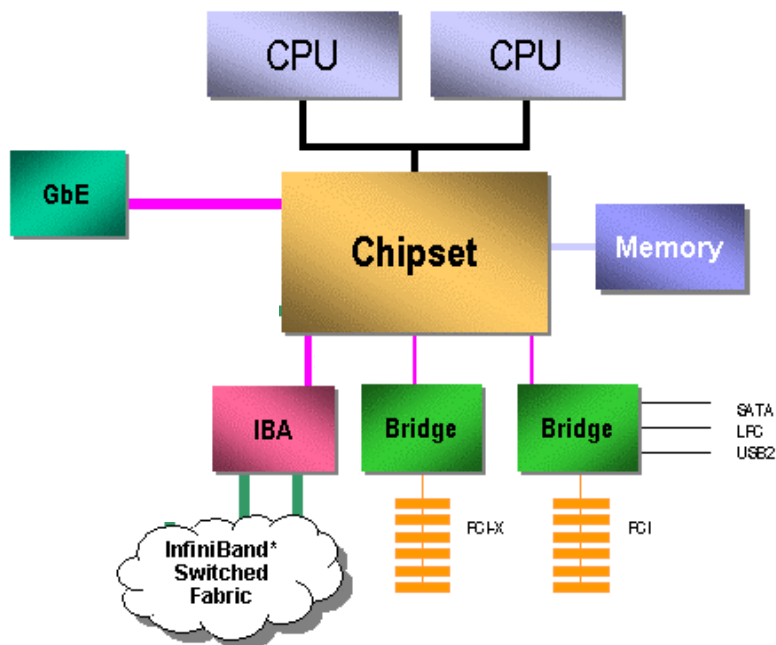
Figure 5. PCI Express-based Server/Workstation System

The networking communications platform could use multiple switches for increased connectivity and Quality of Service (QOS) for differentiation of different traffic types. It too would benefit from a multiple PCI Express links that could be constructed as a modular I/O system.
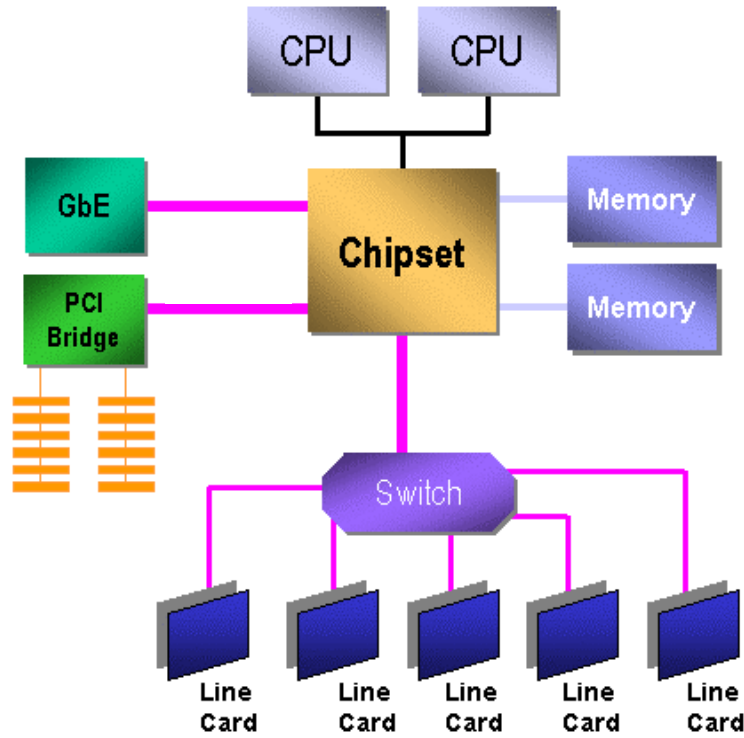

Figure 6. PCI Express-based Networking Communications System

PCI Express Architecture

The PCI Express Architecture is specified in layers as shown in Figure 7. Compatibility with the PCI addressing model (a load-store architecture with a flat address space) is maintained to ensure that all existing applications and drivers operate unchanged. PCI Express configuration uses standard mechanisms as defined in the PCI Plug-and-Play specification. The software layers will generate read and write requests that are transported by the transaction layer to the I/O devices using a packet-based, split-transaction protocol. The link layer adds sequence numbers and CRC to these packets to create a highly reliable data transfer mechanism. The basic physical layer consists of a dual-simplex channel that is implemented as a transmit pair and a receive pair. The initial speed of 2.5 Giga transfers/second/direction provides a 200MB/s communications channel that is close to twice the classic PCI data rate.
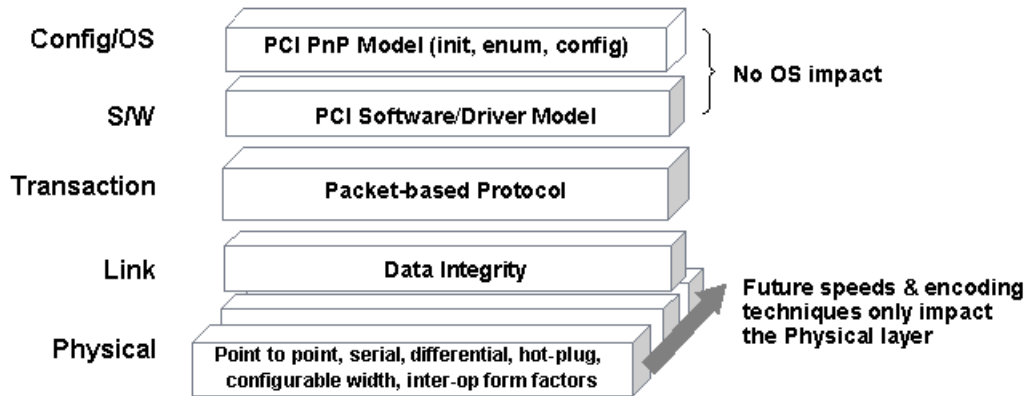

Figure 7. The PCI Express Architecture is specified in layers

The remainder of this section will look deeper into each layer starting at the bottom of the stack.

Physical Layer

The fundamental PCI Express link consists of two, low-voltage, differentially driven pairs of signals: a transmit pair and a receive pair as shown in Figure 8. A data clock is embedded using the 8b/10b encoding scheme to achieve very high data rates. The initial frequency is 2.5 Giga transfers/second/direction and this is possible to increase with silicon technology advances of up to 10 Giga transfers/second/direction (the theoretical maximum for signals in copper). The physical layer transports packets between the link layers of two PCI Express agents.
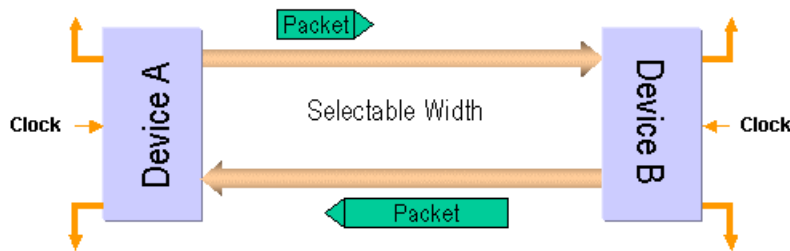

Figure 8 A PCI Express link uses transmit and receive signal pairs

The bandwidth of a PCI Express link may be linearly scaled by adding signal pairs to form multiple lanes. The physical layer supports x1, x2, x4, x8, x12, x16 and x32 lane widths and splits the byte data as shown in Figure 9. Each byte is transmitted, with 8b/10b encoding, across the lane(s). This data disassembly and re-assembly is transparent to other layers.

During initialization, each PCI Express link is set up following a negotiation of lane widths and frequency of operation by the two agents at each end of the link. No firmware or operating system software is involved.

The PCI Express architecture comprehends future performance enhancements via speed upgrades and advanced encoding techniques. The future speeds, encoding techniques or media would only impact the physical layer.
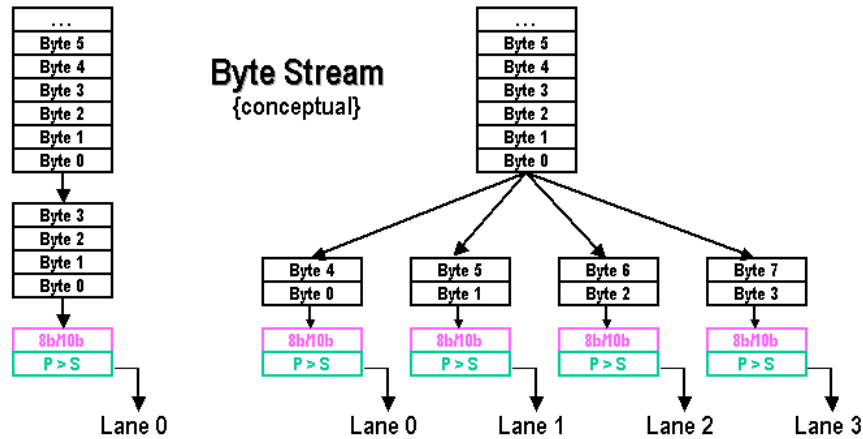

Figure 9. A PCI Express Link consists of one or more lanes

Link Layer

The primary role of a link layer is to ensure reliable delivery of the packet across the PCI Express link. The link layer is responsible for data integrity and adds a sequence number and a CRC to the transaction layer packet as shown in Figure 10.
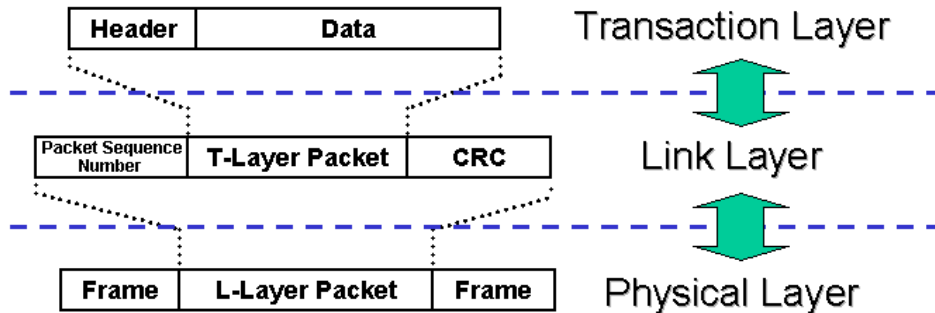

Figure 10. The Link Layer adds data integrity features

Most packets are initiated at the Transaction Layer (next section). A credit-based, flow control protocol ensures that packets are only transmitted when it is known that a buffer is available to receive this packet at the other end. This eliminates any packet retries, and their associated waste of bus bandwidth due to resource constraints. The Link Layer will automatically retry a packet that was signaled as corrupted.

Transaction Layer

The transaction layer receives read and write requests from the software layer and creates request packets for transmission to the link layer. All requests are implemented as split transactions and some of the request packets will need a response packet. The transaction layer also receives response packets from the link layer and matches these with the original software requests. Each packet has a

unique identifier that enables response packets to be directed to the correct originator. The packet format supports 32bit memory addressing and extended 64bit memory addressing. Packets also have attributes such as "no-snoop", "relaxed-ordering" and "priority" which may be used to optimally route these packets through the I/O subsystem.

The transaction layer supports four address spaces: it includes the three PCI address spaces (memory, I/O and configuration) and adds a Message Space. PCI 2.2 introduced an alternate method of propagating system interrupts called Message Signaled Interrupt (MSI). Here a special-format memory write transaction was used instead of a hard-wired sideband signal. This was an optional capability in a PCI 2.2 system. The PCI Express specification re-uses the MSI concept as a primary method for interrupt processing and uses Message Space to support all prior side-band signals, such as interrupts, power-management requests, resets, and so on, as in-band Messages. Other "special cycles" within the PCI 2.2 specification, such as Interrupt Acknowledge, are also implemented as in-band Messages. You could think of PCI Express Messages as "virtual wires" since their effect is to eliminate the wide array of sideband signals currently used in a platform implementation.

Software Layers

Software compatibility is of paramount importance for a third generation local I/O bus. There are two facets of software compatibility; initialization, or enumeration, and run-time. PCI has a robust initialization model wherein the operating system can discover all of the add-in hardware devices present and then allocate system resources, such as memory, I/O space and interrupts, to create an optimal system environment. The PCI configuration space and the programmability of I/O devices are key concepts that are unchanged within the PCI Express Architecture; in fact, all operating systems will be able to boot without modification on a PCI Express-based platform.

The run-time software model supported by PCI is a load-store, shared memory model – this is maintained within the PCI Express Architecture which will enable all existing software to execute unchanged. New software may use new capabilities.

Mechanical Form Factors

The low signal-count of a PCI Express link will enable both an evolutionary approach to I/O subsystem design and a new modular approach that will encourage new system partitioning.

Evolutionary Design:

Initial implementations of PCI Express-based add-in cards will co-exist alongside the current PCI-form factor boards (both full-size and "half-height"). As an example connection; a higher bandwidth link, such as a sixteen-lane connection to a graphics card, will use a new connector placed alongside the existing PCI or AGP connector in the area previously occupied by those type of connectors as shown in Figure 11.
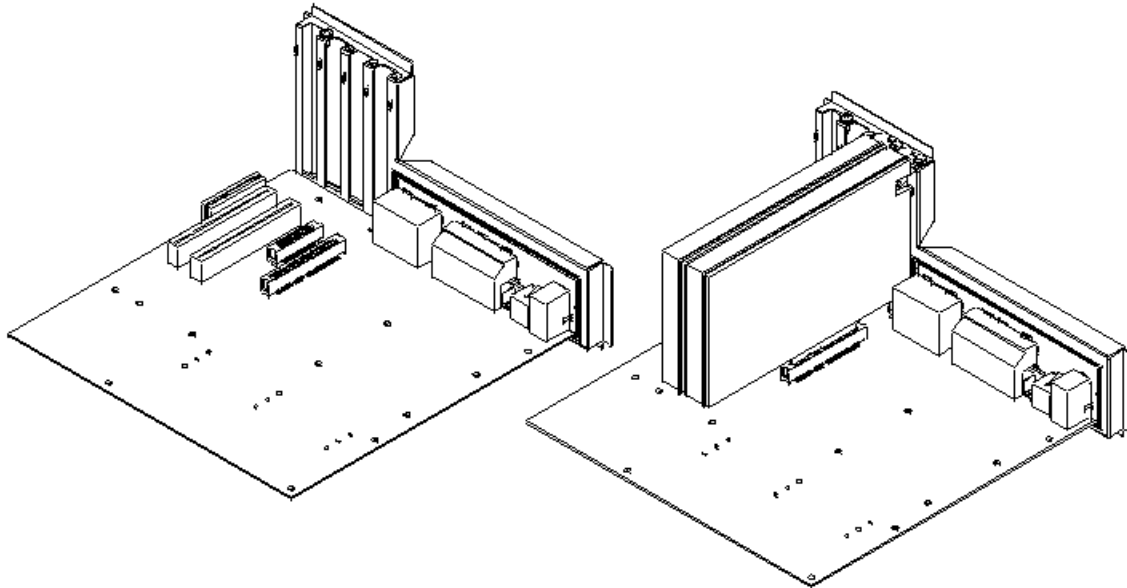

Figure 11. Using an additional connector alongside the PCI connector.

In addition to the base implementations of PCI Express-based add-in cards that replace AGP and PCI cards, other form factors are in progress of being developed. The PCMCIA industry group at http://www.pcmcia.org/ has announced the development of a new specification codenamed NEWCARD* that takes the next step in PC Card evolution. This NEWCARD will offer integrated desktop and mobile external expansion into a single standard. Slated for release later this year, the NEWCARD standard is targeted for OEMs seeking small footprint for thinner mobile designs and sealed systems for desktop designs. NEWCARD is the next evolution of the PC Card, combining a smaller form factor and faster performance with the PC Cards reliability and ease of use. NEWCARD supports hot swappable device-sharing between mobile and desktop devices, similar to how USB devices can be shared. Desktop users will benefit from NEWCARDs plug-n-play ease of use, which eliminates the need to open the chassis to add new features. The new standard is operating-system independent.

Performance Characteristics

PCI Express's differential, point-to-point connection provides a very high speed interconnect using few signals.   Its message space eliminates all prior sideband signals resulting in minimal implementation signals. PCI Express's 100MB/s/pin will translate to the lowest cost implementation for any required bandwidth.

Development Timeline

The Board of Directors approved the PCI Express Specification Rev 1.0 in July 2002. The Specification is available to PCI-SIG members and is currently at Rev 1.0A. The PCI-SIG website is located at www.pci-sig.com. Intel's Developer Network site, established to help the industry is located at http://www.express-lane.org/.

Summary

The PCI Express Architecture meets all of the requirements of a third generation I/O bus. It's advanced features and scalable performance will enable it to become a unifying I/O solution across a broad range of platforms − desktop, mobile, server, communications, workstations and embedded devices. A PCI Express link is implemented using multiple, point-to-point connections called lanes and multiple lanes can be used to create an I/O interconnect whose bandwidth is linearly scalable. This interconnect will enable flexible system partitioning paradigms at or below the current PCI cost structure. PCI Express is software compatible with all existing PCI-based software to enable smooth integration within future systems.