PCI Express and Windows

October 20, 2004

Abstract

This paper provides an overview of the PCI Express interface and describes how PCI Express is supported on the current Microsoft® Windows® family of operating systems. It also provides information for system manufacturers, firmware engineers, and driver developers to create systems and hardware that will take full advantage of the support for PCI Express planned for inclusion in future versions of the Windows operating system, including Windows codenamed "Longhorn."

This information applies for the following operating systems:

Microsoft Windows 2000
Microsoft Windows XP
Microsoft Windows Server™ 2003
Microsoft Windows Server 2003 SP1
Microsoft Windows codenamed "Longhorn"

All PCI specifications referenced in this paper are available on the PCI Special Interest Group (PCI-SIG) Web site at: http://www.pcisig.com/

For the purposes of this paper, the *PCI Local Bus Specification Revision 2.3* is referred to as the "PCI Local Bus Specification," and the *PCI Express Base Specification Revision 1.0a* is referred to as the "PCI Express Base Specification."

It is assumed that the reader has a good understanding of general computer architecture concepts and the PCI bus.

Contents

Introduction

The PCI bus has served the industry well for the past 10 years. During this time, numerous new technologies that place new feature and bandwidth requirements on this I/O bus have been developed. As an aging technology, PCI can no longer handle the demands of the new classes of I/O devices, such as High Definition Audio and Video, Gigabit Ethernet, and next generation storage controllers and devices. Although engineers have performed admirably to extend the usefulness of PCI through their work on the PCI-X bus, it is clearly the right time to introduce a new I/O bus that is more scalable and addresses the demands of next generation hardware and software.

The PCI-SIG has defined a new I/O technology, PCI Express, which over time is expected to replace the current I/O bus technologies PCI, PCI-X, AGP, and CardBus. PCI Express addresses many of the shortcomings of the current I/O bus technologies and provides advanced features that exceed the capabilities of the older technologies.

PCI Express is not a parallel bus like PCI and PCI-X, but rather is a serial interface consisting of individual Links that are shared by only two devices. This serial design has many advantages over the parallel bus design and provides advanced features not offered by PCI or PCI-X.

PCI Express retains full software compatibility with the *PCI Local Bus Specification Revision 2.3*. PCs which have implemented PCI Express can boot and run current Windows operating systems (Windows 2000, Windows XP, and Windows Server 2003) without requiring modifications to the operating system or the device drivers. System manufacturers can take advantage of the higher bandwidth of PCI Express immediately. Firmware engineers can update the ACPI firmware to make some of the advanced features of PCI Express available to the current Windows operating systems. Once available, Windows Longhorn will natively support the advanced features of PCI Express for a deeper, richer user experience.

About PCI Express

PCI Express is a quantum leap in technology over the PCI bus, providing greater bandwidth and an architecture that can be scaled to meet the requirements of next generation hardware and software. This section provides information about the fabric, scalability, and architecture of PCI Express.

PCI Express Fabric

The fundamental building blocks that make up the fabric of the PCI Express interface are Lanes and Links.

- **Lane.** Each Lane represents a dual-unidirectional communications channel between two PCI Express devices. A Lane consists of two pairs of traces that carry low voltage (0.08 to 1.2 volt) signals. Each signal pair carries 2.5 gigabits/second of traffic in one direction, as indicated by the arrows in Figure 1. One signaling pair is used by a device for transmitting data and one pair is used for receiving data. Because Lanes connect only two devices, the signaling pair used by Device A to transmit data is the same pair used by Device B to receive data.
- **Link.** Two PCI Express devices are connected by a Link, and each Link is made up of one or more Lanes. PCI Express supports the following Link widths: x1, x2, x4, x8, x12, x16, and x32 Lanes. Figure 1 shows a x1 Link.

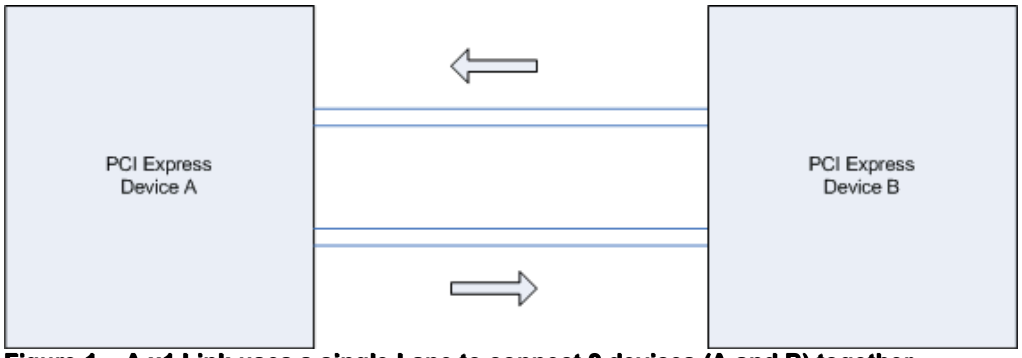

Figure 1 – A x1 Link uses a single Lane to connect 2 devices (A and B) together.

One advantage of the serial architecture of PCI Express is the lower pin count for all PCI Express devices compared to PCI and PCI-X. For example, a x1 Link uses only eight traces (four for data and four for other uses such as power and ground)—significantly fewer than a PCI bus, which uses a minimum of 74 traces. This more pin-efficient design not only reduces manufacturing costs, it also greatly reduces the number of traces that are required on the printed circuit board (PCB), thus simplifying the routing of traces. Additionally, this architecture reduces the electrical issues that limit the signaling rate of PCI and PCI-X.

Scalability

Another advantage of PCI Express over PCI and PCI-X is the amount of bandwidth the former can support while using significantly fewer pins and traces. The PCI Express Specification defines the initial signaling rate for PCI Express as 2.5 gigahertz. This means that each Lane can carry 2.5 gigabits of data simultaneously in each direction using only eight traces.

PCI Express uses 8-bit/10-bit encoding, which means that an extra two bits are added to each byte to embed the clock signal in the data stream. This means that a byte in PCI Express is equal to 10 bits. The bandwidth of a single PCI Express Lane is calculated as follows:

2.5 gigabits/second divided by 10 bits per byte
= 250 megabytes x 2 (each direction on the Link)
= 500 megabytes/second total bandwidth per Lane.

It is interesting to compare the bandwidth of a PCI Express Lane to the bandwidth of the PCI and PCI-X buses. The bandwidth of a PCI or PCI-X bus is calculated as follows:

(Signaling rate in megahertz x width of the bus in bits) / 8 bits per byte
= maximum throughput of bus in megabytes per second.

For PCI and PCI-X, the PCI-SIG has defined signaling rates from 33 MHz to 266 MHz and is working on a specification for a 533 MHz signaling rate. PCI operates at signaling rates of either 33 MHz or 66 MHz, and PCI-X operates at 66 MHz and higher. Both PCI and PCI-X buses support 32- and 64-bit bus widths; however, because of the engineering issues and amount of space required to run the traces for a 64 bit-wide bus, it is used only in the server market segment.

The following table shows the total bandwidth that can be achieved by different PCI and PCI-X bus configurations.

Table 1 – Total bandwidth for PCI and PCI-X

Signaling rate (in MHz)	Bus width (in bits)	Bus type	Total bandwidth (in megabytes/second)

33	32	PCI	132
33	64	PCI-X	264
66	32	PCI and PCI-X	264
66	64	PCI-X	528
133	64	PCI-X	1064
266	64	PCI-X	2128
533	64	PCI-X	4264

You may notice that the bandwidth for a 64-bit wide PCI-X bus running at 66 MHz is higher than a single PCI Express Lane running at 2.5 gigahertz. However, PCI Express can be scaled to meet higher bandwidth requirements by increasing the number of Lanes between the devices.

To further increase the bandwidth of a Link, multiple Lanes can be placed in parallel between two PCI Express devices to aggregate the bandwidth of each individual Lane. As mentioned earlier, Lanes can be added between PCI Express devices in the following increments: x1, x2, x4, x8, x12, x16, or x32.

For example, a typical design for a high-performance desktop PC is expected to implement the following PCI Express Links:

- x16 Link for graphics and other multimedia devices
- x8 Link for high performance chip-to-chip interconnect, network and storage devices
- x4 Link for general chip-to-chip interconnect, network and storage devices
- x1 Link for general I/O

The following table shows the bandwidth of a PCI Express Link for different Lane widths.

Table 2 – Increasing Bandwidth by Adding Lanes

Number of Lanes per Link	Bandwidth (in megabytes/second) in each direction
1	250
2	500
4	1,000 (approximately1 gigabyte)
8	2,000 (approximately 2 gigabytes)
12	3,000 (approximately 3 gigabytes)
16	4,000 (approximately 4 gigabytes)
32	8,000 (approximately 8 gigabytes)

The device that currently uses the most bandwidth on today's PC is the graphics adapter on the 8X AGP bus, which can transfer data at approximately 2 gigabytes/second. When you compare the bandwidth of an x32 PCI Express Link to the bandwidth of an 8X AGP bus, you can see that PCI Express can support the I/O requirements of PCs well into the future.

In addition to adding Lanes to scale the bandwidth, the PCI Express interface is designed so that the signaling rate can also be increased. The PCI-SIG working group is currently working on defining two additional signaling rates, which are expected to be 5 and 10 gigahertz. Once these additional signaling rates are defined, PCI Express system manufacturers will be able to scale the bandwidth of PCI Express Links using any combination of signaling rate and Link width on a

device-by-device basis. These new signaling rates will enable system manufacturers to achieve two to four times the currently defined bandwidth.

PCI Express Architecture

PCI Express is defined in a flexible way that enables system manufacturers to benefit immediately from its scalability while their systems remain backward compatible to PCI and PCI-X. Additionally, PCI Express leverages the form factors and manufacturing processes already used in today's PCs.

The PCI Express architecture consists of the following devices:

- **Root Complex:** The heart of the PCI Express interface is the Root Complex, which is made up of one or more Host Bridges. Each Host Bridge exposes one or more Root Ports, which appear as PCI-to-PCI Bridges to software and can be used to connect other PCI Express devices, such as Endpoints or Switches, to the Root Complex. System CPU(s) and system memory are also connected to the Root Complex, usually through a system bus.
- **PCI Express Slots:** Physical connectors that allow PCI Express adapters to connect to the PCI Express fabric.
- **PCI Express-to-PCI Bridge**: A PCI Express device used to connect PCI devices to the PCI Express fabric.
- **PCI Express-to-PCI-X Bridge**: A PCI Express device used to connect PCI-X devices to the PCI Express fabric.
- **PCI Express Switch**: Provides connectivity to additional PCI Express Endpoints, Bridges, and slots. Each PCI Express Switch consists of one upstream Port and multiple downstream Ports. To software, each Port looks like a virtual PCI-to-PCI Bridge.
- **PCI Express Endpoint**: A device that does not have any downstream devices attached to it and functions as a leaf node in the PCI Express device hierarchy. Examples of PCI Express Endpoints include the graphics processing unit, network, and storage controllers. A PCI Express Endpoint can be integrated directly into the Root Complex, or it can be connected to a PCI Express Root Port or PCI Express Switch.

All PCI Express devices, regardless where they sit in the PCI Express fabric and device hierarchy, communicate with the system CPU(s) and system memory through the Root Complex. Figure 2 represents what is expected to be a typical PCI Express system in a desktop PC. Notice that PCI Express-to-PCI Bridges are used in this design to provide backward compatibility with PCI devices.

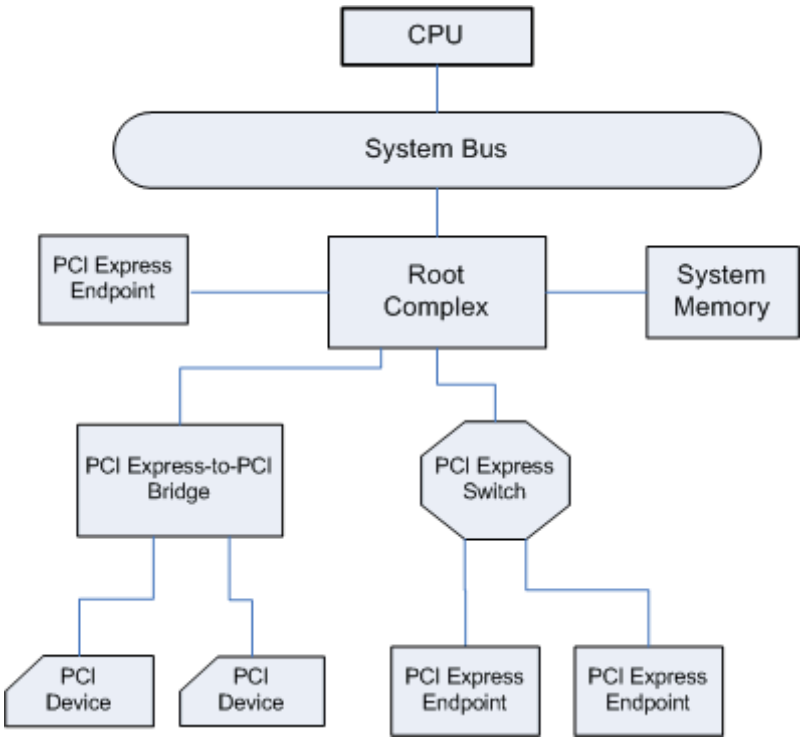

Figure 2. Typical PCI Express interface for a desktop PC

The following figure represents what is expected to be a typical PCI Express system in a server. Notice that PCI Express-to-PCI-X Bridges are used in this design to provide backward compatibility with PCI-X devices.

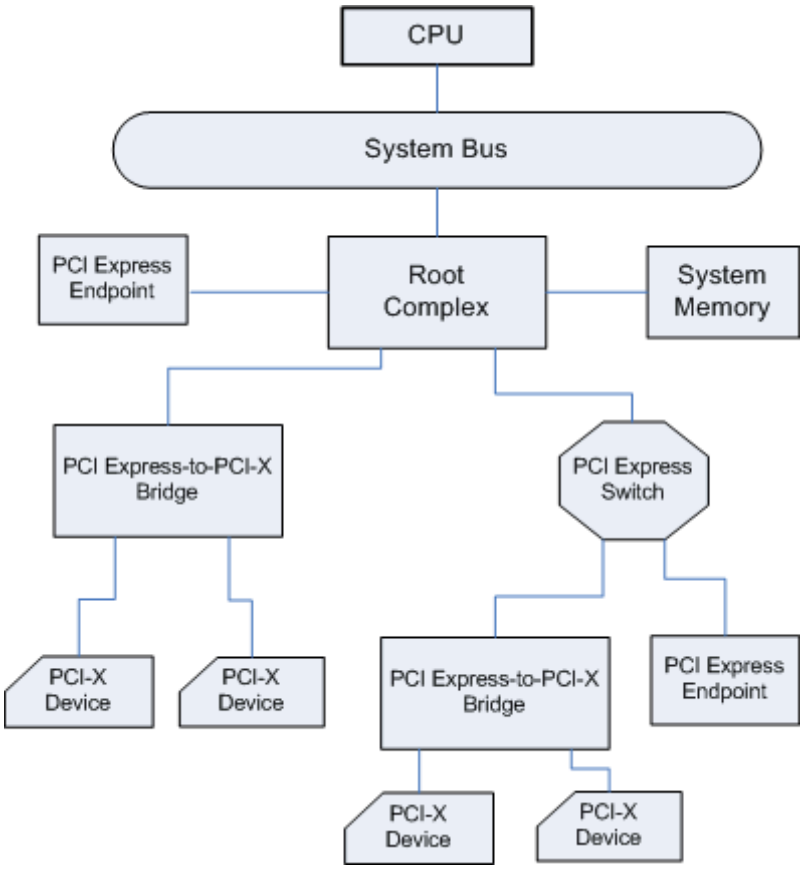

Figure 3 – Typical PCI Express interface with a server

Advantages of PCI Express

This section discusses the advantages of PCI Express compared to PCI, PCI-X, AGP, and CardBus. These advantages include the following:

- Support for multiple market segments and emerging applications
- Cost effective implementation and manufacture
- Compatibility with the PCI software model
- Superior performance and scalability
- Support for new modules

Support for Multiple Market Segments and Emerging Applications

PCI Express is designed to support chip-to-chip connections, board-to-board connections, and hot pluggable devices such as docking stations. PCI Express will be able to replace the following I/O bus technologies currently in use:

- AGP: Used for graphics and multimedia in desktop and mobile systems.
- PCI: Used for general purpose I/O in desktop PCs, servers, and communications equipment, such as routers and network switches.
- PCI-X: Used for high bandwidth and high performance applications primarily in the server market because of the high cost and number of traces required to support the PCI-X devices.

- CardBus (PCMCIA): Used in the mobile market that requires support for hot plug devices and docking stations.

PCI Express is well suited to replace these I/O technologies for the desktop PC, server, mobile, and communications market segments.

Desktop PC Market

Today's desktop PC is not only a personal productivity device; it is a gaming console and video editing workstation as well. These new usage scenarios require tremendous bandwidth which PCI Express is designed to deliver well into the future.

Server Market

The server market requires an I/O technology that delivers reliability, serviceability, and availability. PCI Express strongly delivers these requirements by supporting the following features:

- **Hot plug and removal of devices.** PCI Express simplifies hardware requirements for hot plug functionality, which enables users to replace damaged system devices, such as network interface cards and storage controllers, while the server is running. The PCI-SIG is currently working on defining I/O modules for externally accessible devices so that users can hot plug or hot remove devices without opening the chassis of the server. These I/O modules are described later in this paper.
- **Advanced data integrity and recovery.** PCI Express offers server class error detection, correction, and reporting. This is described in more detail later in this paper.

Mobile Market

PCI Express is designed to support hot plug devices while using less additional circuitry than CardBus, enabling system manufacturers to replace CardBus with PCI Express to support hot pluggable devices and docking stations. Additionally, system manufacturers can use the active state power management (ASPM) feature of PCI Express to reduce the overall system power consumption. The new ASPM feature is described later in this paper.

Because PCI Express natively supports hot plug and advanced power management capabilities, in the future the CardBus controller will no longer be needed, thus reducing the cost of a mobile computer.

Cost Effective Implementation

System manufacturers can build systems using PCI Express at approximately the same or reduced cost levels as the current I/O technologies, thanks to several cost effective advantages in PCI Express implementation.

Simpler Routing on Motherboards

PCI Express delivers more bandwidth than PCI and PCI-X while using approximately one tenth the number of pins and traces. This simpler routing of traces reduces engineering costs and allows the use of cost effective four-layer printed circuit boards (PCBs) for all market segments and form factors. The reduced number of pins for each PCI Express device also simplifies the power delivery solutions.

Same Manufacturing Process

PCI Express was designed to operate at I/O voltage levels compatible with 0.25 microns and future low voltage processes, allowing PCI Express chips to be manufactured using the same silicon process as PCI chips. Because the PCB manufacturing process is the same for PCI Express as for PCI, PCBs using PCI Express can be produced for about the same cost as those manufactured for PCI.

Similar Connectors as PCI

Because PCI Express connectors are very similar to PCI connectors, the manufacturing cost should also be similar. PCI Express connectors may initially cost more to produce due to supply and demand issues. However, because they are smaller than PCI connectors, PCI Express connectors require less raw material to manufacture, and may therefore cost less in the long term than PCI connectors as the demand for PCI Express connectors increases.

Full Compatibility with the PCI Software Model

PCI Express uses the same load/store I/O architecture as PCI and PCI-X. This similarity makes PCI Express fully compatible with the PCI software model. Because of this compatibility, current operating systems that are compatible with the PCI software model can boot and run on PCI Express systems without any change to device drivers or the operating system. However, to take advantage of the new, advanced features of PCI Express, software modification will be necessary.

PCI Express extends the 256-byte Configuration Space of PCI to 4096 bytes while maintaining compatibility with existing PCI enumeration and configuration software. PCI Express accomplishes this by dividing the PCI Express Configuration Space into two regions: the PCI-compatible region (that is, the first 256 bytes) and the extended region (the remaining 3840 bytes). The extended region is useful for complex devices that require large numbers of registers to control and monitor the device.

Note: PCI Express extended Configuration Space is not accessible on current Windows operating systems. Device drivers or the operating system must be updated to use the extended region.

Superior Performance and Scalability

PCI Express offers superior performance and scalability compared to PCI and PCI X.

Performance

PCI Express achieves Link efficiency by using a split transaction protocol and adopting a dual unidirectional Link topology to allow for the simultaneous flow of traffic in both directions on the Link. This split transaction protocol benefits from advanced flow mechanisms, which minimize bottlenecks, contentions, and latencies. This combination of protocol and topology gives PCI Express a performance advantage over PCI and PCI-X.

PCI Express also boosts performance by dedicating bandwidth on each direction of a Link and by performing concurrent cycles. Additionally, PCI Express enables traffic class priority and flow control-based data transfers on the Link, thus reducing overall bus contention issues.

Scalability

The PCI Express interface can be scaled by adding additional Lanes or by increasing the signaling rate. The 2.5 gigabits/second bandwidth for each Lane

means that fewer pins on the chips and fewer traces on the printed circuit board are required to match or surpass the bandwidth of PCI and PCI-X.

Support for New Modules

The PCI-SIG and Personal Computer Memory Card International Association (PCMCIA) are working to define modules that will allow users to hot plug devices without needing to open the chassis of the PC. These modules are intended to ease the installation of devices for the mobile, desktop PC, and server market segments.

Modules for the Mobile and PC Markets

PCI Express supports a new standard defined by PCMCIA called ExpressCard. PCMCIA recommends using ExpressCard as the electrical interface to replace CardBus for mobile systems; it is suitable for use with desktop PCs as well. The ExpressCard interface also contains a USB 2.0 bus interface which allows for the easy migration of USB 2.0 devices to ExpressCard form factors. For these reasons, ExpressCard will replace CardBus and PC Cards as the electrical interface in the long term.

ExpressCard can be used as the new interface for all existing and future devices such as flash media, micro drives, storage controllers, wired/wireless network cards, multimedia devices, and broadband modems. For more information about ExpressCard modules, see http://www.ExpressCard.org.

Modules for Server Markets

The PCI-SIG working group is currently defining an I/O module for the server market segment with two form factors: single wide and double wide. The single wide module is targeted for typical adapter card uses and the double wide module is targeted for more complex adapter cards that require more physical space to implement.

Advanced Features of PCI Express

The PCI Express Specification defines many advanced features which are improvements over the PCI, PCI-X, and AGP buses. This section summarizes the advanced features of PCI Express.

For detailed information about the advanced features of PCI Express, see the *PCI Express Base Specification Revision 1.0a*, available at: http://www.PCISIG.com

Extended Configuration Space

PCI devices are allocated 256 bytes of Configuration Space per function per device on the bus. For PCI Express, devices are allocated 4096 bytes of extended Configuration Space per function per device on the bus. PCI Express allows for 256 buses with each bus containing up to 32 devices and up to 8 functions per device.

If a PCI Express system requires more than 256 buses, then multiple segments can be used to allow the system to access additional buses. A segment is a logical representation of a group of buses.

PCI Express provides an enhanced access mechanism that maps the entire extended Configuration Space into a flat system memory region. This allows software to access the entire extended Configuration Space in a straightforward way instead of the cumbersome CFC/CF8 mechanism used for PCI. For more details about this feature, see “Support with Windows” later in this paper.

Power Management

All features defined in the *PCI Bus Power Management Interface Specification Revision 1.1* continue to be available in PCI Express. PCI Express also provides two advanced power management capabilities, Active State Power Management (ASPM) and slot power budgeting, which go beyond what is available for PCI and PCI-X. Additionally, PCI Express can signal power management events using in-band messages.

Active State Power Management

Active State Power Management is a new PCI Express feature that enables the hardware to engage actively in automatic Link power management. The Link state between two devices in the PCI Express hierarchy can transition from L0 (full on) to an L0s/L1 (idle) state to save power if the Link is not transferring data. The Link power transition is initiated automatically by the device when no traffic is detected on the Link. Once data is available to transfer across the Link, the hardware will bring the Link back to the L0 state.

Note: Device manufacturers are encouraged to design PCI Express devices that take advantage of ASPM in order to consume less power and produce less heat. This is especially important for the mobile and blade server market segments.

Slot Power Budgeting

The PCI Express Specification defines slot power budgeting as a mechanism to be used by software to manage the amount of power that each add-in adapter can use. This enables fine tuning and control of the system's overall power usage. This feature is especially beneficial for mobile systems where battery life is a major concern and for blade servers where thermal issues caused by power dissipation may be problematic.

Power Management Events

PCI Express sends in-band messages over the same Links it uses to transfer data to signal power management events (PME) such as sleep and wake. Because these messages are sent over the same Links as data, PCI Express does not require separate pins on the chips and traces on the PCBs as used in PCI and PCI-X.

Quality of Service

PCI Express defines a new feature, called virtual channels, to provide superior quality of service (QoS) compared to other I/O technologies. Virtual channels can be used to reserve bandwidth on a Link for a particular device. This reserved bandwidth provides isochronous data transfers needed to support time-sensitive applications.

Note: The PCI Express Specification defines the extended virtual channel support as an optional feature. This feature is not supported by current Windows operating systems. Because some chipsets will not support extended virtual channels while other chipsets will only implement them using non-snooped DMA, Windows Longhorn will not support this feature until better rules and guarantees are defined. Microsoft is currently working with hardware manufacturers to define these rules and guarantees.

Hot-Plug PCI Express

PCI Express defines a standard native hot plug model based on the Standard Hot Plug Controller (SHPC), which is a well known and understood industry model used

by PCI and PCI-X. PCI Express is designed to withstand the insertion and removal of hot plug devices with less additional circuitry than PCI and PCI-X. The PCI-SIG and the PCMCIA industry groups are actively defining plug-in modules for hot plugging and hot swapping.

Message Signaled Interrupts (MSI and MSI-X)

PCI Express does not require separate traces for a device to signal interrupts; rather, PCI Express uses in-band messages. This method of signaling interrupts completely eliminates the need for interrupt sharing among devices, which is a clear improvement over the PCI and PCI-X buses.

Furthermore, a device can use different interrupt messages to indicate the individual status of different interrupts. This improves performance by saving a read to the interrupt status register.

Baseline and Advanced Error Reporting

PCI Express signals bus and device errors over the same Links using in-band messages. This signaling scheme is much simpler than PCI and PCI-X because it does not require separate side-band signals.

Like network packets, data packets for PCI Express contain headers with error detection capabilities. PCI Express defines an extensive verification mechanism, using cyclic redundancy checks (CRC) to ensure that data that has been sent from one device is received by the other device without error. If an error is detected, it can be signaled either by using the Completion Status field in the header of a Completion packet, or by using a separate error message packet. The transmitting PCI Express device will be requested to resend the data that is found to be in error.

The PCI Express Base Specification defines correctable and uncorrectable errors for two types of error reporting, baseline and advanced. Baseline error reporting defines the minimum error detecting and reporting capabilities that all PCI Express devices are required to implement. Advanced error reporting is optional and allows for richer bus and device error detection and handling. Both the baseline and the advanced error reporting capabilities provide a vast improvement over the simple PERR and SERR error reporting used in PCI and PCI-X.

The following list shows some examples of the error conditions that PCI Express can detect and report:

- Malformed packets
- Receiver overflow
- Completer aborts
- Completion timeouts
- Flow control
- Unexpected completion

Support with Windows

This section provides information about Microsoft plans to support PCI Express on current and future versions of Windows. For the purposes of this paper, the current versions of Windows are Windows 2000, Windows XP, and Windows Server 2003 (including the latest service pack for each); the future version is Windows Longhorn.

Note: Microsoft does not plan to support PCI Express in previous versions of Windows.

Support in Current Versions of Windows

PCI Express is fully software compatible with *PCI Local Bus Specification Revision 2.3*. PCI Express will support all PCI features used by the current Windows operating systems; that is, applications and drivers that currently work with PCI on Windows® 2000, Windows XP, Windows Server 2003 and later versions do not need to be modified to support PCI Express.

To take advantage of some of the advanced features of PCI Express with the current Windows operating systems, ACPI firmware can be used to enable the following features:

Hot Plugging

ACPI firmware can work in conjunction with current versions of Windows to configure and comprehend PCI Express hot plug events. PCI Express hot plug events are captured at the Root Ports, and a hot plug signal is sent to the General Purpose Event (GPE) block. The GPE block will use the System Control Interrupt (SCI) to notify the operating system of the event. The operating system will then invoke the proper ACPI firmware method to service the hot plug event.

Power Management

PCI Express fully supports *PCI Bus Power Management Interface Specification Revision 1.1* and the current operating systems can take full advantage of this support.

PME. ACPI firmware can work in conjunction with current versions of Windows to configure and comprehend PCI Express PME events. PCI Express PME messages are captured at the Root Ports and a PME signal is sent to the GPE block. The GPE block will notify the operating system of the event via the System Configuration Interrupt (SCI). The operating system will then invoke the proper ACPI firmware method to service the event.

ASPM. The ACPI firmware can enable this feature before the operating system boots. After the initial setting is complete, ASPM is handled by the hardware and is fully automatic.

Slot Power Budgeting. ACPI firmware can set up slot power budgeting for each add-in adapter before booting an operating system.

For complete details about these advanced features, see the following WinHEC 2004 slide presentations available at:
http://www.microsoft.com/whdc/winhec/pres04-tech.mspx

- *Implementing PCI Express on the Current Windows Operating Systems (Part 1)*
- *Implementing PCI Express on the Current Windows Operating Systems (Part 2)*

Support in Windows Longhorn

Microsoft plans to provide native support for many of the advanced features of PCI Express in Windows Longhorn. Native support in the operating system will reduce the overall engineering and support cost for system manufacturers and will help to provide a common user experience.

For complete details, see the WinHEC 2004 presentation titled *PCI Express and Windows Longhorn* available at:
http://www.microsoft.com/whdc/winhec/pres04-tech.mspx

Support in Windows Longhorn is planned for the following PCI Express advanced features:

Extended Configuration Space

Support multiple segments as well as access to the 4096 bytes of extended Configuration Space per function for each device on the bus using memory mapped configuration access.

Hot Plugging

Support native PCI Express Hot Plug for ExpressCards (surprise removal), Server I/O Modules (non-surprise removal), and Edge Cards (normal PCI Express plug-in adapters, non-surprise removal). ACPI firmware support is no longer required. This will reduce ACPI firmware development, test, and debugging cost for the OEMs. It will also provide a common hot plug user experience across different platforms.

MSI/MSI-X

Provide MSI/MSI-X support with single or multiple messages per device.

Power Management

Windows Longhorn will provide full support for *PCI Bus Power Management Interface Specification Revision 1.1*.

PME. Windows Longhorn will provide native PME support. ACPI firmware support is no longer required, which will reduce ACPI firmware development, test, and debugging cost for OEMs. It will also provide a common PME user experience across different platforms.

ASPM. Windows Longhorn will support the configuration of ASPM for hardware devices depending on the user-specified power model. ASPM can increase battery life for mobile systems as well as reduce power consumption and thermal issues for desktop systems and blade servers.

Advanced Error Logging and Reporting: Windows Longhorn will provide native support for this feature to enhance reliability, accessibility, and serviceability (RAS) for servers. RAS enables system administrators to monitor the health of the server system, and it assists in the diagnosis of hardware problems. This reduces downtime and service costs for the customer.

Other general PCI support that Microsoft plans to implement in Windows Longhorn include:

I/O Reduction. Itanium-based and x86-based systems have only 64K of I/O space, which is a scarce resource used by both Endpoints and PCI-to-PCI Bridges. Since each PCI-to-PCI Bridge can consume 4K or more of I/O resources, the available I/O space can quickly become depleted. This can be especially problematic for PCI Express systems where Switches, with upstream and downstream Ports

represented by virtual PCI-to-PCI Bridges, are used to fan out PCI Express devices and slots. The I/O reduction feature enables Endpoints and PCI-to-PCI Bridges to turn off their unneeded I/O BARs, thus saving I/O resources. The saved I/O resources can be redistributed to other Endpoints and PCI-to-PCI Bridges that really need them.

For more information about IO Reduction, see *I/O Resource Usage Reduction*, available at:
http://www.microsoft.com/whdc/winhec/pres04-tech.mspx

Multi-level Resource Rebalance. Multi-level resource rebalance is closely associated with hot plug technology. The multi-level resource rebalance feature enables the operating system to revoke and redistribute the resources (I/O, memory, and interrupt) on the bus upon a change event, such as a device hot insertion or removal, in the hierarchy of the bus. This is a much more efficient and flexible resource allocation mechanism compared to the static resource allocation mechanism implemented in the current Windows operating systems.

For more information about multi-level resource rebalance, see *PCI Multi-level Rebalance in Windows Longhorn*, available at:
http://www.microsoft.com/whdc/winhec/pres04-tech.mspx

Resources and More Information

Windows Logo Program Issues

New “Designed for Windows” logo program requirements are in place for both *Microsoft Windows Logo Program System and Device Requirements Versions 2.2* and *3.0*. System manufacturers, firmware engineers, device manufacturers, and driver developers need to review and understand this information to ensure the proper implementation of PCI Express hardware and software for the Windows operating systems.

For additional details, see the WinHEC 2004 presentation titled *PCI Express and Windows Longhorn*, available at:
http://www.microsoft.com/whdc/winhec/pres04-tech.mspx

Call to Action

System manufacturers, firmware engineers, device manufacturers, and driver developers should review the information in this paper and develop hardware and devices that take advantage of PCI Express support in Windows operating systems.

Feedback

To provide feedback about this paper, please send e-mail to **pciesup@microsoft.com**.

References

Specifications

PCI Bus Power Management Interface Specification Revision 1.1
http://www.pcisig.com/specifications/conventional/

PCI Express Base Specification Revision 1.0a
http://www.pcisig.com/specifications/pciexpress/base/

PCI Local Bus Specification Revision 2.3
http://www.pcisig.com/specifications/conventional/

ExpressCard
http://www.microsoft.com/whdc/system/bus/ExpressCard/default.mspx
http://www.ExpressCard.org/

Power Management on WHDC
http://www.microsoft.com/whdc/system/pnppwr/powermgmt/default.mspx

Windows Logo Program for Hardware
http://www.microsoft.com/whdc/winlogo/default.mspx

WHQL Test Specifications, HCTs, and testing notes
http://www.microsoft.com/whdc/hwtest/pages/specs.mspx

Glossary

The following terms are used in this paper:

Active State Power Management (ASPM)
A hardware-autonomous mechanism used to reduce the power consumed by a PCI Express Link through hardware management rather than software management.

Device
A logical device that corresponds to a PCI device Configuration Space. This can be either a single or multi-function device. Examples of devices are: PCI Express-to-PCI Bridges, PCI Express-to-PCI-X Bridges, PCI Express Switches, and PCI Express Endpoints.

Downstream
The relative position of a device on the bus in relation to another device and the Root Complex. For example, if device B is on the secondary side of device A, then device B is further from the Root Complex than device A and is said to be downstream from device A.

Endpoint
A device with a type 00h Configuration Space header.

ExpressCard
New PCMCIA-defined standard for electrical interface to replace CardBus for mobile systems; also suitable for use with desktop PCs. The ExpressCard interface contains both a x1 PCI Express and a USB 2.0 bus interface. The USB 2.0 bus interface allows for the easy migration of USB 2.0 devices to ExpressCard form factors.

Extended Configuration Space
A 4096-byte region of Configuration Space allocated for the configuration of advanced features per device function for a PCI Express device. PCI Express extends this region beyond the 256-byte Configuration Space used by PCI by dividing the PCI Express Configuration Space into two regions: the PCI-compatible region (that is, the first 256 bytes) and the extended region (the remaining 3840 bytes). The extended region is useful for complex devices that require large numbers amounts of registers to control and monitor the device. PCI Express maintains compatibility with the existing PCI enumeration and configuration software in the first 256 bytes.

Fabric
A fabric is composed of point-to-point Links that interconnect a set of components.

Lane
A set of differential signal pairs, one pair for transmission and one pair for reception.

Link
The collection of two Ports and their interconnecting Lanes. A Link is a dual simplex communications path between two components. A Link consists of one or more physical Lanes.

Packet
A fundamental unit of information that is transferred across a Link. A packet consists of a header that, in some cases, is followed by a data payload.

Peripheral Component Interconnect Special Interest Group (PCI-SIG)
Formed in 1992, the PCI-SIG is the industry organization chartered to develop and manage the PCI, PCI-X, and PCI Express standard.

Port
1. Logically, an interface between a component and a PCI Express Link.

2. Physically, a group of transmitters and receivers located on the same chip that defines a Link.

Quality of Service (QoS)
Attributes affecting the bandwidth, latency, jitter, relative priority, and so on, for differentiated classes of traffic.

Receiver
The component that receives packet information across a Link.

Reliability, Accessibility, and Serviceability (RAS)
Enables system administrators to monitor the health of the server system and assists in the diagnosis of hardware problems. This reduces downtime and service costs for the customer.

Request
A packet used to initiate a transaction sequence. A request includes operation code and, in some cases, address, data, length, or other information.

Root Complex
The heart of the PCI Express interface, is the Root Complex, which is made up of one or more Host Bridges. Each Host Bridge exposes one or more Root Ports, which appear as PCI-to-PCI Bridges to software, and can be used to connect other PCI Express devices, such as Endpoints or Switches, to the Root Complex. System CPU(s) and system memory are also connected to the Root Complex, usually through a system bus.

Root Port
A PCI Express Port in a Root Complex that maps a portion of the hierarchy through an associated virtual PCI-to-PCI Bridge.

Segment
A logical grouping of up to 256 buses that the CPU addresses using the same system specific base address.

Slot Power Budgeting
PCI Express mechanism used by software to manage the amount of power that each add-in adapter can use. Enables fine tuning and control of the system's overall power usage.

System Bus
The bus between the CPU and the core chipset. The system bus is also called the CPU local bus.

Transmitter
A component that sends packet information across a Link.

Upstream

The relative position of a device in relation to another device and the Root Complex. For example, if device A is on the primary side of device B, then device A is closer to the Root Complex than device B and is said to be upstream from device B.