The Emergence of PCI Express* in the Next Generation of Mobile Platforms

Mohammad Kolbehdari, Digital Enterprise Group, Intel Corporation
David Harriman, Digital Enterprise Group, Intel Corporation
Altug Koker, Mobility Group, Intel Corporation
Seh Kwa, Mobility Group, Intel Corporation
Brad Saunders, Mobility Group, Intel Corporation

Index words: PCI Express* architecture, ExpressCard*, mini-card, GMCH, ICH, SIOM, DLLP, TLP, PHY, power management (PM), reliability, high-end graphics, PC built on Intel® Centrino™ mobile technology

ABSTRACT

The PCI Express* architecture, both as a unified foundation of I/O, graphics, and networking interconnections, and as a preeminent building block on chip-to-chip, board-to-board, and system-to-system, is widely adopted by multiple market segments in the computing and communication industries. PCI Express architecture is a state-of-the-art serial interconnect technology that keeps pace with recent advances in processor and memory subsystems. From its initial release at 0.8 V, 2.5 GHz, the PCI Express technology has evolved to the general-purpose interconnect of choice for a wide range of applications, including graphics, storage, networking, etc. The PCI Express architecture retains the familiar PCI software and configuration interfaces for seamless migration and adoption in desktop and mobile PC platforms, enterprise servers and workstations, and, increasingly, a wide range of communication and embedded systems.

The PCI Express technology addresses requirements from multiple market segments in the computing and communication industries, and it supports chip-to-chip, board-to-board, and adapter solutions at an equivalent or lower cost than existing PCI designs. Currently, PCI Express architecture supports a 2.5 GT/s signaling rate that yields 500 MB/s bandwidth per lane and a maximum bandwidth of 16 GB/s in a 32-lane configuration. Consistent with the expected cadence of I/O performance progression, the next generation of the PCI Express interconnect will support a signaling rate of 5 GT/s, doubling the performance of the existing links.

The PCI Express interconnect provides numerous architectural improvements over existing I/O technologies. It defines a native hot-plug scheme, enables aggressive power management, provides advanced Reliability, Availability, and Serviceability (RAS) and Quality of Service (QoS) features, and simplifies PCB layouts. In this paper, we discuss the unique, universal capabilities and values of PCI Express technology emerging in the next generation of mobile platforms. We focus on PCI Express architecture, power management, and mobile applications such as graphics, networking, and form factors including the ExpressCard* module as well as future form factors such as the PCI Express Wireless Card and the PCI Express Mini Card. Finally, we cover the next generation of mobile PC platforms built on the Intel® Centrino™ mobile technology.

INTRODUCTION

The PCI Express technology has emerged as the platform I/O solution of choice in the computing and communications industries. Some significant features of the PCI Express technology are also geared towards the next generation of mobile platform designs especially the Active State Power Management (ASPM) capability that enables aggressive power management within the link

* Other brands and names are the property of their respective owners.

® Intel is a registered trademark of Intel Corporation or its subsidiaries in the United States and other countries.

™ Centrino is a trademark of Intel Corporation or its subsidiaries in the United States and other countries.

layer as defined in the PCI Express Base specification. Support for ASPM varies by form factor; however, all mobile electromechanical specifications require ASPM support by PCI Express applications. In this paper, we focus on PCI Express applications, PCI Express architecture and protocols, as well as PCI Express form factors for next-generation mobile platforms and high performance, such as ExpressCard module defined by the Personal Computer Memory Card International Association (PCMCIA) industry group and the PCI Express Mini Card defined by the PCI Special Interest Group (PCI-SIG).

Additionally, the next-generation Intel Centrino mobile technology platform contains a variant of the Intel 915 Express as its "root complex." This chipset is comprised of a PCI Express x16 link optimized for high-performance graphics applications chipsets, the Graphics Memory Controller Hub (GMCH), and the I/O Controller Hub (ICH), which provide four PCI Express x1 links for general-purpose I/O applications. These latter links can be used to implement a combination of PCI Express Mini Card sockets, slots for ExpressCard modules, and other PCI Express devices on the system board that enable communications, storage, or other peripheral applications. If additional ports are necessary to implement a higher-end platform, OEMs may add a PCI Express switch to the ICH to increase I/O fanout. To facilitate a smooth, gradual transition for legacy devices, the conventional, multi-drop PCI bus continues to be supported via the ICH.

PCI EXPRESS ARCHITECTURE

Architectural Overview

The PCI Express architecture uses familiar software and configuration interfaces, the conventional PCI bus architecture, by providing a new high-performance physical interface and numerous new and enhanced capabilities that are built into a framework that retains software compatibility with the existing conventional PCI infrastructure. The enhanced capability is new and different from the legacy, parallel PCI bus architecture and supports scalable link widths in 1-, 2-, 4-, 8-, 16- and 32-lane configurations.

PCI Express power management is built on the PCI Power Management (PCI-PM) software architecture defined for conventional PCI. Additional PCI Express specific power-management capabilities are defined that further extend power manageability by allowing direct hardware control of power states with entry and exit latencies that are low enough to be effectively invisible to software.

PCI Express supports a native hot-plug architecture. Hot-plug support requirements vary by form factor and are documented in the various PCI Express electromechanical specifications. The hot-plug "Toolkit" is defined in the PCI Express base specification to ensure a consistent, common interface for system software to correctly manage hot-plug operations.

PCI Express architecture supports numerous Reliability, Availability, and Serviceability (RAS) features such as error detection and reporting that can be detected using an End-to-end Cyclic Redundancy Check (ECRC) and a Cyclic Redundancy Check (CRC). Erroneous packets are corrected, and the reporting and logging of error conditions is considerably expanded. Mechanisms such as traffic service differentiation, including architecturally defined mechanisms for system control of arbitration, are also provided.

Support for multiple form factors was considered from the outset of the architecture definition. In addition to a card form factor (similar to the existing mainstream PCI cards/slots), there are other form factors defined such as the PCI Express Mini Card, also similar to the existing conventional Mini PCI form factor for notebooks, the ExpressCard standard, meant to replace the CardBus PC Card* form factor, a wireless card form factor designed to fit in the lid of notebook computers, an enterprise-class hot-plug module commonly called a Server I/O Module (SIOM), and several mezzanine cards and blade modules. Additional form factors are being defined.

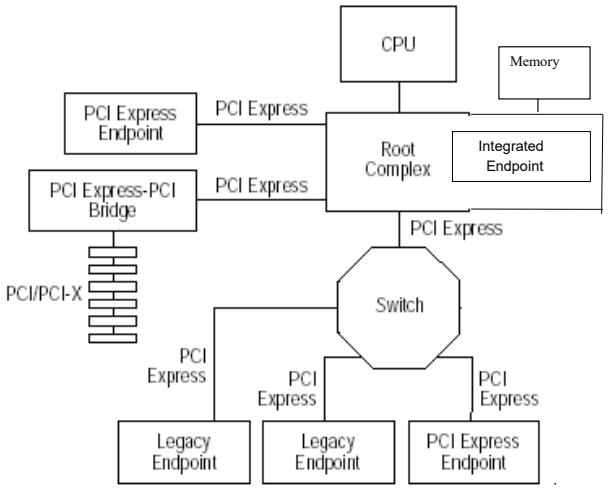

Figure 1: PCI Express architecture system

Figure 1 shows an example of the PCI Express architecture system. There are six types of PCI Express devices. Endpoints include devices such as network and disk interfaces, and they are categorized as either PCI Express Native endpoints or Legacy endpoints. The

* Other brands and names are the property of their respective owners.

Legacy endpoint category is intended for existing designs that have been migrated to PCI Express, and it allows legacy behaviors such as use of I/O space and locked transactions. PCI Express endpoints are not permitted to require the use of I/O space at runtime and must not use locked transactions. By distinguishing these categories, it is possible for a system designer to restrict or eliminate legacy behaviors that have negative impacts on system performance and robustness. PCI Express/PCI Bridges allow older PCI devices to be connected to PCI Express-based systems. Since PCI Express is a point-to-point interconnect, switches are used to increase connectivity. The Root Complex (RC), which includes the Graphics Memory Controller Hub (GMCH) and the I/O Controller Hub (ICH), is at the top of a PCI Express hierarchy, and it connects PCI Express to the CPU and main memory.

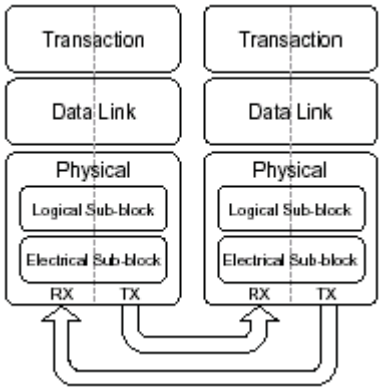

Figure 2: PCI Express link and layered architecture

Figure 2 shows two components connected by PCI Express and highlights the layered structure of the PCI Express interface at the upper component. It should be noted that the layering structure shown represents the way PCI Express is described, but is not an implementation requirement. The Physical Layer (PHY) initialized the link between the two components on a link, and it manages low-level aspects of data transfer and power management. The Data Link Layer (DLL) provides reliable data transfer service to the Transaction Layer (TL) and also a lower overhead communication mechanism for link management of flow control and power. Data packets generated and consumed by the DLL are called Data Link Layer Packets (DLLPs). The TL generates and consumes data packets used to implement load/store data transfer mechanisms and also manages the flow control of those packets between the two components on a link. Data packets generated and consumed by the TL are called Transaction Layer Packets (TLPs).

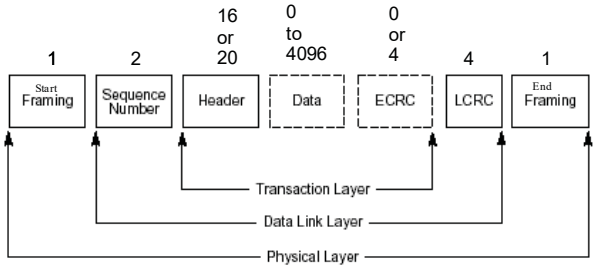

Figure 3: Representation of a Transaction Layer Packet as transmitted over a link

Figure 3 shows how a TLP generated by the TL is modified as it flows through the DLL and PHY for transmission. The Header describes the type of packet and includes information needed by the receiver to process the packet, including any needed routing information, and it is either 12 bytes or 16 bytes in size. Some TLPs include data that are included following the Header. Up to 4096 bytes of data can be included in a TLP, although in a specific platform the maximum size may be limited to a smaller value. The number of bytes of data is always a multiple of 4. The TL may append a 4-byte ECRC, which will travel with the TLP throughout the PCI Express hierarchy "end-to-end." The packet formed by the TL is delivered to the DLL that adds a Sequence Number and a local CRC called a "LCRC." Both of these are used by the receiving DLL to determine that the TLP was received without corruption and without packet loss. Finally, the TLP is delivered to the PHY, which converts it from a sequence of 8-bit bytes to a sequence of 10-bit symbols and adds framing symbols at the start and end. This sequence of symbols is then transmitted across the link to the other component, which checks and disassembles the TLP using a process that mirrors the assembly process at the transmitter.

We now consider in more detail the layered structure of a PCI Express interface.

Electrical Signaling

PCI Express preserves the PCI and PCI-X* architecture with additional new PHY electrical sub-block features as follows:

- *Embedded clock*. Unlike PCI and PCI-X, the clock is not an external data or strobe, but it is embedded in the data signal itself. This new capability enables PCI Express to reduce the clock timing skew.
- *Differential link*. This is the difference in the D+ and D- signals. It forms the signal amplitude and has the

* Other brands and names are the property of their respective owners.

advantage of low crosstalk, low Inter Symbol Interference (ISI), improved Electromagnetic Compatibility (EMC), low common mode, and shared reference ground, among others.

- Full-duplex transmitting with independent transmitter and receiver signal links.
- *AC coupling.* This isolates each transmitter and receiver enabling both hot-plug and hot-swap capabilities and also decodes the transmitter and receiver common mode voltage.
- *8B/10B encoding scheme*. This significantly improves EMC and electrical signal performance. This feature enables the frequency band range limit by encoding each 8-bit byte data to one of two defined 10-bit codes given in the specification. The numbering disparity for each symbol is controlled by this feature maintaining the maximum difference between the number of zeros or ones transmitted across multiple symbols at a delta of 2. This difference keeps a DC balance at both the transmitter and the receiver.
- *Data scrambling with a polynomial equation*. This transfers an 8-bit data byte to a different data byte value that reduces Electromagnetic Interference (EMI) radiation and emission.
- De-emphasis on functionality.
- Polarity inversion and lane reversal.

PCI Express architecture is attributed with ports, links, lanes, and detect. A port is a group of transmitters and receivers on a component that constitutes a link when connected to another component. A line is one set of transmit and receive differential pairs. Link manages an intelligent training sequence to detect and establish deterministic links between component ports. Detect is a feature when an upstream device, RC, or north bridge polls each transmit lane to detect an RC time constant that indicates that a load is present. Once the upstream device detects a load, it initiates a polling process to determine whether the load is an active device or a test load. Figure 4 shows the concepts of ports, links, and lanes.

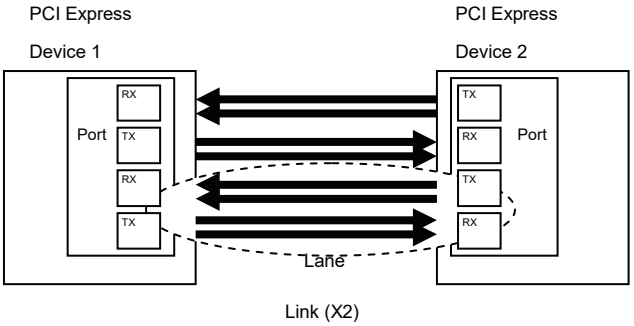

Figure 4: The concept of ports, links, and lane

The PCI Express base specification [1] requires AC coupling between the transmitter and receiver. It also requires each end of the link to be on-die terminated with a 100 Ohm differential. In addition, the PCI base specification requires each transmitter (TX) component to use on-die equalization or de-emphasis with a typical value of 3.5 dB +/- 0.5 dB. The PCI base specification defines the Unit Interval (UI) of 400 ps +/- 300 parts per million (ppm). The PCI Express timing specification defines eye diagrams, one at the transmitter end and one at the receiver end. Also the PCI Express Card Electromechanical Specification [2] defines a specific eye diagram with given height and width for the add-in card and system board. The PCI-base specification allows certain insertion and return losses to ensure that the transmitter signal has adequate amplitude at the receiver signal pin. It also defines a rise and fall time of 0.125 UI or 50 PS. For all the electrical and AC specifications see the PCI Express base specification and also the PCI Express Card Electromechanical Specification [1-2].

Protocol

The PHY Electrical sub-block, described above, is controlled by the Logical sub-block that implements link initialization and low-level control functions as well as packet framing and, for multi-lane links, distribution of packet information across the lanes. The Link Training and Status State Machine (LTSSM) is responsible for establishing the link between two components at the PHY level, including the width of a multi-lane link. When establishing a link, the LTSSM starts in the Detect state, moving to the Polling state when another component is found to be present on the link. Once the two components have established a data transfer connection, they enter the Configuration state where the configuration of the link itself is negotiated between the two components. Finally the L0 state is entered to establish normal link operation. The LTSSM is also capable of re-establishing the link if it is "lost" due to an electrical disturbance or a power state transition, using the Recovery state. It also manages link power state transitions at the PHY level using the L0s, L1, and L2 states.

Depending on system clocking, it is possible that the transmitter and receiver will operate at slightly different clock rates. The PHY uses a periodic transmission of a specific sequence of symbols to allow the receiver to compensate for any drift between the two components.

In a multi-lane link it is also necessary for the receiver to compensate for any differences in skew between the multiple lanes of the link. This is done by transmitting specific sequences of symbols simultaneously on all lanes such that the receiver can detect and compensate for any differences in the receipt of the sequences on a lane-by-lane basis. In this way, the other layers see the received data as all arriving at the same time.

Once the physical link is established by the PHY, the DLLs and TLs of the two components will then exchange flow control information to allow the transmission of TLPs between the two components. This initial exchange of flow control information is managed by the Data Link Control and Management State Machine of the DLL. Once both components have received the required information, TLP communication starts. The DLL implements a retry mechanism that stores transmitted TLPs until they are acknowledged by the receiver, presenting the view of a reliable link to the TL.

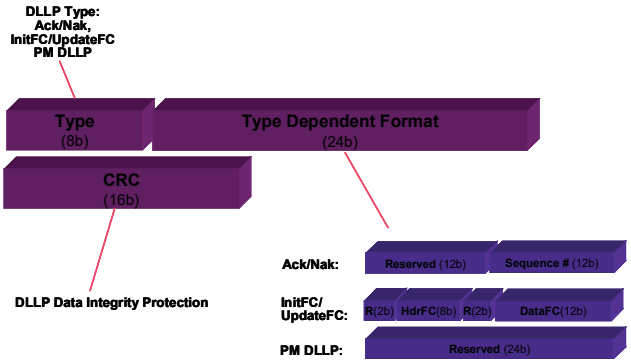

Figure 5: Data Link Layer Packet (DLLP) format

The retry mechanism, continued flow control information exchange, and certain power-management protocols use DLLPs, of which there are 15 specific types defined. Figure 5 shows the format of the 6-byte body of a DLLP, not including the framing information added by the transmitting PHY and removed by the receiving PHY. The first byte describes the type of the DLLP. The format of the following three bytes depends on the type of the DLLP. The final two bytes are a 16-bit CRC that allows the receiver to determine that a DLLP was received without corruption. DLLPs that are corrupted in transit are simply discarded by the receiver, and all protocols using DLLPs ensure that such loss will be corrected by a subsequent transmission.

There are two types of TLPs used by the TL: Request TLPs, and Completion TLPs, simply called "Requests" and "Completions," respectively.

A schematic representation of the Header of a Request is shown in Figure 6. Note that this figure does not represent the actual bit positions of the bits in the Header, but is simplified to highlight the functionality of the fields. The Fmt and Type fields encode the type of layout of the TLP, for example to encode a Configuration Write or a Memory Read. The Length field encodes the length of data included with the TLP, or the amount of data requested to be returned for a Read Request. The TC and Attr fields are used for traffic service differentiation. The TD field indicates TLP includes an ECRC. The EP field indicates that a TLP includes data that are "poisoned" i.e., known by the transmitter to be erroneous. The Requestor ID is used to route a Completion, if needed, back to the Requestor, and the Tag is used by the Requestor to distinguish different Completions for multiple outstanding Requests. The field labeled "BE/Msg. Code" is used for Configuration, I/O, and Memory Requests to encode the byte enables for the Request, and for Message Requests to encode the Message type. The remainder of the Header is used for routing information, either an address in Memory or I/O space, or the unique ID of a specific device as used for a Configuration Request.

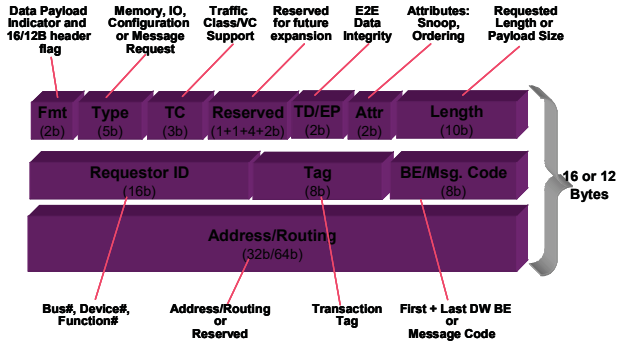

Figure 6: Representation of the format of a request TLP

Completion Headers are similar to Request Headers. Figure 7 shows a schematic representation of a Completion Header. Completions always use the 12-byte Header size, because full address information is not required. Descriptions of fields not present in the Request Header follow.

The Status field communicates to the Requester that its Request was completed successfully or with some type of error. In addition to a Successful Completion status, there are two error statuses: Unsupported Request and Completer Abort, which correspond to Master Abort and Target Abort in conventional PCI. There is also a

completion status used only in association with an initial configuration Request to a device (Configuration Retry status), which indicates that the Completer requires more time before it is able to service the Configuration request. Completion headers include two ID fields: the Requestor ID and Tag. These are the same as the Requestor ID and Tag values supplied with the corresponding Request. The Completer ID identifies the entity that serviced the Request to generate the Completion, and uses the same format as the Requestor ID. Byte Count and Lower Address can be used by Requestors to simplify processing of data returned in Completions for Memory Read requests.

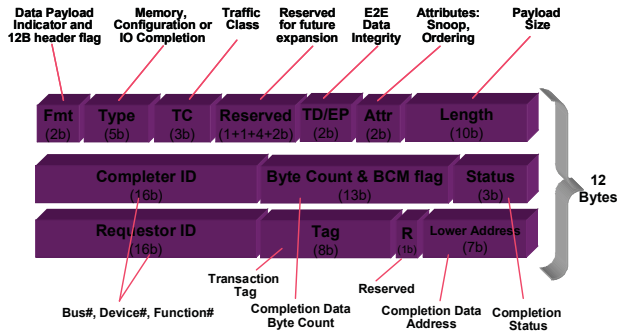

Figure 7: Representation of the format of a Completion TLP

PCI Compatibility and New Software Features

PCI Express devices include the same Configuration headers as corresponding conventional PCI devices. System software that comprehends conventional PCI works with PCI Express devices using these configuration mechanisms. For systems with updated system software, PCI Express expands the capabilities available to software to provide for native hot-plug support, improved data integrity, and error-reporting mechanisms, all of which are covered in this section, as well as more advanced power management, which is covered in the following section. For additional new capabilities not covered in this paper, refer to the PCI Express Specification [1]. PCI Express provides a new memory-mapped configuration access mechanism to replace the windowing mechanism used for conventional PCI in PC architecture systems, and it provides each device with an expanded configuration space of 4 KB compared to 256 B in conventional PCI.

PCI Express defines a register interface to support a "toolbox" of hot-plug capabilities that can be used by different form factors to allow different usage models with a consistent system software interface. Each element of the toolbox that a form factor may specify as a required or optional element of hot-plug support for that form factor has an architected capability bit associated with the port connected to the slot or connector. Table 1 lists the supported elements that form factors may employ to support a particular hot-plug usage model, although it should be noted that in the mobile environment only the Surprise Remove capability is used.

Table 1: Elements of the PCI Express hot-plug "toolbox"

Attention Button	Used to request hot-plug operations. Note: A software user interface may be used instead or in addition.
Attention Indicator	Indicates slot and/or adapter that requires user attention.
Power Indicator	Indicates that the slot has power.
Electromechanical Interlock	Prevents removal of an adapter from a slot under system software control.
Manually operated Retention Latch (MRL) Sensor	Allows software detection of MRL operation by the user, for example to indicate that the user is about to remove an adapter.
Power Controller (software controlled)	Allows software control of slot power.
Hot-Plug Surprise	Indicates that an adapter removal without advanced software notification is permitted by this form factor.

If an element is included, it must be used in a manner consistent with certain defined usages.

The ExpressCard form factor permits surprise removal of the adapter, and does not include a software-controlled power controller or support for the button/indicator user interface. The toolkit approach allows a common system software implementation to support all PCI Express hot-pluggable form factors.

POWER MANAGEMENT

PCI Power Management is an optional capability for conventional PCI devices, but is required for PCI Express devices. The PCI power management states are mapped to PCI Express link states to allow existing system software to power-manage PCI Express Links without having to comprehend the actual link states.

Additionally, the PCI Express specifications define an Active State Power Management (ASPM) capability that enables aggressive power management operating within the link layer. Support of ASPM is form-factor specification dependent, and all of the mobile platform

form-factor specifications call its implementation out as a mandatory feature of all PCI Express end-point links.

Device Power States

The Advanced Control and Power Interface (ACPI) specification defines component device power states (D-states) that allow the platform to establish and control power states for the component ranging from fully ON to fully OFF (drawing no power) and various in-between levels of power-saving states, annotated as D0-D3. Similarly, PCI Express defines a series of link power states (L-states) that work specifically within the link layer between the component and its upstream PCI Express port (typically in the host chipset). For a given component D-state, only certain L-states are possible as detailed below.

- *D0 (Fully on)*: The device is completely active and responsive during this D-state. The link may be in either L0 or a low-latency idle state referred to as L0s. Minimizing L0s exit latency is paramount for allowing frequent entry into L0s while facilitating performance needs via a fast exit. The next lower link power state, L1 state, may be achieved either by hardware-based ASPM or by requesting the link to enter L1 after the OS places the downstream device in a D1-D3 state.
- *D1 and D2*: There is no universal definition for these intermediate D-states. In general, D1 is expected to save less power but preserve more device context than D2. L1 state is the required link power state in both of these D-states.
- *D3 (Off)*: Primary power may be fully removed from the device ($D3_{cold}$), or not removed from the device ($D3_{hot}$). $D3_{cold}$ maps to L2 if auxiliary power is supported on the device with wake-capable logic, or to L3 if no power is delivered to the device. Sideband WAKE# mechanism is recommended to support wake-enabled logic on mobile platforms during the L2 state. $D3_{hot}$ maps to L1 to support clock removal on mobile platforms.

Table 2 summarizes the mapping from D-states to L-states for a PCI Express link.

Table 2: Mapping from D-states to L-states

Downstream Component D-state	Permissible Upstream Component D-state	Permissible L-state
D0	D0	L0, L0s, or L1
D1	D0-D1	L1
D2	D0-D2	L1
$D3_{hot}$	D0-$D3_{hot}$	L1
$D3_{cold}$	D0-$D3_{cold}$	L2 or L3

Active State Power Management

ASPM is the hardware-based capability to power-manage the PCI Express link. Only L0s and L1 are used during ASPM.

- *L0s*: This link state is a very low exit latency (<1 µs) link state intended to reduce power wastage during short intervals of logical idle between link activities. The power-saving opportunities during this state include, but are not limited to, most of the transceiver circuitry as well as the clock gating of at least the link layer logic. Devices must transition to L0s independently on each direction of the link.
- *L1*: This link state is a low exit latency (~2-4 µs) link state that is intended to reduce power when the device becomes aware of a lack of outstanding requests or pending transactions. The power-saving opportunities during this state include, but are not limited to, shutdown of almost all the transceiver circuitry, clock gating of most PCI Express architecture logic, and shutdown of the PLL. CLKFREQ# can be opportunistically de-asserted during L1 to allow for platform reference clock gating.

In a multi-lane PCI Express link, the detection of L0 or L1 exits can be communicated through lane #0 of a configured link. Also note that L1 exit signaling does not have to be derived from Phase Locked Loop (PLL). (If this were not the case, the opportunity to shut down the PLL during L1 would be lost.) More aggressive low-power implementations may consider the minimization of leakage power during L1 state.

PCI Express Link Power States

Configuration of devices into D-states will automatically cause the PCI Express links to transition to the appropriate L-states. Refer to Table 2 for the mapping.

- *L2/L3 Ready:* This link state prepares the PCI Express link for the removal of power and clock. The

device is in the $D3_{hot}$ state and is preparing to enter $D3_{cold}$. The power-saving opportunities for this state include, but are not limited to, clock gating of all PCI Express architecture logic, shutdown of the PLL, and shutdown of all transceiver circuitry.

- *L2*: This link state is intended to comprehend $D3_{cold}$ with Aux power support. Sideband WAKE# signaling is recommended to cause wake-capable devices to exit this state. The power-saving opportunities for this state include, but are not limited to, shutdown of all transceiver circuitry except detection circuitry to support exit, clock gating of all PCI Express logic, and shutdown of the PLL as well as appropriate platform voltage and clock generators.
- *L3 (link off):* Power and clock are removed in this link state, and there is no Aux power available. To bring the device and its link back up, the platform must go through a boot sequence where power, clock, and reset are reapplied appropriately.

L2/L3 Ready and L2 are the link states with the lowest power but longest exit latency whereas L0 is the link state with lower power but short exit latency.

Pipelining and Buffering to Minimize Power and Maximize Performance

Techniques to improve packet scheduling and ensure effective utilization of bandwidth are also essential to a power-optimized architecture. Pipelining packets effectively and providing additional buffering are effective methods to reduce power while sustaining higher levels of performance. Two scenarios are described below as illustrated in Figure 8.

- Scenario (A) shows an "ask for one, get one" approach of utilizing a PCI Express link. One Completion is submitted after its Request is issued. Flow Control (FC) update and Acknowledge packets alternate, assuming no non-recoverable link errors.
- Scenario (B) illustrates the streaming of multiple Requests and subsequently multiple Completions before FC update and Acknowledge packets are issued. Note that the transmitting lane of the Requester is idle more often in scenario (A) than scenario (B), thus presenting longer periods for residence in lower link power states, provided efficient power-management features are implemented. Longer idle periods allow for power down of circuitry that would require non-zero energizing and synchronization time. Scenario (B) also allows the Completer to submit an FC update packet and an Acknowledgement packet after a few TL packets have been transmitted. Since the Completer is aware of the pending transactions, it is able to minimize the overhead of FC update and Acknowledge packets to maximize bandwidth utilization.

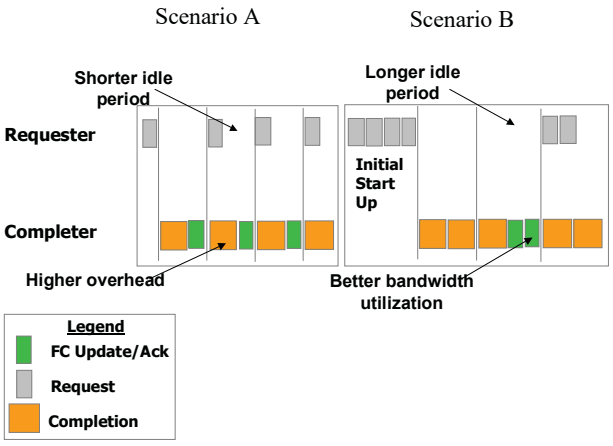

Figure 8: Scenarios to illustrate scheduling and bandwidth utilization impact

PCI EXPRESS MOBILE APPLICATIONS

PCI Express provides a high-performance, adaptable interface technology for the interconnection of applications peripheral to the platform's processor and memory subsystem. For mobile platforms, three key application areas are being initially deployed with PCI Express including the platform graphics and two add-in card formats for extending and upgrading the platform.

Similar to what is happening in the desktop PC space; PCI Express-based graphics are being established as the follow-on technology to existing Accelerated Graphics Port AGP-based graphics and is natively supported in the Intel host chipset (915PM). Also within the host chipset, the PCI Express I/O interconnect covers support for moving peripherals from the traditional PCI bus to PCI Express ports exposed via new add-in card connectors or alternatively available to devices to be integrated into the system board.

In the remainder of this section we focus on the two new add-in form-factors that are enabled by PCI Express technology.

PCI Express MiniCard Electromechanical (CEM) Form-Factor

The PCI Express Mini CEM add-in card shown in Figure 9 provides for expanding and upgrading the communications capabilities of the mobile platform. Replacing the conventional Mini PCI add-in card, this form factor is roughly half the size and is based on PCI

Express x1 and USB 2.0. The specification of add-in card and socket definition is given in the PCI-SIG.

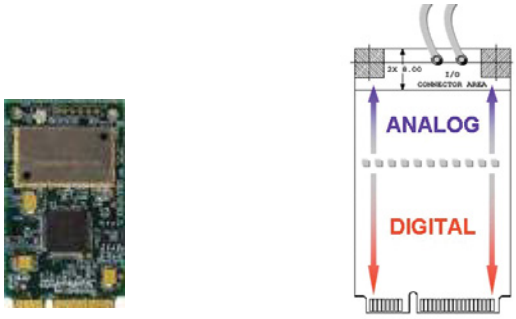

Figure 9: PCI Express mini CEM form factor

Both wired and wireless communications applications are comprehended by the Mini CEM specification. Examples include Local Area Network (LAN, e.g., 10/100/1000 Mbps Ethernet), Wide Area Network (WAN, e.g., V.90/V.92 modem), Wireless-LAN (W-LAN, e.g., 802.11b/g/a), Wireless-WAN (W-WAN, e.g., cellular data), and Wireless-Personal Area Network (W-PAN, e.g., Bluetooth*). Although not specifically considered, other applications may also find their way to this form factor.

BTO/CTO Form Factor

The PCI Express Mini Card specifically targets addressing system manufacturers' needs for build-to-order (BTO) and configure-to-order (CTO) applications rather than providing a general end-user-replaceable module. This form factor has characteristics more typical of an "embedded" application including the platform integration of the media interfaces such as communications connectors or wireless antennas.

The Mini Card and host socket is based on a single 52-pin card-edge type connector for its system interfaces. The host system connector is similar to a Small Outline Dual In-line Memory Module (SO-DIMM) connector and is modeled after the Mini PCI Type III connector but without side retaining clips, rather relying on two card-retention mounting points at the opposite end of the card. The placement of I/O connectors on a Mini Card is at the end opposite of the system connector. Depending on the application, one or more connectors may be required to provide for cabled access between the card and media interfaces such as LAN and modem line interfaces and/or RF antennas.

With both PCI Express x1 and USB 2.0 interfaces defined for the Mini Card and given the BTO/CTO nature of the application, the OEM has the option to only accommodate one of these interfaces in a given platform or socket. The OEM also dictates what I/O connector interfaces are built into the platform. Card vendors are obligated to work closely with each OEM to ensure that the application that they are designing will function properly in the target socket. Due to the mobile focus for this form factor, all of the advanced power-management features of these interfaces are expected to be used aggressively including support for PCI Express ASPM.

Application-Specific Interface Features

The PCI Express Mini CEM specification defines a number of system interface features helpful in the implementation of communications applications. These include defining a Light Emitting Diode (LED) indicator interface, a wireless radio transmitter disable control, remote UIM/SIM socket support, and System Management Bus (SMBus). Most of these features are specific to wireless applications.

The LED interface provides for three separate indicators, one each associated with WLAN, WWAN, and WPAN technologies. A recommended usage model covers indications for various states of radio operation including on, off, searching/associating, and transmit/receive activity. To support emerging requirements for controlling wireless RF transmitters in areas where interference may be a potential hazard, such as on commercial aircraft, a wireless disable signal is provided for so that the OEM can implement an on/off switch in the platform that directly controls the radio's ability to transmit.

For WWAN applications, an ISO standard UIM/SIM socket interface is defined to allow the OEM to support the SIM card needs for GSM/GPRS network account security and management. Due to the small size of the Mini Card, SIM card support on the radio card may not be practical.

The ExpressCard Standard Form Factor

Following the hot plug-and-play usage model successfully established by CardBus PC Card modular add-in cards for mobile platforms, the PCMCIA industry group has developed a PCI Express-based replacement technology released as the ExpressCard Standard. This new technology replaces conventional parallel buses for I/O devices with scalable, high-speed, serial interfaces, either PCI Express x1 for high-performance applications, or USB 2.0 to take advantage of the wide range of USB

* Bluetooth is a trademark owned by its proprietor and used by Intel Corporation under license.

silicon already available. Whichever interface is used; the end user experience is the same.

Module Form Factors

The ExpressCard Standard is designed to deliver high-performance, modular expansion to both desktop and mobile platforms in a lower cost, smaller form factor. Users are able to add a wide variety of applications including memory, wired and wireless communications, multimedia, and security features by inserting ExpressCard modules into compliant systems. At roughly half the size and lighter than today's PC Card, ExpressCard products also leverage the proven advantages of PC Card technology, including reliability, durability, and expansion flexibility while offering improved performance.

There are two standard formats of ExpressCard modules shown in Figure 10: the ExpressCard/34 module (34 mm x 75 mm) and the ExpressCard/54 module (54 mm x 75 mm). Both module formats are 5 mm thick, the same as the Type II PC Card, and at a standard length of 75 mm, are 10.6 mm shorter than a standard PC Card. A common, 26-pin, beam-on-blade style connector designed for high durability and reliability is used for both module formats, and the corresponding host connector accommodates the insertion of either module. The host slot for the ExpressCard/54 module features a novel guidance mechanism that also supports ExpressCard/34 modules by steering the narrower module into the connector socket. In any host system that implements multiple slots, all slots provide equivalent I/O interface functionality.

Figure 10: ExpressCard modules

The two ExpressCard module sizes give system manufacturers greater flexibility than in the past. While the ExpressCard/34 device is better suited to smaller systems, the wider ExpressCard/54 module can accommodate applications that do not physically fit into the narrower ExpressCard/34 form factor such as Smart Card readers, CompactFlash readers, and 1.8" disk drives. With its extra space for components and for spreading thermal energy, the ExpressCard/54 module format is also a natural choice for higher-performance and first-generation applications. The specification also allows for extended module formats to integrate features such as LAN and phone line connectors, or antennas for wireless cards into products.

Hot-Plug Functionality and Power Management

Leveraging PCI Express and USB built-in support for hot-plug functionality, ExpressCard technology is designed to allow users to install and remove modules at anytime, without having to switch off their systems. By relying on the auto-detection and configuration of the native I/O buses, ExpressCard technology can be implemented on a host system without an external slot controller. A device to control power to the slot is required, based on a simple, wired, module presence detection scheme.

Both PCI Express and USB natively support features that enable module applications to be placed in very low power states while maintaining the ability to detect and respond to wake-up requests. For example, an ExpressCard application can receive network messages via a wireless communications module even while the PC is in a sleep state. Additionally, PCI Express reference clock on/off control during ASPM L1/L2 states is supported. Effective use of these features is the key to creating high-performance applications that are both power and thermally efficient.

ExpressCard technology will typically require less electrical power than products built on previous PC Card standards although the slot power specification does provide adequate peak current to meet the needs of applications such as wireless transmitters. Independent of the amount of power drawn from the host system, the ExpressCard Standard also specifies thermal power dissipation at a maximum of 2.1 W for modules. Thermal dissipation limits are specific to the heat released within the slot discounting the energy that is dissipated in module extensions outside the confines of the slot.

User Benefits

Users will be able to identify modules and host systems that are compliant with the ExpressCard Standard by looking for the licensed ExpressCard logo, an energetic rabbit signifying mobility, fast performance, and ease-of-use, either directly on the product, in product documentation, or in the ExpressCard Compliant Products Directory available online at expresscard.org (see Figure 11). The compliance program intends to ensure interoperability between ExpressCard modules and systems using a two-step process consisting of self-compliance testing against a comprehensive requirements checklist and formal interoperability testing between modules and host systems. For product manufacturers, the compliance process also covers key ingredients such as

the connectors and power switch circuits used in building ExpressCard modules and host systems.

Figure 11: ExpressCard logo

By supporting both PCI Express and USB in all compliant module slots, ExpressCard technology brings new functionality to computer users not found in today's PC Card. The technology delivers a consistent, easy, reliable, and non-threatening way to connect devices into their systems. ExpressCard modules can be plugged in or removed at almost any time. With an expected strong adoption in desktop systems, driven by the desire to move these platforms to a sealed-box expansion model, users will find that they will be able to move modules between their desktop and mobile platforms so that they can easily share useful resources and transfer data between the platforms.

PCI EXPRESS MICROARCHITECTURE IN MOBILE PLATFORM

In next-generation mobile platforms, the Intel 915 family of chipsets, the GMCH and ICH incorporate a PCI Express-based interconnect. These two hubs form the RC of the platform. This means they connect a host CPU/memory subsystem to a PCI Express hierarchy. The GMCH contains a PCI Express x16 port that has been designed and optimized for high-end graphics usage. There are also 4x1 general-purpose PCI Express ports that that can be connected to form the rest of the PCI Express fabric such as PCI Express Mini CEM sockets, ExpressCard slots, and system board PCI Express devices that cover communications and storage applications.

PCI Express fabric in next-generation mobile platform microarchitecture implements multiple "virtual channels" that are described in the PCI Express specification-SIG. They are used in order to provide a reliable service for both bandwidth and latency. These virtual channels are controlled and assigned to different peripherals on the platform. Depending on the system needs, they are designed to utilize the system resources, CPU, and memory between I/O connections.

Due to the high-speed serial nature of the PCI Express links on the next-generation mobile platform, power management became one of the focus areas. The microarchitecture of the RC implements aggressive power-control policies to keep the consumed current to a minimum while performing to system needs.

PCI Express Microarchitecture Partitioning

The PCI Express architecture is specified in layers as shown in Figure 2. Compatibility with the PCI addressing model load-store architecture with a flat address space is maintained to ensure that all existing applications and drivers operate unchanged. The PCI Express configuration uses standard mechanisms as defined in the PCI Plug-and-Play specification. The software layers generate read and write requests that are transported by the TL 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 PHY consists of a full-duplex channel that is implemented as a transmit pair and a receive pair for each lane. The initial speed of 2.5 GHz (250 MHz internally) results in 2.5 Gb/s/direction that provides a 250 MB/s communications channel in each direction (500 MB/s total).

PCI Express uses packets to communicate information between components. Packets are formed in the TLs and DLLs to carry the information from the transmitting component to the receiving component. As the transmitted packets flow through the other layers, they are extended with additional information necessary to handle packets at those layers. At the receiving side the reverse process occurs, and packets get transformed from their PHY representation to the DLL representation and finally (for TLPs) to the form that can be processed by the TL of the receiving device.

Power Management Policies

Beyond supporting the already defined Active Link Power states, the next-generation mobile platform implements architecture-specific, power-saving options and policies. These policies in some cases take full advantage and aggressively pursue ASPM states as well as take advantage of notebook form factors.

Aggressive L0s Entry

PCI Express links on the next-generation mobile Platform pursue a quick entry to low-power ASPM states, especially while deciding on L0 to L0s transition. With this aggressive entry policy, in order to minimize the latency issues, the links also implement a very quick exit time in order to service the incoming requests. This policy is observed to be not only helpful in keeping the link power low, but also in packetizing the traffic to increase the ASPM state residency times.

Compressive Core Controls for Analog Logic
Due to the high-speed nature of PCI Express and clock recovery logic, most power on PCI Express links is being consumed at the interface level where analog controls reside. The next-generation mobile platform implements comprehensive controls from the core logic to minimize the use of power both during the active states as well as ASPM states. Some of these controls are outlined below:

- *Single Squelch Detection*: In the next-generation platform, built on Intel Centrino mobile technology, both during L0s and L1 ASPM states, PCI Express link controls enable single-lane squelch detection for exits. This feature reduces the unnecessary use of squelch detection logic across the links and saves significant power for wide PCI Express links.
- *Low-Power Circuits*: In order to save power during active states, PCI Express interfaces use optionally low-voltage swings with reduced drive strengths. For close interface links it is proven to produce significant power savings where the driver current is reduced to half (from 20 mA to 10 mA).

Clock Gating Microarchitecture
The Intel 915 family of chipsets in next-generation mobile platforms uses aggressive clock-gating policies and domains to minimize unnecessary switching across the micro-architecture. In previous microarchitectures, clock gating has been a proven method to reduce power by eliminating unnecessary toggles. For next-generation mobile platforms, we expanded the methodology by grouping different sections of the microarchitecture into separate clocking domains. These domains include

- Separate clocking for different layers.
    - Transaction Layer clocking.
    - Link Layer clocking.
    - Physical Layer clocking .
        - Logical Sub Block.
        - Electrical Sub Block.
- Separate Clocking for RX (receiver) and TX (transmitter).

All the listed clocking sections of the microarchitecture are controlled via a list of indicators that signals whether the block in question is needed or not.

All these power-saving features are included in the next-generation mobile platform to increase the battery life and reduce the thermal envelope on the system. They are implemented in such a way that keeps the balance between a very high-performance I/O connection system, like the PCI Express, and a very power-conscious mobile platform.

SUMMARY

The emergence of PCI Express technology as a unified foundation of I/O, graphics, and networking interconnections for the next generation of mobile platforms built on Intel Centrino mobile technology has been presented. PCI Express architecture as a state-of-the-art platform I/O solution of choice has been discussed with many values such as high performance, high-bandwidth scalability, easy networking connections, improved power management, interoperability, network reliability, native hot-plug capability, and software compatibility. In this paper we focused on areas such as integration of PCI Express technology into the next generation of mobile platforms, by describing the PCI Express architecture and protocols, PCI Express power management, advanced RAS and QoS features, PCI Express mobile applications such as graphics, networking, and form factors including ExpressCard standard, the PCI Express Mini Card as well as future form factors specifications such as the PCI Express Wireless Card and the PCI Express Cable. We also focused on ASPM capability that enables aggressive power-management within the link layer as defined in the PCI Express Base Specification. The next-generation Intel Centrino mobile platform based on the Intel 915 family Express chipsets has also been described with integration of PCI Express technology.

ACKNOWLEDGMENTS

We would like to acknowledge the encouragement and help that we received from our managers, particularly Valerio Jim, Ajay Bhatt, and Bala Cadambi. Additional thanks go out to individuals who reviewed the paper and provided valuable feedback: Ramin Neshati, Dave Puffer, and Aditya Sreenivas.

REFERENCES

[1] PCI Express Base Specification, 1.0a. Portland, OR: PCI-SIG.

[2] PCI Express card Electromechanical (CEM) Specification, Portland OR: PCI-SIG.

AUTHORS' BIOGRAPHIES

Mohammad Kolbehdari is a senior staff hardware engineer in the Digital Enterprise Group. He received B.Sc., M.Sc., and Ph.D. degrees in Electrical Engineering from Communication University Tehran, Iran, University of Mississippi Oxford MS, and Temple University, Philadelphia, Pennsylvania in 1986, 1991, and 1994,

respectively. He conducted research in the EM lab at Temple University prior to joining the Department of Electronics, Carleton University, Ottawa, Canada as a research associate faculty member from 1994 to 1998. He joined Intel in 1998 and has been involved in many product development teams including the Pentium® 4 FSB and PCI Express electrical designs. His areas of interest are EM modeling, power delivery, signal integrity, I/O serial differential design, system-level architecture, platform initiatives, and clock jitter characterization and measurements. His e-mail is mohammad.kolbehdari at intel.com.

David Harriman is a senior staff platform architect in the Digital Enterprise Group. He develops and promotes next-generation platform technologies, focusing on I/O interconnects. He has been a leading contributor to the PCI Express specification since its inception. He joined Intel in 1993, and he worked on the development of the first desktop chipset for the Pentium Pro processors and on the first AGP chipset. He was one of the primary developers of the Intel Hub architecture. He received a Bachelor's degree with dual majors in Electrical Engineering and Computer Science from the University of Wisconsin, Madison in 1993. He holds more than 20 patents relating to chipsets and interconnects. His e-mail is david.harriman at Intel.com.

Altug Koker is a senior staff engineer in the Mobility Group. He received a B.Sc. degree in Electrical Engineering from Bilkent University, Ankara Turkey in 1995 and an M.Sc. degree in Electrical and Computer Engineering from State University of New York at Buffalo in 1997. Altug conducted research on micro electronics while he was at Bilkent University, Ankara prior to joining the Department of Electronics and Computers at the State University of New York at Buffalo where he researched the design of high-performance computers and networks. He joined Intel in 1997 and has been involved in many chipset design projects including the initial Accelerated Graphics Ports (AGP), Pentium III and 4 FSB, and PCI Express ports. His e-mail is altug.koker at intel.com.

Seh Kwa is a staff platform architect in the Mobility Group of Intel Corporation. He joined Intel in 1998 as part of EPG 460GX chipset team before focusing on low-power Intel architecture and power management for the last four years. Prior to joining Intel, he was involved in low-power Digital Signal Processing as well as communication development at two startup/IPO companies. He received B.S.E.E., B.S.C.S, M.S.E.E. and D.Sc. E.E. degrees from Washington University. His e-mail is seh.kwa at intel.com.

Brad Saunders is a senior mobile systems architect focusing on I/O chipset definition in Intel Corporation's Mobility Group. Brad also leads the PCMCIA industry group responsible for the ExpressCard Standard and is the technical editor for both the PCI Express Mini and Wireless Form-Factor (WFF) electro-mechanical specifications within the PCI-SIG. Brad came to Intel Corporation as part of the Xircom, Inc. acquisition in early 2001, where he had spent two years in Xircom's chief technology office. Prior to that, he spent 21 years working in various fields of communications at Rockwell International, ranging from secure communications for defense applications to analog modems for mobile systems. Brad received his B.S.E.E. degree from the University of California, Irvine in 1976. His e-mail is brad.saunders at intel.com.

® Pentium is a registered trademark of Intel Corporation or its subsidiaries in the United States and other countries.