PCI Express* Architecture Power Management

Rev 1.1

SEH KWA AND DEBRA T. COHEN – INTEL CORPORATION

White Paper - November 8, 2002

Information in this document is provided in connection with Intel products. No license, express or implied, by estoppel or otherwise, to any intellectual property rights is granted by this document. Except as provided in Intel's Terms and Conditions of Sale for such products, Intel assumes no liability whatsoever, and Intel disclaims any express or implied warranty, relating to sale and/or use of Intel products including liability or warranties relating to fitness for a particular purpose, merchantability, or infringement of any patent, copyright or other intellectual property right. Intel products are not intended for use in medical, life saving, or life sustaining applications. Intel may make changes to specifications and product descriptions at any time, without notice.

Intel
Research &
Development

PCI EXPRESS* ARCHITECTURE POWER MANAGEMENT

November 2002 Rev 1.1

ABSTRACT

This paper presents power management guidelines for PCI Express links on Intel-based Mobile platforms. It describes the mapping from platform sleeping states and device power states to link power states, including the procedure to support Mobile-specific S1/POS and CPU C3/C4 scenarios. Device and platform power saving opportunities are identified for each link power state. L1 entry policy is also recommended to optimize device power. Several power optimization techniques are described, including minimizing flow control updates and acknowledgement packets to improve bandwidth efficiency, and pipelining packets to increase opportunities for active state link power management. These power management guidelines enable architectural innovation to achieve power-optimized interconnect performance.

Contents

Figures

Tables

Revision History

Rev.	Description	Date
1.0	Initial Release	September 2002
1.1	Corrected the guidance for mapping link states to Mobile S1/POS in section 2.1.	November 2002

1. Introduction

1.1 Purpose of the Document and Target Audience

This document is a collection of guidelines and recommendations for Intel-based Mobile PC platforms with PCI Express interconnect technology. The intent is to provide information that will improve PCI Express architecture enabled notebook performance while at the same time minimizing power consumption to reduce thermal problems and to maximize battery life. This is by no means intended to be a comprehensive description of all possible optimizations. It is also not intended to be used as a standard or specification.

The target audience for this document is architects, engineers, and system developers whose role is to develop devices or systems incorporating PCI Express technology.

1.1.1 Scope

This document focuses on power management guidelines for the PCI Express architecture. Specifically, it addresses the PCI Express architecture's L-states (link power states) under ACPI-defined S-states (system sleeping states) and D-states (device power states). Processor C-states (power states) and P-states (performance states) are considered in certain discussions but not elaborated upon since PCI Express architecture connectivity is not expected to include processors.

This document focuses on Intel-based Mobile implementations using Microsoft Windows* family operating systems, and does not address usage models under other operating environments.

1.2 References

PCI Express Base Specification Rev. 1.0

PCI Express Card Electromechanical Specification Rev. 1.0

PCI Local Bus Specification, Rev. 2.3

PCI Hot-Plug Specification, Rev. 1.1

PCI Standard Hot-Plug Controller, Rev. 1.0

PCI-to-PCI Bridge Architecture Specification, Rev. 1.1

PCI Power Management Interface Specification, Rev. 1.1

Advanced Configuration and Power Interface Specification, Rev. 2.0

1.3 Glossary

8bit/10bit The data encoding scheme used in the PCI Express Physical Layer
IBM Journal of Research and Development, Vol 27, #5, Sept 1983 "A DC-Balanced, Partitioned-Block 8B/10B Transmission Code" by Widmer and Franaszek

asserted The active logical state of a conceptual or actual signal.

Beacon 30 kHz–500 MHz signal used to exit L2.

Completion A Packet used to terminate, or to partially terminate, a Sequence is referred to as a *Completion*. A Completion always corresponds to a preceding Request, and in some cases includes data.

De-asserted Refers to the inactive logical state of a conceptual or actual signal.

Downstream Downstream refers either to the relative position of an interconnect/system element (Link/device) as something that is farther from the Root Complex, or to a direction of information flow, i.e., when information is flowing away from the Root Complex. For a Downstream Sequence, the Request flows Downstream and any Completions flow Upstream.

Hierarchy The Hierarchy defines the I/O interconnect topology supported by the PCI Express architecture.

Hierarchy Domain A PCI Express Hierarchy is segmented into multiple fragments by the Root Complex that sources more than one PCI Express interface. These sub-hierarchies are called Hierarchy Domains.

Lane A set of differential signal pairs, one pair for transmission and one pair for reception. A by-N Link is composed of N lanes.

Link A dual-simplex communications path between two components. The collection of two Ports and their interconnecting Lanes.

Packet A fundamental unit of information transfer consisting of a header that, in some cases, is followed by a Data Payload.

Port In a logical sense, an interface associated with a component, between that component and a PCI Express Link. In physical terms, a group of transmitters and receivers physically located on the same chip that define a Link.

Receiver The component receiving Packet information across a Link.

Request A Packet used to initiate a Sequence is referred to as a Reques*t.* A Request includes some operation code, and, in some cases, it includes address and length,
data, or other information.

Root Complex An entity that includes a Host Bridge and one or more Root Ports.

Root Port A PCI Express Port, on a Root Complex, that maps a portion of the PCI Express interconnect Hierarchy through an associated virtual PCI-PCI Bridge.

Symbol A 10 bit quantity produced as the result of 8bit/10bit encoding.

Symbol Time The period of time required to place a Symbol on a Lane (ten times the Unit Interval).

Transceiver The physical transmitter and receiver pair on a single chip.

Transmitter The component sending Packet information across a Link is the *Transmitte*r.

Unit Interval, UI Given a data stream of 1010… pattern, the Unit Interval is the value measured by averaging the time interval between voltage transitions, over a time interval long enough to make all intentional frequency modulation of the source clock negligible.

Upstream *Upstream* refers either to the relative position of an interconnect/system element (Link/device) as something that is closer to the Root Complex, or to a direction of information flow, i.e., when information is flowing towards the Root Complex. For an Upstream Sequence, the Request flows Upstream and any Completions flow Downstream.

2. Link states in relation to sleeping states and device power states

This section relates the expected behaviors of link L-states to system S-states and device D-states.

2.1 Sleeping states

ACPI S-states provide several sleeping options between the ACPI Working (S0) and Soft-off (S5) states. When the platform is in S0, it appears to be "on." When the platform is in one of the S1-S5 states, it appears to be "off." Only the operating system (OS) can move the platform from S0 to one of S1-S5, and will do so based on user input such as a user "Stand by" or "Hibernate" command or a predefined inactivity time-out. Only an OS-enabled wake event that is enabled prior to S1-S5 entry can move the platform out of S1-S5 to S0. Examples of possible wake events are power button activation, keystroke, mouse movement, real-time clock alarm, power management event (PME#) from wake-enabled devices, and platform-specific events such as thermal over-temperature or low battery alarm.

The various S-states have different entry/exit latencies characteristics; the lower-numbered S-states latency penalty is smaller than the higher-numbered S-states. The following relates the possible L-states for a PCI Express interconnect on a mobile platform.

- S0 (Working): This is the active system state in which the platform appears to be "on." During this state, CPU can enter low power states. PCI Express has provided support to comprehend CPU's low power states of C3/C4 through flow control credits. Consumption of flow credits by downstream devices becomes an indication of bus mastering events on the platform. The power management controller on the platform can register this indication. Processor power management control logic can use this indication to decide whether to enter the low power C3/C4 state. In addition, this indication may also cause the processor to exit from these low power states.

- S1/POS, S1 (Microsoft* OS Stand By): These are the lowest wake latency sleeping states. In these states, no system context is lost (processor or chipset), and hardware maintains all system contexts. This type of Stand By may be implemented either as Mobile S1/POS ("Power On Suspend"), or as Desktop S1:

  - Mobile platforms typically implement S1/POS with all clocks stopped, processor in Deep Sleep or Deeper Sleep clock control state, all devices powered on, and system context saved in system memory using self-refresh mode. Endpoint devices should be placed in the $D3_{hot}$ state, and PCI Express links should be in the L2 state. Section 2.3 below provides a guideline on how to configure the PCI Express link of a device in $D3_{hot}$ during S1/POS.

  - Desktop platforms typically implement S1 with all clocks running, and the processor in Sleep or Stop Grant state. As in the Mobile S1/POS state, all devices are left powered on, and system context is saved in system memory using self-refresh mode. Depending on system configuration, devices can be in D1-$D3_{hot}$ states, and PCI Express links can be in L0s or L1 state. Note that the L-state of the links closer to the root complex must be lower-numbered (providing faster exit latencies) than the L-state of the links closer to the endpoints in the PCI Express hierarchy.

- S2: S2 is not supported in current Intel Mobile platforms.

- S3 (Microsoft OS Stand By) and S4 (Microsoft OS Hibernate): S3 is a low wake latency sleeping state in which all device contexts are lost except system memory. S4 is the lowest power, longest wake latency state, in which system context is saved to hard disk. In S3 and S4, only wake-enabled logic remains powered on. Depending on the device configuration with respect to Aux power support, the L-state of the PCI Express link may be either L2 or L3. L2/L3 Ready state may be used to stage the link in preparation for power and clock removal. Either in-band beacon or sideband WAKE# can be used to support wake-enable devices. However, sideband WAKE# mechanism is recommended on mobile platforms during these sleeping states.

- S5 (Microsoft OS Soft-off): In this state, the OS does not save any context and requires a complete OS boot when the system wakes. Since ACPI 2.0 allows wake events during this state, L2 and L3 are feasible states for any PCI Express link. The L2/L3 Ready state may be used as a transitional state to prepare the link for power and clock removal.

Table 1 below summarizes the relation between S-states and L-states for PCI Express link.

Table 1 Relationship between S-states and L-states

S-state	**Permissible L-state**
S1	L0s or L1
S1/POS	L2 (see section 2.3)
S3	L2 or L3, where L2/L3 Ready is a transitional state
S4	L2 or L3, where L2/L3 Ready is a transitional state
S5	L2 or L3, where L2/L3 Ready is a transitional state

2.2 Device power states

ACPI defines D-states D0-D3. The following relates the possible L-states for a PCI Express link to the D-states of the relevant components.

- D0 (Fully-on): The device is completely active and responsive during this D-state. The link may be in L0 or L0s. Assuming L0s usage is enabled for the link, active state power management in each port’s transmitters will transition the appropriate link direction to L0s after an idle period in the range of 25%-100% of the opposing port’s reported L0s exit latency. Minimizing L0s exit latency therefore allows for frequent entry into L0s while facilitating performance needs via a fast exit. L1 state may be achieved either by hardware-based active state power management or by requesting the link to enter L1 after the OS placing the downstream device in D1-D3 state.

- D1 and D2: There is no universal definition for these 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 aux power is supported on the device with wake-capable logic, or L3 if no power is delivered to the device. Although in-band beacon is the default mechanism, sideband WAKE# mechanism is recommended to support wake-enabled logic on mobile platforms during L2 state. D3$_{hot}$ maps to L1 as part of S1/POS to support clock removal on mobile platforms, or L2/L3 Ready as a transitional state to stage the preparation of power and clock removal.

Table 2 below 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 or L2/L3 ready
D3$_{cold}$	D0-D3$_{cold}$	L2 or L3

2.3 Support of S1/POS on Intel mobile platforms

S1/POS on Intel mobile platforms supports the possibility of only stopping the system clocks. In this case, the device must stay in D3 to comprehend clock removal during S1/POS.

- The device will be configured for D3 and must complete the PME_Turn_Off and PME_TO_Ack handshake with upstream device.

- The device must initiate a PM_Enter_L23 DLLP to prepare for clock removal immediately. It must shut down all appropriate circuitry (e.g. PLL) after receiving a PM_Receive_Ack DLLP.

- Exit signaling from the power management controller in the root complex or sideband WAKE# signaling from the endpoint would initiate the exit sequence from S1/POS.

This process will repeat upwards through the PCI Express hierarchy domain to the root complex as the platform prepares itself for clock removal.

3. Power saving opportunities during link power states

This section describes power saving opportunities during various L-states, mainly at the device level. Platform level opportunities are identified where appropriate.

3.1 Active state power management

Active state power management is the hardware capability to power-manage the PCI Express link. Only L0s and L1 are used during active state power management.

- L0s: This link state is a very low exit latency link state intended to reduce power wastage during short intervals of logical idle between link activities. L0s support is required, and, assuming L0s usage is enabled for the link, active state power management in each port's transmitter must transition the appropriate link direction to L0s after an idle period in the range of 25%-100% of the opposing port's reported L0s exit latency. Acknowledgment of this entry is implicit. 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. Minimizing L0s exit latency optimizes performance/power considerations. For example, innovation on the clock recovery mechanism would help to reduce the number of fast-training sequences required and hence minimize L0s exit latency. It is strongly recommended that Mobile devices optimize their implementation to minimize the number of fast-training sequences for synchronization during L0s exit.
- L1: This link state is a low exit latency link state that is intended to reduce power when the device becomes aware of a lack of outstanding requests or pending transactions. Although the PCI Express base specification rev1.0 defines L1 support to be optional, it is required for mobile platforms in order to optimize battery life and thermal design power constraints. If L1 entry is rejected, the link must transition to L0s. L1 entry policy is not mandated in the PCI Express base specification; however, to promote innovation, this document discusses a few approaches to optimizing L1 usage. The power saving opportunities during this state include, but are not limited to, shutdown of most of the transceiver circuitry, clock gating of most PCI Express architecture logic, and shutdown of the PLL.

In the presence of a multi-lane PCI Express link, it is recommended that the signaling and detection of L0s or L1 exit can be communicated through lane #0 of a configured link. The other lanes can then begin synchronization and training within one or two symbol time of the initial event on lane #0.

It is also worthy to note that L1 exit signaling does not have to be derived from 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.

3.2 Device 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. The hardware is responsible for this autonomous mapping of D-states to L-states. Unlike active state power management, all L-states are comprehended for D-state mapping. Since L0s and L1 are already discussed above in section 3.1, the following will focus on the L2/L3 Ready, L2, and L3 states:

- 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 endpoint device must initiate a PME message packet after signaling either in-band beacon or sideband WAKE# to exit this state. The power saving opportunities for this state include, but are not limited to, clock gating of all PCI Express architecture logic, and shutdown of the PLL, and shutdown of all transceiver circuitry except in-band beacon or sideband WAKE# detection 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 L0s is the link state with lower power but short exit latency.

3.3 L1 entry policy

The policy governing L1 entry is interesting and presents innovation opportunities for product differentiation. There are at least two factors to consider when deriving the algorithm for L1 utilization:

- Achieving a meaningful number of entries into L1, where each entry would result in a length of time spent in this link state that is equal to or greater than the worst case L0s exit latency between the two connected devices.
- Minimizing power per performance (such as watt/byte transferred) through a combination of L0, L0s and L1.

Here are two examples of algorithmic-based policy. They are by no means exhaustive.

- An operational approach defines the interval between new packet arrivals by means of a rule, such as the Generic Cell Rate Algorithm (GCRA) in CCITT recommendation I.371 using Virtual Scheduling (VS).
- A statistical approach determines the running average of time spent in L1, such that L1 entry happens when there has been no outstanding request or pending transaction for the duration that is a fraction of this running average. The fraction is selected with consideration of either or both factors described above, and may be adaptive over time.

4. Minimize power, maximize performance

Techniques to improve packet scheduling and ensure effective utilization of bandwidth are also essential to a power-optimized architecture. Two scenarios are described below as illustration.

- Scenario (A) in Figure 1 shows an "ask for one, get one" approach of utilizing a PCI Express link. One 256-byte completion is submitted after its request is issued. Flow control update and acknowledge packets alternate, assuming no non-recoverable link errors.
- Scenario (B) in Figure 1 illustrates the streaming of multiple requests and subsequently multiple completions before flow control 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 a flow control (FC) update packet and an acknowledgement packet after a few transaction layer 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 1 Scenarios to illustrate scheduling and bandwidth utilization impact

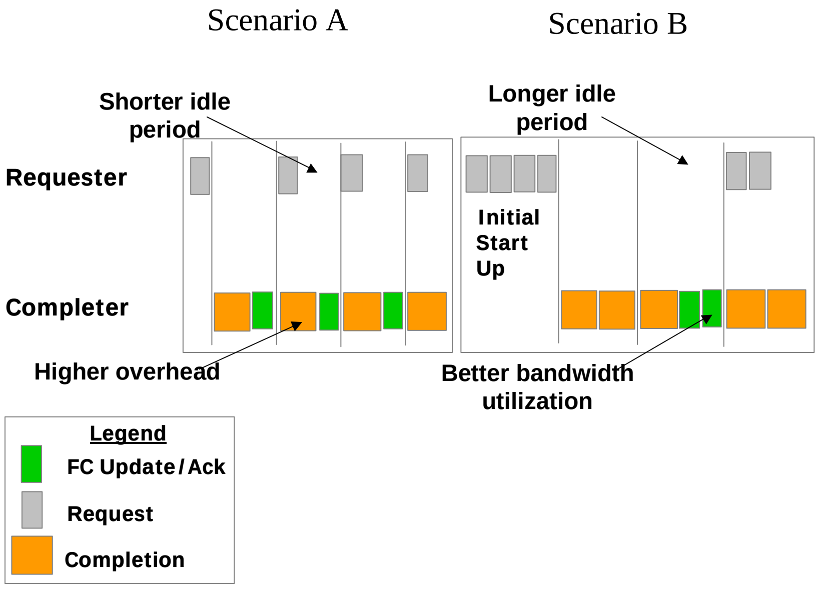


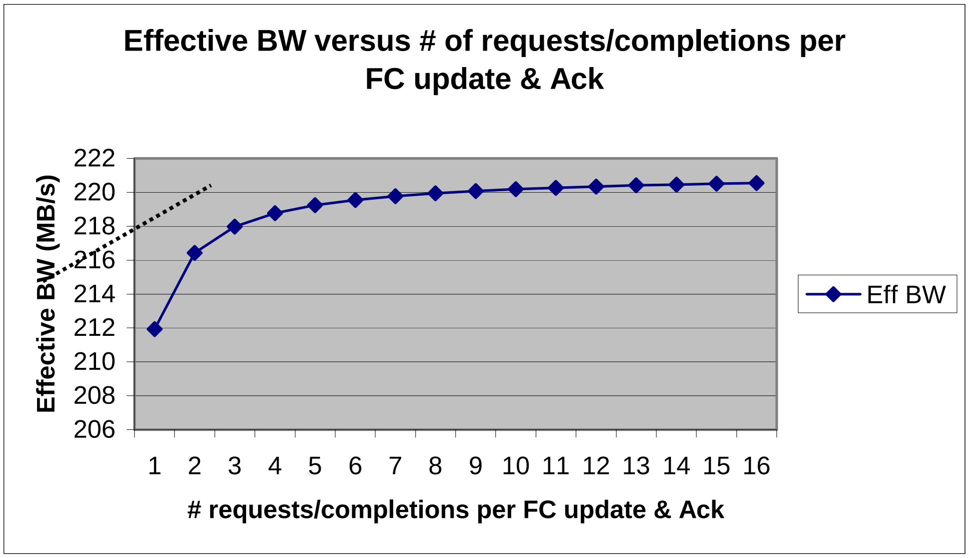
Figure 2 below shows a plot of the effective bandwidth of a x1 PCI Express link as a function of number of requests/completions per flow control update and acknowledge packets. The sequential streaming of N requests and completions has to be sustained in order to achieve this effective bandwidth. The graph also accounts for 8b/10b encoding as well as a 32-byte overhead and 256-byte payload per packet.

Streaming just one request and completion per flow control update and acknowledge limits the bandwidth to less than 212 MB/s on a x1 PCI Express link. Increasing the number of requests and completions would eventually reach a bandwidth of greater than 220 MB/s. However, increasing the number of consecutive requests/completions per FC update and Ack is not "free", since as the number of consecutive requests/completions without a corresponding FC update and Ack increases, additional hardware must be provided to account for the possibility of errors requiring the requests/completions to be re-transmitted. Therefore, the designer must make a tradeoff between hardware complexity and performance optimization. Based on the return-on-investment analysis of Figure 2, the optimal number of requests and completions per flow control update and acknowledge is 3.

Streaming of multiple packets to improve bandwidth utilization is meaningful even for applications with low bandwidth demands if they are implemented to take advantage of the clustering. Not only it improves bandwidth utilization, it provides both the requester and the completer more opportunities of extended idle periods to perform link power management.

Of course this is just one example of many possibilities in an interesting but complicated scheduling and bandwidth utilization puzzle.

Figure 2 Effective bandwidth as a function of number of requests/completions per FC update/Ack



Two observations are important at this point:

1. This model assumes that the data initiator and consumer do not present a bottleneck to performance. In fact, if the data producer and consumer do present a bottleneck (e.g. via access latencies), a different optimal streaming model may pertain, in which the choice of the number of requests and completions may be selected to minimize the bottleneck.

2. This model also assumes there is sufficient buffering capability (or flow control credits) to allow for a ping-pong style of servicing completions and retiring requests. Both the requester and completer must throttle the packets they send such that they operate at an optimal performance point and avoid the inefficiency of under-utilizing (e.g. sending one at a time) or saturating the available flow control credits.

5. Summary

PCI Express is a new technology that requires innovative thinking in order to maximize performance and minimize power. Poor implementations that waste bandwidth, power, or both will find themselves burdening performance as well as thermal constraints, battery life, and power delivery.

- Take advantage of the power management capabilities provided by PCI Express to deliver low power solutions. PCI Express ensures ease of transition by providing ACPI based power management.
- Aggressively transition to L0s-L2 to power manage at both device and platform level. There are many innovation opportunities to provide best-of-class performance/power, such as low-power state entry policies and synchronization circuitry for exit from low-power states.

Optimization in scheduling and pipelining will not only improve bandwidth utilization, but also provide extended idle periods for the link to enter low-power states.

Want More Info on PCI Express?

Intel Developer Network for PCI Express: http://developer.intel.com
PCI-SIG Web Site: http://www.pcisig.com