Device Guidelines for PCI Express* Technology Extensions

Jaya Jeyaseelan,
Client Platform Architect, Intel Corporation

David Fair,
Enterprise IHV Enabling Manager, Intel Corporation

PCIS002
Agenda

• Impact of devices on platform power
• Introduction to the power management extensions – LTR and OBFF
• Implementation guidelines for typical devices on mobile and desktop platforms
• Implementation guidelines for typical devices on server and workstation platforms
• Summary
• Next steps
Agenda

- Impact of devices on platform power
- Introduction to the power management extensions – LTR and OBFF
- Implementation guidelines for typical devices on mobile and desktop platforms
- Implementation guidelines for typical devices on server and workstation platforms
- Summary
- Next steps
Power Impact of Device Activity

- Frequent and random device activity bringing platform components out of low power states can have significant power impact
  - e.g. 100 bus master transactions per second = ~200mW

Opportunity to reduce platform power by aligning device activity

PCI Express* WLAN device activity on Intel® Core™ 2 Duo platform; Source: Intel Corporation
Platform Power Savings Opportunity

- Usage Analysis: Typical Mobile Platform in S0 state is ~90% idle
- Even when idle, platform components are kept in high power state to meet the service latency requirements of devices
  - Increasing latency results in buffer overflows and failures for some devices

Average power scenario on running MM07 (Office* Productivity suite)
On an Intel® Core™i5 platform running Windows 7*

Idle workload power (~8W – 10W)

Source: Intel Corporation

Opportunity to reduce platform power for an idle workload
Agenda

• Impact of devices on platform power
• Introduction to the PCI Express* power management extensions – LTR and OBFF
• Implementation guidelines for typical devices on mobile and desktop platforms
• Implementation guidelines for typical devices on server and workstation platforms
• Summary
• Next steps
Power vs. Response Latency (Mobile)

Variable service latency indication from devices required for aggressive, yet robust power management.
Latency Tolerance Reporting (LTR)

**LTR Mechanism**
- LTR message (TLP) sent by device dynamically as a function of workload
  - Smaller values during active workloads
  - Larger value when idle

**LTR Benefits**
- Dynamic power-performance tradeoffs
- Robust power management
  - Platform enters lower power/longer latency states only when devices can tolerate it
Impact of LTR on Platform Idle Power

Power impact is higher since future platforms will have lower idle floor.

When a device doesn’t support LTR, platform latency will be set to ~20usec.

% increase in Platform Idle Power

- When a device supports LTR:
  - 20us: 160%
  - 30us: 140%
  - 60us: 120%
  - 300us: 100%
  - 500us: 80%
  - 1000us: 60%

- When a device doesn’t support LTR:
  - 20us: 0%
  - 30us: 20%
  - 60us: 40%
  - 300us: 60%
  - 500us: 80%
  - 1000us: 100%

LTR Values

Lower Idle Power and Increased Battery Life

Data from power model for Client Notebook
Source: Intel Corporation

+ This data is for illustration purposes only & actual data will be available as platforms become available
+ All products, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice

Crucial that all devices support LTR for maximum power savings
LTR - Device Implementation

Select the mechanism appropriate for your device

- **Hardware driven LTR**
  - When device latency requirements change frequently
  - For devices like Network devices which asynchronously receive data from network

- **Software guided LTR**
  - When device latency requirements change infrequently
  - For slave devices which execute work items scheduled by software
  - May define an MMIO register in device, a write to which would trigger an LTR message
A new LTR value is in effect no later than the value sent in the previous LTR.

Latency guidance based on buffer size and utilization.

LTR = 60us
LTR_Active_Idle = 400usec

Latency 400usec
Optimized Buffer Flush/Fill (OBFF)

**OBFF Mechanism**
- Indicates optimal windows for bus mastering and interrupt activity
  - Intel chipsets will drive WAKE# at the root complex for OBFF

**Optimal Windows**
- **Active Window** – Platform fully active. Optimal for bus mastering and interrupts
- **OBFF Window** – Platform memory path available for memory reads and writes
- **Idle Window** – Platform is in low power state
Impact of OBFF on Platform Power

- Meaningful increase in deep package C-state residency for server platforms
- For Client platforms, baseline C-state residency is high. OBFF gives additional power savings of $\geq 200$mW on DMA scenarios (network/disk file transfers)
  - Still significant for battery operated mobile devices
  - Savings increase when idle floor is lower

Data from prototype in Intel® Xeon® Processor 5500 based server platform running Linux* 2.6.18
Source: Intel Corporation

Increase in platform low power state residency
Active State Power Management (ASPM)

• Support ASPM on all devices
  - Power savings per link add up across the platform
  - L1 entry policy is device driven. Ensure you don’t re-enter L1 state when waiting for platform to come out of low power state

• Support Dynamic Clock Control (CLKREQ#) on Mini-Card and ExpressCard form factors
  - “CLKREQ# asserted” to achieve fast exits between data bursts
  - “CLKREQ# de-asserted” when device is pervasively idle
    ☀ Host reference clock savings (~60mW per clock) & device PLL savings (~100mW per device)

• Recommended link exit Latencies
  - L0s: 256ns
  - L1: <5us
  - L1 w/ PLLs off: <=20us (Not recommended for devices on servers)
    ☀ Higher link exit timings can cause host CPU thread stall on MMIO accesses
  
*Support “CLKREQ# de-asserted” when pervasively idle*
Agenda

• Impact of devices on platform power
• Introduction to the power management extensions – LTR and OBFF
• Implementation guidelines for typical devices on mobile and desktop platforms
• Implementation guidelines for typical devices on server and workstation platforms
• Summary
• Next steps
# LTR Recommendation for Client Devices

<table>
<thead>
<tr>
<th>Devices</th>
<th>LTR_Active</th>
<th>LTR_Act_Idle</th>
<th>LTR_IDLE</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>WLAN</td>
<td>60usec</td>
<td>300usec (minimum)</td>
<td>LTR_No_Req (unassociated)</td>
<td>Device Initiated</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LTR_MaxPlatLat (associated and radio off)</td>
<td></td>
</tr>
<tr>
<td>Ethernet LAN (1Gb or lower)</td>
<td>60usec</td>
<td>300usec (minimum)</td>
<td>LTR_MaxPlatLat (LPI mode)</td>
<td>LPI – Low Power Idle mode in IEEE 802.3az standard</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Graphics</td>
<td>60usec</td>
<td>Optional</td>
<td>LTR_MaxPlatLat</td>
<td>Can be SW guided</td>
</tr>
<tr>
<td>Client Storage (e.g. memory card reader)</td>
<td>60usec</td>
<td>Optional</td>
<td>LTR_MaxPlatLat</td>
<td>Can be SW guided</td>
</tr>
</tbody>
</table>

+ BIOS programs LTR Extended Capability Structure field with LTR_MaxPlatLat (1msec)

+ These numbers are preliminary. Monitor the following link for updates: [http://developer.intel.com/technology/pciexpress/devnet/index.htm](http://developer.intel.com/technology/pciexpress/devnet/index.htm)

**LTR value below 60usec will result in increased platform power**

- Request lower values only when necessary – for short durations
Agenda

• Impact of devices on platform power
• Introduction to the power management extensions – LTR and OBFF
• Implementation guidelines for typical devices on mobile and desktop platforms
• Implementation guidelines for typical devices on server and workstation platforms

• Summary
• Next steps
LTR Recommendation for Server Devices

<table>
<thead>
<tr>
<th>Devices</th>
<th>LTR_Active</th>
<th>LTR_Act_Idle</th>
<th>LTR_Idle</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethernet LAN (1Gb or lower)</td>
<td>60usec</td>
<td>300usec (minimum)</td>
<td>LTR_MaxPlatLat (LPI mode)</td>
<td>Device Initiated</td>
</tr>
<tr>
<td>≥10 Gb (Ethernet, FibreChannel, or InfiniBand)</td>
<td>50usec</td>
<td>Optional in 1st gen platforms</td>
<td>LTR_MaxPlatLat (LPI mode)</td>
<td>Device Initiated</td>
</tr>
<tr>
<td>SATA Controller</td>
<td>50usec</td>
<td>Optional</td>
<td>LTR_MaxPlatLat</td>
<td>Can be SW guided</td>
</tr>
<tr>
<td>SAS Controller</td>
<td>50usec</td>
<td>Optional</td>
<td>LTR_MaxPlatLat</td>
<td>Can be SW guided</td>
</tr>
</tbody>
</table>

Any LTR value below 50usec will result in increased platform power

- Request lower values only when performance is needed – for short durations
- These numbers are preliminary. Monitor the following link for updates: [http://developer.intel.com/technology/pciexpress/devnet/index.htm](http://developer.intel.com/technology/pciexpress/devnet/index.htm)
Use of Device Buffer Watermarks
Example: Buffer Flushes for Receive Path on NIC

- On-chip buffers *must* be flushed when the “high-water mark” (HWM) is hit to prevent over-runs
  - Defined down from the top of the buffer by at least LTR + the PCIe L1 exit latency
- Depending on application, buffers likely will be flushed based on various time-out situations
  - For example, preventing “stale” data
- “Low-water marks” (LWM) are recommended to coalesce data for more efficient bus usage
  - LWM effectively sets a minimum size for a transmission (to host) block
Implementation Guidelines for OBFF
Example: Buffer Flush

<table>
<thead>
<tr>
<th>OBFF State</th>
<th>Behavior</th>
<th>Device timers</th>
<th>LWM</th>
<th>HWM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTIVE</td>
<td>• Normal operation – send data and interrupts</td>
<td>• On transition to ACTIVE, flush all buffered data and interrupts</td>
<td>• Send data once LWM is reached</td>
<td>• NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Complete the flush before starting device timers and local buffering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OBFF (BUFFERED)</td>
<td>• Align deferrable interrupts</td>
<td>• On timer expiry, flush buffered data and interrupts</td>
<td>• Send data once LWM is reached</td>
<td>• NA</td>
</tr>
<tr>
<td></td>
<td>• Perform critical transactions as necessary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IDLE</td>
<td>• Coalesce deferrable data and interrupts</td>
<td>• Device exposes timer toggle capability bit to enable/disable timers in OBFF IDLE (Details in whitepaper*)</td>
<td>• NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Perform critical transactions as necessary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Device exposes timer toggle capability bit to enable/disable timers in OBFF IDLE (Details in whitepaper*)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Details to follow in future whitepaper on Intel® Developer Network for PCI Express*
  Architecture @ http://developer.intel.com/technology/pciexpress/devnet/index.htm
Agenda

• Impact of devices on platform power
• Introduction to the power management extensions – LTR and OBFF
• Implementation guidelines for typical devices on mobile and desktop platforms
• Implementation guidelines for typical devices on server and workstation platforms

• Summary
• Next steps
Summary

- Impact of devices (even low power ones) on platform power is significant
- The PCI Express* Power Management Extensions provide a framework for reducing the impact of devices on platform power
- Framework creates a cooperative power management model and all devices in the ecosystem must support the model for maximum impact
- We have seen a lot of interest in these power management extensions and highly encourage you to implement them
  - Opportunity for devices to differentiate and claim platform power reductions
Agenda

• Impact of devices on platform power
• Introduction to the power management extensions – LTR and OBFF
• Implementation guidelines for typical devices on mobile and desktop platforms
• Implementation guidelines for typical devices on server and workstation platforms
• Summary
• Next steps
Next Steps

• Download the PCI Express* 3.0 spec from the PCI-SIG Website for reference on LTR and OBFF

• IHVs:
  – Start architecting devices with a view towards using the Interconnect Bus extensions
  – Work with your OEMs/ODMs to understand requirement and timeline
  – Contact Intel representative to get the enabling tools and collateral
Additional Sources of Information on This Topic

• Other sessions:
  – PCIS001: PCI Express* 3.0 Technology: Logical PHY Considerations for Intel® Platforms
  – PCIS003: PCI Express* 3.0 Technology: Enabling Tools and Updates to the PIPE 3.0 Specification
  – EBLS001: Interconnect Bus Extensions for Energy-Efficient Platforms
  – PCIQ001: PCI Express* 3.0 Q&A

• More Web-based info:
  – http://www.pci-sig.org
  – Intel® Developer Network for PCI Express* Architecture
  – ‘Energy-efficient platform devices’ whitepaper
    ❇️ http://www.intel.com/technology/mobility/notebooks.htm
Legal Disclaimer

- 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 may make changes to specifications and product descriptions at any time, without notice.
- All products, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.
- Intel, processors, chipsets, and desktop boards may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.
- Code names featured are used internally within Intel to identify products that are in development and not yet publicly announced for release. Customers, licensees and other third parties are not authorized by Intel to use code names in advertising, promotion or marketing of any product or services and any such use of Intel's internal code names is at the sole risk of the user.
- Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.
- Intel, Core, Intel Sponsors of Tomorrow, and Intel Sponsors of Tomorrow logo and the Intel logo are trademarks of Intel Corporation in the United States and other countries.
- *Other names and brands may be claimed as the property of others.
- Copyright ©2010 Intel Corporation.
Risk Factors

The above statements and any others in this document that refer to plans and expectations for the second quarter, the year and the future are forward-looking statements that involve a number of risks and uncertainties. Many factors could affect Intel’s actual results, and variances from Intel’s current expectations regarding such factors could cause actual results to differ materially from those expressed in these forward-looking statements. Intel presently considers the following to be the important factors that could cause actual results to differ materially from the corporation’s expectations. Demand could be different from Intel’s expectations due to factors including changes in business and economic conditions; customer acceptance of Intel’s and competitors’ products; changes in customer order patterns including order cancellations; and changes in the level of inventory at customers. Intel operates in intensely competitive industries that are characterized by a high percentage of costs that are fixed or difficult to reduce in the short term and product demand that is highly variable and difficult to forecast. Additionally, Intel is in the process of transitioning to its next generation of products on 32nm process technology, and there could be execution issues associated with these changes, including product defects and errata along with lower than anticipated manufacturing yields. Revenue and the gross margin percentage are affected by the timing of new Intel product introductions and the demand for and market acceptance of Intel’s products; actions taken by Intel’s competitors, including product offerings and introductions, marketing programs and pricing pressures and Intel’s response to such actions; defects or disruptions in the supply of materials or resources; and Intel’s ability to respond quickly to technological developments and to incorporate new features into its products. The gross margin percentage could vary significantly from expectations based on changes in revenue levels; product mix and pricing; start-up costs, including costs associated with the new 32nm process technology; variations in inventory valuation, including variations related to the timing of qualifying products for sale; excess or obsolete inventory; manufacturing yields; changes in unit costs; impairments of long-lived assets, including manufacturing, assembly/test and intangible assets; the timing and execution of the manufacturing ramp and associated costs; and capacity utilization. Expenses, particularly certain marketing and compensation expenses, as well as restructuring and asset impairment charges, vary depending on the level of demand for Intel’s products and the level of revenue and profits. The majority of our non-marketable equity investment portfolio balance is concentrated in the flash memory market segment, and declines in this market segment or changes in management’s plans with respect to our investment in this market segment could result in significant impairment charges, impacting restructuring charges as well as gains/losses on equity investments and interest and other. Intel’s results could be impacted by adverse economic, social, political and physical/infrastructure conditions in countries where Intel, its customers or its suppliers operate, including military conflict and other security risks, natural disasters, infrastructure disruptions, health concerns and fluctuations in currency exchange rates. Intel’s results could be affected by the timing of closing of acquisitions and divestitures. Intel’s results could be affected by adverse effects associated with product defects and errata (deviations from published specifications), and by litigation or regulatory matters involving intellectual property, stockholder, consumer, antitrust and other issues, such as the litigation and regulatory matters described in Intel's SEC reports. An unfavorable ruling could include monetary damages or an injunction prohibiting us from manufacturing or selling one or more products, precluding particular business practices, impacting our ability to design our products, or requiring other remedies such as compulsory licensing of intellectual property. A detailed discussion of these and other factors that could affect Intel’s results is included in Intel’s SEC filings, including the report on Form 10-Q for the quarter ended March 27, 2010.

Rev. 5/7/10
Backup Slides
LTR Semantics for Reads and Writes

- **Latency values are applicable only to leadoff cycles**
  - Defined as the first memory transaction of potentially multiple memory transactions that will occur in quick (<5us) succession

- **For Read requests**
  - Latency is measured as delay from read request TLP to the receipt of the completion

- **For Write requests**
  - If the device issues one or more write requests such that it cannot issue another write request due to Flow Control backpressure, the latency is measured from the transmission of TLP that exhausts FC credit to the receipt of the DLLP returning more credits
Example: WLAN Device

Use of device PM states to give Latency guidance

Latency information with Wi-Fi Power Save

**Use of device PM states to give Latency guidance**
Device Buffering and OBFF

- Implement sufficient device buffering to maximize energy efficiency
  - Will reduce frequent host memory accesses and allow for OBFF activity alignment
  - Allow for a minimum of ~300usec of inactivity between bursts. Recommend ~1msec for greater energy efficiency when performance permits

- Classify device-initiated transactions: critical vs. deferrable
  - Perform critical transactions as necessary
  - Align deferrable transactions to platform activity
    - No deferrable accesses in Idle Window
    - Memory accesses in OBFF Window
    - Interrupts and Memory accesses in Active Window

Source: Intel Corporation