Disclaimer

**Presentation Disclaimer**: All opinions, judgments, recommendations, etc. that are presented herein are the opinions of the presenter of the material and do not necessarily reflect the opinions of the PCI-SIG®.
Agenda

- PCI Express® Layered Model
- Physical Layer Examples
  - Power Management Examples
  - Link Training Examples
  - Equalization Examples
  - M-PCIe™
- Data Link Layer Examples
  - Flow Control Examples
- Transaction Layer Examples
  - Completion Timeout
- PCIe® Storage Examples
  - Different PCIe Storage Technologies
    - NVMe
    - PQI/SCSIe
    - AHCI/SATA Express
PCI EXPRESS LAYERED MODEL
PCI Express Layered Model

Application

Software
- Ethernet, NVMe, SOP, AHCI, SATA…

Transaction

Device Configuration and Control
Data Transfer to/from Memory

Data Link

Management of Packets:
Flow Control and ACK/NAK Protocol

Physical

Link Training

Logical Sub Block

Electrical Sub Block

Electrical (Analog) Signalling
What Can I See at Each Layer?

- **Application**
  - It is all inside the PCI Express Payload, could include: NVMe commands, Ethernet Frames etc.

- **Transaction**
  - TLPs: Config Rd/Wr, Mem Rd/Wr, IO Rd/Wr, Messages

- **Data Link**
  - DLLPs: InitFC, UpdateFC, ACK, NAK PM_Enter_L1 etc.....

- **Physical**
  - Logical Sub Block
  - Electrical Sub Block
  - TS1, TS2 all Ordered Sets
  - Eye Diagrams, analog waveforms
How do Two Devices Talk?

Root Complex

Application

Transaction

Data Link

Physical

Logical Sub Block

Electrical Sub Block

End Point

Application

Transaction

Data Link

Physical

Logical Sub Block

Electrical Sub Block

RX

TX

RX

TX
What about M-PCle?

- “A PCI Express Link consists of a PCIe PHY as defined in chapter 4 whereas the M-PCle Link consists of M-PHY, a physical layer specification developed by MIPI Alliance.”
## M-PCle Trace Capture

<table>
<thead>
<tr>
<th>Packet</th>
<th>Status</th>
<th># Packets</th>
<th>Time Delta</th>
<th>Time Stamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>SC</td>
<td>2</td>
<td>0.000 ms</td>
<td>0000 010 000 000 s</td>
</tr>
<tr>
<td>1</td>
<td>SC</td>
<td>2</td>
<td>0.000 ms</td>
<td>0000 020 000 000 s</td>
</tr>
<tr>
<td>2</td>
<td>SC</td>
<td>2</td>
<td>0.000 ms</td>
<td>0000 030 000 000 s</td>
</tr>
<tr>
<td>3</td>
<td>SC</td>
<td>2</td>
<td>0.000 ms</td>
<td>0000 040 000 000 s</td>
</tr>
<tr>
<td>4</td>
<td>SC</td>
<td>2</td>
<td>0.000 ms</td>
<td>0000 050 000 000 s</td>
</tr>
<tr>
<td>5</td>
<td>SC</td>
<td>2</td>
<td>0.000 ms</td>
<td>0000 060 000 000 s</td>
</tr>
<tr>
<td>6</td>
<td>SC</td>
<td>2</td>
<td>0.000 ms</td>
<td>0000 070 000 000 s</td>
</tr>
<tr>
<td>7</td>
<td>SC</td>
<td>2</td>
<td>0.000 ms</td>
<td>0000 080 000 000 s</td>
</tr>
<tr>
<td>8</td>
<td>SC</td>
<td>2</td>
<td>0.000 ms</td>
<td>0000 090 000 000 s</td>
</tr>
<tr>
<td>9</td>
<td>SC</td>
<td>2</td>
<td>0.000 ms</td>
<td>0000 100 000 000 s</td>
</tr>
<tr>
<td>10</td>
<td>SC</td>
<td>2</td>
<td>0.000 ms</td>
<td>0000 110 000 000 s</td>
</tr>
</tbody>
</table>

- **RRAP Tr 0**: 24-3196
- **RRAP Tr 1**: 25-4117
- **RRAP Tr 2**: 713-817
- **RRAP Tr 3**: 741-813
- **DL Packets**: 610-306
- **DL Packets**: 624-216
- **DL Packets**: 626-243
- **DL Packets**: 634-151

---

**PCI-SIG Developers Conference**

Copyright © 2015, PCI-SIG, All Rights Reserved
M-PCle Layers

- **RRAP Protocol**

- **Link Training**

- **Flow Control Initialization and Transaction Layer**
M-PCIe Form Factor

- M-PCIe may have different form factors
  - M.2
  - Embedded
M.2 Card Technology

- M.2 Card Format defines:
  - Socket 2: SATA or x2 PCI Express fits modules with the “B” key for SSD, cache
  - Socket 3: x4 PCIe up to 4 GB/s fits modules with the “M” key for ultimate performance SSD or cache

- PCIe SSD with both “B” and “M” key fit into both Socket 2 and 3 hosts using only two PCIe lanes in Socket 3 hosts

- Applications
  - Notebooks, Ultrabooks & Desktops
  - Tablets
  - Servers
  - Portable Gaming Devices
  - Devices that require SSDs
  - Portable Mobile Devices
Which Layer?

- Transaction layer problems mostly due to legacy carry over (at present)
- Data link layer problems can be the most difficult to detect in development
- Physical layer issues for protocol are mostly related to the LTSSM
“PCle 1.0, 2.0, 3.0???”

- What is the difference?
  - These are different revisions of the specification
  - They do not necessarily mean a device can support higher speeds
  - The changes go beyond the speed capability and include ECNs and other updates

- We tend to use Gen 1, Gen 2, Gen 3 synonymously with 2.5GT/s, 5GT/s and 8GT/s
  - A PCle 1.0a or 1.1 device supports 2.5GT/s
  - A PCle 2.0 device must support 2.5GT/s and can support 5GT/s
  - A PCle 3.0 device must support 2.5GT/s and can support up to 5GT/s and 8GT/s

- How does this apply to the layered model?
“PCle 1.0, 2.0, 3.0???”

- The most significant change in specification revisions was to add higher speed capabilities.
- However, there are many other ECNs that were approved after the release of a specification that are rolled into newer releases.
- For example, there was a 2.1 revision of the specification that included ECNs proposed after the release of the 2.0 specification.
- The PCI Express 3.1 Base Specification was released.
“PCle 1.0, 2.0, 3.0???”

- There are many changes in Configuration Space that can result in differing test results.
- The way PCI-SIG tests devices is not the same for a PCle 1.1 device and PCIe 3.0 device that only supports 2.5GT/s.
How does this fit into the layered model?

Compliance testing is split out into sections:

- **Configuration Space**
  - Using PCIECV tool from PCI-SIG

- **System BIOS Testing**
  - Using PCIEPT tool from PCI-SIG

- **Link and Transaction Layer**
  - Transaction Layer testing
  - Data Link Layer testing
  - Testing Link Training

- **Electrical**
  - Link Equalization/De-emphasis testing
  - Transmitter Signal Quality
  - Receiver Jitter Tolerance
PHYSICAL LAYER EXAMPLES

- Application
- Transaction
- Data Link

Physical
- Logical Sub Block
- Electrical Sub Block
My Device Does Not Show in Device Manager

- Is the Link Active?
- Firstly do a snapshot trigger
  - If the link is active, you will see Update_FC packets and SKP ordered sets
  - If the link is not active, you may see TS1 or TS2 ordered sets
  - If the link is unstable, you may see Update_FC and SKP ordered sets, with occasional TS1/TS2 sequences as the link goes to Recovery
- Bottom Line: If the link is not active, there is no way the device manager will see the device
Is the Device Seen by the System?
First Check the Electrical Characteristics

- Is the Signal Integrity of the Link good on both sides
- Signal Integrity problems can show in a whole manner of ways, from instable links to devices just not showing up
TS1 and TS2 Ordered Sets would indicate that the link is not in L0
The state machine rules for configuring and operating a PCIe Express Link are defined in the following sections.

**4.2.6.1. Detect**

The Detect state machine is shown in Figure 4-12.

- Transmitter is in an Electrical Idle state.
  - Note: The DC common mode voltage is not required to be within specification.
  - 2.5 GT/s data rate is selected as the frequency of operation. If the frequency of operation was 2.5 GT/s data rate on entry to this state, the LTSSM must stay in this state for at least 1 ms, during which the frequency of operation must be changed to the 2.5 GT/s data rate.
  - Note: This does not affect the advertised data rate in the TS1 and TS2 Ordered Sets.
- LinkUp = 0b (status is cleared).
- The directed_speed_change variable is set to 0b. The upconfigure_capable variable is reset to 0b. The idle_to_sck_transion variable is reset to 0b. The select_deemphasis variable must be set to either 0b or 1b based on platform specific needs for the Downstream component and identical to the Selectable De-emphasis bit in the Link Control 2 register for the Upstream component.
  - Note: since these variables are defined with the 2.0 specification, pre-2.0 devices would not implement these variables and will always take the path as if the directed_speed_change and upconfigure_capable variables are constantly reset to 0b and the idle_to_sck_transion variable is constantly set to 1b.
- The next state is Detect Active after a 12 ms timeout or if Electrical Idle is broken on any Lane.

**4.2.6.1.2. Detect.Active**

- The Transmitter performs a Receiver Detection sequence on all un-configured Lanes that can form one or more Links (see Section 4.3.1.8 for more information).
- Next state is Polling if a Receiver is detected on all un-configured Lanes.
- Next state is Detect Quiet if a Receiver is not detected on any Lanes.

---

**Figure 4-11: Main State Diagram for Link Training and Status State Machine**

---
PCIe Link Training Debug

- LTSSM state changes appear as 1000’s of ordered sets
Example: Link is Not Running at Maximum Speed

Do both sides advertise 5GT/s or 8GT/s – Was a speed change initiated?
Dynamic Equalization

- The PCI Express 3.0 Specification allows for 8.0GT/s data rate
- In order to establish a link reliably at 8GT/s, the protocol allows for dynamic equalization of the link during the speed change in the Recovery state
- This is a 4 phase process and happens in the Recovery.Equalization substate
Recovery Substate Machine

Figure 4-26: Recovery Substate Machine
Potential Equalization Problems

- Link does not get to 8GT/s
- Link gets to 8GT/s but is unstable
- Equalization fails because of wrong coefficients
- Does the equalization process complete?
- Protocol Analyzer is the only way to get visibility
How Do I Debug Equalization?

- Use the EC field as a trigger – this will help navigate the equalization
- Compress the Training Sequences helps digest the process – there are thousands of them!
POWER MANAGEMENT EXAMPLES

Application

Transaction

Data Link

Physical

Logical Sub Block

Electrical Sub Block
L0s

- It can be challenging to debug L0s problems, it is important to set things up properly
  - The exit latency from L0s is very short
  - Give the analyzer the best chance of locking by disabling auto detection such as speed, link width and polarity
  - Make sure the analyzer is set up properly with L0s turned off and ensure clean captures
ASPM – L0s
L1 and L1 Substates

- L1 is a more aggressive form of power management than L0s, and the L1 Substates ECN takes that further

- There are 2 mechanisms to enter L1 - ASPM L1 and PCIPM L1

  ✓ Each have a different entry mechanism
  ✓ PCIPM uses a config write from the RC to the PMSCR register on the device to initiate the transition to L1
  ✓ ASPM L1 the downstream component indicates the desire to enter the L1 state, and sends PM_Active_State_Request_L1 to the root, the root acknowledges with PM_Request_ACK DLLP
ASPM

- Advanced post processing shows how many times each state was entered
L1 Substates

- A device indicates its support of L1 substates in the configuration space
- L1 PM Substates is considered enabled on a Port when any combination of the ASPM L1.1 Enable, ASPM L1.2 Enable, PCI-PM L1.1 Enable and PCI-PM L1.2 Enable bits associated with that Port are Set
L1 Substates – L1.1

Upstream Port Initiated Exit from L1.1

Downstream Port Initiated Exit from L1.1
L1 Substates – L1.2

Upstream Port Initiated Exit from L1.2

Downstream Port Initiated Exit from L1.2
How Does it Look on a Protocol Analyzer?

L1 Entry Request

Link Enters Electrical Idle

CLKREQ# Deasserted

CLKREQ# Asserted

Link Enters Recovery
Failed Entry To L1

Figure 5-6: L1 Transition Sequence Ending with a Rejection (L0s Enabled)
L1 Transition Sequence

Figure 5-7: L1 Successful Transition Sequence
DATA LINK LAYER EXAMPLES
Data Link Layer Issues

- Some problems are related to flow control mechanisms
  - Most DLL FC problems will not be discovered by the current compliance tests
  - Many DLL FC problems are not discovered until after a product is released
  - Problems may only show up with specific combination of Root and End Point that is not tested during compliance testing
  - Data Link Layer problems often result in performance issues
Detecting Credit Errors

<table>
<thead>
<tr>
<th>Packet</th>
<th>VC ID</th>
<th>HdrFC</th>
<th>DataFC</th>
<th>CRC 16</th>
<th>Idle</th>
<th>Time Stamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1404</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0xA2ED</td>
<td>0.000 ns</td>
<td>0025.257 597 748 s</td>
</tr>
<tr>
<td>1405</td>
<td>0</td>
<td>2</td>
<td>16</td>
<td>0xADE5</td>
<td>40.000 ns</td>
<td>0025.257 597 780 s</td>
</tr>
<tr>
<td>1406</td>
<td>0</td>
<td>8</td>
<td>16</td>
<td>0x315A</td>
<td>48.000 ns</td>
<td>0025.257 597 820 s</td>
</tr>
<tr>
<td>1407</td>
<td>0</td>
<td>8</td>
<td>8</td>
<td>0xD3FA</td>
<td>96.000 ns</td>
<td>0025.257 597 860 s</td>
</tr>
<tr>
<td>1415</td>
<td>0</td>
<td>3</td>
<td>17</td>
<td>0x5DAF</td>
<td>27.668 µs</td>
<td>0025.257 597 900 s</td>
</tr>
</tbody>
</table>

Look here!
Data Link Layer Problems

- Flow control overflow
- Why is a packet getting a NAK?
  - Set up trigger on NAK and look what happened before
TRANSACTION LAYER EXAMPLES
Completion Timeout Example

- This is a transaction layer problem, but how would it exhibit itself on the link?
  - System may have “Blue Screened”
  - Need to look for a request, and then set a timer. If a completion is not sent before the timer expires, then possible completion timeout condition may be occurring
  - Set the trigger close to the end of the buffer, to capture what happened before the event

- The link may be still in L0 state, but TLP traffic has stopped
PCI EXPRESS STORAGE
PCIe Storage Devices

- Where does this fit into the layered model
  - NVMe
  - PQI/SOP
  - AHCI/SATA Express

- How do I debug?

- “I don’t care about PCI Express!!!”

- Compliance Programs for Storage are outside of the scope of PCI-SIG, however PCI-SIG is impacted by form factor changes
PCI Express Storage

Application

Transaction

Data Link

Physical
  Logical Sub Block
  Electrical Sub Block

Software
  – Ethernet, NVMe, SOP, AHCI, SATA…
PCle Storage Devices

- There are several different technologies emerging for PCI Express based storage
- These take advantage of the increasing speeds of Flash Memory storage and the next generations of NAND technology
- PCI Express provides a solid, scalable, tested platform to take advantage of these higher speeds
- The principle of storage over PCI Express is the same for all technologies but the software and implementations are very different
- These technologies have history in legacy storage technologies
How Do I Debug These Products?

- It may take a different way of thinking to use a PCI Express analyzer for debugging a storage problem.
- The storage protocols reside within the PCI Express payload data, and requires further decoding.
- It is possible that the only interface to these devices is PCI Express.
- New form factors are emerging also, SFF8639, M.2 for example.
PCIe Storage

- The underlying PCIe traffic has an impact on performance of the Storage layers
- Queue balancing
- Possibility of multiple devices over one link
  - Devices behind a switch
  - Virtualized devices
  - Built in scalability of technologies such as NVMe
The NVMe Commands are actually PCI Express Memory Reads and Writes to specific addresses.
NVMe Example - Continued

The diagram shows a Trace View of a storage system using NVMe protocol. The table displays various entries with columns for Time Stamp, Time Delta, Item, Type H⇒D, Type D⇒H, Address, and Data. Each entry represents a transaction in the system, with details such as time, type of data transfer, and the address involved.

The data indicates transactions such as `NVM 111` with types `CQHDBL` and `SQ_CMD`, and addresses like `00000002 1EC00000`. The data size is also noted, ranging from 1 dword to 1024 dworms.

The Trace View interface includes options for recording, generating reports, viewing settings, and other functionalities related to protocol analysis.
Serial ATA Example

SATA Commands on a PCI Express Link
The ATA Commands are actually PCI Express Memory Reads and Writes to specific addresses.
PCI Express Storage

- The packets or frames are all embedded into the payload of the PCI Express transactions.
- The storage technologies are typically quite simple at the hardware level and leave software to carry out most of the work.
- A PCI Express analyzer with additional decodes for storage protocols is required for visibility on the bus.
- Bear in mind that storage devices may or may not reside on standard PCI Express (CEM) form factors.
Summary and Conclusions

- The easiest way to debug PCI Express problems is to determine which layer is showing the problem.
- Using the correct tool will make the job much easier.
- Use a “Snapshot” trigger first of all to assess what is going on.
- Ensure lower level (electrical) layers are compliant and within spec before debugging protocol problems.
- Remember the PCI-SIG compliance testing is not a substitute for thorough validation, it is a useful sanity check.
Thank you for attending the PCI-SIG Developers Conference Israel 2015

For more information please go to www.pcisig.com