PCI Express (PCIe)
Overview

Peter Onufryk
Sr. Director, Product Development
PMC-Sierra
NVMe and PCIe Layering

- **NVMe**
  - Driver interface (registers)
  - Queuing interface
  - Command set
  - Command processing model

- **PCIe**
  - Reliable memory read/write transactions
  - Discovery and configuration
  - Switching & routing
  - Physical layer
PCIe Characteristics

- Specification defined by PCI-SIG
  - www.pcisig.com
- Packet based protocol over serial links
  - Software compatible with PCI and PCI-X
  - Reliable in-order packet transfer
- High performance and scalable from consumer to enterprise
  - Scalable link speed (2.5 GT/s, 5.0 GT/s, 8.0 GT/s)
  - Scalable link width (x1, x2, x4, …. x32)
- Primary application is as an I/O interconnect
  - Not a CPU interconnect
  - Some multi-host applications (NTB, MR-IOV)
  - Some outside the box applications (PCIe cable)
PCIe and Server Architecture

*Source - Intel
PCle is Everywhere

SERVERS
- Mid-range / High End
- Small Business
- Blade Server

STORAGE
- SAN Switch
- NAS
- RAID System

COMMUNICATIONS
- Enterprise Switch / Router
- Wireless
- Metro Switch / Router

CONSUMER
- Mobile
- Entertainment
- Imaging

Flash Memory Summit 2013
Santa Clara, CA
PCIe Fabric Topology
Data Transfers
PCle and NVMe
PCIe Layers

- PCI Architecture
  - Transaction Layer
  - Data Link Layer
  - Physical Layer

PCIe Link
Physical Layer

- **Scalable Speed**
  - Gen1 – 2.5 GT/s
  - Gen2 – 5.0 GT/s
  - Gen3 – 8.0 GT/s

- **Scalable Width**
  - x1, x2, x4, x8, x12, x16, x32

- **Encoding**
  - 8b10b: 2.5 GT/s & 5.0 GT/s
  - 128b/130b: 8 GT/s
## PCIe Performance

<table>
<thead>
<tr>
<th>Generation</th>
<th>Raw Bit Rate</th>
<th>Bandwidth Per Lane Each Direction</th>
<th>Total x16 Link Bandwidth#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gen 1</td>
<td>2.5 GT/s</td>
<td>~ 250 MB/s</td>
<td>~ 8 GB/s</td>
</tr>
<tr>
<td>Gen 2</td>
<td>5.0 GT/s</td>
<td>~ 500 MB/s</td>
<td>~ 16 GB/s</td>
</tr>
<tr>
<td>Gen 3</td>
<td>8 GT/s</td>
<td>~ 1 GB/s</td>
<td>~ 32 GB/s</td>
</tr>
</tbody>
</table>

# Link bandwidth in each direction x 2 (full duplex)

*Source – PCI-SIG PCI Express 3.0 FAQ*
Data Link Layer

- **Primary Function**
  - Reliable exchange of Transaction Layer Packets (TLPs) between the two components of a Link

- **Other Functions**
  - Initialization (flow control credits)
  - Power management
  - Track and report link state to transaction layer (i.e., DL_up & DL_Down)
Transaction Layer

- **Primary Function**
  - Assembly and disassembly of Transaction Layer Packets (TLPs) for read and write transactions

- **Other Functions**
  - Event signaling (e.g., interrupts, power management, errors)
  - Management of TLP credit based flow control
## Address Spaces

<table>
<thead>
<tr>
<th>Address Space</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Memory</strong></td>
<td>Data transfer to/from memory mapped locations</td>
</tr>
<tr>
<td></td>
<td>- 64-bit memory address space</td>
</tr>
<tr>
<td><strong>I/O</strong></td>
<td>Data transfer to/from IO-mapped locations</td>
</tr>
<tr>
<td></td>
<td>- 32-bit I/O space</td>
</tr>
<tr>
<td><strong>Configuration</strong></td>
<td>Device Function configuration &amp; setup</td>
</tr>
<tr>
<td></td>
<td>- 16-bit configuration space</td>
</tr>
<tr>
<td><strong>Message</strong></td>
<td>Event signaling &amp; general purpose messaging</td>
</tr>
</tbody>
</table>
Functions and Configuration Space

- **Function** – an addressable entity in configuration space
  - Architecturally visible (i.e., can be discovered and configured)
  - Capable of issuing requests and generating completions

![Diagram showing traditional and alternative routing-ID interpretation (ARI) ID]

Traditional ID: 8 5 3

Alternative Routing-ID Interpretation (ARI) ID: 8 8
Type 0 and Type 1 Functions

- **Type 0 Functions**
  - Function 0
  - Single Function Device

- **Type 1 Functions**
  - PCI-to-PCI Bridge

Multi-Function Device
Function Config. Space Registers

PCI Express Extended Configuration Space

PCI Express Extended Capability

Advanced Error Reporting

PCI Express Extended Capability

MSI Capability Structure

PCI Express Capability Structure

PCI Compatible Configuration Space Header

PCI Configuration Space
PCle Switches and I/O Fan-out

Root Complex

PCle Switch

Endpoint

PCle Switch

Endpoint

PCle Switch

Endpoint

PCle Switch

Endpoint
PCIe Switch

Physical View

Logical View

Type 1 Function
PCI-to-PCI Bridge

Virtual PCI Bus

Type 1 Function
PCI-to-PCI Bridge

Type 1 Function
PCI-to-PCI Bridge
Non-Transparent Bridge (NTB)
PCIe Multi-Path Usage Model

Flash Memory Summit 2013
Santa Clara, CA
PCIe Overview Summary

- PCIe is ubiquitous
- PCIe provides a scalable interface for SSDs
  - Scalable link width and speed
- PCIe is not a bottleneck
  - Highest performance standard I/O attach point
- PCIe switches provide I/O fan-out
  - Allows multiple SSDs to be connected to a Root Port
- NVMe works within the standard PCIe framework
  - Allows use of off-the-shelf Root Complexes and Switches
- PCIe may be used to connect multiple hosts to an SSD