PCI Express (PCIe) Overview

Peter Onufryk
Sr. Director, Product Development
PMC-Sierra

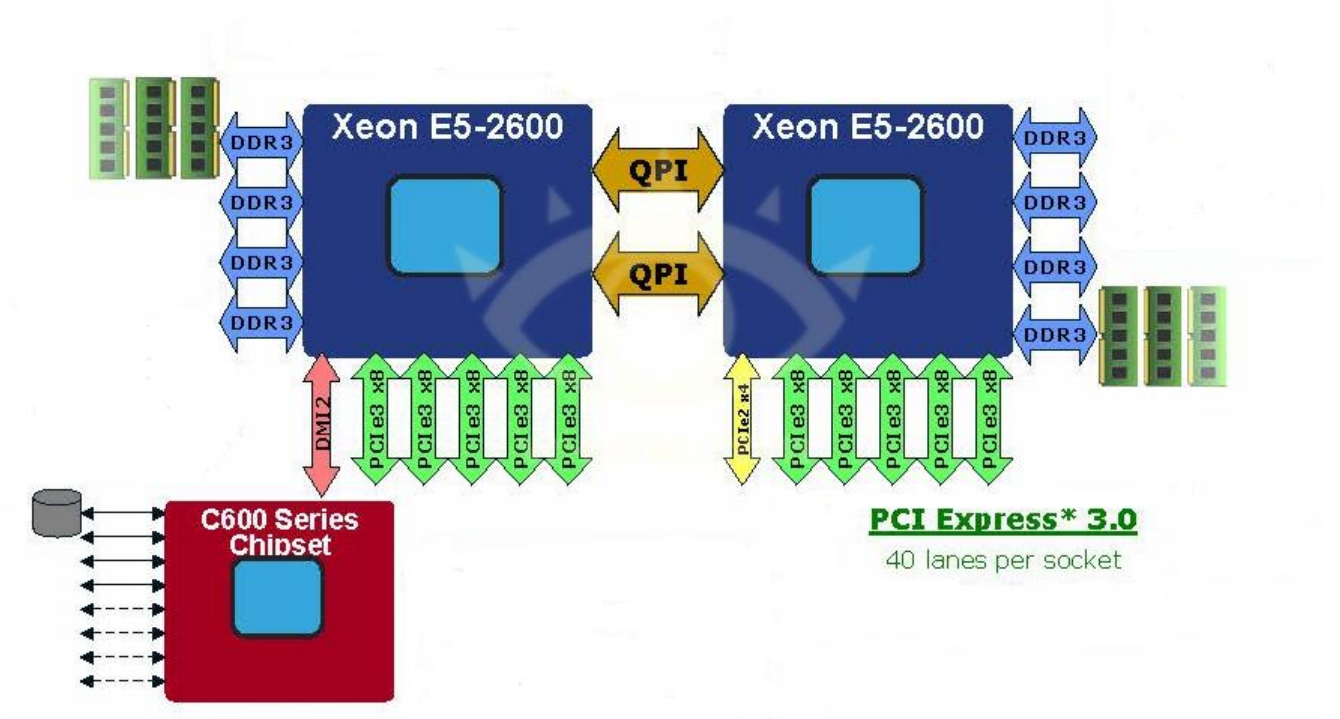
NVMe and PCIe Layering

- 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

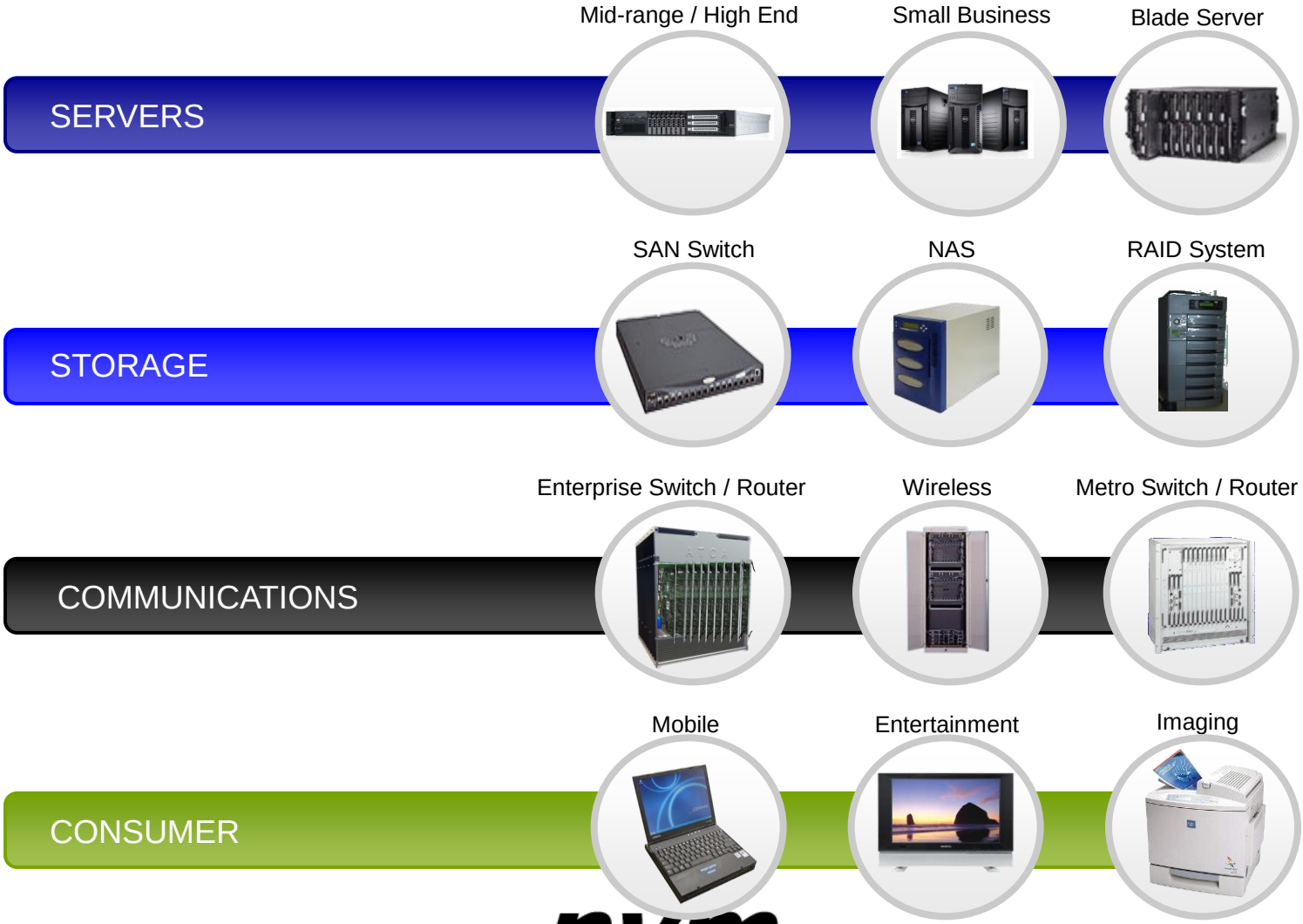
- Specification defined by PCI-SIG
  - www.pcisig.com
- Packet based protocol over serial links
  - Software compatible with PCI and PCI-X
  - Reliable in-order packet transfer
- High performance and scalable from consumer to enterprise
  - Scalable link speed (2.5 GT/s, 5.0 GT/s, 8.0 GT/s)
  - Scalable 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

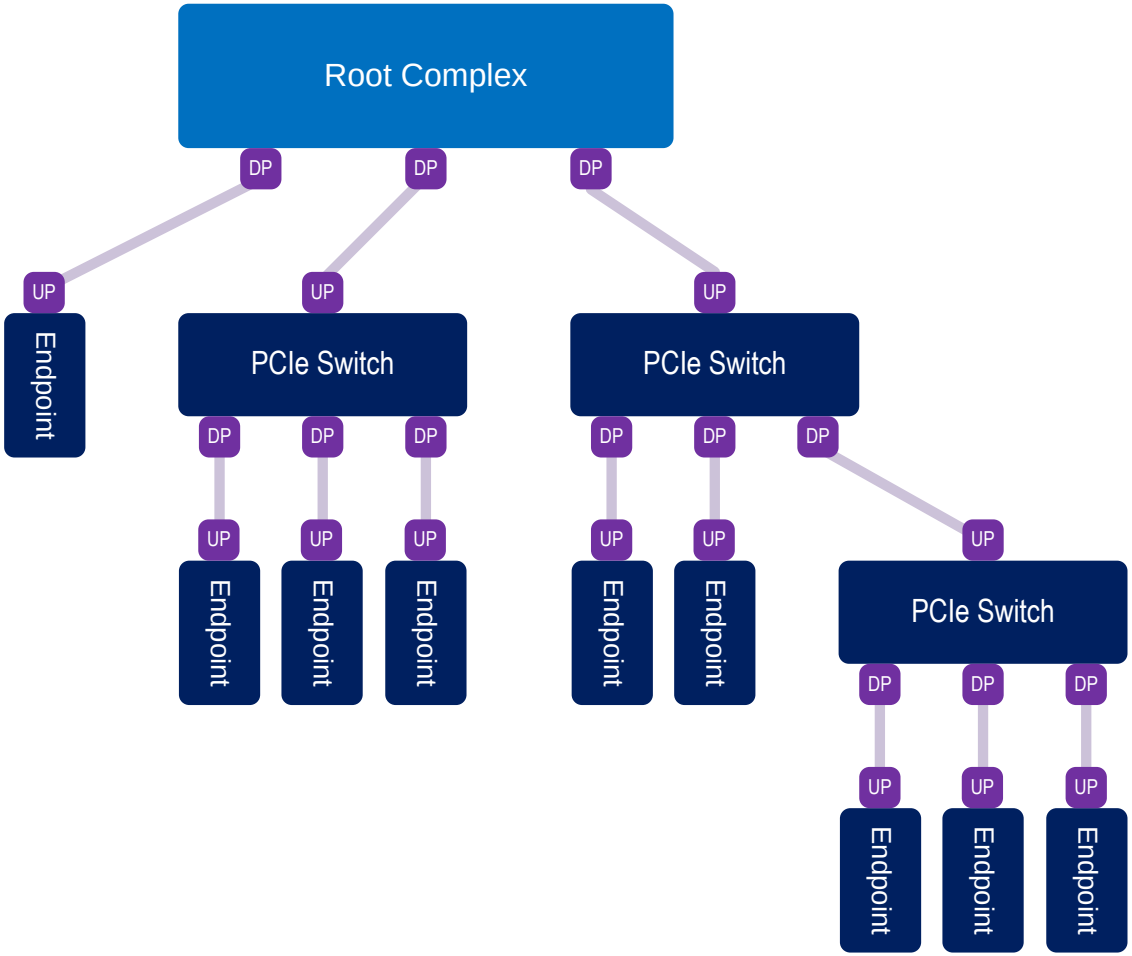


Source - Intel

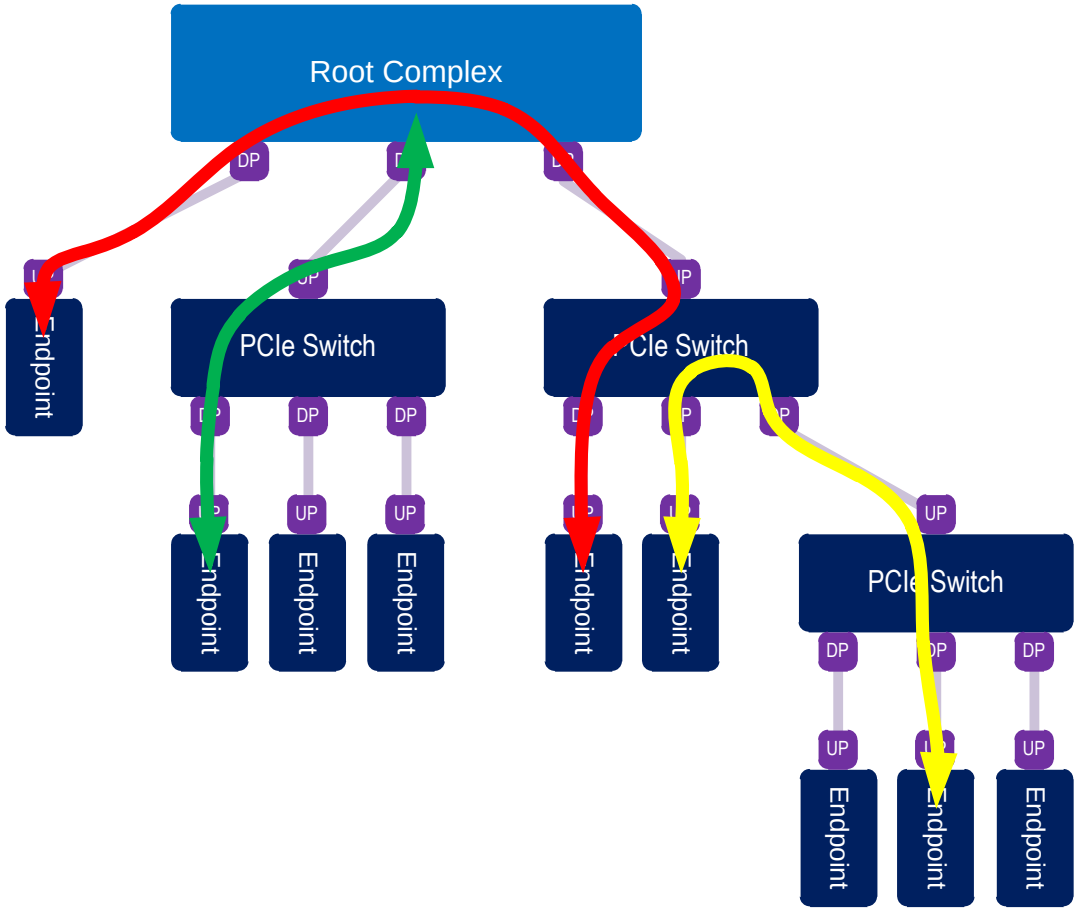
PCIe is Everywhere



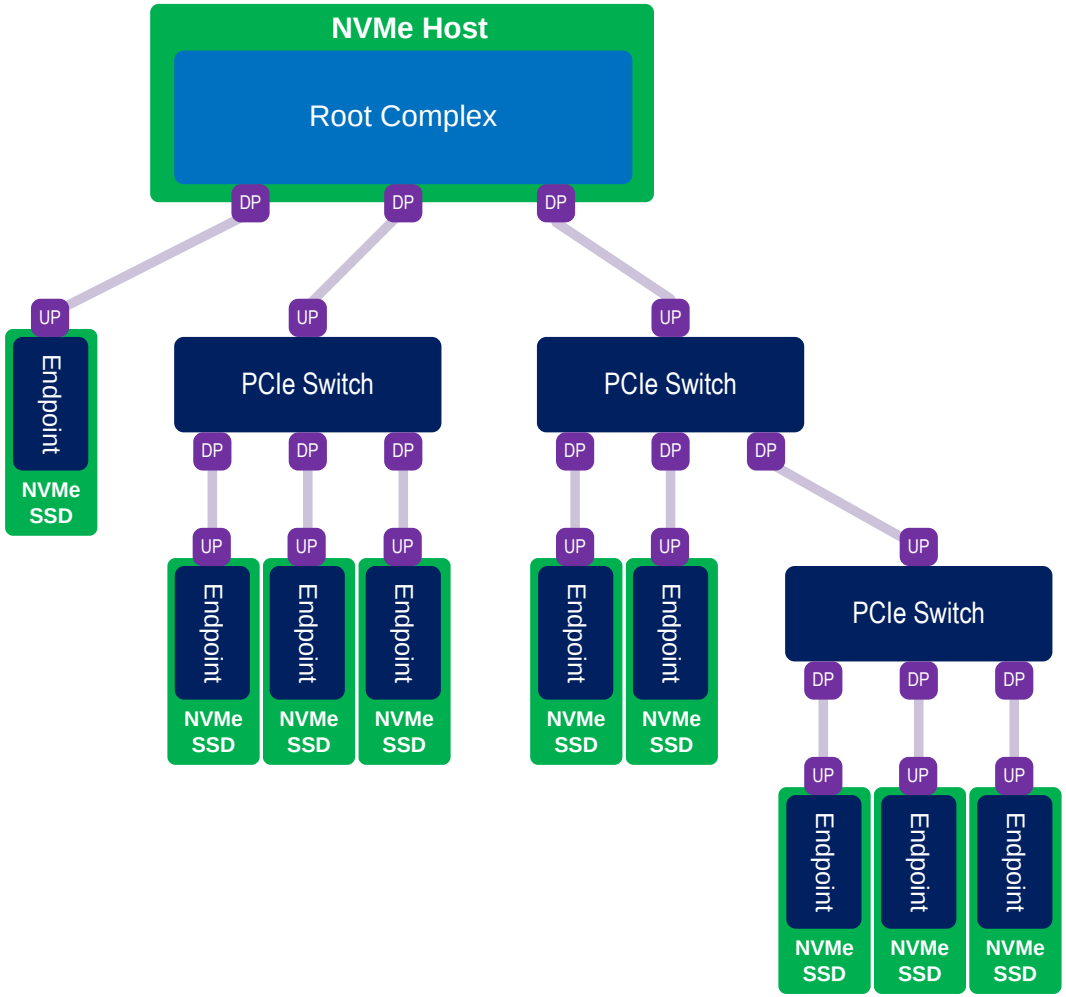
PCIe Fabric Topology



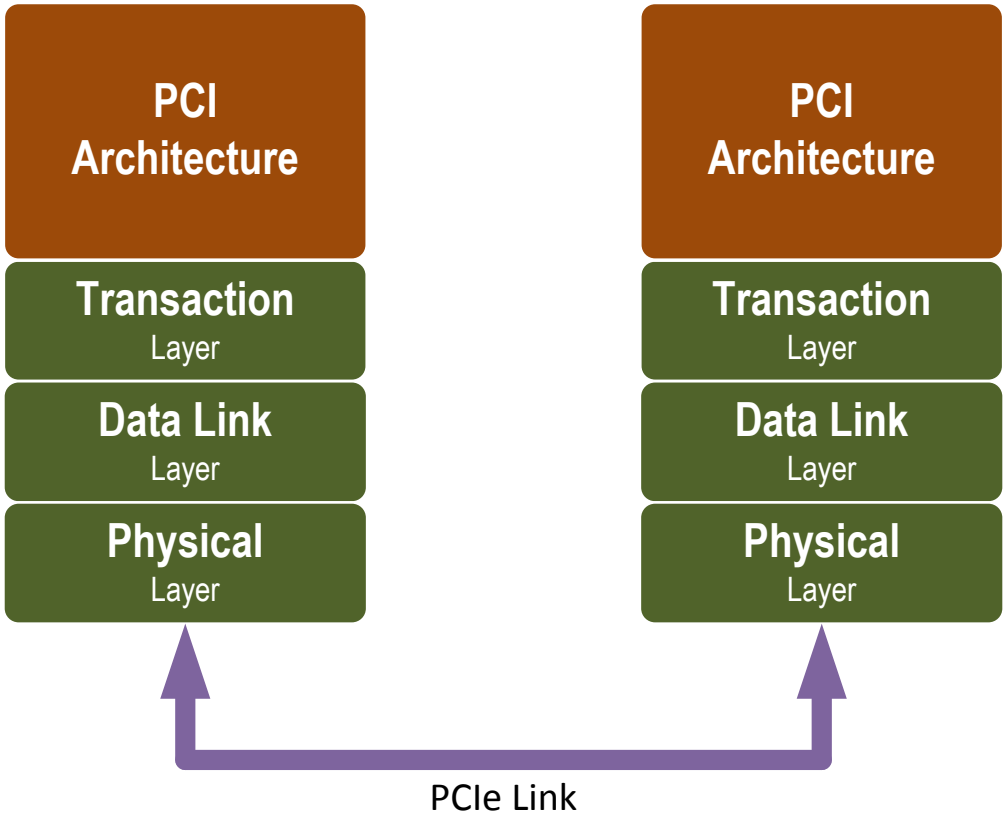
Data Transfers



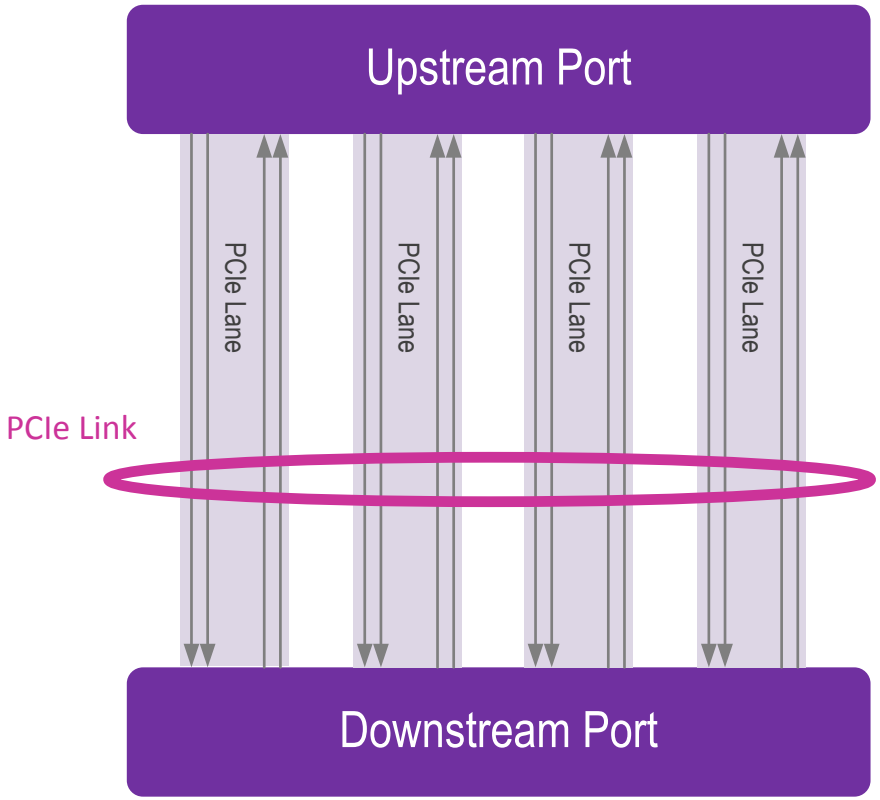
PCIe and NVMe



PCIe Layers


PCI Architecture
Transaction Layer
Data Link Layer
Physical Layer
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

Generation	Raw Bit Rate	Bandwidth Per Lane Each Direction	Total x16 Link Bandwidth#
Gen 1	2.5 GT/s	~ 250 MB/s	~ 8 GB/s
Gen 2	5.0 GT/s	~500 MB/s	~16 GB/s
Gen 3	8 GT/s	~ 1 GB/s	~ 32 GB/s

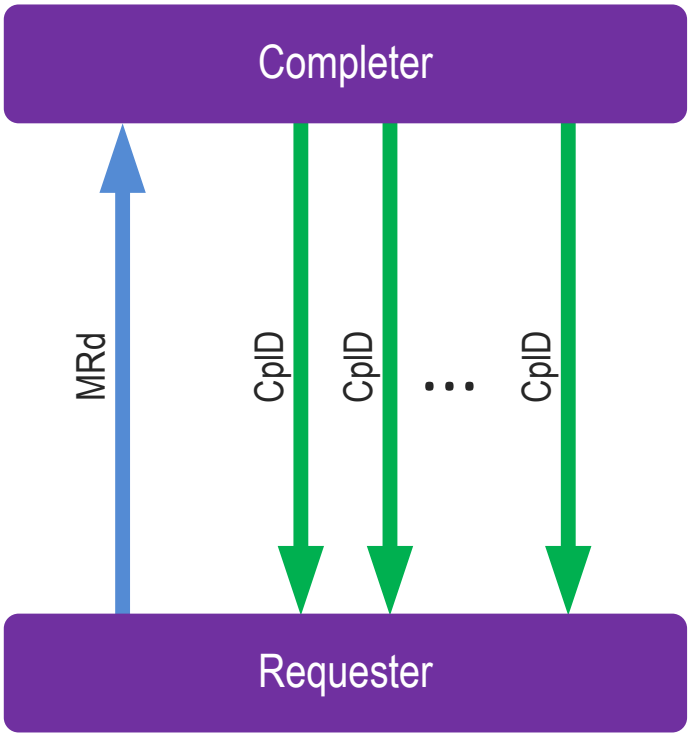
 # Link bandwidth in each direction x 2 (full duplex)

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



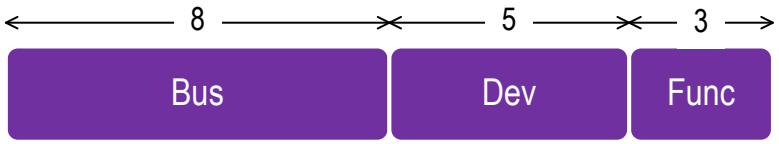
- Primary Function
  - Assembly and disassembly of Transaction Layer Packets (TLPs) for read and write transactions
- Other Functions
  - Event signaling (e.g., interrupts, power management, errors)
  - Management of TLP credit based flow control

Address Spaces

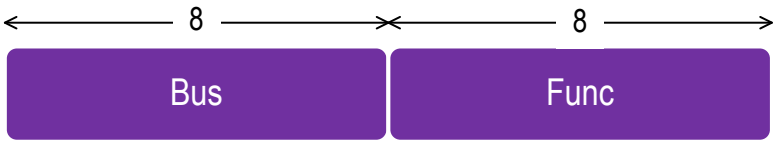
Address Space	Function
Memory	Data transfer to/from memory mapped locations - 64-bit memory address space
I/O	Data transfer to/from IO-mapped locations - 32-bit I/O space
Configuration	Device Function configuration & setup - 16-bit configuration space
Message	Event signaling & general purpose messaging

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

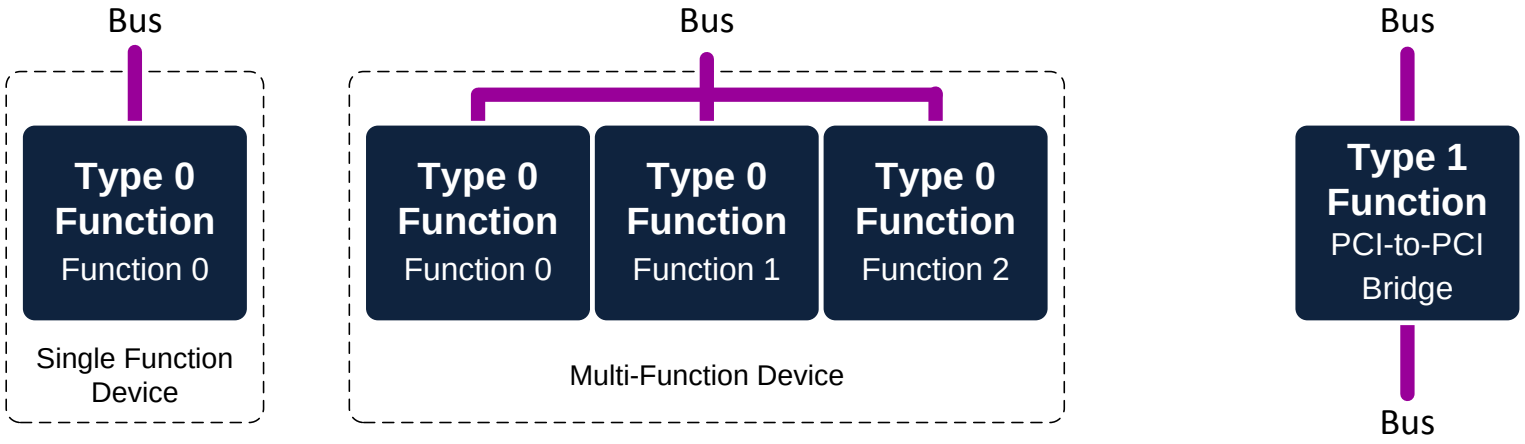


Traditional ID

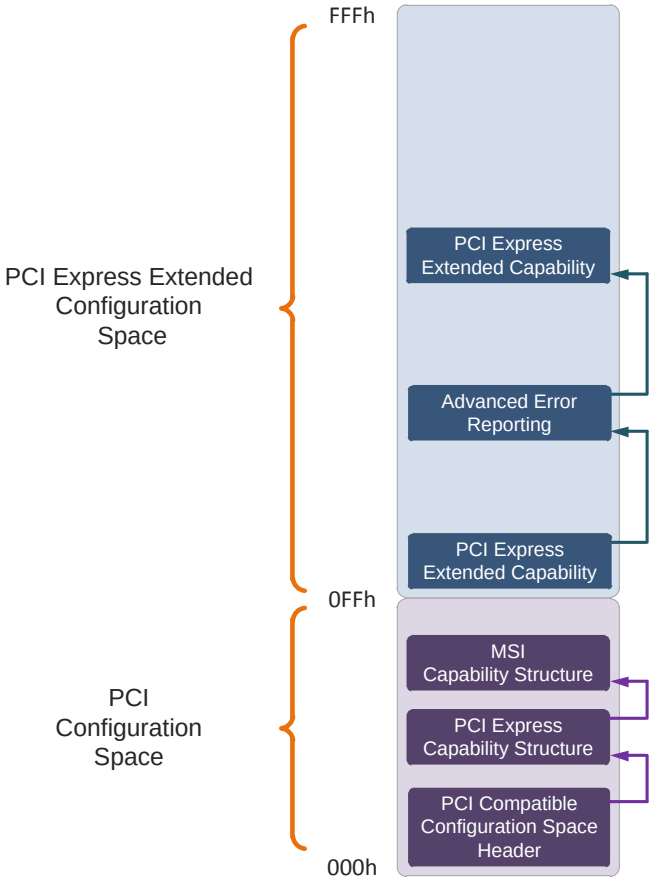


Alternative Routing-ID Interpretation (ARI) ID

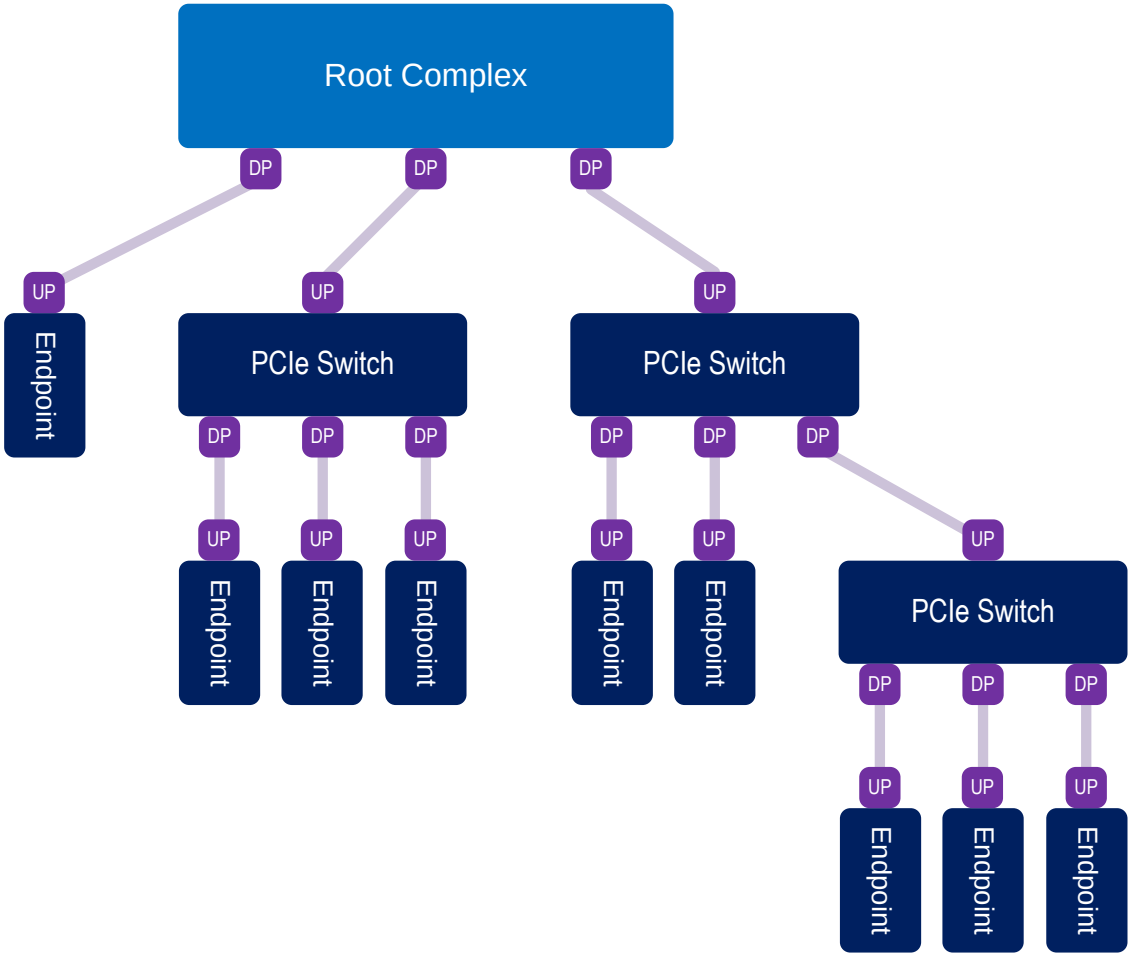
Type 0 and Type 1 Functions



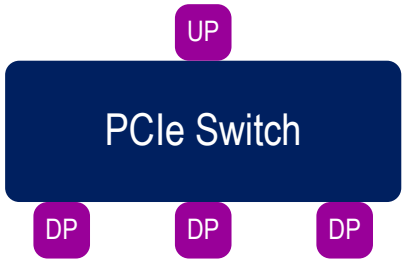
Function Config. Space Registers



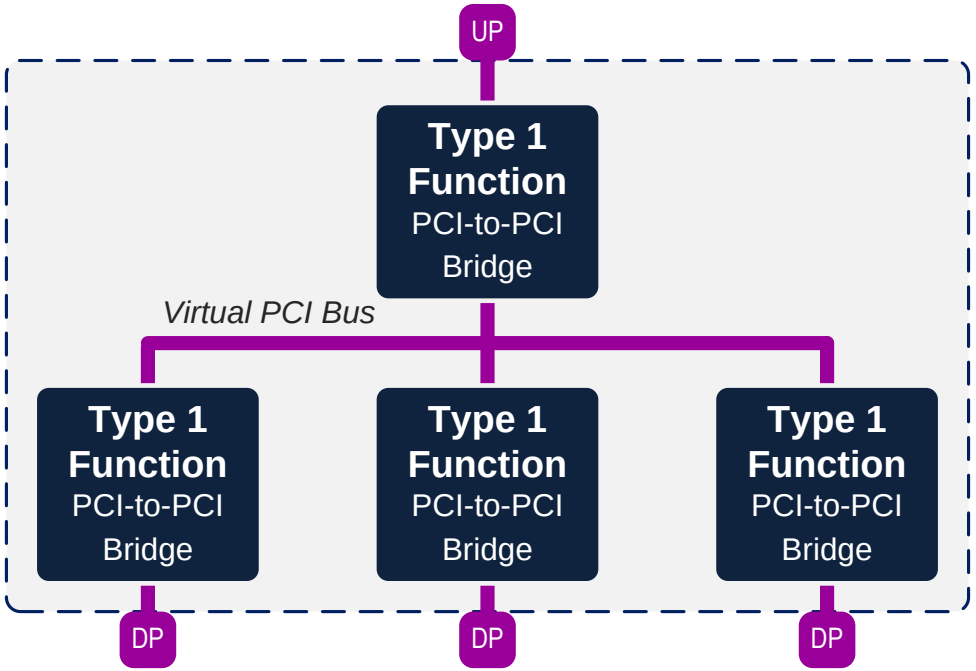
PCIe Switches and I/O Fan-out



PCIe Switch

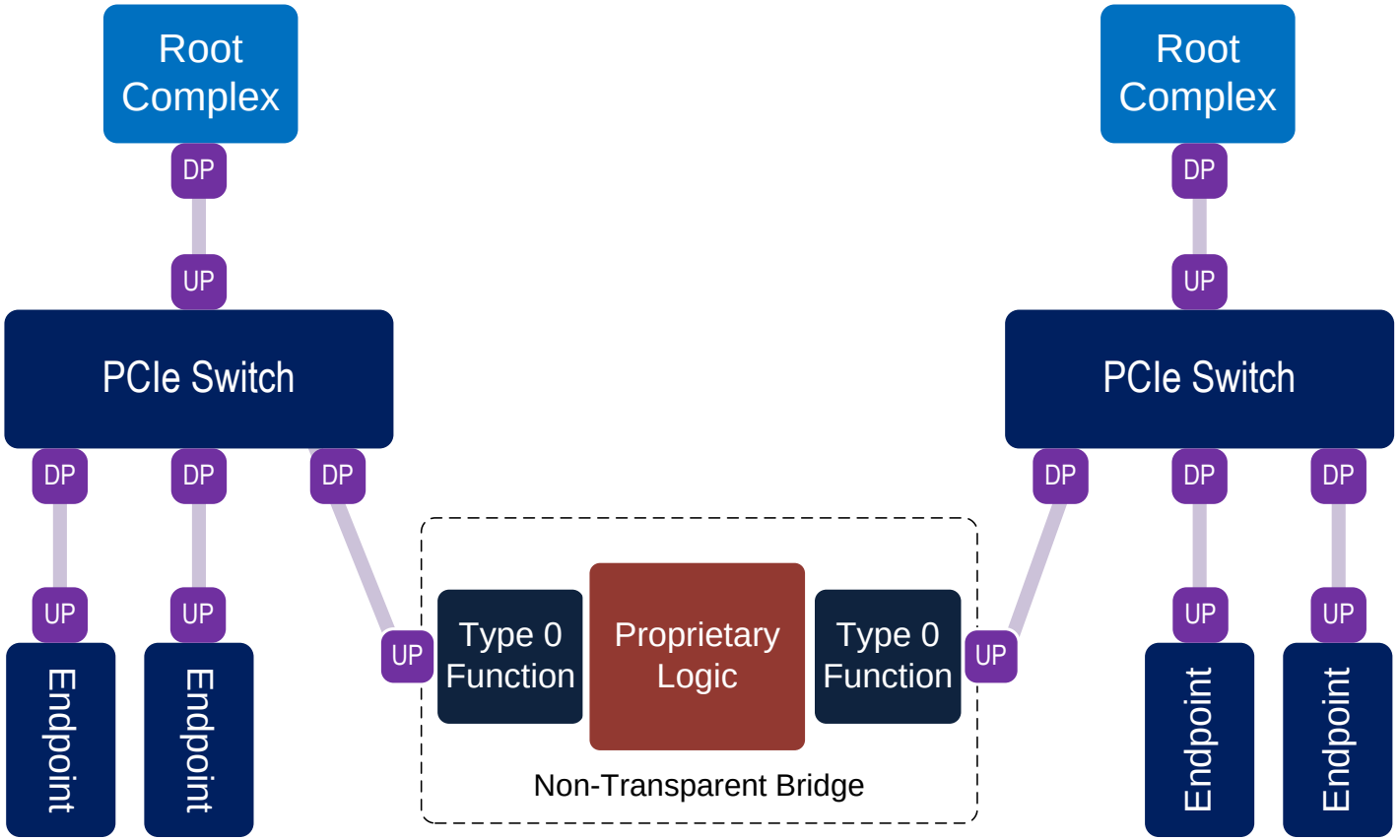


PCIe Switch
Physical View

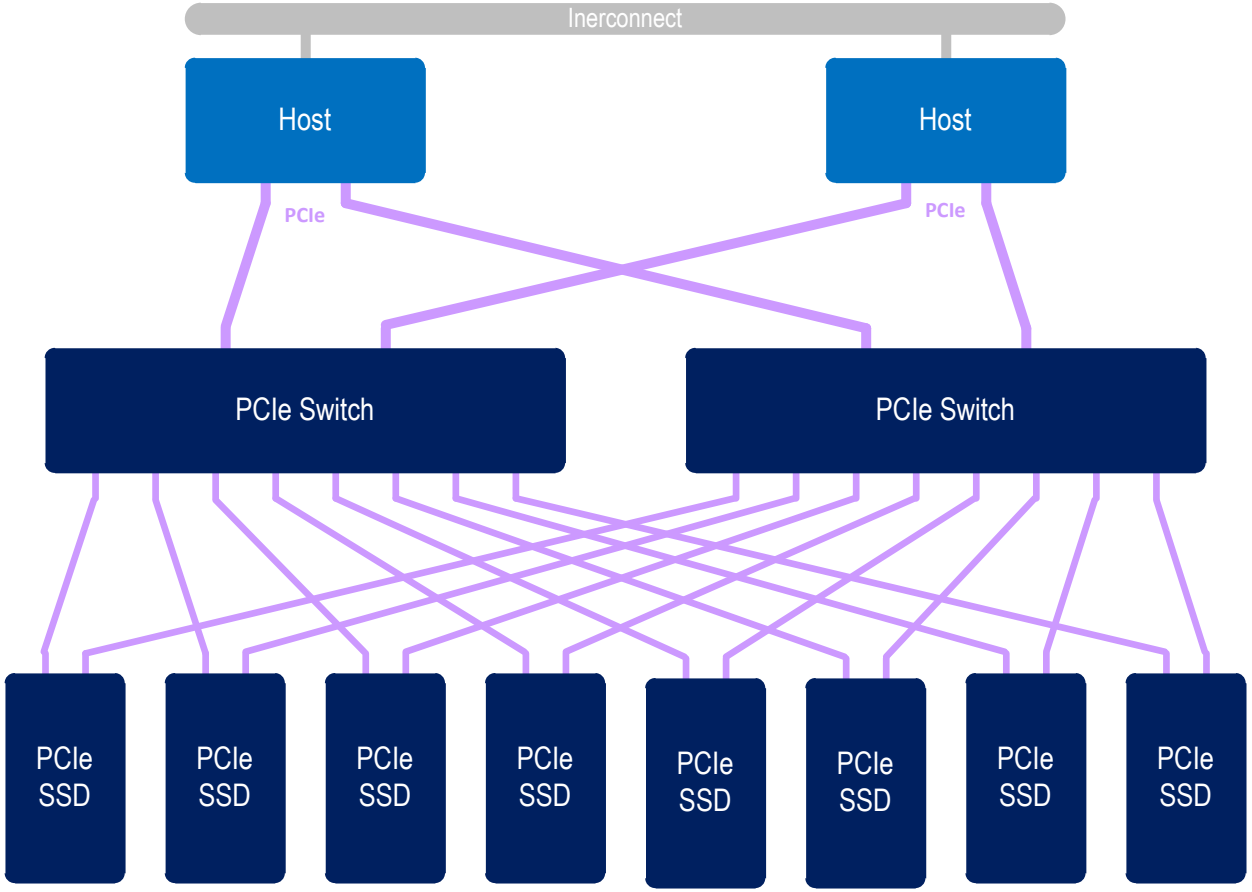


PCIe Switch Logical View

Non-Transparent Bridge (NTB)



PCIe Multi-Path Usage Model



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