PCI Express and Its Interface to Storage Architectures

Ron Emerick
Oracle Corporation

Abstract

PCI Express and Its Interface to Storage Architectures

- PCI Express Gen2 and Gen3, IO Virtualization, FCoE, SSD and PCI Express Storage Devices are here. What are PCI Express Storage Devices – Why do you care? This session describes PCI Express, Single Root and the implications on FCoE, SSD, PCI Express Storage Devices and impacts of all these changes on storage connectivity, storage transfer rates. The potential implications to the Storage Industry and Data Center Infrastructure will also be discussed.

  - Knowledge of PCIe Architecture, PCIe Roadmap, System Root Complexes
  - Expected Industry Roll Out of latest IO Technology and required Root Complex capabilities
  - Implications and Impacts of FCoE, SSD and NVMe Devices to storage Connectivity
  - What Might the Data Center Look Like?

Agenda

- IO Architectures
  - PCI Express is Here to Stay
  - PCI Express Tutorial
  - New PCI Express based architectures
  - How does PCI Express work
- IO Evolving Beyond the Motherboard
  - Serial Interfaces
    - InfiniBand, 10 GbE, 40 GbE, 100 GbE
    - PCIe IO Virtualization
  - Review of PCI Express IO Virtualization
  - Impact of PCI Express on Storage

I/O Architecture

- PCI provides a solution to connect processor to IO
  - Standard interface for peripherals – HBA, NIC etc
  - Many man years of code developed based on PCI
  - Would like to keep this software investment
- Performance keeps pushing IO interface speed
  - PCI/PCI-X 33 Mhz, 66 Mhz to 133 Mhz
  - PCI-X at 266 Mhz released
    - Problems at PCI-X 512 Mhz with load and trace length
- Parallel interfaces are almost all replaced
- Parallel PCI has migrated to serial PCI Express

PCI Express Introduction

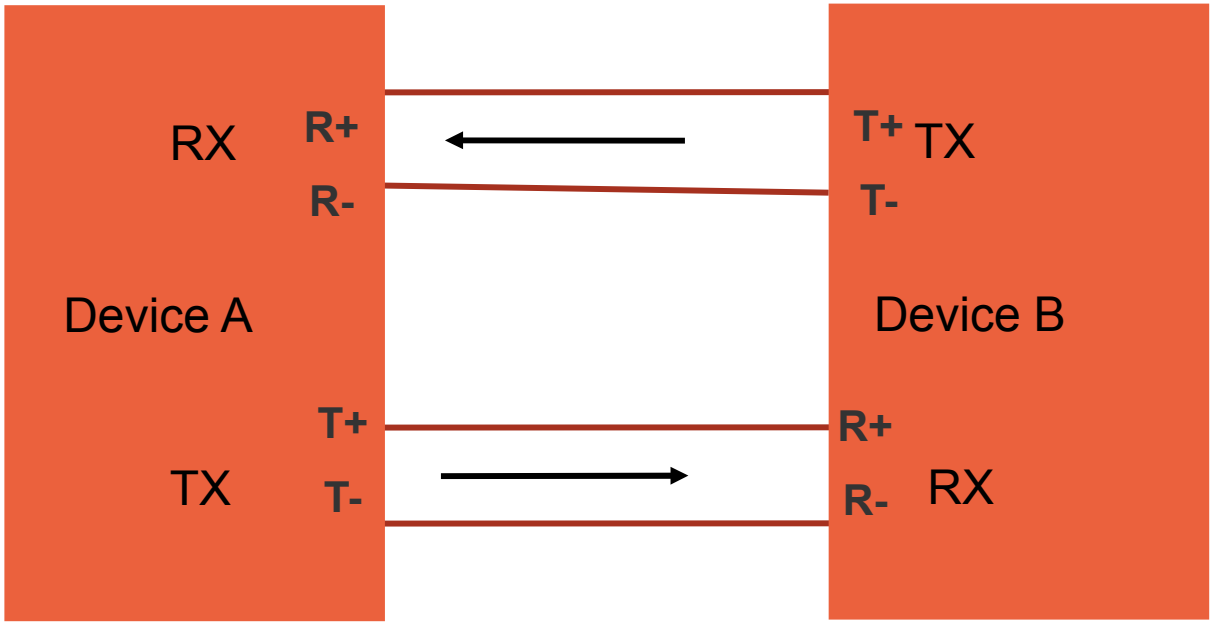
- PCI Express Architecture is a high performance, IO interconnect for peripherals in computing/ communication platforms
- Evolved from PCI and PCI-X TM Architectures
  - Yet PCI Express architecture is significantly different from its predecessors PCI and PCI-X
- PCI Express is a serial point- to- point interconnect between two devices (4 pins per lane)
- Implements packet based protocol for information transfer
- Scalable performance based on the number of signal Lanes implemented on the interconnect

PCI Express Overview

- Uses PCI constructs
  - Same Memory, IO and Configuration Model
  - Identified via BIOS, UEFI, OBP
  - Supports growth via speed increases
- Uses PCI Usage and Load/ Store Model
  - Protects software investment
- Simple Serial, Point- to- Point Interconnect
  - Simplifies layout and reduces costs
- Chip- to- Chip and Board-to-Board
  - IO can exchange data
  - System boards can exchange data
- Separate Receive and Transmit Lanes
  - 50% of bandwidth in each direction

PCIe What's A Lane

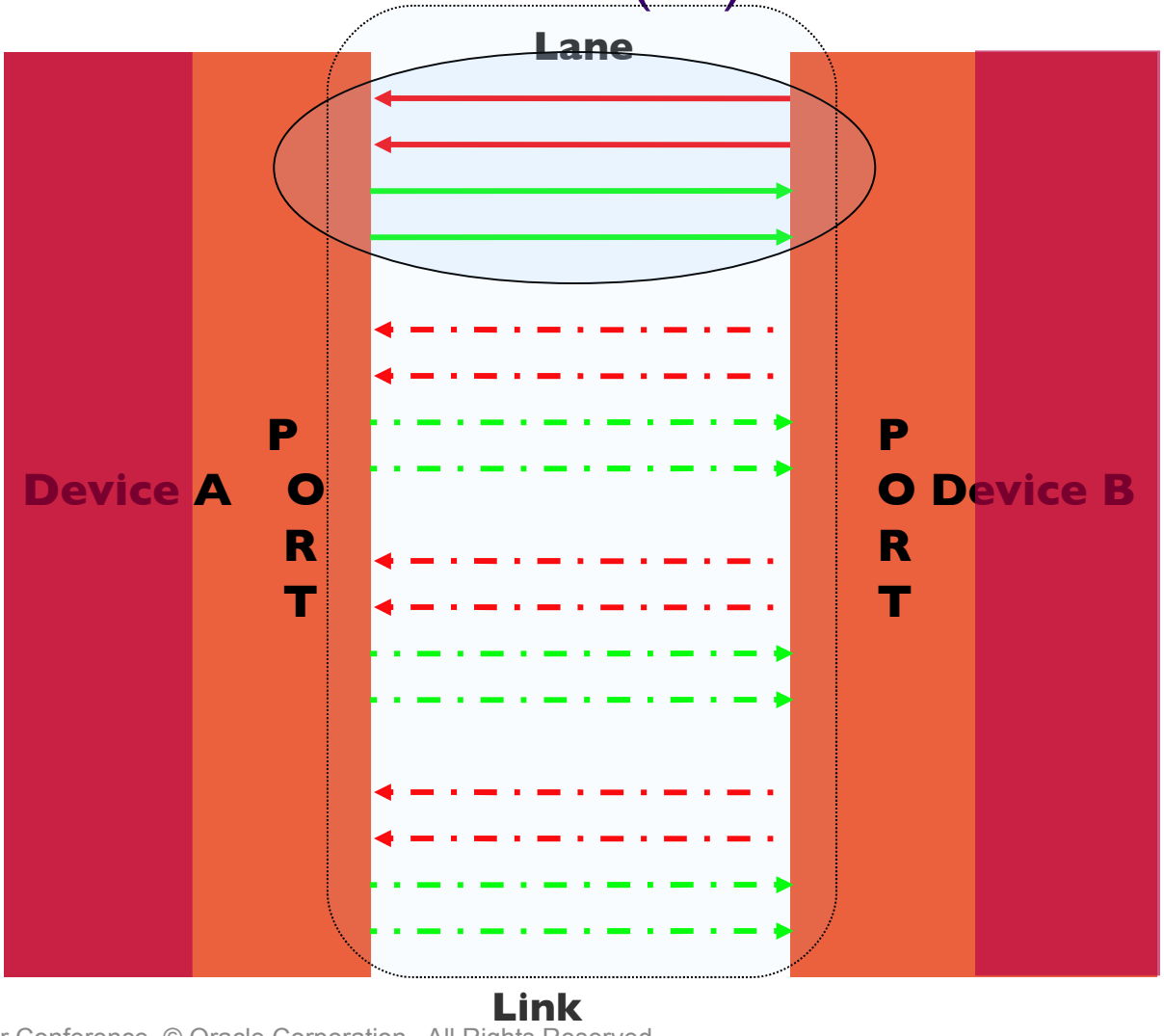
Point to Point Connection Between Two PCIe Devices



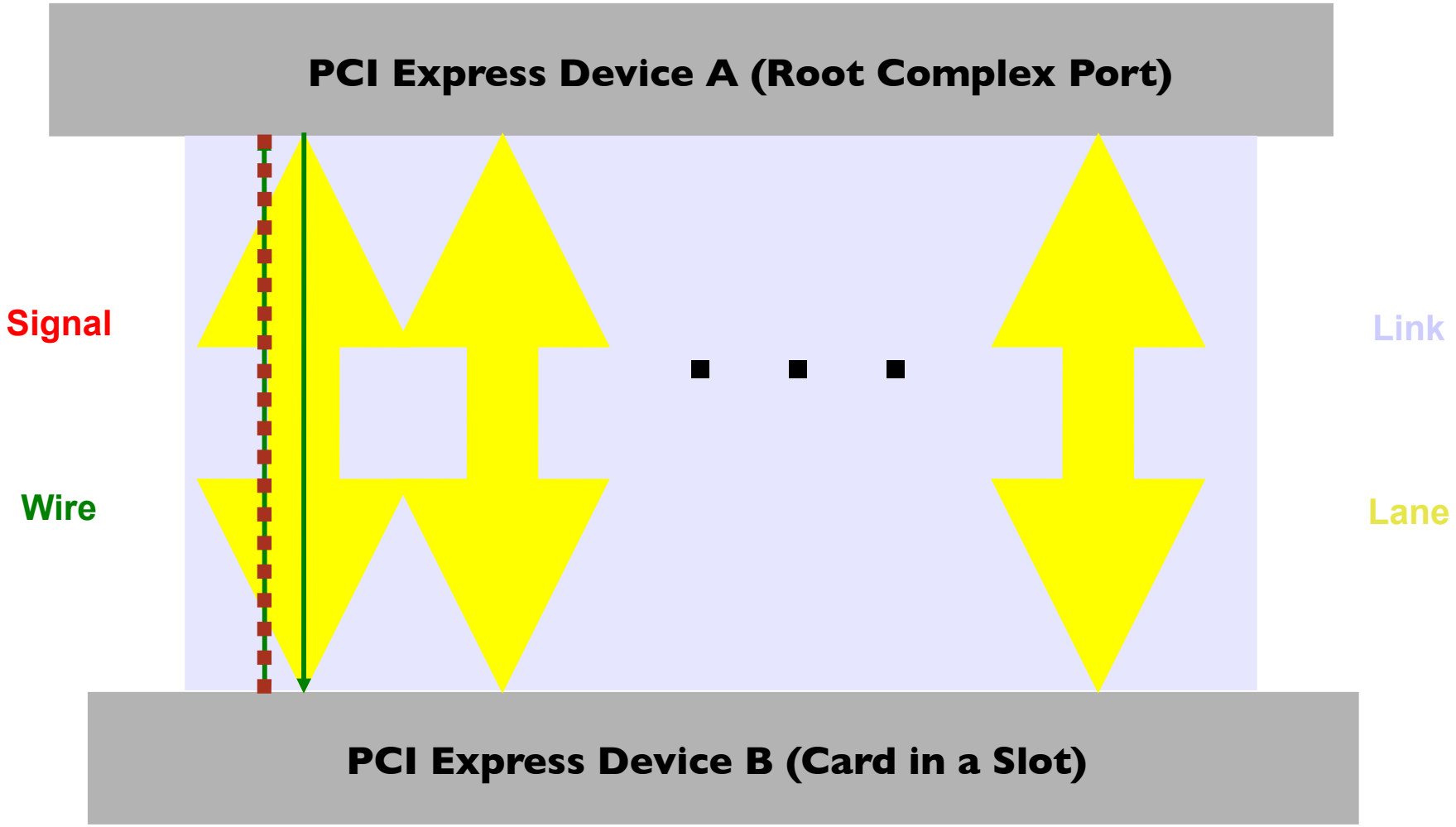
This Represents a Single Lane Using Two Pairs of Traces, TX of One to RX of the Other

PCIe – Multiple Lanes

Links, Lanes and Ports – 4 Lane (x4) Connection



PCI Express Terminology



Transaction Types

Requests are translated to one of four types by the Transaction Layer:

- Memory Read or Memory Write
  - Used to transfer data to or from a memory mapped location. Protocol also supports a locked memory read transaction variant.
- IO Read or IO Write
  - Used to transfer data to or from an IO location
  - These transactions are restricted to supporting legacy endpoint devices.

Transactions Types (cont)

Requests can also be translated to:

- Configuration Read or Configuration Write:
  - Used to discover device capabilities, program features, and check status in the 4KB PCI Express configuration space.
- Messages
  - Handled like posted writes. Used for event signalling and general purpose messaging.

PCI Express In Industry

- PCIe Gen 1.1 Shipped in 2005
  - Approved 2004/2005
    - Frequency of 2.5 GT/s per Lane Full Duplex (FD)
    - Use 8/10 Bit Encoding => 250 MB/s/lane (FD)
    - 2.5 GT @ 1 bit/T * 8/10 encoding / 8 bit/byte = 250 MB/s FD
    - PCIe Overhead of 20% yields 200 MB/s/lane (FD)
    - x16 High Performance Graphics @ 50W (then 75W)
    - x8, x4, x1 Connector (x8 is pronounced as by 8)
- PCIe Gen 2.0 Shipped in 2008
  - Approved 2007
    - Frequency of 5.0 GT/s per Lane
    - Doubled the Theoretical BW to 500 MB/s/lane 4 GB per x8
    - Still used 8/10 bit encoding
    - Support for Genesco features added (details later)
    - Power for x16 increased to 225W

PCI Express Activities

- PCIe Gen 3.0
  - Approved in 2011
  - Frequency of 8 GT/s per Lane
  - Uses 128/130 bit encoding / scrambling
  - Nearly Doubled the Theoretical BW to 1000 MB/s/lane
  - Power for X16 increased to 300W
- PCIe Gen 4.0
  - Work groups are active
  - Frequency of 16 GT/s per Lane
  - Uses 128/130 bit encoding / scrambling
  - Target Dates for Spec is late 2014

PCI Express Throughput

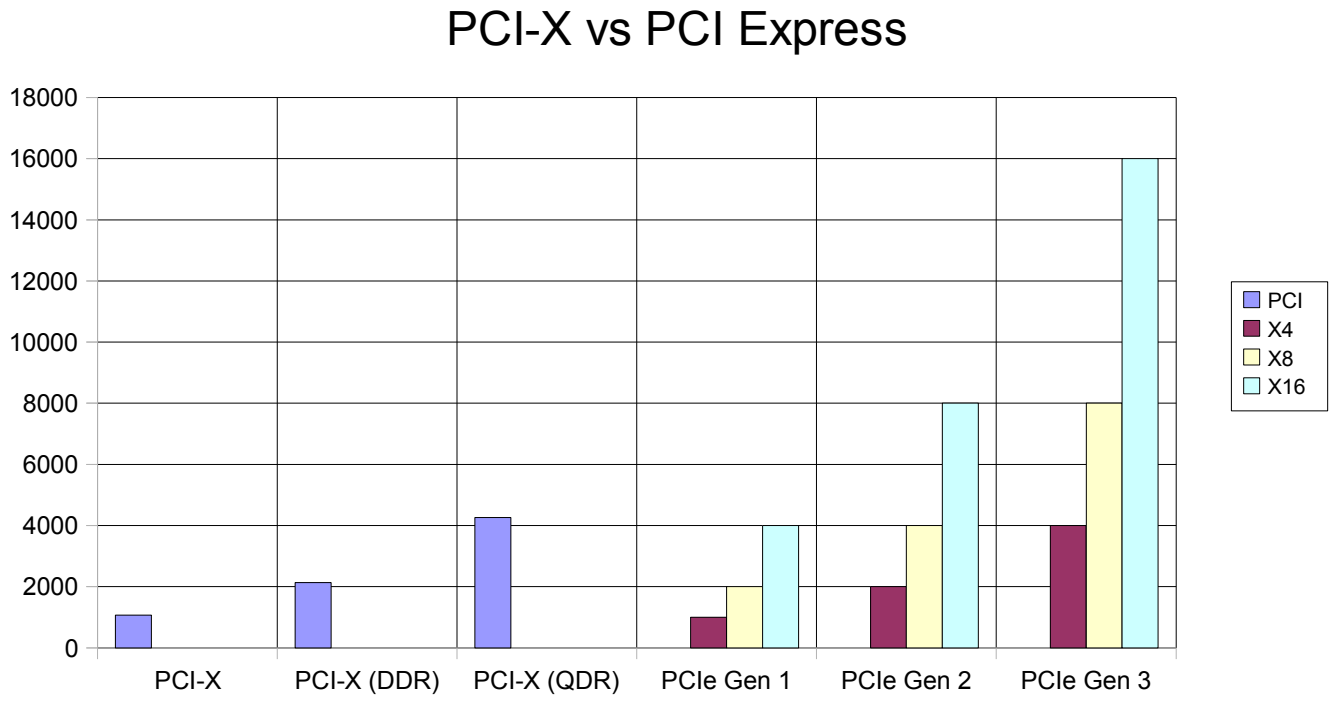
Link	Width	X1	X2	X4	X8	X16	X32
Gen1	2004	0.5	1	2	4	8	16
Gen2	2007	1	2	4	8	16	32
Gen3	2010	2	4	8	16	32	65
Gen4	2014?	4	8	16	32	64	128

- Assumes 2.5 GT/sec signalling for Gen1
- Assumes 5 GT/sec signalling for Gen2
  - 80% BW available due to 8 / 10 bit encoding overhead
- Assumes 8 GT/sec signalling for Gen3

Aggregate bandwidth implies simultaneous traffic in both directions

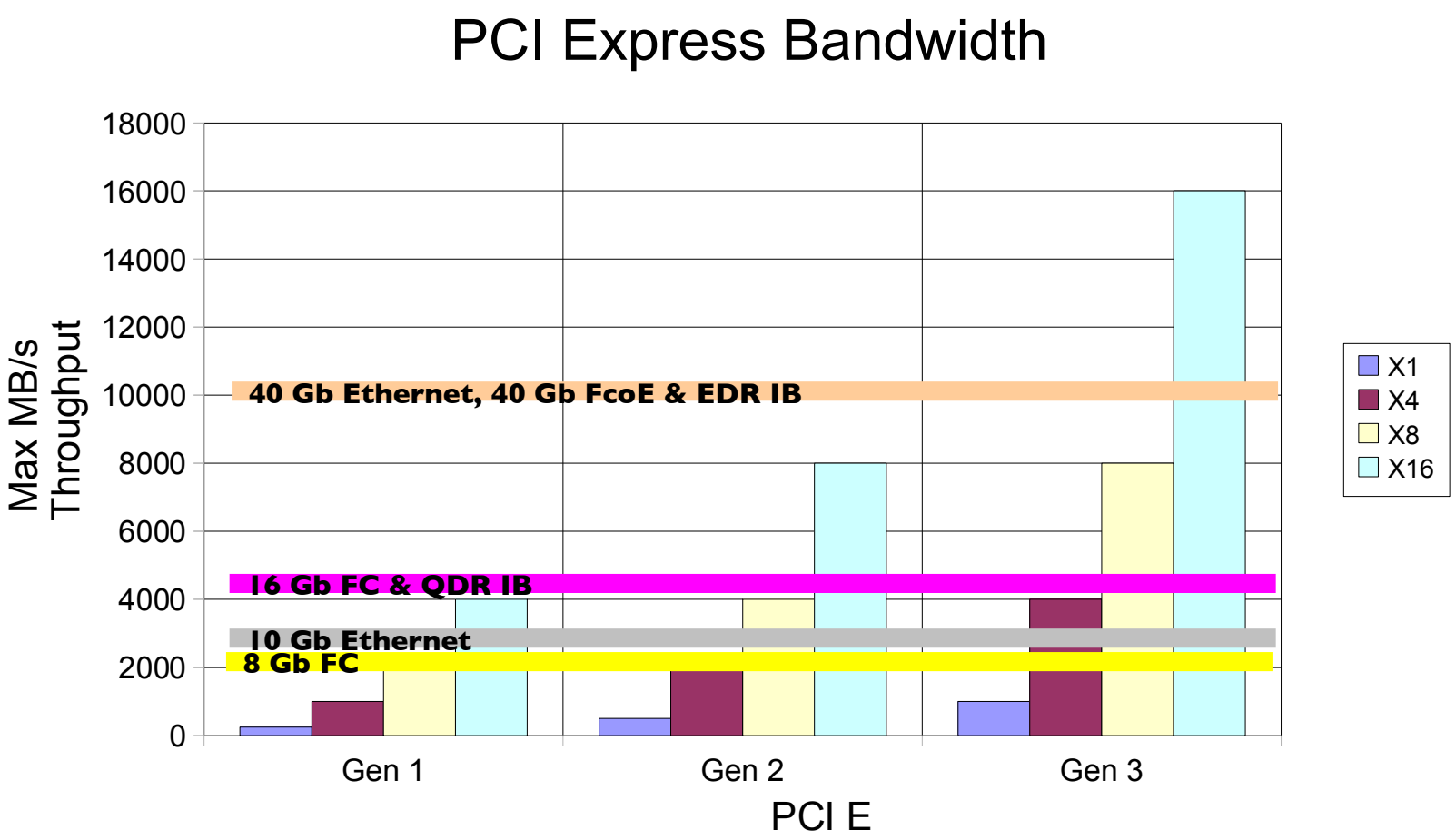
PCI-X vs PCIe Throughput

How does PCI-X compare to PCI Express?



- PCI-X QDR maxs out at 4263 MB/s per leaf
- PCIe x16 Gen1 maxs out at 4000 MB/s
- PCIe x16 Gen3 maxs out at 16000 MB/s

IO Bandwidth Needs



Sample PCI Express Topology

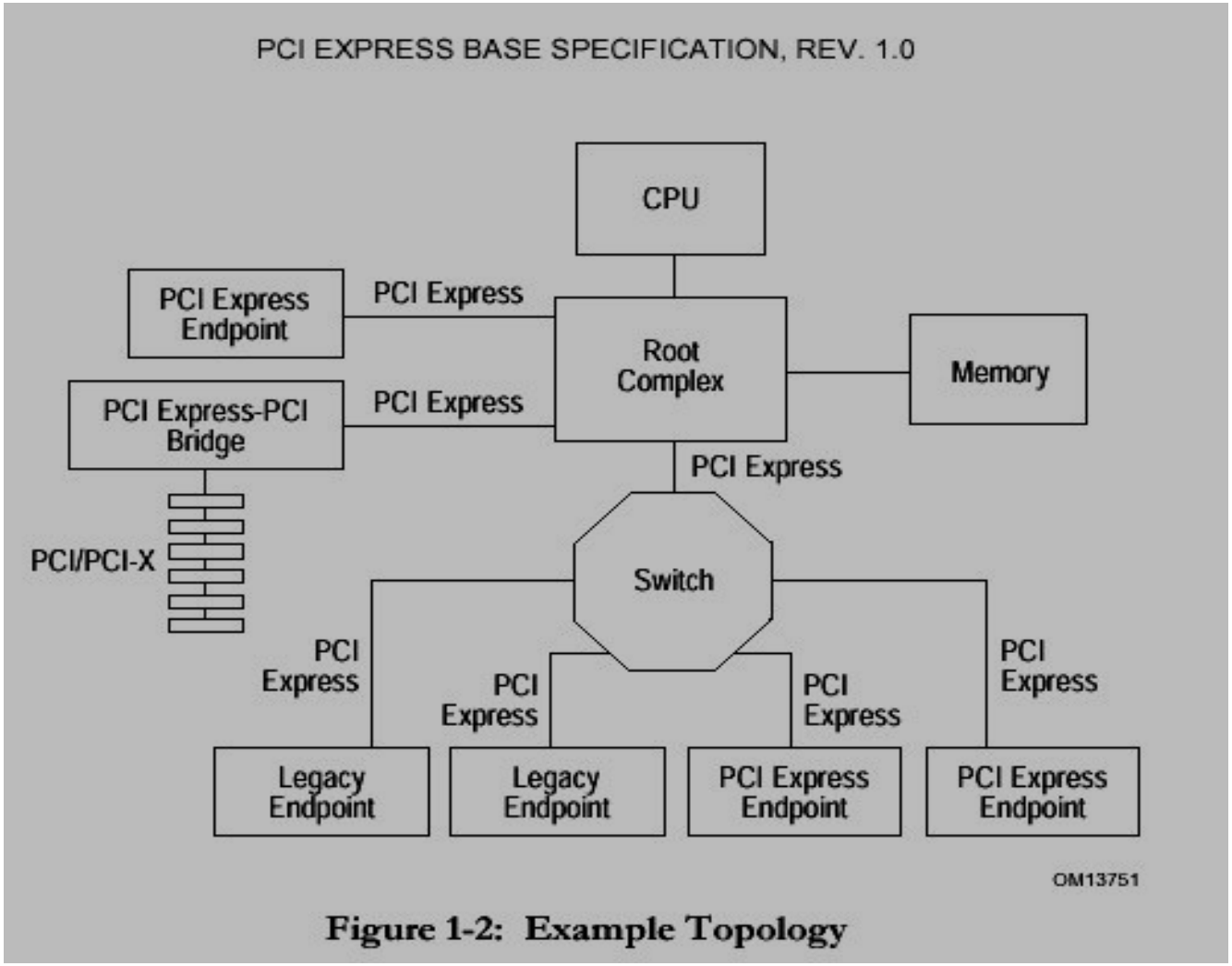


Figure 1-2: Example Topology

Benefits of PCI Express

- Lane expansion to match need
  - x1 Low Cost Simple Connector
  - x4 or x8 PCIe Adapter Cards
  - x16 PCIe High Performance Graphics Cards
- Point- to- Point Interconnect allows for:
  - Extend PCIe via signal conditioners and repeaters
  - Optical & Copper cabling to remote chassis
  - External Graphics solutions
  - External IO Expansion
- Infrastructure is in Place
  - PCIe Switches and Bridges
  - Signal Conditioners

New IO Interfaces

- High Speed / High Bandwidth
  - Fibre channel – 8 Gb, 16 Gb, 32 Gb , FCoE
    - Storage area network standard
  - Ethernet – 10Gb, 40Gb, 100Gb
    - Provides a network based solution to SANs
  - InfiniBand - QDR, FDR, EDR
    - Choice for high speed process to processor links
    - Supports wide and fast data channels
  - SAS 2.0, 3.0 (6 Gb, 12 Gb)
    - Serial version of SCSI offers low cost storage solution
  - SSDs
    - Solid State Disk Drive (SFF 8639 connedtor)
    - Solid State PCIe Cards (NVMe)

Evolving System Architectures

- Processor speed increase slowing
  - Replaced by Multi-core Processors
    - Sixteen-core here, 32 core coming
  - Requires new root complex architectures
- Root Complexes are at PCIe Gen3
  - Multiple slots at X16 Gen3 and X8 Gen3
  - PCI SIG is working on Gen4 (16 GT/S)
- Interface speeds are increasing
  - Ethernet moving from GbE to 10G, FC from 8 Gb to 16 Gb, Infiniband is now QDR with FDR and EDR coming
    - Single applications struggle to fill these links
    - Requires applications to share these links

  40 Gb / 100 Gb on the horizon

Drivers for New IO Architectures

- Solid State Storage
  - SAS Flash Cards
  - PCIe NVMe Flash Cards
  - NVMe drives in 2.5" formfactor
    - Shared IO support N+1 redundancy for IO, power and cooling
    - Remotely re-configurable solutions can help reduce operating cost
    - Hot plug of cards and cables provide ease of maintenance
  - Less cable types to manage
  - Reduction in types of experts
    - System Admins are expensive
      - Especially IB and SAN

What Do We Do???

- What We Need
  - Shared IO support N+1 redundancy for IO, power and cooling
  - Remotely re-configurable solutions can help reduce operating cost
  - Hot plug of cards and cables provide ease of maintenance
- Possible Solutions
  - Blade centres from multiple vendors
  - Storage and server clusters
  - PCI Express IOV allows commodity I/O to be used
  - Root Complexes with multiple X4 point to point connection
  - Flexible slots

Share the IO Components

PCIe IOV Provides this Sharing

- Root Complexes are PCIe
  - Closer to CPU than 10 GbE or IB
  - Requires Root Complex SW Modifications
- Based Upon PCI SIG Standards
- Allows the Sharing of High Bandwidth, High Speed IO Devices

Storage Device Capabilities

Storage Interfaces

- Common External Storage Interfaces

  All things network

  - NAS, iSCSI, FcoE devices (all ethernet based)
  - FC devices (normally SAS backend)
  - IB devices
- Common Internal Storage Interfaces
  - SAS or SATA to Internal HDD/SSD
  - SAS to PCIe SAS Flash Card
  - NVMe interface to NVMe Flash card

PCIe Card Possibilities

All Devices are limited by the Speed of the Flash Memory Controller on the Card.

- SAS Flash Cards
  - SAS Controller is 6 or 12 Gb/s (750 or 1500 MB/s) per SAS lane
  - Most interfaces are 4 lanes of SAS with one lane to each Flash module
- NVMe PCIe Gen3 X8
  - Interface is 64 Gb/s (8000 MB/s) per Device

IO Path

- Application Program Issues an IO

  OS determines the Target of the IO and Encapsulates the IO

  Possible Targets:

  NAS, iSCSI across network

  SCSI & Network

  FCoE across Network/SAN

  SCSI, FC & Network

  FC across SAN

  SCSI & FC

  SAS internal PCIe Flash Card, HDD or SSD, external JBOD or Array

  SAS

  PCIe NVMe Flash Card

  Block Read & Write or DMA Read or Right

Hard Drive Capabilites

- SAS 6 Gb/s

  Interface is 6 Gb/s (750 MB/s) per SAS lane

Speed	Latency	Peak Perf	Sustained
10,000	2.9-3.0 ms	600 MB/s	168-202 MB/s
15,000	2.0 ms	600 MB/s	152-202 MB/s

- SATA 6 Gb/s

  Interface is 6 Gb/s (750 MB/s) per SATA lane

Speed	Latency	Peak Perf	Sustained
5400	4.2 - 5.6 ms	300 MB/s	100 MB/s
7200	4.16 ms	600 MB/s	125-180 MB/s

SSD 2.5” Drive Capabilites

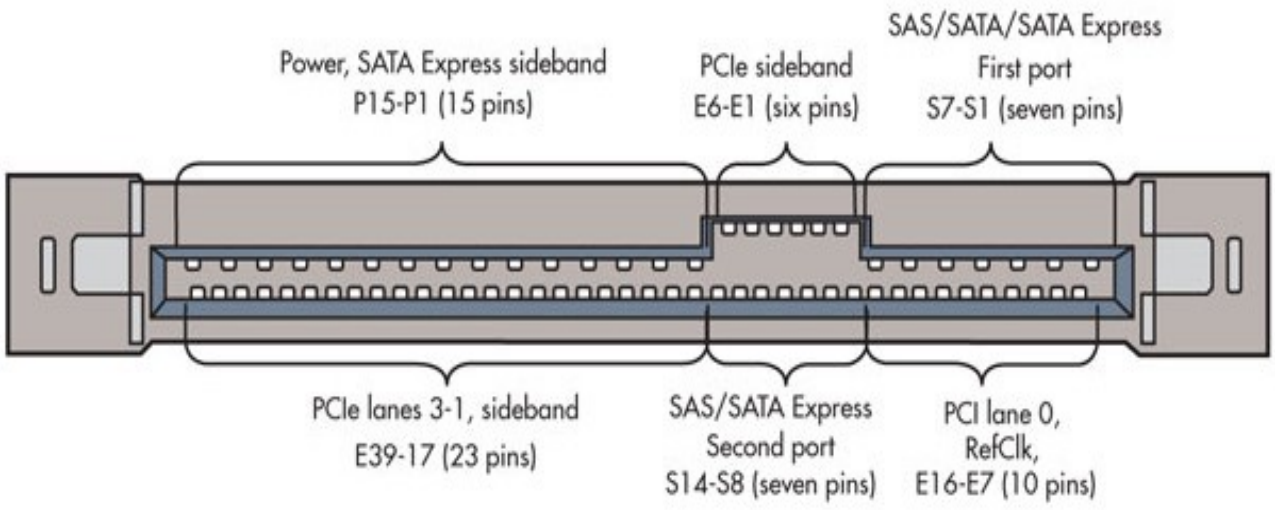
- SAS 12 Gb/s
  - Interface is 12 Gb/s (1500 MB/s) per SAS lane
- SATA 6 Gb/s
  - Interface is 6 Gb/s (750 MB/s) per SATA lane
- NVMe PCIe Gen3 X4
  - Interface is 32 Gb/s (4000 MB/s) per Device

New HDD/SSD Connector

SFF-8639 Drive Backplane Connector

Introduction to the new SFF8639 Connector for Backplanes. SFF 8639 is an extension of the 8482 SAS Connector which has higher signal quality requirements (to support 12Gb/s SAS and gen 3 PCIe) and adds support for SATA Express and 4 lanes of MultiLink SAS or PCI Express.

The connector has a total of 6 High Speed signal paths, strangely each specification only ever uses up to 4 of them at any time.



SFF-8639 Drive Connector

The standard defines a backplane connector that is designed to support the following interfaces:

Single Port SATA - All existing SATA Drives as defined by the SATA Specification Revision 3.1

Two Port SATA Express - new SATA Express drives as defined in the soon to be ratified SATA Express Specification

Dual Port SAS - All existing dual port SAS Drives up to 6Gb/S (SFF8482) and new 12Gb/S Drives (SFF8680)

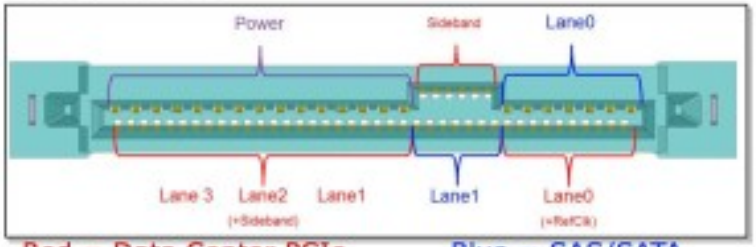
MultiLink SAS - new 4 lane SAS Drives (SFF8630)

PCI Express - PCI Express disk drives with up to 4 lanes of PCI Express data (SFF8639,SSD Form Factor V1.0)

Does SATA Express Work?

Trend: Two Connectors

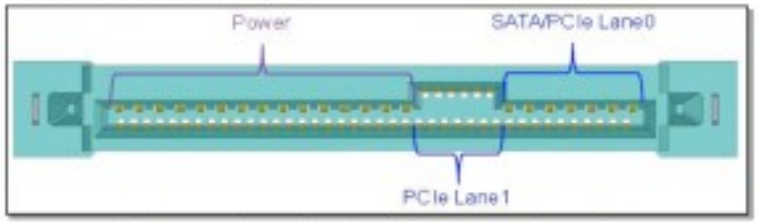
SFF-8639



Red = Data Center PCIe Blue = SAS/SATA

- 4x PCI Express* unleashes performance SSDs
- Cables require RefClk and six high speed signal lanes, shield?
- Increased device attach flexibility between SATA / SAS* / PCI Express
- Decreased system flexibility with directly wired SATA and PCI lanes

SATA* Express



Blue = Muxed single SATA or 2x PCIe

- 2x PCI Express limits SSD performance, still nearly 2x SATA
- Cost optimized w/ two high speed lanes, no RefClk, likely no shield
  - Requires clockless drive w/ SSC, in PCI SIG now
- Increased system flexibility supporting muxed SATA and PCI
- "It fits, but doesn't work!" risk

SFF-8639: Optimized for performance and device flexibility
SATA Express: Focus on rapid low cost platform transition

What Your Next Data Center Might Look Like

Data Center in 2014-2016

- Root complexes are PCIe 3.0
  - Integrated into CPU
  - Multiple Gen3 x16 from each socket (exposing more x16 as slots)
  - Multicast and Tunneling
  - PCIe Gen4 in 2016 and beyond
- Networking
  - Dual ported Optical 40 GbE (capable of FCoE, iSCSI, NAS)
  - 100 GbE Switch Backbones by 2016
  - Quad 10 Gbase-T and Quad MMF (single ASIC)
  - Dual/Quad Legacy GbE Copper and Optical
  - Dual ported FDR & EDR InfiniBand for cluster, some storage
- Graphics
  - x16 Single/Dual ported Graphics cards @ 300 W (when needed)

Data Center in 2014-2016 (2)

- Storage Access
  - SAS 3.0 HBAs, 8 and 16 port IOC/ROC
  - 16/32 Gb FC HBAs pluggable optics for single/dual port
  - Multi-function FC & CNA (converged network adapters) at 16/32 Gb FC and 40 Gb FCoE
- Storage will be:
  - Solid State Storage
    - SSS PCIe Cards, 1 ru trays of FLASH DIMMS
    - SSS in 2.5" drive formfactor following all current disk drive support models
  - 2.5" and 3.5" 10K RPM SAS (capacities up to 2 TB to 4 TB)
  - 2.5" and 3.5" SATA 2.0 Drives (capacities 1 TB to 8 TB)
  - SAS 3.0 Disk Arrays Front Ends with above drives
  - 16/32 Gb FC Disk Arrays with above drives
  - FDR/EDR IB Storage Heads with above drives

Data Center in 2016-2020

- Storage Access
  - SAS 4.0 HBAs, 8 and 16 port IOC/ROC by 2017
  - 32 Gb FC HBAs pluggable optics for single/dual port
  - Multi-function FC & CNA (converged network adapters) at 32 Gb FC and 40 Gb FCoE, 100 Gb FCoE Possiblities
- Storage will be:
  - Solid State Memory Controllers occupy System Board Memory Controller Locations – FLASH Memory DIMM on the MB
  - Solid State Storage
    - SSS PCIe Cards, 1 ru trays of FLASH DIMMS
    - SSS in 2.5" and 3.5" drive formfactor following all current disk drive support models
  - 2.5" and 3.5" 10K/15K RPM SAS (capacities up to 2 to 4 TB)
  - 2.5" and 3.5" SATA 2.0 Drives (capacities 500 GB to 4 TB)
  - SAS 4.0 Disk Arrays Front Ends with above drives
  - 16/32 Gb FC Disk Arrays with above drives
  - FDR/EDR IB Storage Heads with above drives

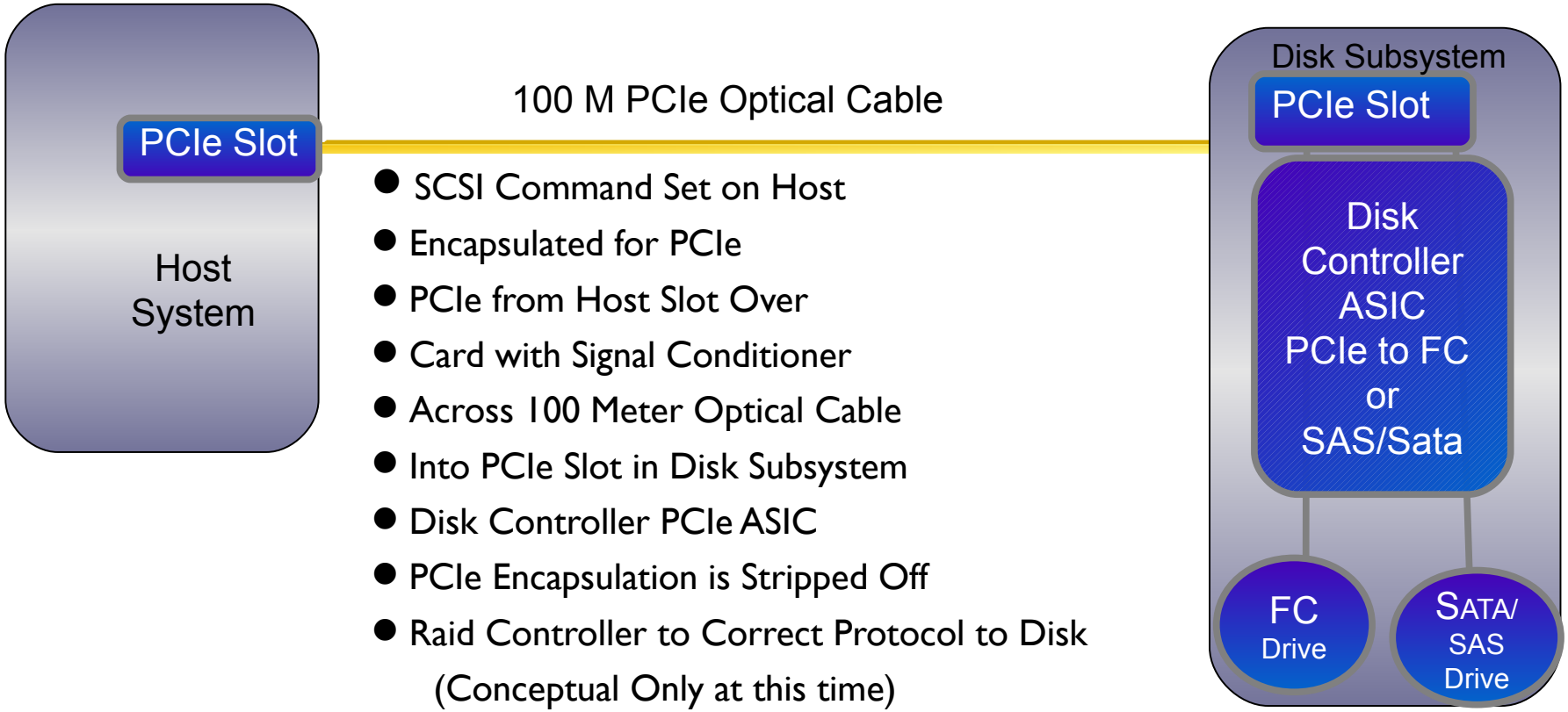
Data Center in 2025

- CPUs Dedicated to Different Activities
  - IO CPU contains All IO Interfaces
  - Memory CPU contians DRAM and Flash Memory Controllers
  - High Speed Socket to Socket Interfaces at 100 GT/s
  - Graphics CPUs are external to the Chassis Connected via above 100 GT/s Interface
- IO & Storage Interfaces
  - Predominate IO Interface is 100 / 400 Gb Ethernet
  - HDR IB in limited deployment
  - All Storage Controllers are attached via Ethernet or IB
  - Backside Storage remains SAS/FC

Data Center in 2025

- Storage Access
  - SAS 4.0 HBAs, 8 and 16 port IOC/ROC by 2017
  - 32 Gb FC HBAs pluggable optics for single/dual port
  - Multi-function FC & CNA (converged network adapters) at 32 Gb FC and 40 Gb FCoE, 100 Gb FCoE Possiblities
- Storage will be:
  - Solid State Memory Controllers occupy System Board Memory Controller Locations – FLASH Memory DIMM on the MB
  - Solid State Storage
    - SSS PCIe Cards, 1 ru trays of FLASH DIMMS
    - SSS in 2.5" and 3.5" drive formfactor following all current disk drive support models
  - 2.5" and 3.5" 10K/15K RPM SAS (capacities up to 2 to 4 TB)
  - 2.5" and 3.5" SATA 2.0 Drives (capacities 500 GB to 4 TB)
  - SAS 4.0 Disk Arrays Front Ends with above drives
  - 16/32 Gb FC Disk Arrays with above drives
  - FDR/EDR IB Storage Heads with above drives

Future Storage Attach Model



- SCSI Command Set on Host
- Encapsulated for PCIe
- PCIe from Host Slot Over
- Card with Signal Conditioner
- Across 100 Meter Optical Cable
- Into PCIe Slot in Disk Subsystem
- Disk Controller PCIe ASIC
- PCIe Encapsulation is Stripped Off
- Raid Controller to Correct Protocol to Disk

(Conceptual Only at this time)

Glossary of Terms

PCI – Peripheral Component Interconnect. An open, versatile IO technology. Speeds range from 33 Mhz to 266 Mhz, with pay loads of 32 and 64 bit. Theoretical data transfer rates from 133 MB/ s to 2131 MB/ s.

PCI-SIG - Peripheral Component Interconnect Special Interest Group, organized in 1992 as a body of key industry players united in the goal of developing and promoting the PCI specification.

IB – InfiniBand, a specification defined by the InfiniBand Trade Association that describes a channel-based, switched fabric architecture.

Glossary of Terms

Root complex – the head of the connection from the PCI Express IO system to the CPU and memory.

HBA – Host Bus Adapter.

IOV – IO Virtualization

Single root complex IOV – Sharing an IO resource between multiple operating systems on a HW Domain

Multi root complex IOV – Sharing an IO resource between multiple operating systems on multiple HW Domains

VF – Virtual Function

PF – Physical Function