PCI Express and Its Interfaces to Flash Storage Architectures

Ron Emerick, Oracle Corporation

SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

  NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

Abstract

PCI Express and Its Interfaces to Flash Storage Architectures

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

Abstract

This session will provide the attendee with:

- Knowledge of PCI Express Architecture
  - System Root Complexes relationship to 'Slots'
- IO Virtualization connectivity possibilities to Storage
- Types of Flash Connectivity available
- Implications and Impacts of Flash and PCIe Storage Devices to Storage Connectivity
- What comes after this – who knows?

Agenda

Knowledge of PCI Express Architecture

PCI Express Roadmap

System Root Complexes and IO Virtualization.

Flash Storage Device Capabilities

Expected Industry Roll Out of latest IO Technologies and required Root Complex capabilities.

What Does the Data Center look like in the next 10 years?

PCI Express Introduction

- PCI Express Architecture is a high performance, IO interconnect for peripherals in computing/ communication platforms
- Evolved from PCI and PCI-X™ 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

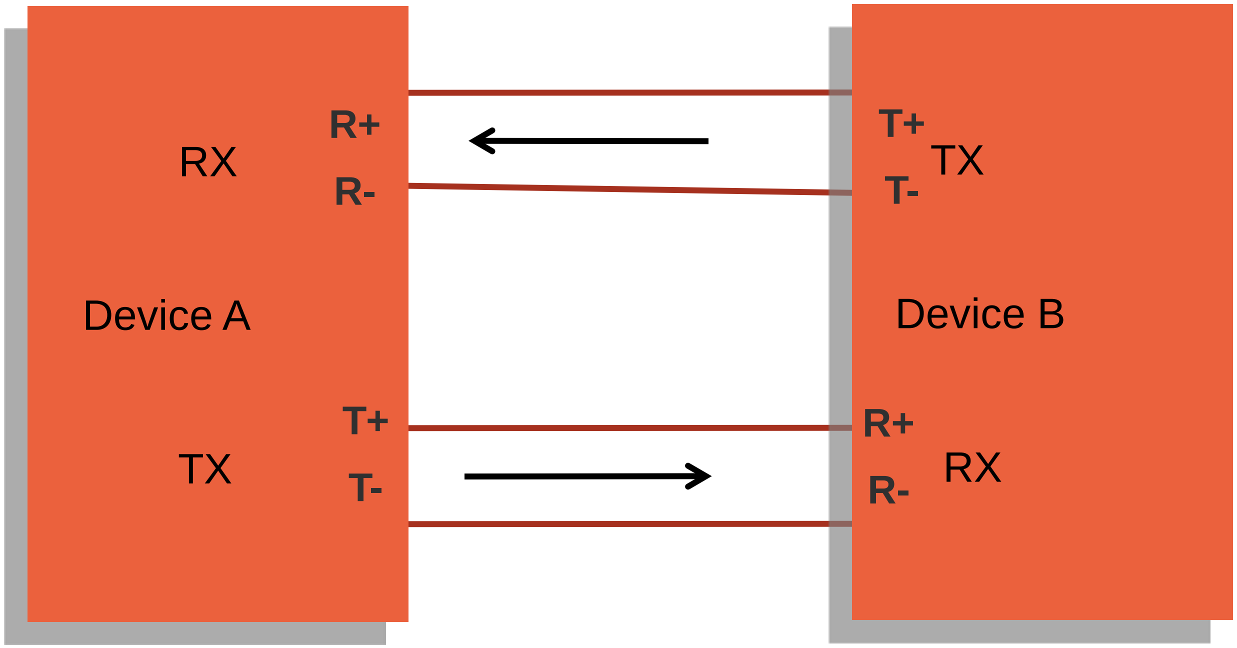
Transactions Types

Requests can also be translated to:

- 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.
- 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.

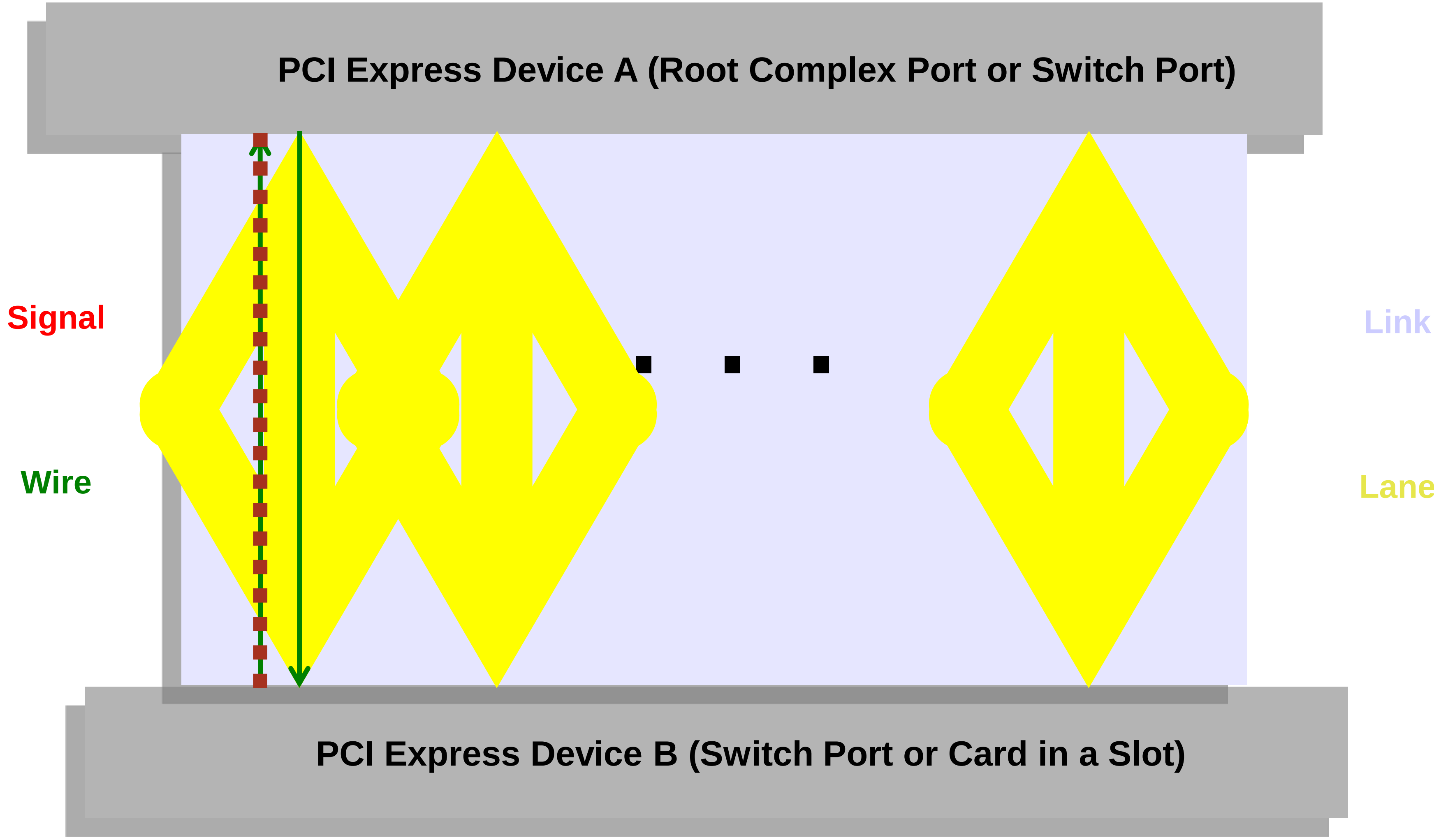
PCIe What's A Lane

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

PCI Express Terminology



Agenda

Knowledge of PCI Express Architecture

PCI Express Roadmap

System Root Complexes and IO Virtualization.

Flash Storage Device Capabilities

Expected Industry Roll Out of latest IO Technologies and required Root Complex capabilities.

What Does the Data Center look like in the next 10 years?

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

Aggregate bandwidth implies simultaneous traffic in both directions

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)
    - 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
    - Use 8/10 Bit Encoding => 500 MB/s/lane (FD)
    - 5 GT @ 1 bit/T * 8/10 encoding / 8 bit/byte = 500 MB/s FD
    - PCIe Overhead of 20% yields 400 MB/s/lane (FD)
    - Power for x16 increased to 225W

Current PCI Express Activities

- PCIe Gen 3.0
  - Approved in 2011
    - ›Frequency of 8.0 GT/s per Lane
    - ›Nearly Doubled the Theoretical BW to ~1000 MB/s/lane
    - ›Uses 128/130 bit encoding / scrambling
    - ›8 GT @ 1 bit/T * 128/130 encoding / 8 bit/byte = 984 MB/s FD
    - ›PCIe Overhead of 20% yields 1574 MB/s/lane (FD)
    - ›Power for x16 still TBD
- External expansion
  - Cable work group is active
- PCIe IO Virtualization (SR / MR IOV)
  - Architecture allows shared bandwidth

Current PCI Express Activities

- PCIe Gen 4.0 is being worked

  Base Specification is at 0.5, should be at 1.0 early 2016

  - ›Frequency of 16.0 GT/s per Lane
  - ›Nearly Doubled the Theoretical BW to ~2000 MB/s/lane
  - ›Uses 128/130 bit encoding / scrambling
  - ›16 GT @ 1 bit/T * 128/130 encoding / 8 bit/byte = 1968 MB/s FD
  - ›PCIe Overhead of 20% yields 1574MB/s/lane (FD)
  - ›Power for x16 increased to 300W
    (250 W via additional connector)

- Concerns
  - Trace Length – Skew & Jitter
  - Solution – retimer, redriver, …

Agenda

Knowledge of PCI Express Architecture

PCI Express Roadmap

System Root Complexes and IO Virtualization.

Flash Storage Device Capabilities

Expected Industry Roll Out of latest IO Technologies and required Root Complex capabilities.

What Does the Data Center look like in the next 10 years?

Sample PCI Express Topology

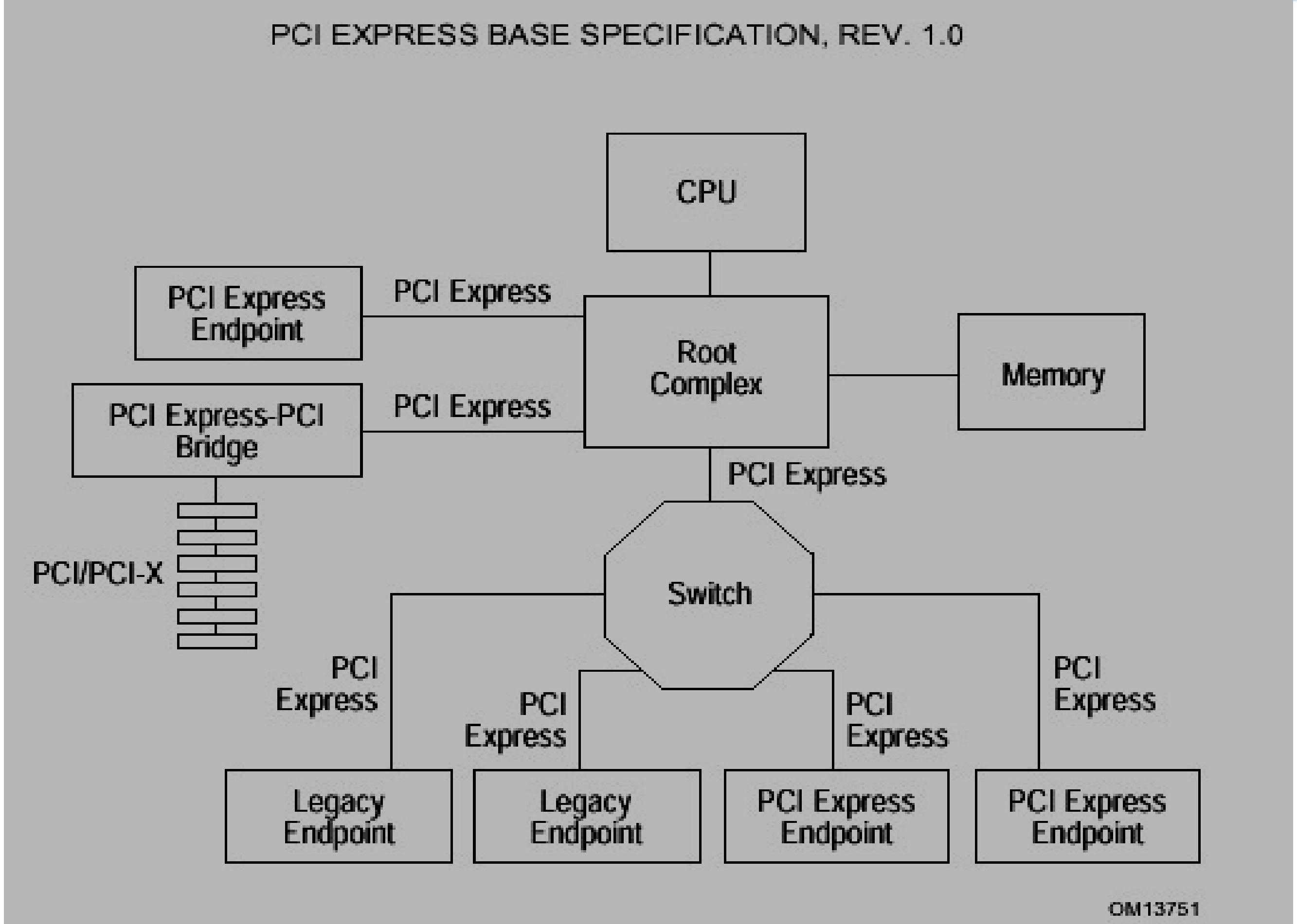


Figure 1-2: Example Topology

Important IOV Terms

IOV – IO Virtualization

- Single root complex IOV – Sharing an IO resource between multiple System Images on a single HW Domain
- Multi root complex IOV – Sharing an IO resource between multiple System Images on multiple HW Domains

SI – System Image (Operating System Point of View)

Multi-resource IO Device – An IO Device with resources that can be allocated to Individual SIs. (Quad port GbE, one port to each of four Sis.)

Shareable IO Device – A resource within an IO Device that can be shared by multiple SIs. (A port on a IO Device that can be shared.)

VF – Virtual Function

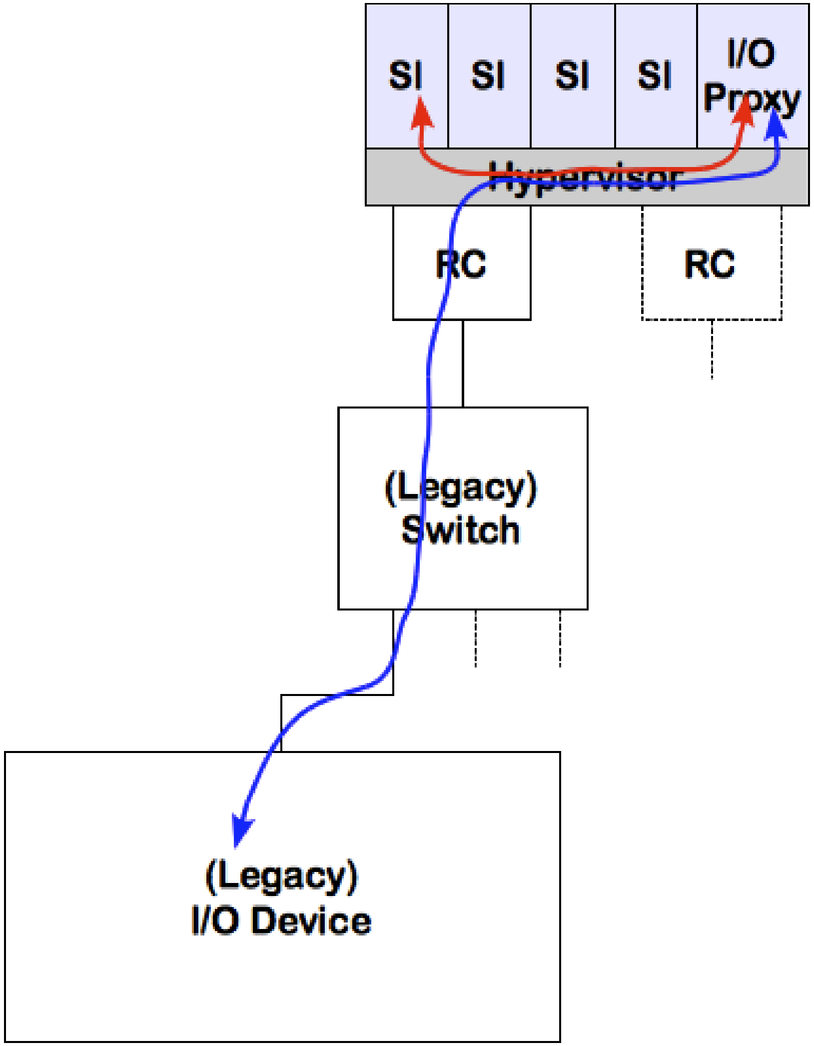
PF – Physical Function

IO Virtualization in Data Center

SNIA Education

- SR IOV works well in Multi-core/Multi-socket Systems
  - Best with multiple High Bandwidth PCIe Slots (Gen3 & Gen4)
- System Runs as a Single HW Domain
  - Running Multiple SW VMs
  - VMs Share the SR IOV IO Devices per the System Administrator
  - Allows High Bandwidth Devices to be shared among multiple VMs
- Much Better usage of IO Devices
  - Multiple VMs are sharing the IO Devices
  - Reduces the number of IO Devices of the same type that are needed
  - Allows for a larger variety of IO Devices that can be installed in a System

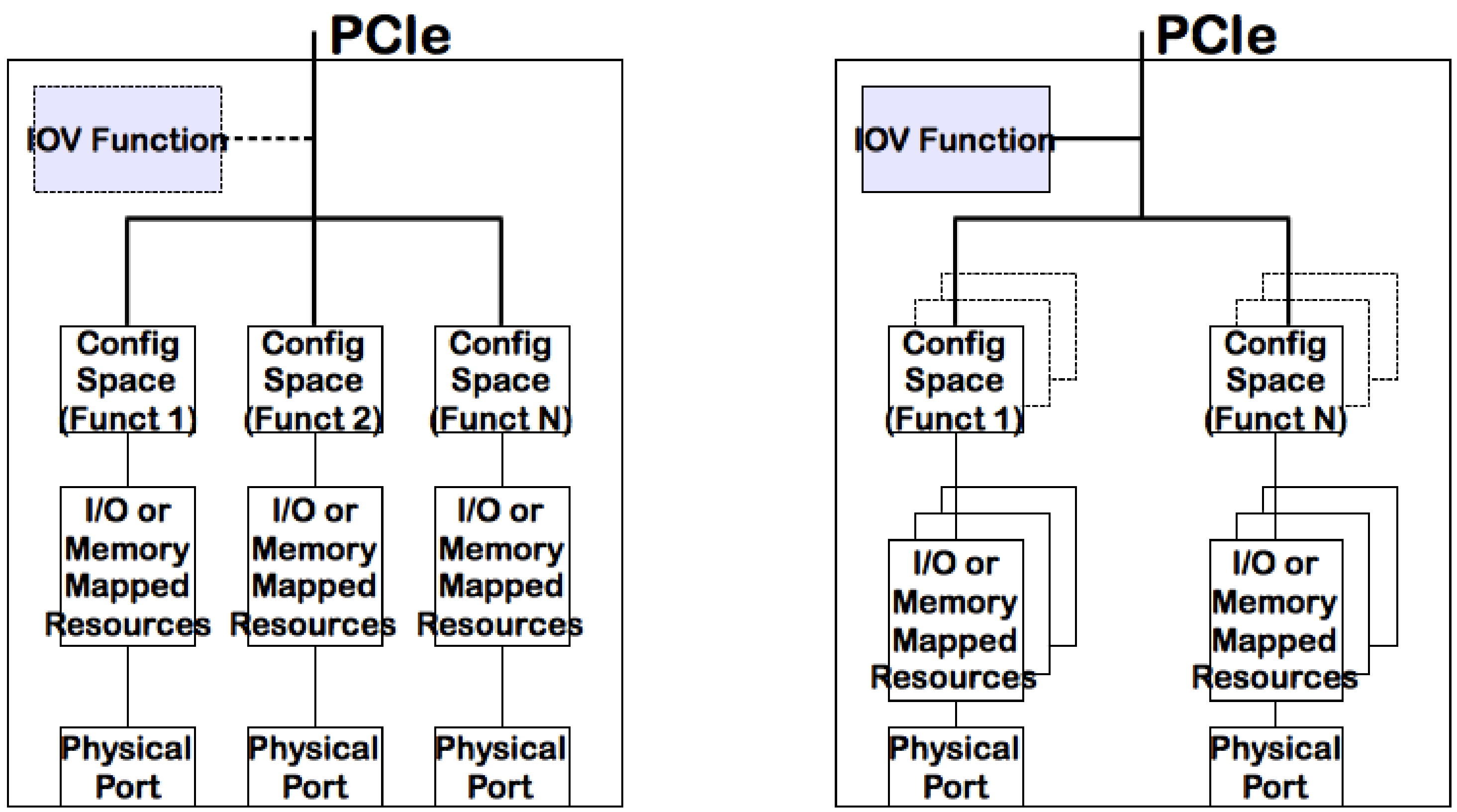
Multi-SI with IO Proxy



I/O Proxy Usage Model:

- Single Host/Multi SI
- Single or multi-function I/O devices
- I/O Proxy performs full bus probing and owns all I/O functions. I/O Proxy performs I/O on behalf of SI's

Multi-Resource IO Device



Multi-resource I/O Device

Shareable I/O Resources (within Multi-resource I/O Device)

Multi-Resource IO Device

Each System Image (SI) is allocated a full set of resources all the way to the Physical Port for the resources that they are using.

There are separate 'IOV Functions' to be used to control the physical attributes of the device. Such as chip level reset.

Single Root IOV

- Before Single Root IOV the Hypervisor was responsible for creating virtual IO adapters for a Virtual Machine

- This can greatly impact Performance
  - Especially Ethernet but also Storage (FC & SAS)

- Single Root IOV pushes much of the SW overhead into the IO adapter
  - Remove Hypervisor from IO Performance Path

- Leads to Improved Performance for Guest OS applications

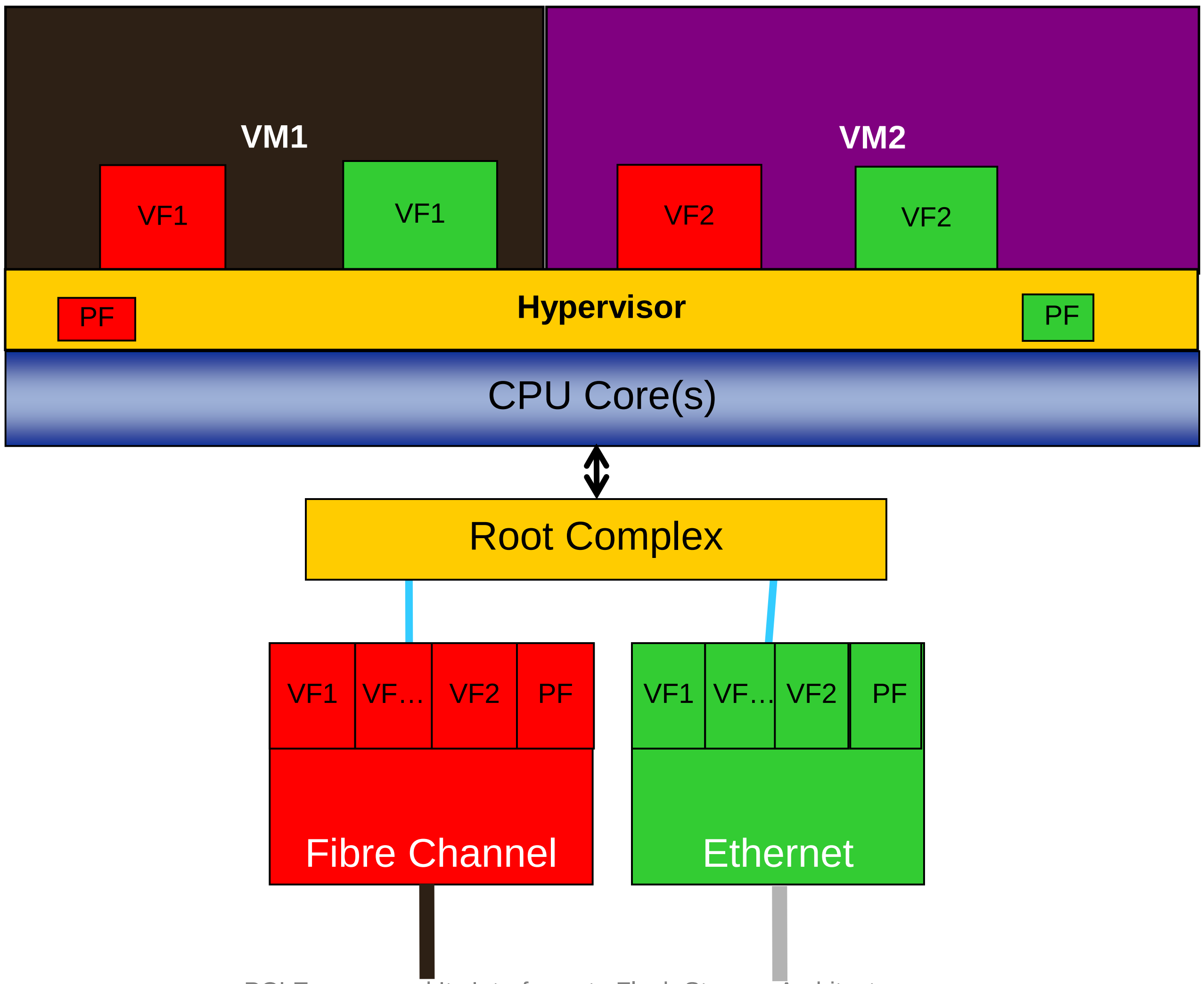
IO Provisioning

- Flexibility
  - Scaling of Compute and IO Resources
- Industry Standard Solution
  - Low Cost Low Profile Adapters
  - No Impact on Existing OS PCI Device Drivers
- Investment Protection
  - Independent CPU and IO Resource Upgrade Paths
  - Ability to move IO Resources to Next Generation Platforms (HW vendor support required)

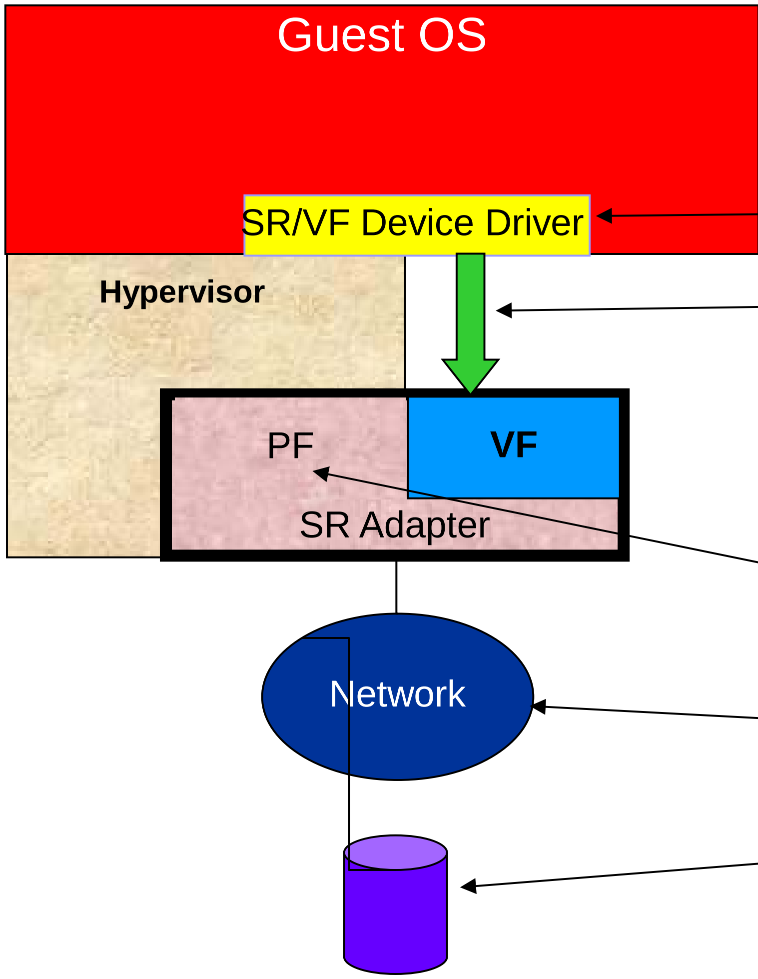
Physical and Virtual Functions

- IO Devices have at least one Physical Function
  - In control domain
  - Multiple Virtual Functions (up to 256)
  - Assigned to Virtual Domains via Control Domains
  - Hardware perspective – device is shared by the Virtual Domains
- Virtual Functions in Virtual Domains
  - Behave like dedicated device functions
  - OS perspective – they own the device

PCI-SIG Single Root



IO via SR Virtualization



SR Adapter Specific Driver

I/Os go directly to adapter via VF

Hypervisor configures VF via PF

PF is visible to Host

Targets are seen as Targets to Guest OS Device Driver

How does SR IO Virtualization Work

- IO Devices have Physical Functions each with Virtual Functions
- Virtual Functions are Dedicated (mapped) to Virtual Machines
  - Control Domain (VM) maps the Virtual Functions to Virtual Machines
  - Virtual Machine uses the Virtual Function as though it is the Hardware device
  - Virtual Machine issues the IO to the Virtual Function
  - Physical Function Understands the Virtual Function Map
  - Physical Function performs the IO on behalf of the Virtual Function
  - Physical Function returns the IO response to the Correct Virtual Function
  - Virtual Machine has just completed an IO
- The Above Scenario is Successfully completed by multiple Virtual Machines to the Same Physical Device via Mapped Virtual Functions.

SR IOV Devices

- More Practical

  All things network

  - 10 GbE, 40 GbE, FCoE, iSCSI, GbE Devices
  - IB devices
  - Ability to share these resources across without Hypervisor (Virtual Machine involvement)
- Less Practical
  - Fibre Channel, SAS and PCIe SSS Cards
  - Already have the ability to map LUNs to Virtual Machines

SR IOV - What and When

SNIA Education

- SR IOV Capable PCIe Endpoints
  - Most Ethernet Controllers are SR IOV capable now or will be soon
- SR IOV Capable Systems
  - Several vendors have SR IOV capable Operating Systems available.
  - PCIe Gen3 Root Complexes are capable of supporting SR IOV

Agenda

Knowledge of PCI Express Architecture

PCI Express Roadmap

System Root Complexes and IO Virtualization.

Flash Storage Device Capabilities

Expected Industry Roll Out of latest IO Technologies and required Root Complex capabilities.

What Does the Data Center look like in the next 10 years?

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
  - PCIe - NVMe interface to NVMe Flash card

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

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 X4
  - Interface is 32Gb/s (4000 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

Issues with Accessing Flash

- How do you share the PCIe device with multiple Guests
- How do Flash devices get shared
- External storage gets shared via a network
- Earlier I mentioned 'SR IOV' PCIe devices
  - Allows multiple Guest domains to access the same PCIe device, thus the same Flash device

Agenda

Knowledge of PCI Express Architecture

PCI Express Roadmap

System Root Complexes and IO Virtualization.

Flash Storage Device Capabilities

Expected Industry Roll Out of latest IO Technologies and required Root Complex capabilities.

What Does the Data Center look like in the next 10 years?

Flash Memory Solutions

- SAS Based SSDs
  - SAS 2.0 SSDs shipping now
- SAS Flash Memory Cards
  - SAS 2.0 cards have been shipping for over 2 years
  - SAS 3.0 cards are shipping
- NVMe PCIe SFFs
  - Currently shipping
- NVMe PCIe Cards
  - Currently shipping
- M.2 Sata and M.2 PCIe
  - Currently shipping or will ship soon

Agenda

Knowledge of PCI Express Architecture

PCI Express Roadmap

System Root Complexes and IO Virtualization.

Storage Device Capabilities

Expected Industry Roll Out of latest IO Technologies and required Root Complex capabilities.

What Does the Data Center look like in the next 10 years?

Data Center in 2016-2020

- Root complexes are PCIe 4.0
  - Integrated into CPU
  - Multiple Gen3 X16 or Gen4 x8 from each socket
  - 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 2018
  - Quad 10 Gbase-T and Quad MMF (single ASIC)
  - Dual/Quad Legacy GbE Copper and Optical
  - Dual ported EDR InfiniBand for cluster, some storage
- Graphics
  - X16 Single/Dual ported Graphics cards @ 450 W (when needed)

  Most Graphics solutions will be external to the server/workstation due to Power and Cooling Requirements

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 5.0 HBAs, 8 and 16 port IOC/ROC by 2021
  - 64/128 Gb FC HBAs pluggable optics for single/dual port
  - Multi-port 40/100/400 Gb FCoE
- 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 4 to 8 TB)
  - 2.5" and 3.5" SATA 2.0 Drives (capacities 2 to 8TB)
  - SAS 5.0 Disk Arrays Front Ends with above drives
  - 64/128Gb FC Disk Arrays with above drives
  - EDR/HDR IB Storage Heads with above drives

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.

IOV – IO Virtualization

Single root complex IOV – Sharing an IO resource between multiple System Images on a HW Domain

Multi root complex IOV – Sharing an IO resource between multiple System Images on multiple HW Domains

SI – System Image (Operating System Point of View)

VF – Virtual Function

PF – Physical Function

Attribution & Feedback

The SNIA Education Committee would like to thank the following individuals for their contributions to this Tutorial.

Authorship History

Name/Date of Original Author here:

Updates:

- **Ron Emerick / March 2012**
- **Ron Emerick / August 2012**
- **Ron Emerick//July 2014**

Additional Contributors

Joel White
David Kahn

Please send any questions or comments regarding this SNIA Tutorial to ***tracktutorials@snia.org***