Education

PCI Express IO Virtualization Overview

Ron Emerick, Oracle Corporation

Author: 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.

© 2012 Storage Networking Industry Association. All Rights Reserved.

Abstract

- PCI Express IO Virtualization Specifications working with System Virtualization allowing multiple operating systems running simultaneously within a single computer system to natively share PCI Express Devices. This session describes PCI Express, Single Root and Multi Root IO Virtualization. The potential implications to Storage Industry and Data Center Infrastructures will also be discussed.

© 2012 Storage Networking Industry Association. All Rights Reserved.

Abstract

This tutorial will provide the attendee with:

- Knowledge of PCI Express Architecture and Performance Capabilities, System Root Complexes and IO Virtualization.
- The ability for IO Virtualization to change the use of IO Options in systems.
- IO Virtualization connectivity possibilities in the Data Center (via PCI Express).

© 2012 Storage Networking Industry Association. All Rights Reserved.

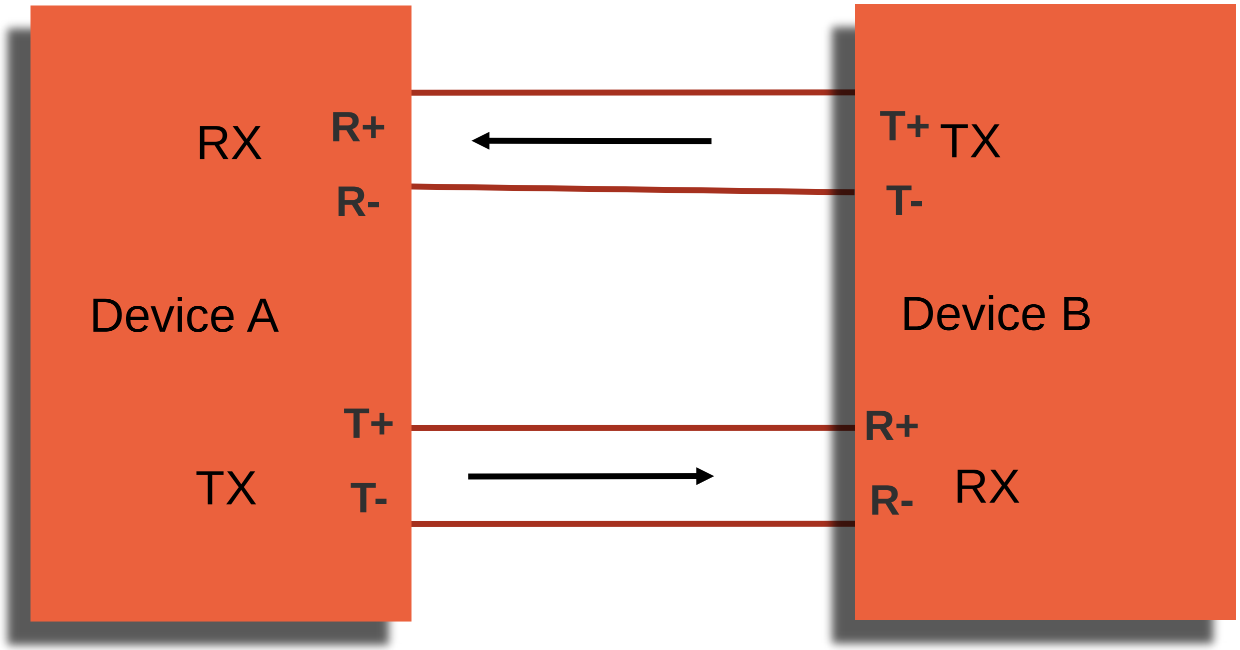
PCI Express Introduction

- 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

© 2012 Storage Networking Industry Association. All Rights Reserved.

PCIe What's A Lane

Point to Point Connection Between Two PCIe Devices

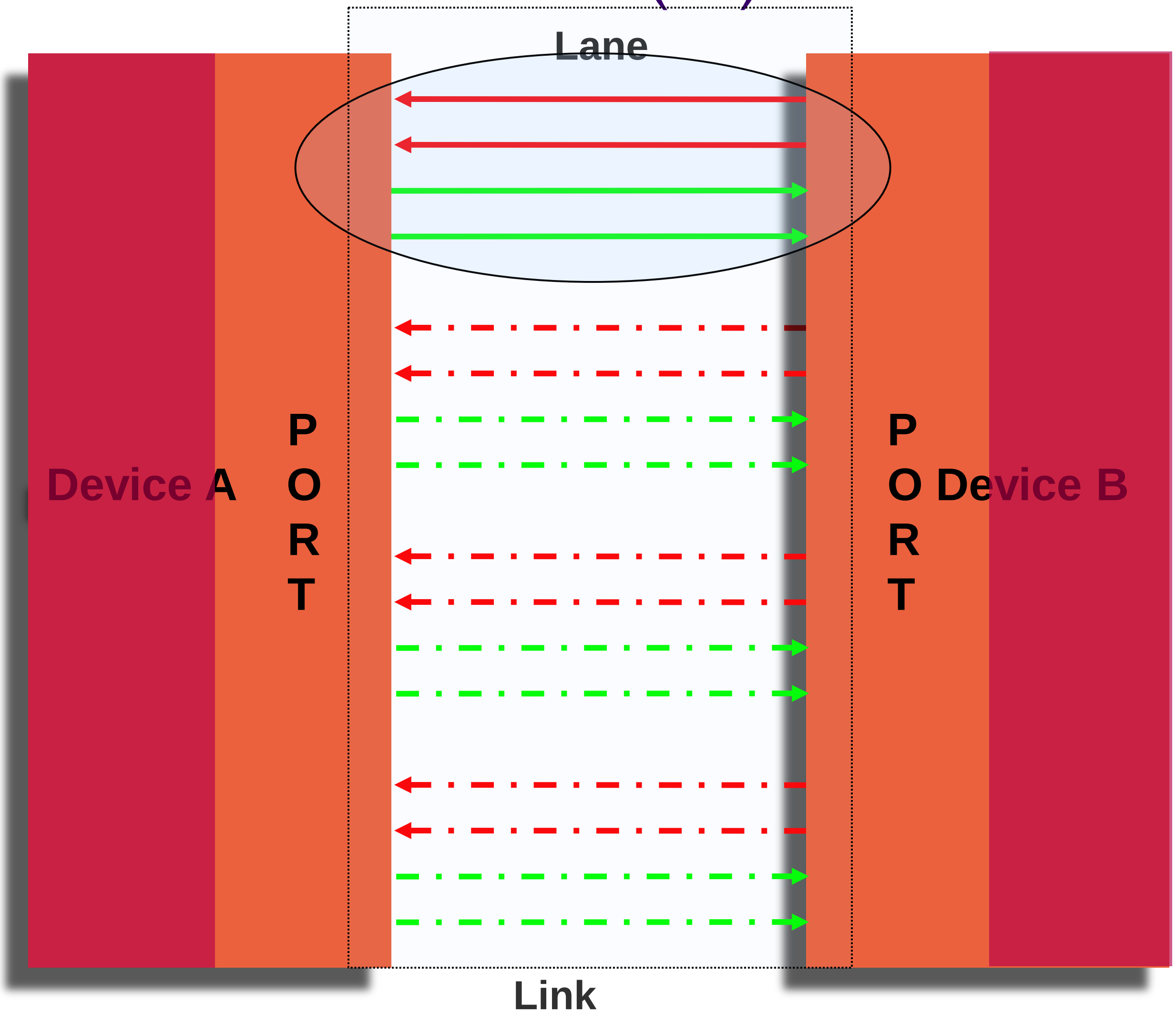


This Represents a Single Lane Using Two Pairs of Traces, TX of One to RX of the Other

© 2012 Storage Networking Industry Association. All Rights Reserved.

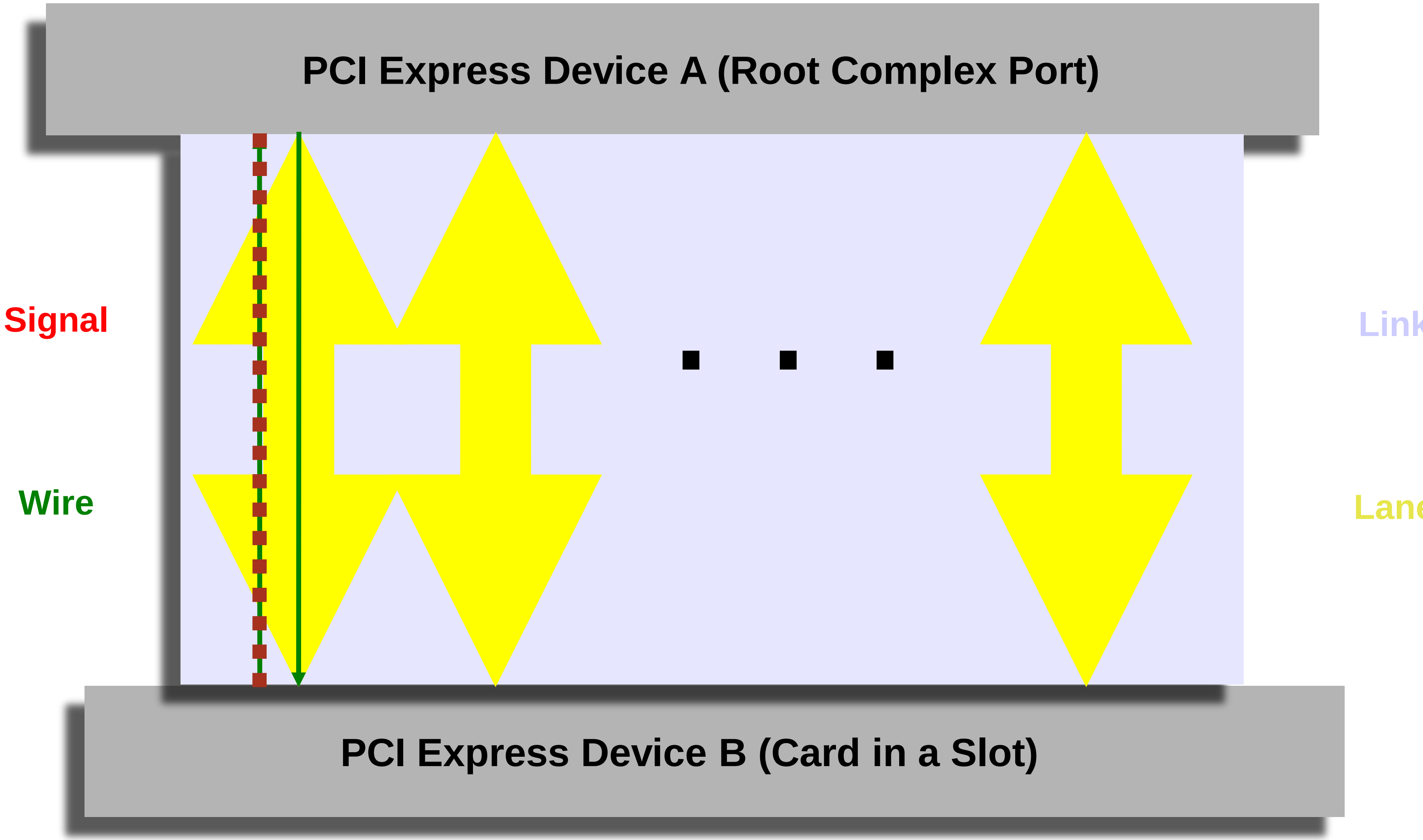
PCIe – Multiple Lanes

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



© 2012 Storage Networking Industry Association. All Rights Reserved.

PCI Express Terminology



© 2012 Storage Networking Industry Association. All Rights Reserved.

PCI Express Throughput

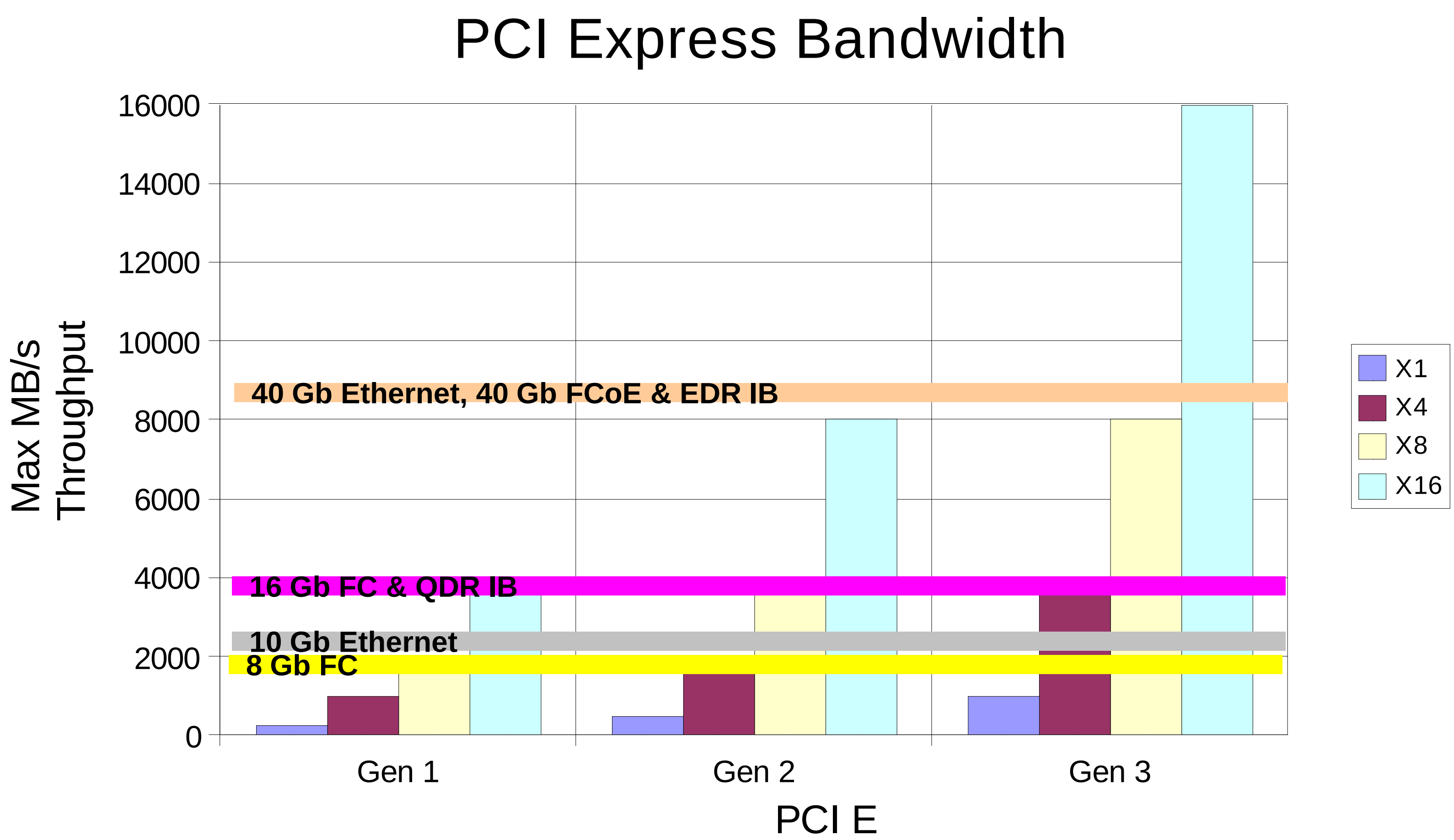
Link Width		X1	X2	X4	X8	X16	X32
Aggregate BW (Gbytes/s)	Gen1 (2004)	0.5	1	2	4	8	16
	Gen2 (2007)	1	N/A	4	8	16	32
	Gen3 (2010)	2	N/A	8	16	32	64

- 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
Peak bandwidth is higher than any bus available

© 2012 Storage Networking Industry Association. All Rights Reserved.

IO Bandwidth Needs



© 2012 Storage Networking Industry Association. All Rights Reserved.

PCI Express In Industry

Education SNIA

- 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

© 2012 Storage Networking Industry Association. All Rights Reserved.

Current PCI Express Activities

- PCIe Gen 3.0
  - Approved in 2011
    - Frequency of 8.0 GT/s per Lane
    - Uses 128/130 bit encoding / scrambling
    - Nearly Doubled the Theoretical BW to 1000 MB/s/lane
    - Support for Genesco features included

      Standard for Co-processors, Accelerators, Encryption, Visualization, Mathematical Modelling, Tunnelling
    - Power for x16 increased to 300W (250 W via additional connector)
- External expansion
  - Cable work group is active
- PCIe IO Virtualization (SR / MR IOV)
  - Architecture allows shared bandwidth

© 2012 Storage Networking Industry Association. All Rights Reserved.

Important IOV Terms

Education SNIA

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.)

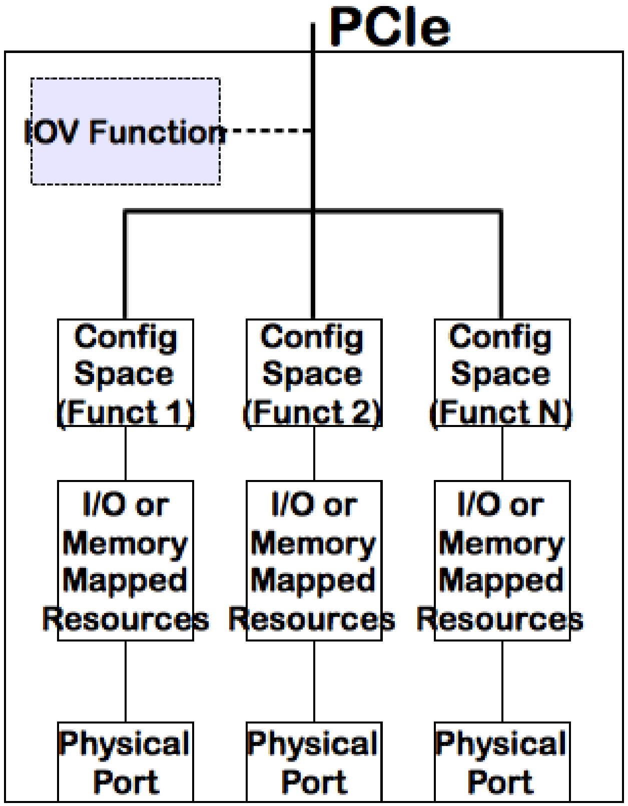
Shareable IO Device – A resource within an IO Device that can be shared by multiple SIs. (A port on a IO Device that can be shared.)

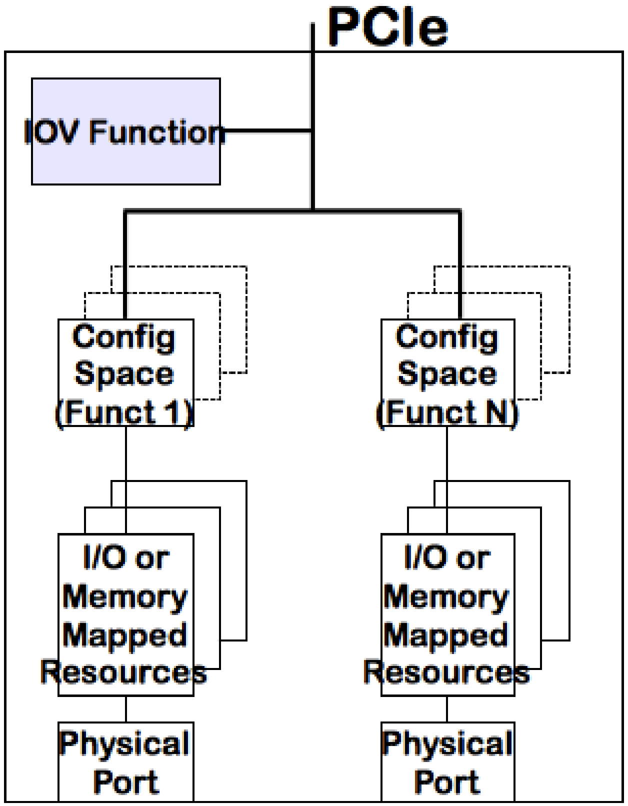
VF – Virtual Function

PF – Physical Function

© 2012 Storage Networking Industry Association. All Rights Reserved.

Multi-Resource IO Device





Multi-resource I/O Device

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

© 2012 Storage Networking Industry Association. All Rights Reserved.

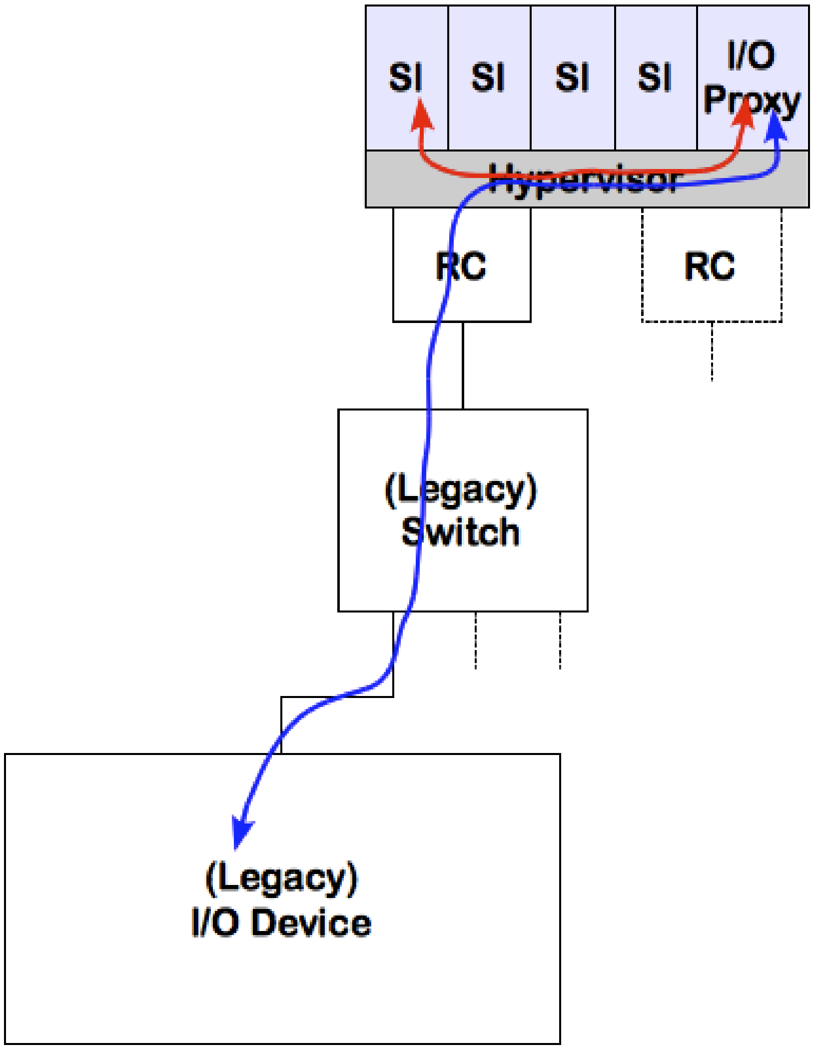
Multi-Resource IO Device

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.

© 2012 Storage Networking Industry Association. All Rights Reserved.

Multi-SI with IO Proxy

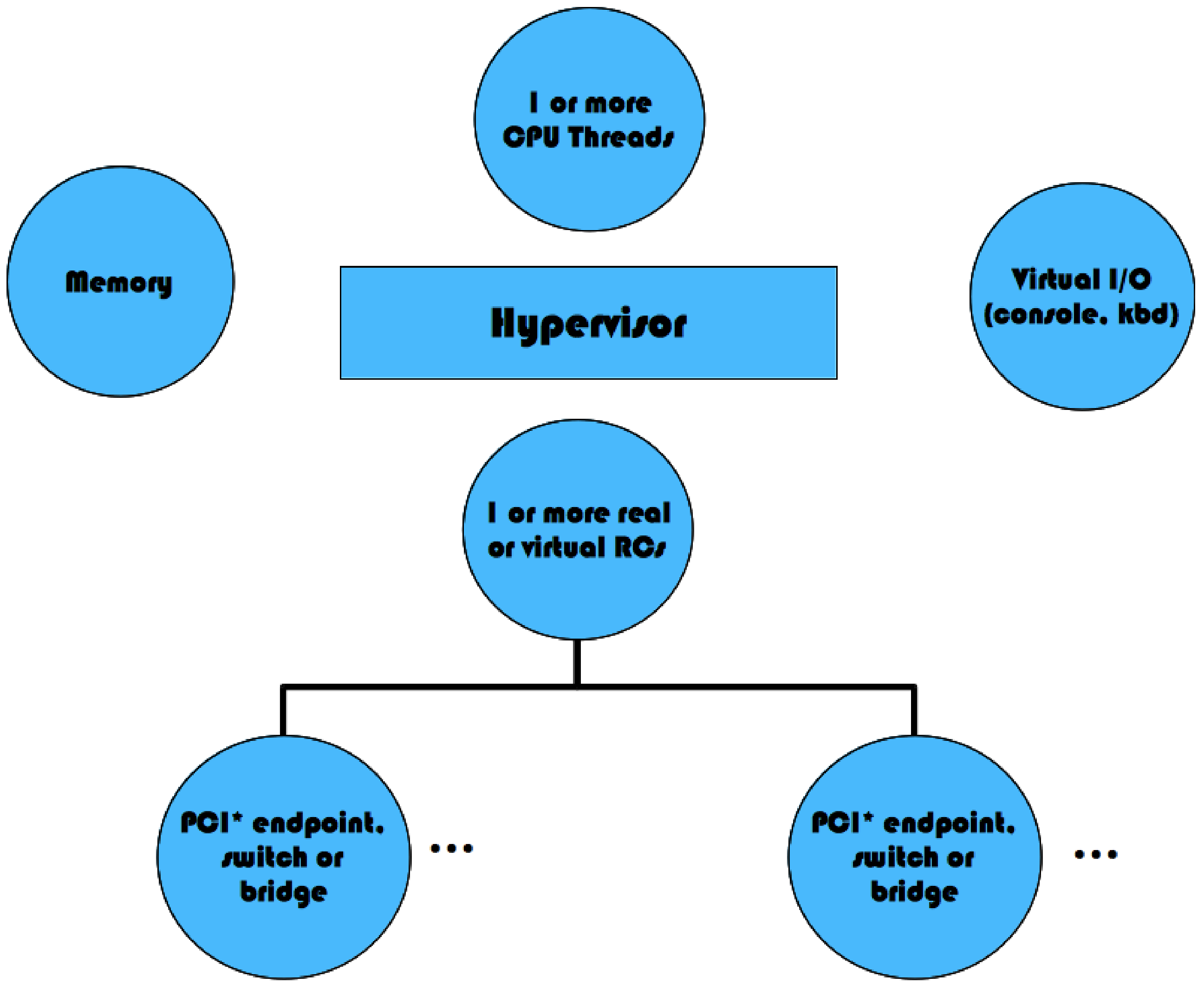


I/O Proxy Usage Model:

- 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

© 2012 Storage Networking Industry Association. All Rights Reserved.

System Image View of HW



© 2012 Storage Networking Industry Association. All Rights Reserved.

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

© 2012 Storage Networking Industry Association. All Rights Reserved.

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)

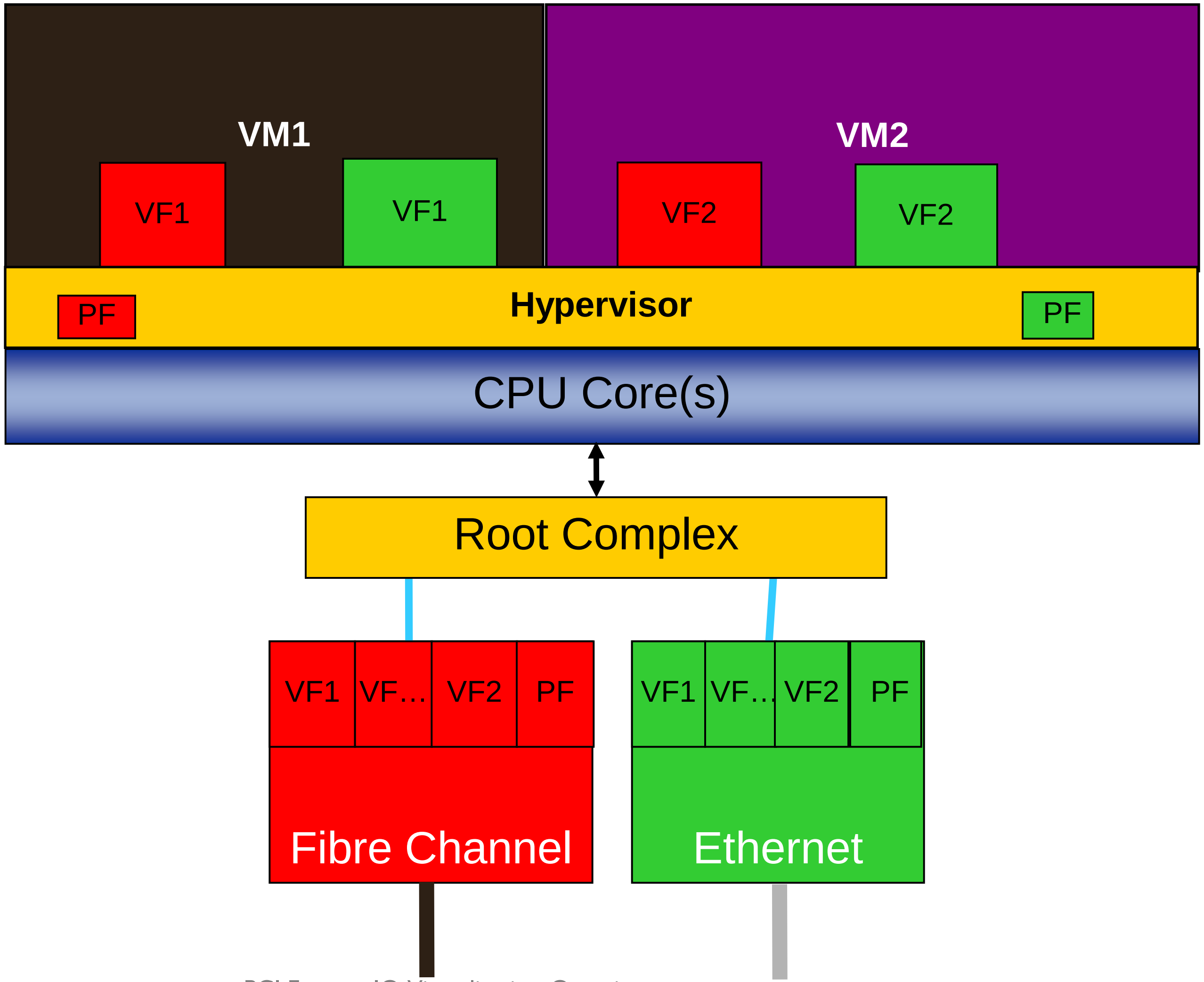
© 2012 Storage Networking Industry Association. All Rights Reserved.

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

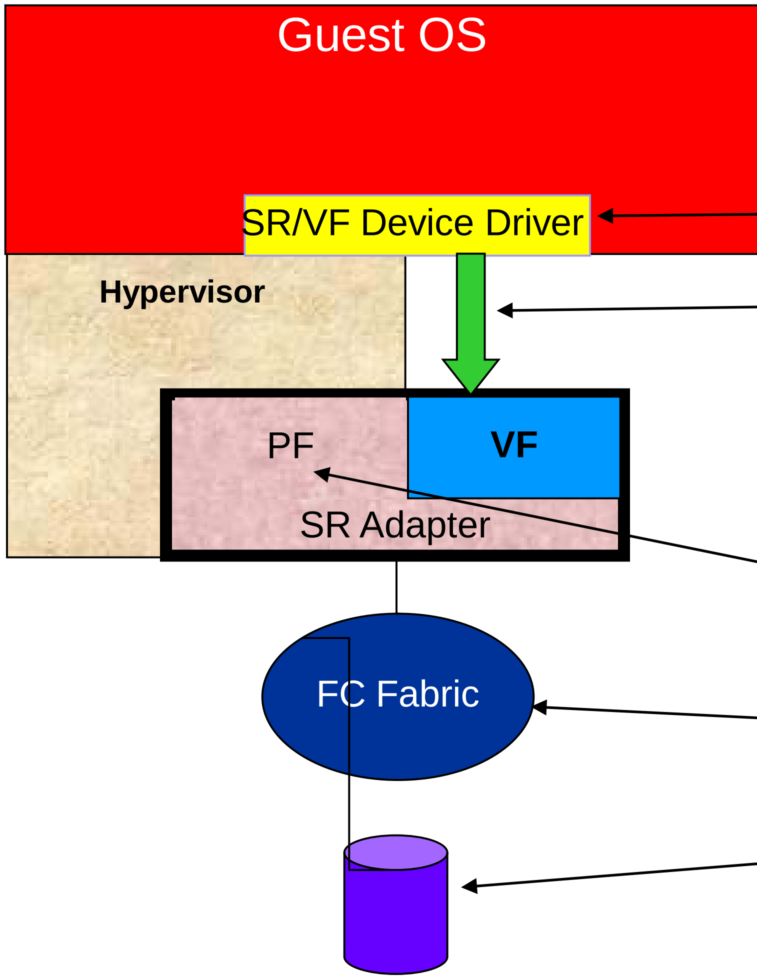
© 2012 Storage Networking Industry Association. All Rights Reserved.

PCI-SIG Single Root



© 2012 Storage Networking Industry Association. All Rights Reserved.

Fibre Channel & SR Virtualization



SR Adapter Specific Driver (fabric aware)

I/Os go directly to adapter via VF

Hypervisor configures VF via PF

Fabric is visible to Host

Fibre Channel LUNs are seen as LUNs to Guest OS Device Driver

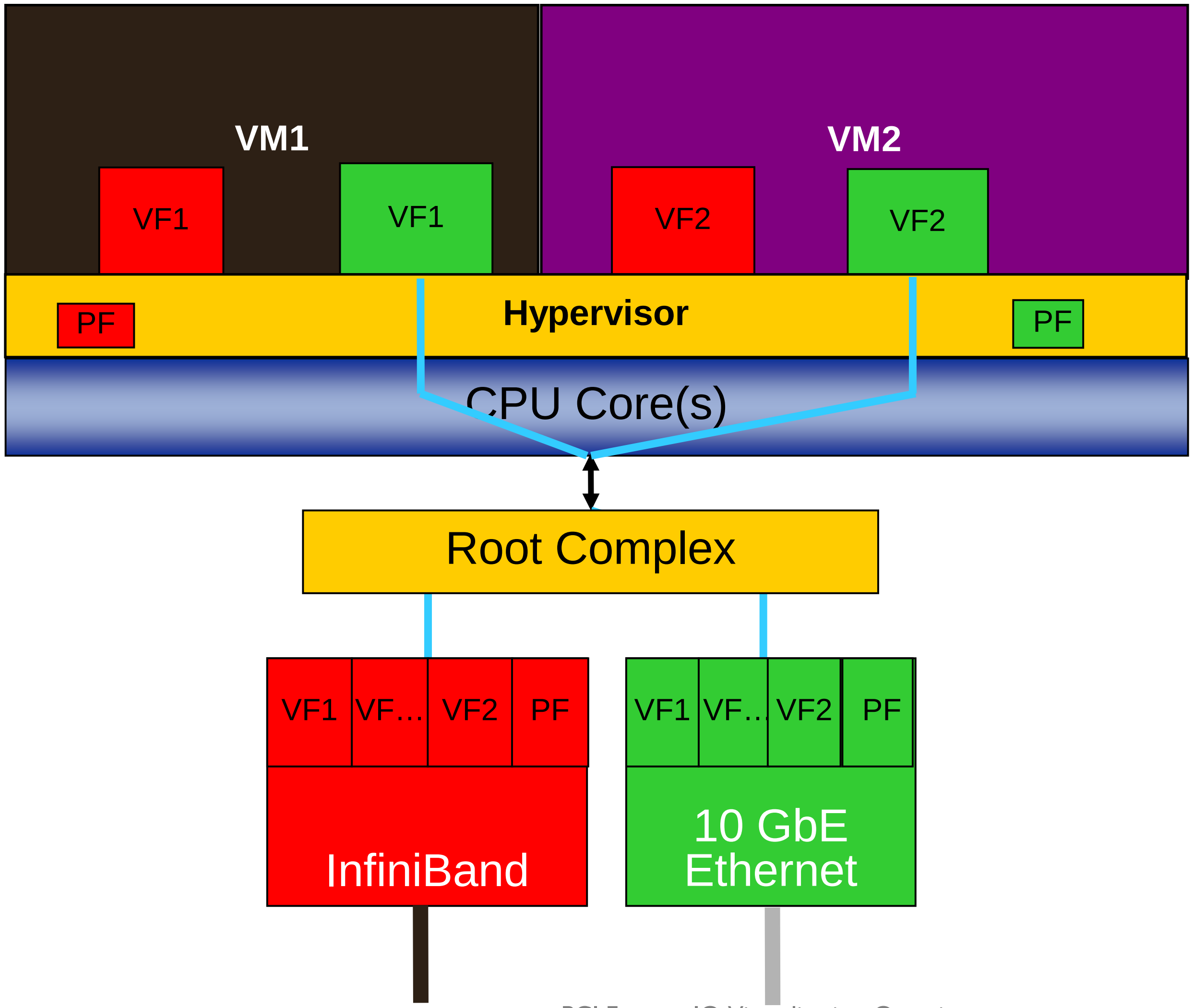
© 2012 Storage Networking Industry Association. All Rights Reserved.

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.

© 2012 Storage Networking Industry Association. All Rights Reserved.

How SR-IOV Works



- 10 GbE Controller Receives a Packet for VF2
- Controller determines it goes to VF2 in VM2
- Controller sends the packet to VM2
- VM2 delivers the packet to correct application via Ethernet SW stack

- Application on VM1 sends Ethernet transaction via VF1 to Ethernet Client over 10 GbE
- VM1 sends out through the Ethernet Controller

- Packet comes in for VF1
- Controller determines that the packet goes to VF1 in VM1
- Controller sends the packet to VM1
- VM1 delivers the packet to correct application via Ethernet SW stack

© 2012 Storage Networking Industry Association. All Rights Reserved.

IO Virtualization in Data Center

- 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

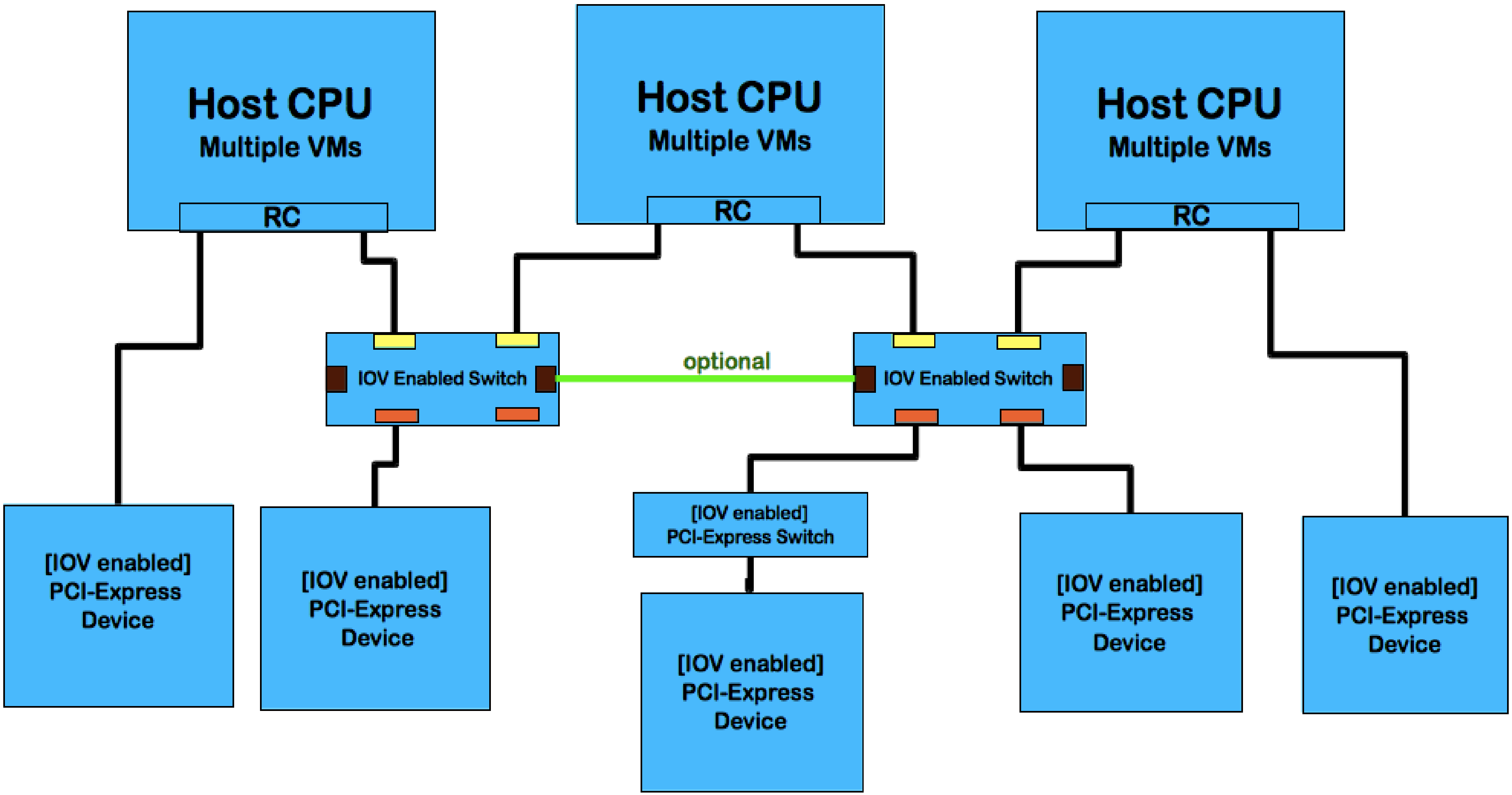
© 2012 Storage Networking Industry Association. All Rights Reserved.

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

© 2012 Storage Networking Industry Association. All Rights Reserved.

Multi-Root IO Virtualization



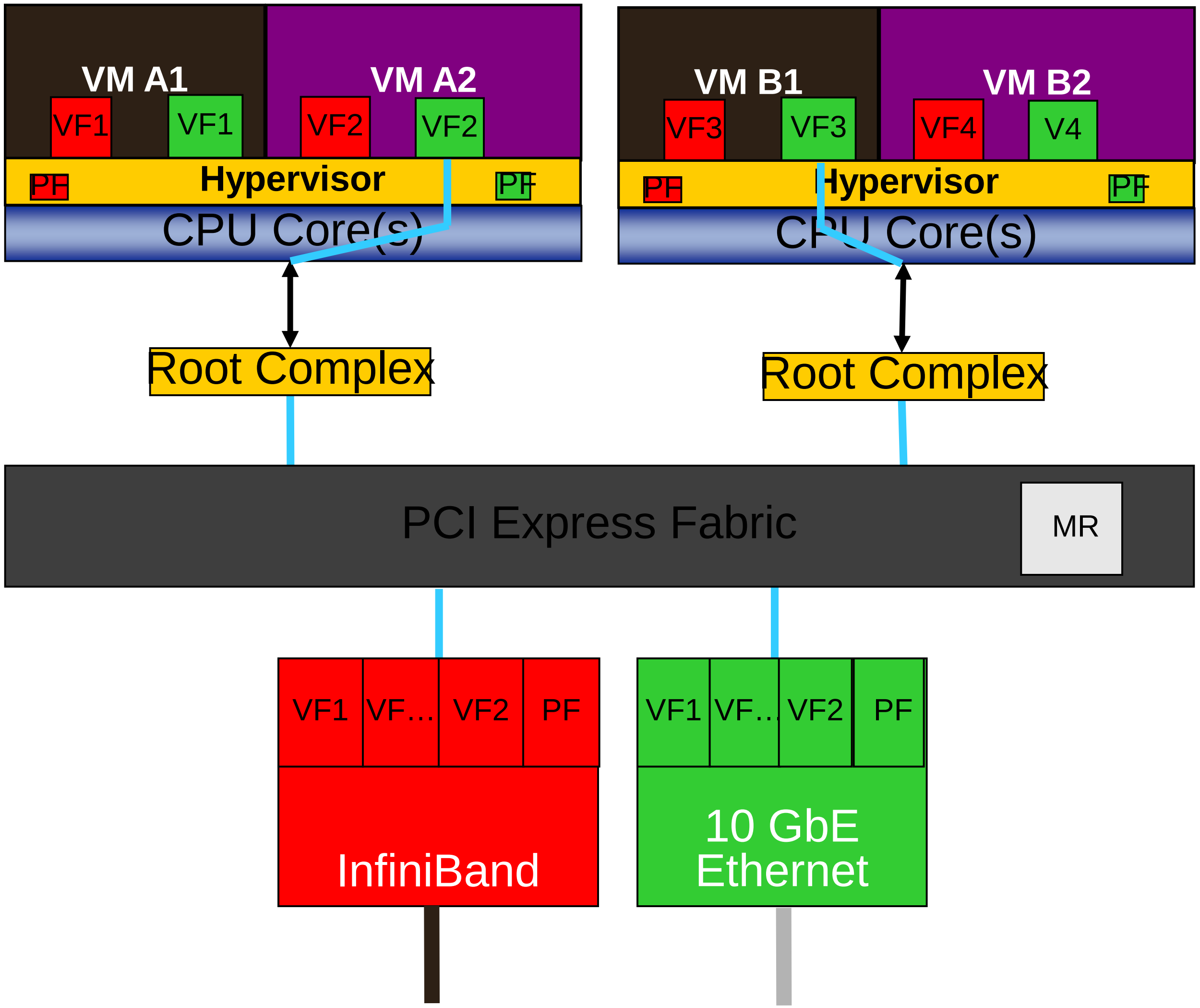
© 2012 Storage Networking Industry Association. All Rights Reserved.

How Does MR IO Virtualization Work

- Multiple Hardware Domains utilizing same IO Endpoints
- IO Devices have Physical Functions each with Virtual Functions
- Virtual Functions are Dedicated (mapped) to Virtual Machines within Hardware Domains
  - External MR Manager 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.

© 2012 Storage Networking Industry Association. All Rights Reserved.

How MR-IOV Works



- 10 GbE Controller Receives a Packet for VF2
- Controller determines it goes to VF2 in VM A2
- Controller sends the packet to VM A2
- VM A2 delivers the packet to correct application via Ethernet SW stack

- Application on VM1 sends Ethernet transaction via VF3 to Ethernet Client over 10 GbE
- VM B1 sends out through the Ethernet Controller

- Packet comes in for VF3 B1
- Controller determines that the packet goes to VF3 in VM B1
- Controller sends the packet to VM B1
- VM B1 delivers the packet to correct application via Ethernet SW stack

© 2012 Storage Networking Industry Association. All Rights Reserved.

MR IO Virtualization

- MR IOV Status
  - Best Fitted for Blade Environment
  - Some Top of Rack Implementations Available
  - Requires All External IO Devices to Support MR IOV
- Harder than SR IOV
  - Multiple HW Domains are Sharing IO Devices
  - IO Devices are external to Hosts
  - MR Manager Controls all MR Devices
  - Must present Device Present to MR Host even when offline

© 2012 Storage Networking Industry Association. All Rights Reserved.

Hands-On LAB

COMPUTERWORLD SNIA SNW

Unified Storage Infrastructure HOL and Unified Storage IP Solutions HOL

© 2012 Storage Networking Industry Association. All Rights Reserved.

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.

© 2012 Storage Networking Industry Association. All Rights Reserved.

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

© 2012 Storage Networking Industry Association. All Rights Reserved.

Q&A / Feedback

Many thanks to the following individuals for their contributions to this tutorial.

- SNIA Education Committee

Joel White
David Khan
Ron Emerick

Send any questions or comments on this presentation to SNIA: tracktutorials@snia.org

© 2012 Storage Networking Industry Association. All Rights Reserved.