



Education

PCI Express Impact on Storage Architectures and Future Data Centers

Ron Emerick, Sun Microsystems

- The material contained in this tutorial is copyrighted by the SNIA.
- 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.

PCI Express Impact on Storage Architectures and Future Data Centers

➤ PCI Express Gen2/Gen3, IO Virtualization, FCoE, D are here or coming soon. This session describes PCI Express, Single Root and Multi Root IOV and the implications on FCoE, SSD 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. This tutorial will provide the attendee with:

- Knowledge of PCIe Architecture, PCIe Roadmap, System Root Complexes, and IO Virtualization
- Expected Industry Roll Out of latest IO Technology and required Root Complex capabilities
- Implication and Impacts of FCoE, SSD and IO to Storage Connectivity
- IO Virtualization Connectivity possibilities in the Data Center (via PCIe)

➤ 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, GbE & 10 GbE
 - > PCIe IO Virtualization
- ◆ Review of PCI Express IO Virtualization
- ◆ Impact of PCI Express on Storage

Changing 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 PCI speed**
 - ◆ Moved from 32bit/ 33Mhz to 64bit/ 66Mhz, then
 - ◆ PCI-X introduced to reduce layout challenges
 - > PCI-X 133Mhz well established
 - > Problems at PCI-X 266Mhz with load and trace lengths
- **Parallel interfaces gradually being replaced**
 - ◆ ATA to SATA (PATA is going away)
 - ◆ SCSI to SAS
- **Move parallel PCI to serial PCI Express**

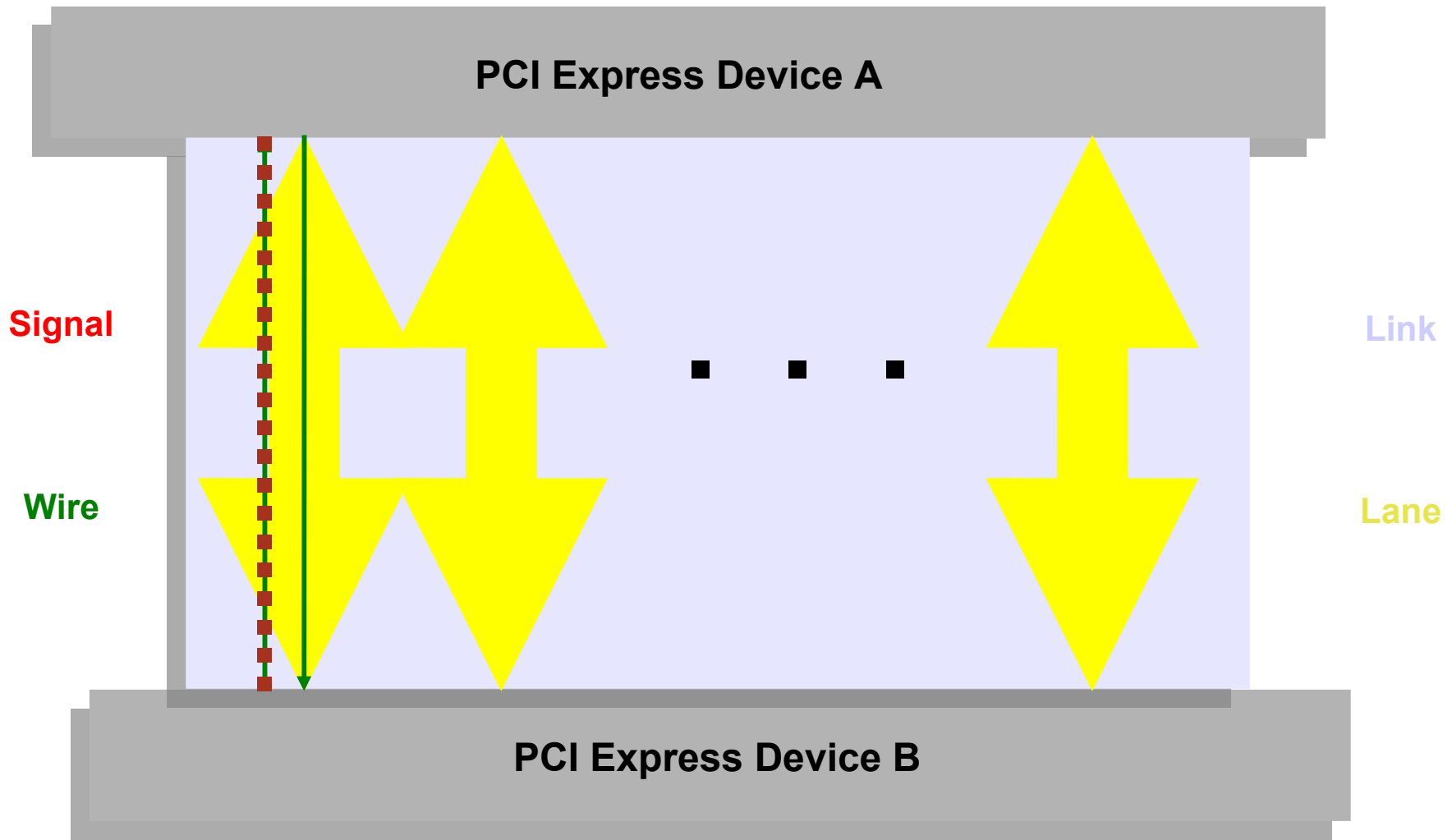
PCI Express Introduction

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

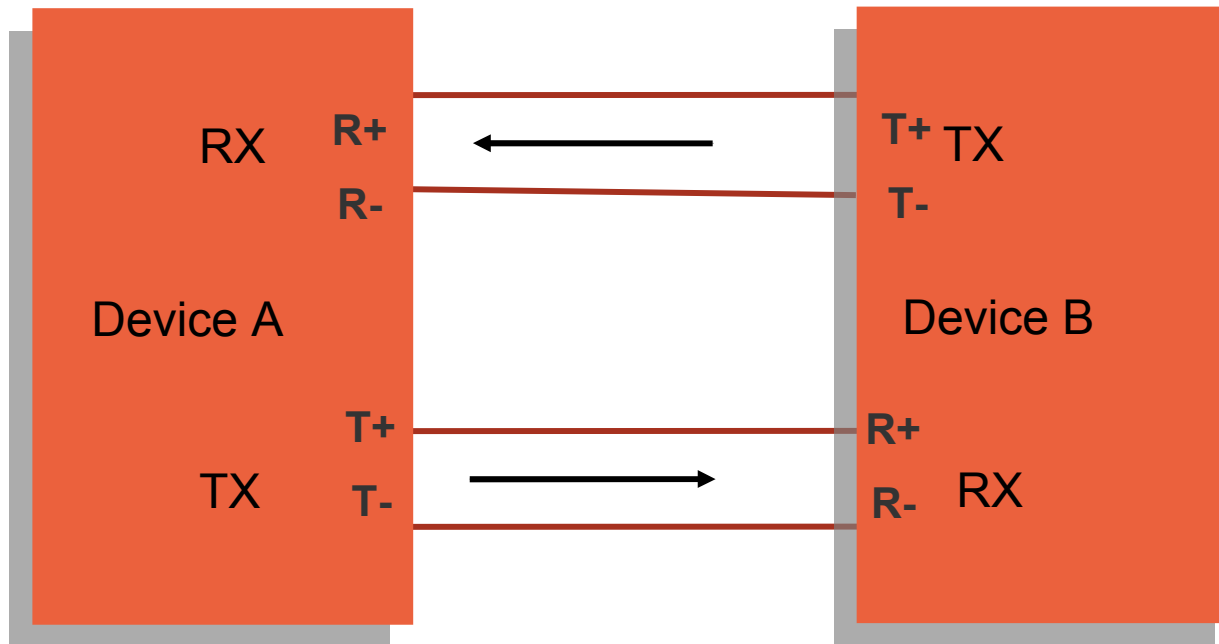
- ▶ **Uses PCI constructs**
 - ◆ Same Memory, IO and Configuration Model
 - ◆ Supports growth via speed increases
- ▶ **Uses PCI Usage and Load/ Store Model**
 - ◆ Protects software investment
- ▶ **Simple Serial, Point- to- Point Interconnect**
 - ◆ Simplifies layout and reduces costs
- ▶ **Chip- to- 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

PCI Express Terminology



PCIe What's A Lane

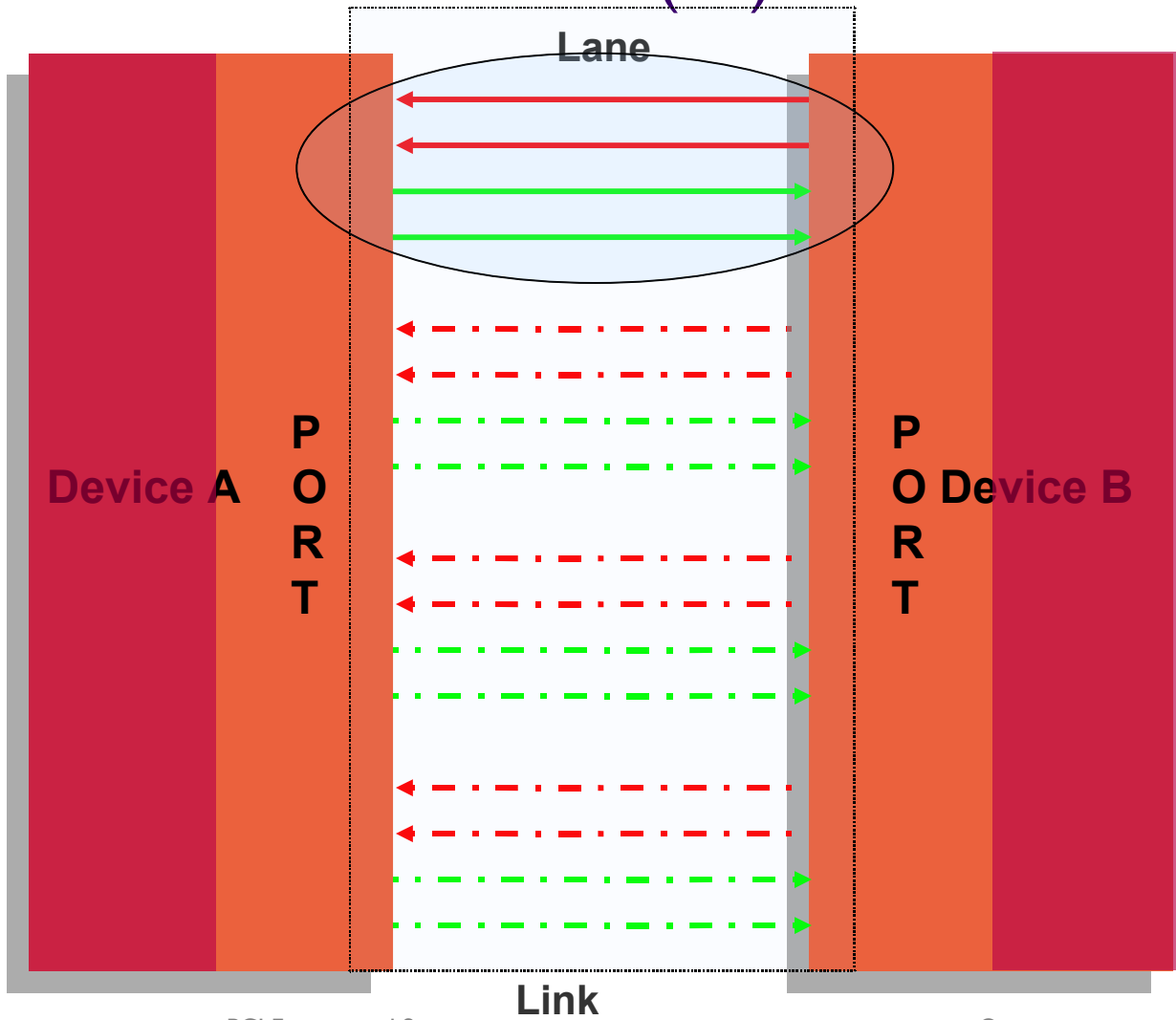
Point to Point Connection Between Two PCIe Devices



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



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.

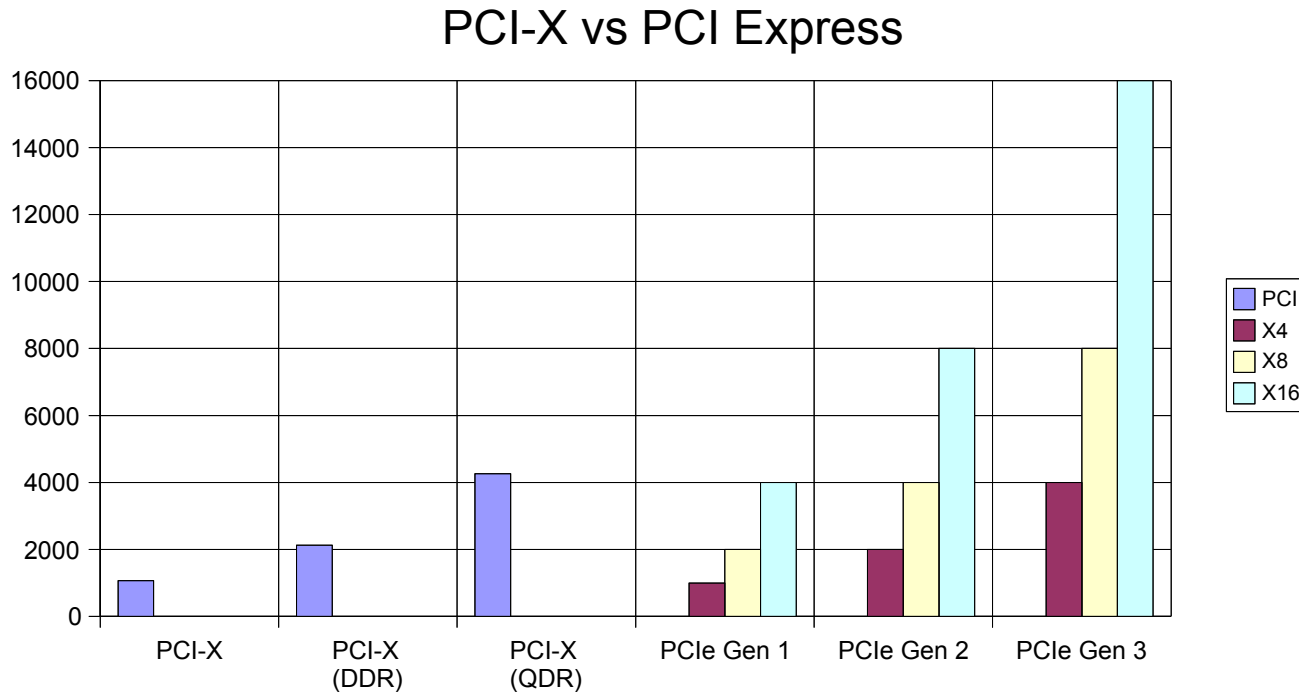
PCI Express Throughput

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

Aggregate bandwidth implies simultaneous traffic in both directions
Peak bandwidth is higher than any bus available

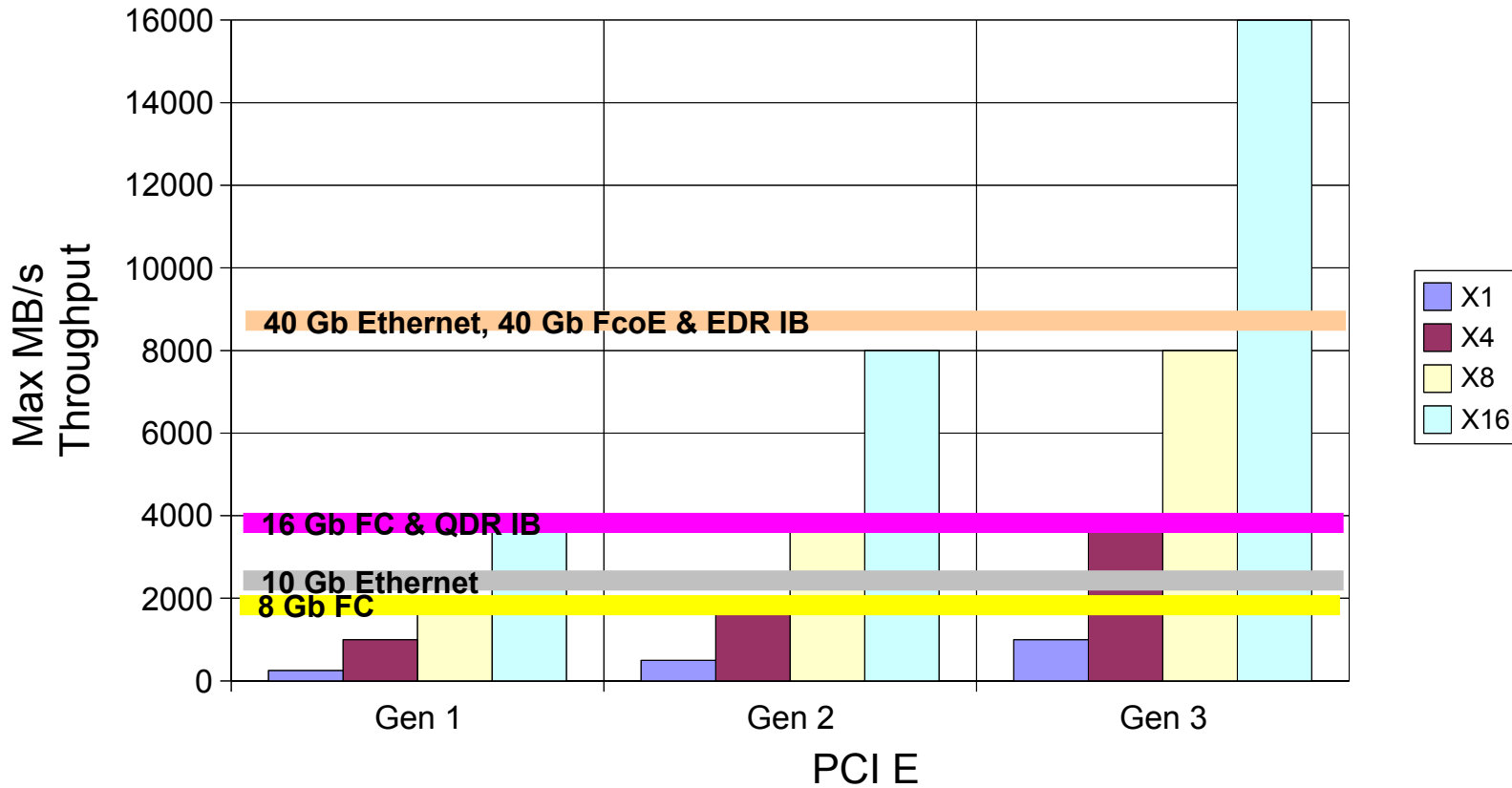
How does PCI-X compare to PCI Express?



➤ PCI-X QDR maxs out at 4263 MB/ s per leaf

➤ PCIe x16 Gen1 maxs out at 4000 MB/ s

PCI Express Bandwidth



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:

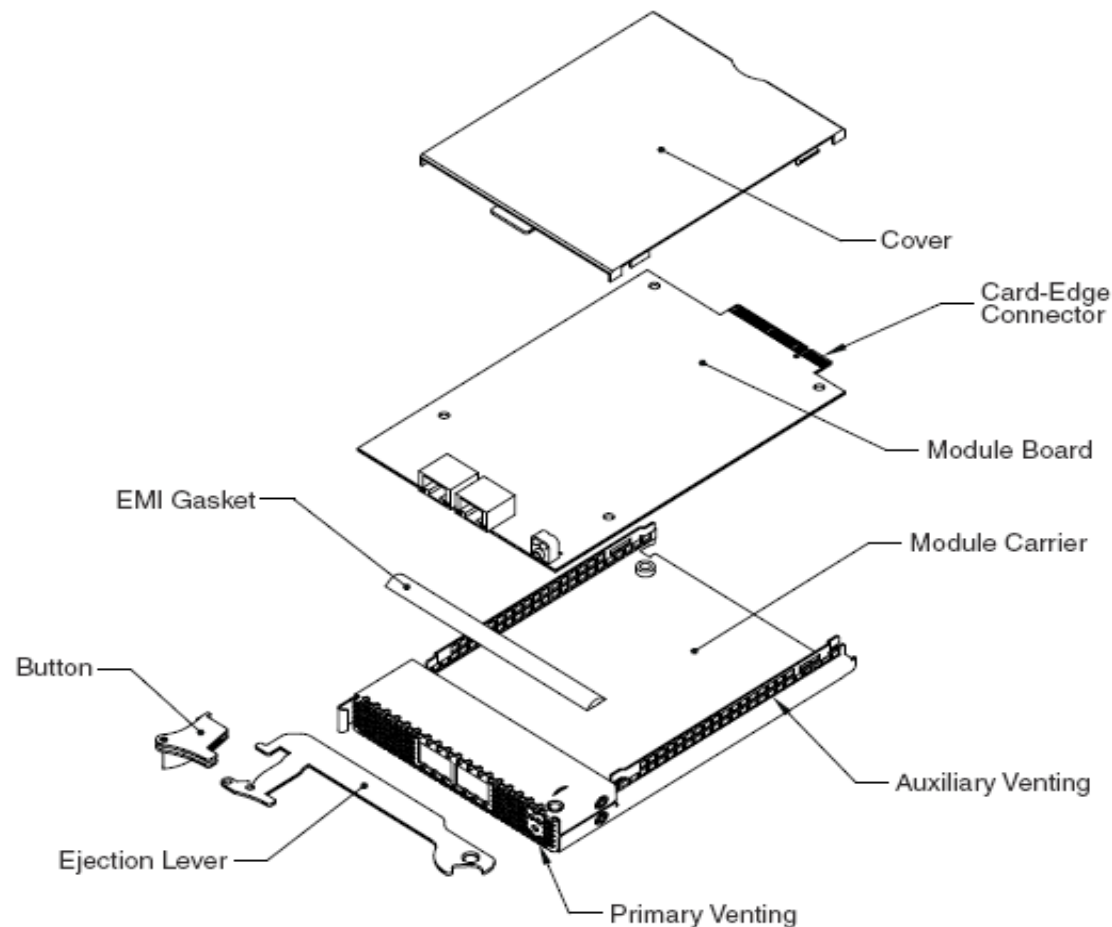
- ◆ Extend PCIe via signal conditioners and repeaters
- ◆ Optical & Copper cabling to remote chassis
- ◆ External Graphics solutions
- ◆ External IO Expansion

➤ Infrastructure is in Place

- ◆ PCIe Switches and Bridges
- ◆ Signal Conditioners

Express Module (EM)

- ◆ **Developed by the PCI-SIG (Initially Server IO Modules)**
 - ◆ Fully compatible with latest PCI Express specification
 - ◆ Designed to support future generations of PCI Express
- ◆ **Adds the necessary Hot Plug hardware and software**
- ◆ **Commodity pricing model using standard PCI Express silicon and ½ size card**
- ◆ **PCIe EM Products available today providing:**
 - ◆ SAS Internal/ external
 - ◆ 4 Gb FC External
 - ◆ GbE External
 - ◆ 10 GbE External
 - ◆ IB External



PCI Express In Industry

➤ PCIe Gen 1.1 Shipped in 2005

- ◆ Desktop Systems
 - > x16 High Performance Graphics
 - > x1 Low Cost Simple Connector
- ◆ Blades with Express Modules, Embedded Switches,...
- ◆ Servers with multiple x4 and x8 connectors

➤ PCIe Gen 2.0

- ◆ Desktop systems added Gen2 x16 slots in Q4 2007
- ◆ Servers shipping slots 2009
- ◆ Adoption is slower than Gen 1

➤ Cards Available

- ◆ x4, x8 cards - 10 GbE, Dual/Quad GbE, 4 Gb FC, SAS, IB
- ◆ x16 High Performance Graphics @ 150 W

Recent PCI Express Changes

- Power increase for Graphics Cards to 300 Watts
- Performance roadmap
 - ◆ Gen 2.0 Doubled to 5Gbits/ sec (DDR) with 8 / 10bit encoding
 - ◆ Gen 3.0 Doubles again to 8Gbits/sec (no 8 / 10bit encoding)
 - ◆ Gen 3.0 Base Spec at 0.5, 1.0 expected end of 2010
- External expansion
 - ◆ Copper cable and connector specified, Optical cable available
 - ◆ Cableing Spec 2.0 at .5 now
- Geneseo enhancements to PCIe 2.0
 - ◆ Standard for co-processors, accelerators
 - ◆ Encryption, visualization, mathematical modelling
- PCIe IO Virtualization (SR/ MR IOV)
 - ◆ Architecture allows shared bandwidth

- Processor speed increase slowing
 - ◆ Replaced by Multi-core Processors
 - Quad-core here, 8 and 16 core coming
 - ◆ Requires new root complex architectures
- Requires high speed interface for interconnect
 - Minimum 10Gb data rates
 - Must support backplane distances
 - Bladed systems
 - Single box clustered processors
 - Need backplane reach, cost effective interface to IO
- Interface speeds are increasing
 - Ethernet moving from GbE to 10G, FC from 4 Gb to 8 Gb, Infiniband is now QDR with EDR coming
 - Single applications struggle to fill these links
 - Requires applications to share these links

Drivers for New IO Architectures

- **High Availability Increasing in Importance**
 - Requires duplicated processors, IO modules and interconnect
 - Use of shared virtual IO simplifies and reduces costs and power
 - 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
 - PCI Express Modules with IOV enable this
- **Growth in backplane connected blades and clusters**
 - Blade centres from multiple vendors
 - Storage and server clusters
 - Storage Bridge Bay hot plug processor module
 - PCI Express IOV allows commodity I/O to be used
- **Options**
 - Use an existing IO interface like 10GbE/ Infiniband
 - Enhance PCI Express
 - Find a New IO interface

New IO Interfaces

- High Speed / High Bandwidth
 - Fibre channel – 8 Gb, 16 Gb, FCoE
 - Storage area network standard
 - Ethernet – 10Gb, 40Gb
 - Provides a network based solution to SANs
 - InfiniBand - QDR, 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 Formfactor
 - Solid State PCIe Cards
 - Solid State Iru Trays of Flash

Using the IO

- Redundancy Requires TWO of Everything
- Single Applications Don't Use All the BW of:
 - Fibre channel – 8 Gb, 16 Gb, FCoE
 - Ethernet – 10Gb, 40Gb
 - InfiniBand - QDR, EDR
 - SAS 2.0, 3.0 (6 Gb, 12 Gb)
- Single System Images May Not Use All the BW

What Do We Do?

Share the IO Components

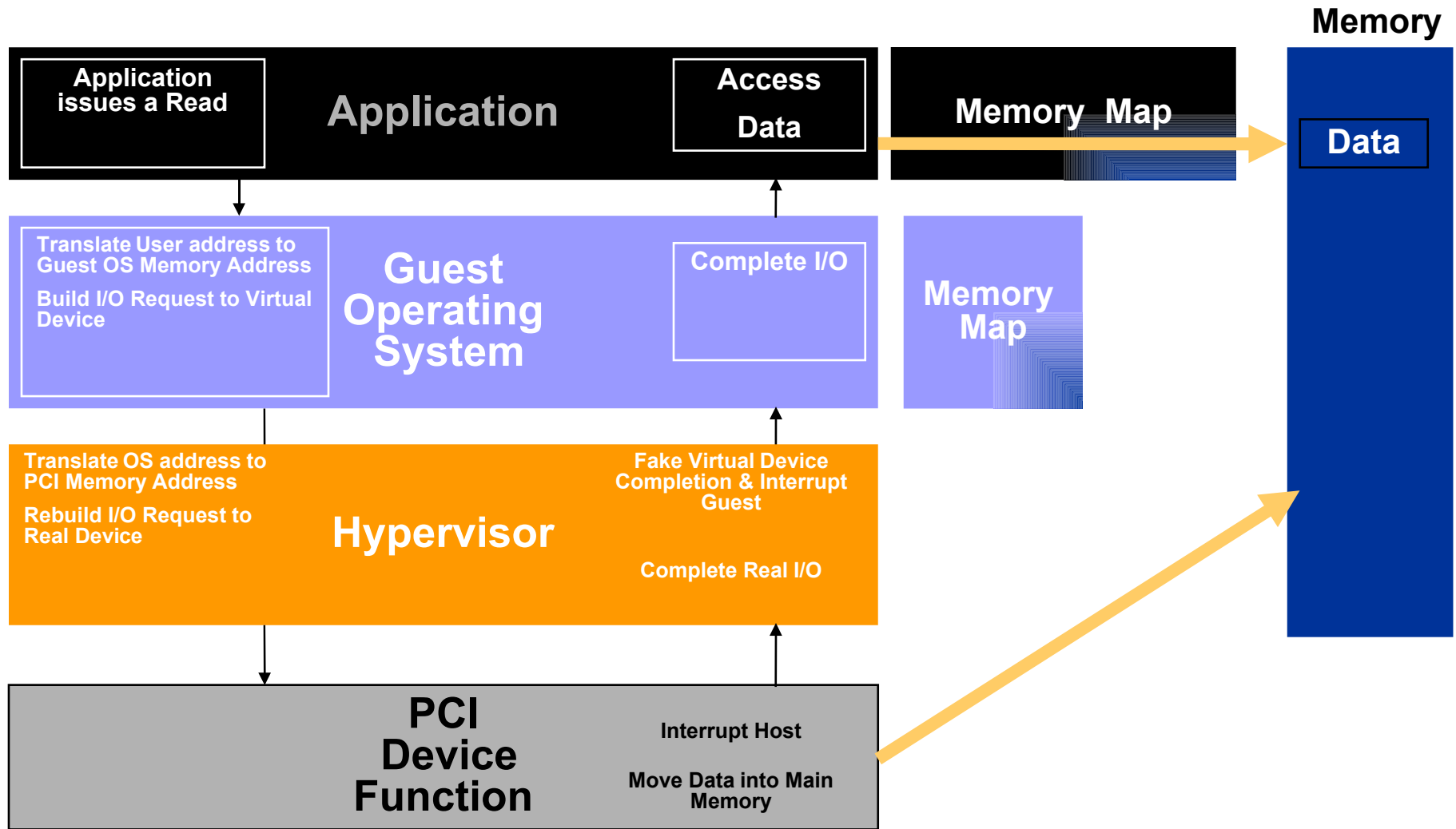
PCIe IOV Provides this Sharing

- 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

Single Root IOV

Better IO Virtualization for Virtual Machines

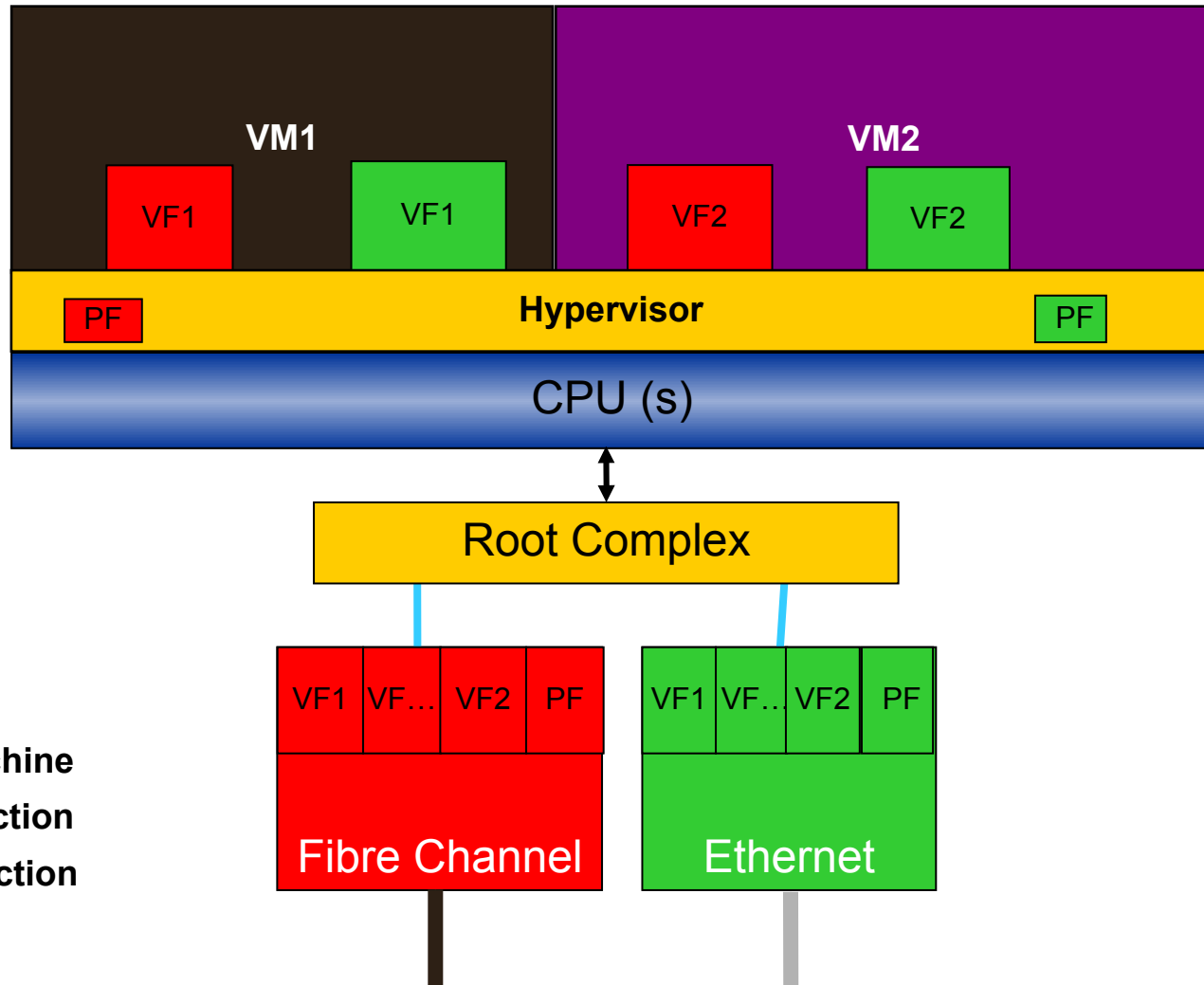
System I/O with a Hypervisor



Single Root IOV

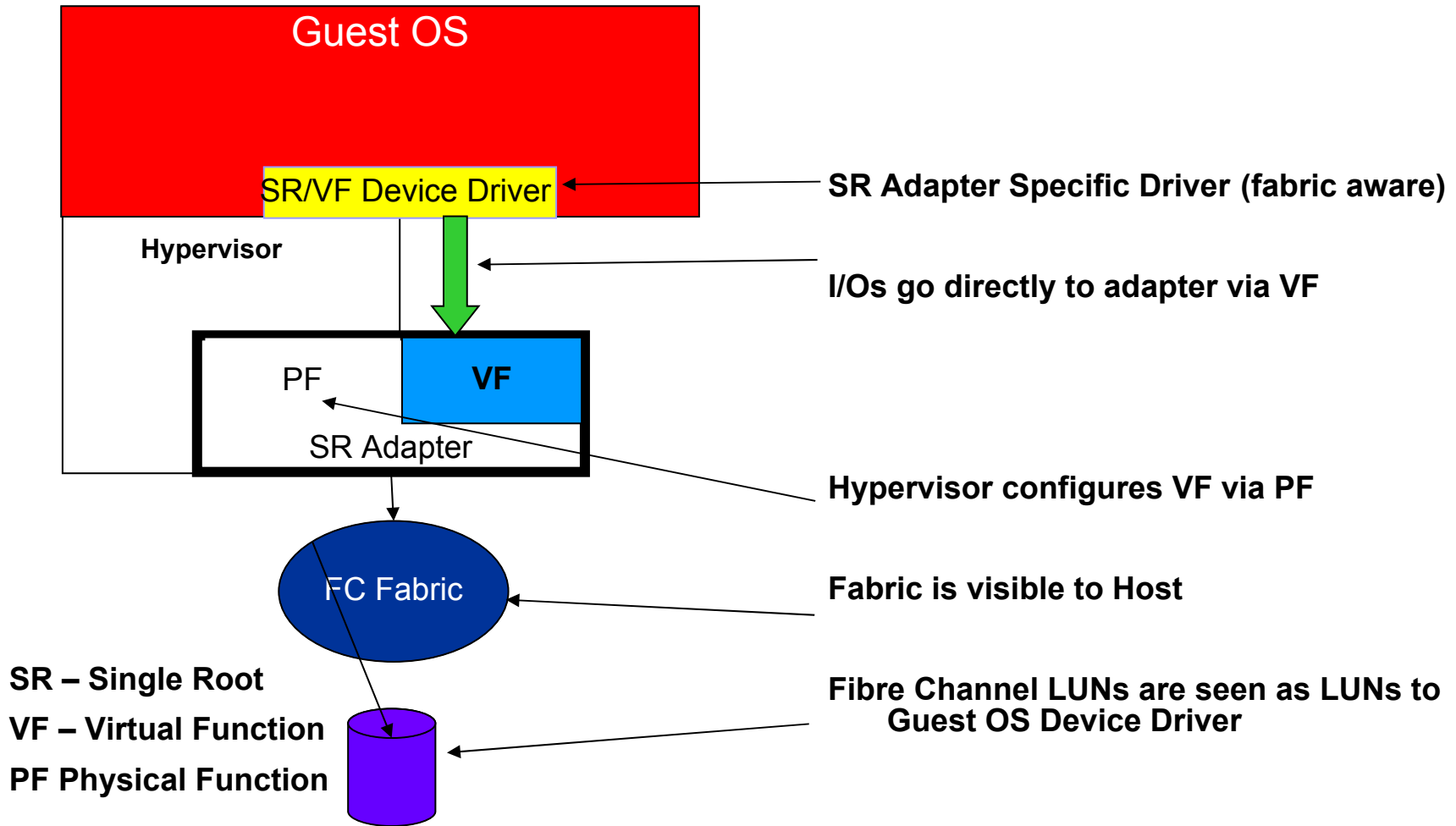
- Before Single Root (SR) IOV the Hypervisor was responsible for creating virtual IO adapters for a Virtual Machine
 - This can greatly impact Performance
 - Especially Ethernet but also Storage (FC & SAS)
- Single Root IOV pushes much of the SW overhead into the IO adapter
 - Remove Hypervisor from IO Performance Path
- Leads to Improved Performance for Guest OS applications

PCI-SIG Single Root



VM – Virtual Machine
VF – Virtual Function
PF Physical Function

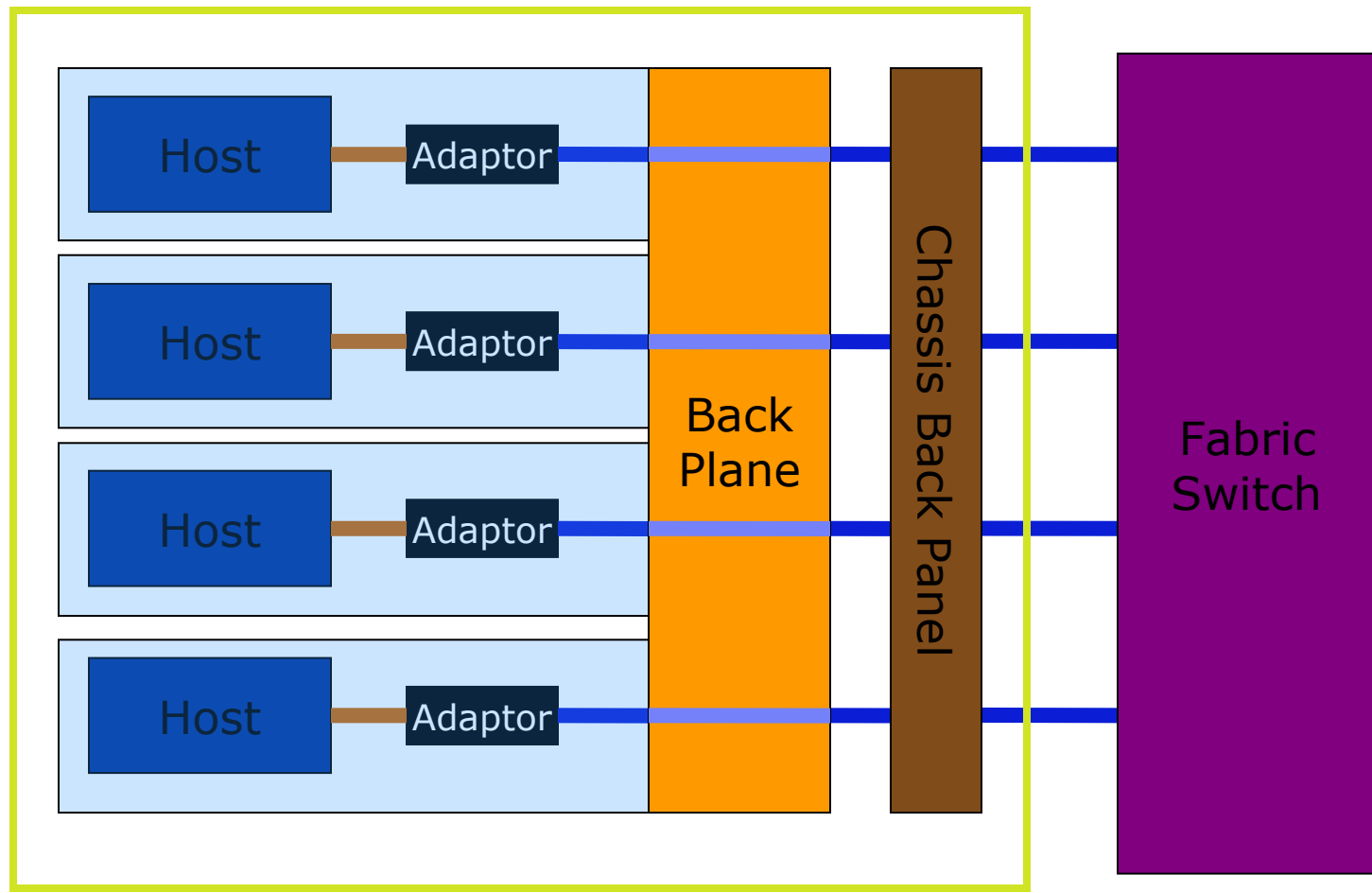
Fibre Channel & SR Virtualization



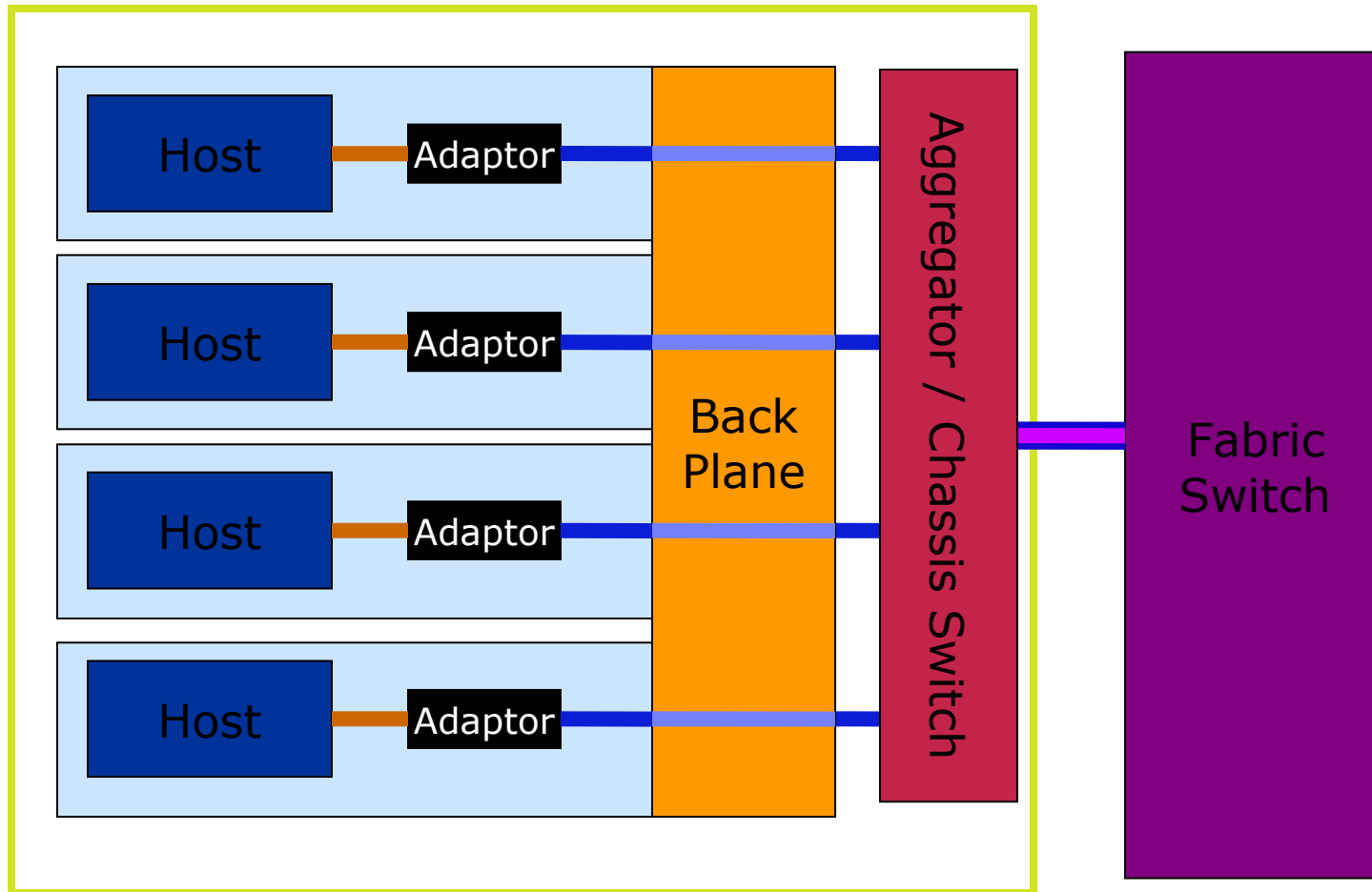
MultiRoot IOV

Virtualizing an IO adapter for multiple Hosts

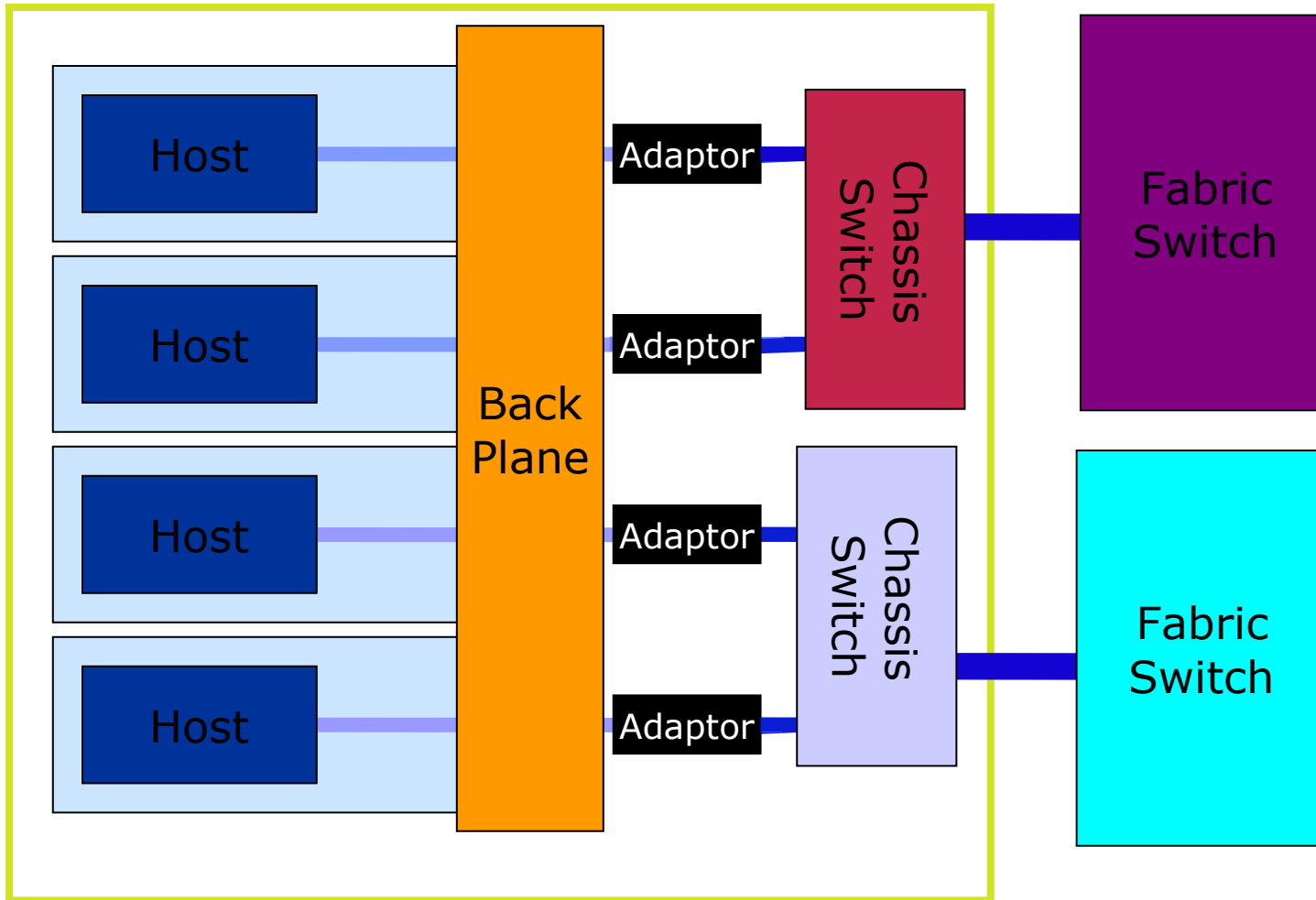
It Started With Pass Thru



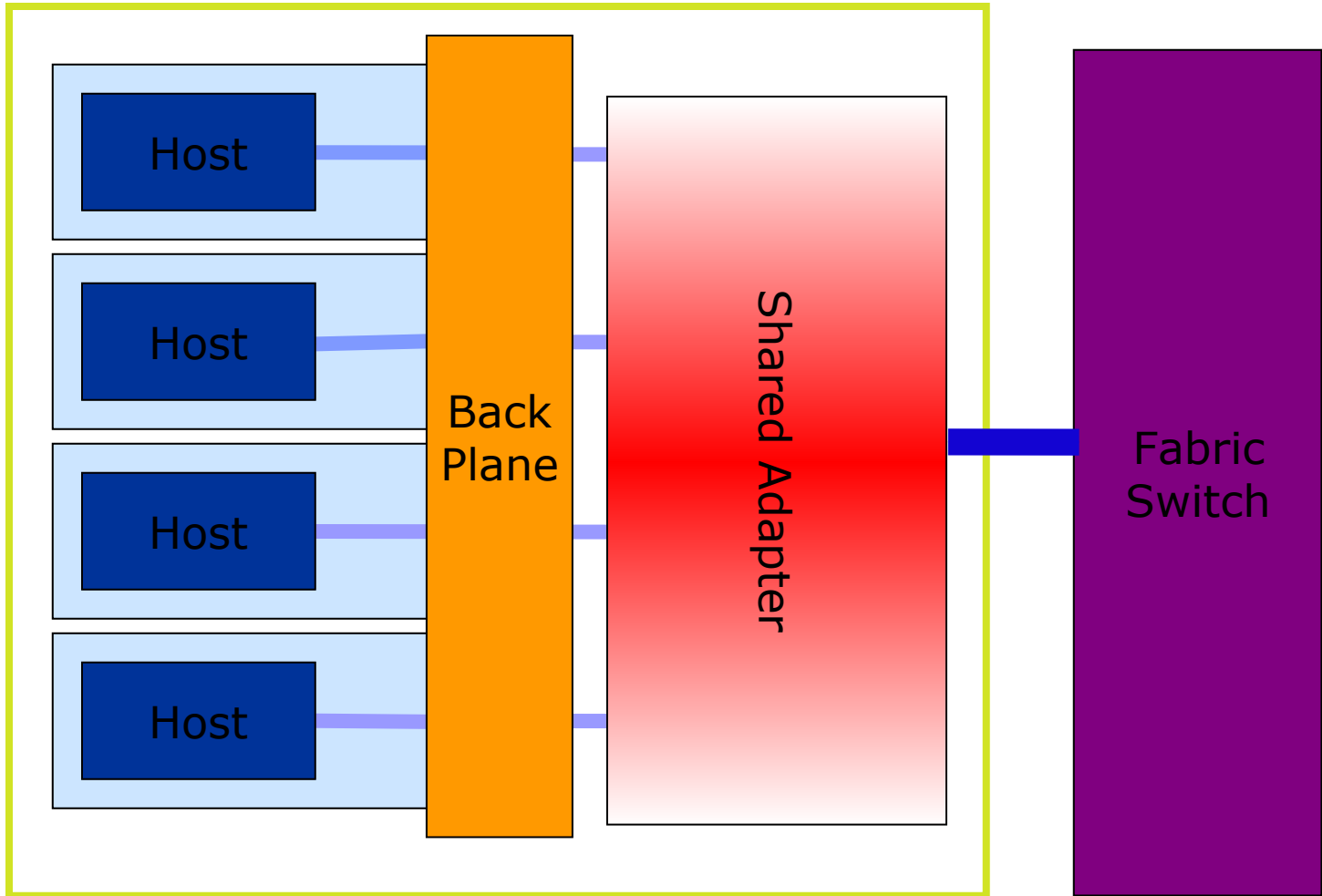
Chassis Switching to Aggregation



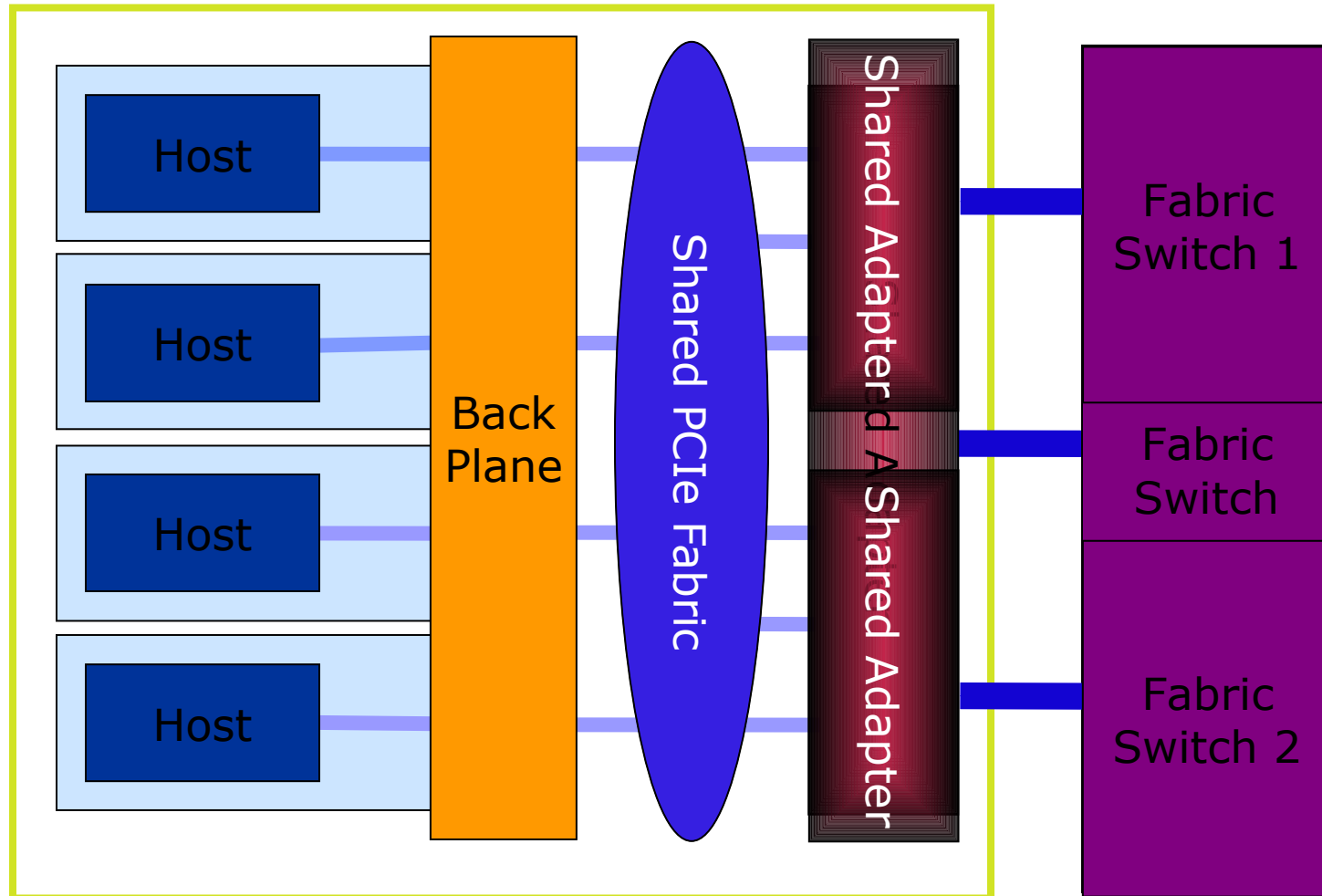
Stretch the PCIe Bus Across the Backplane



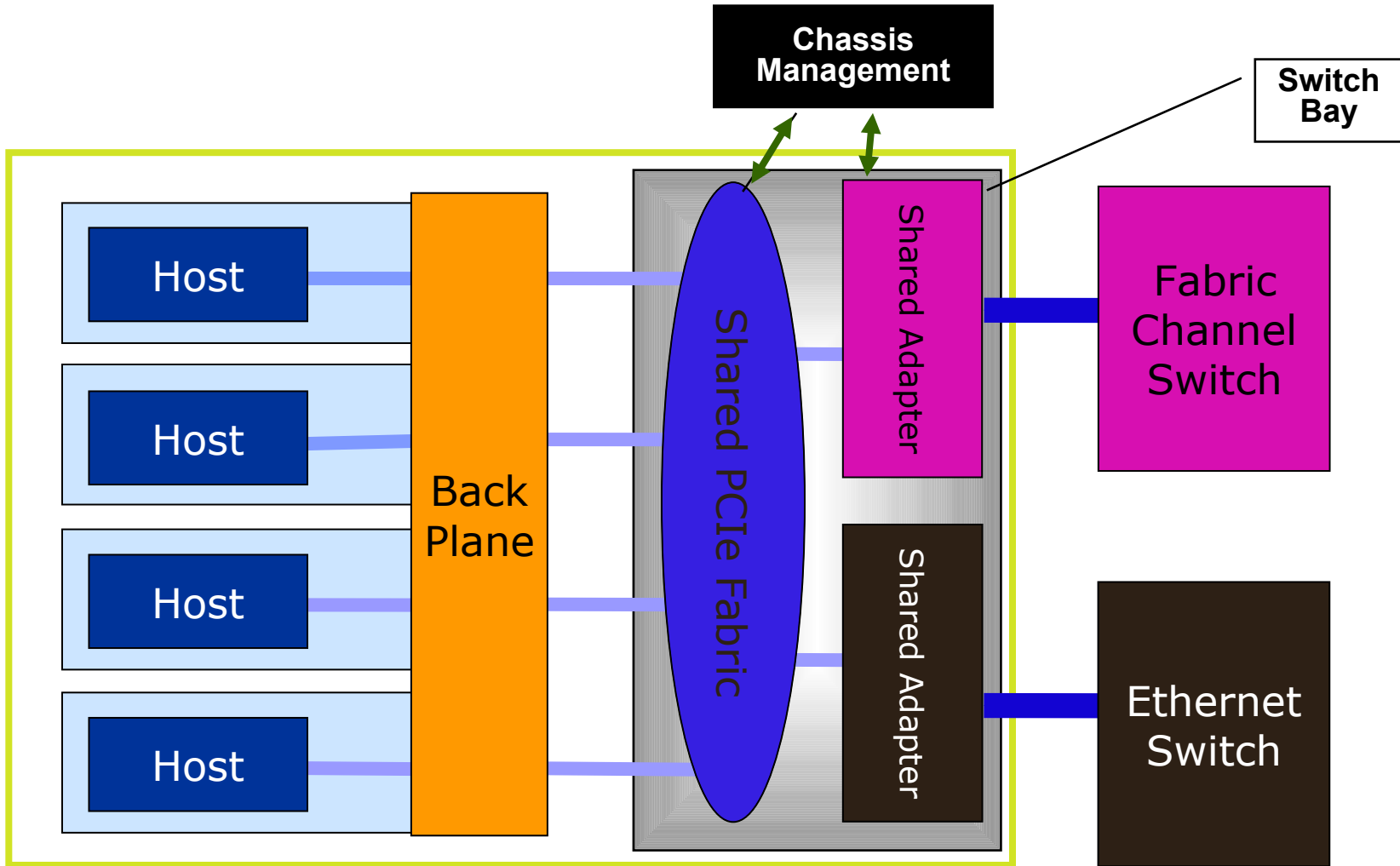
Merge the Adapter and Aggregator SNIA



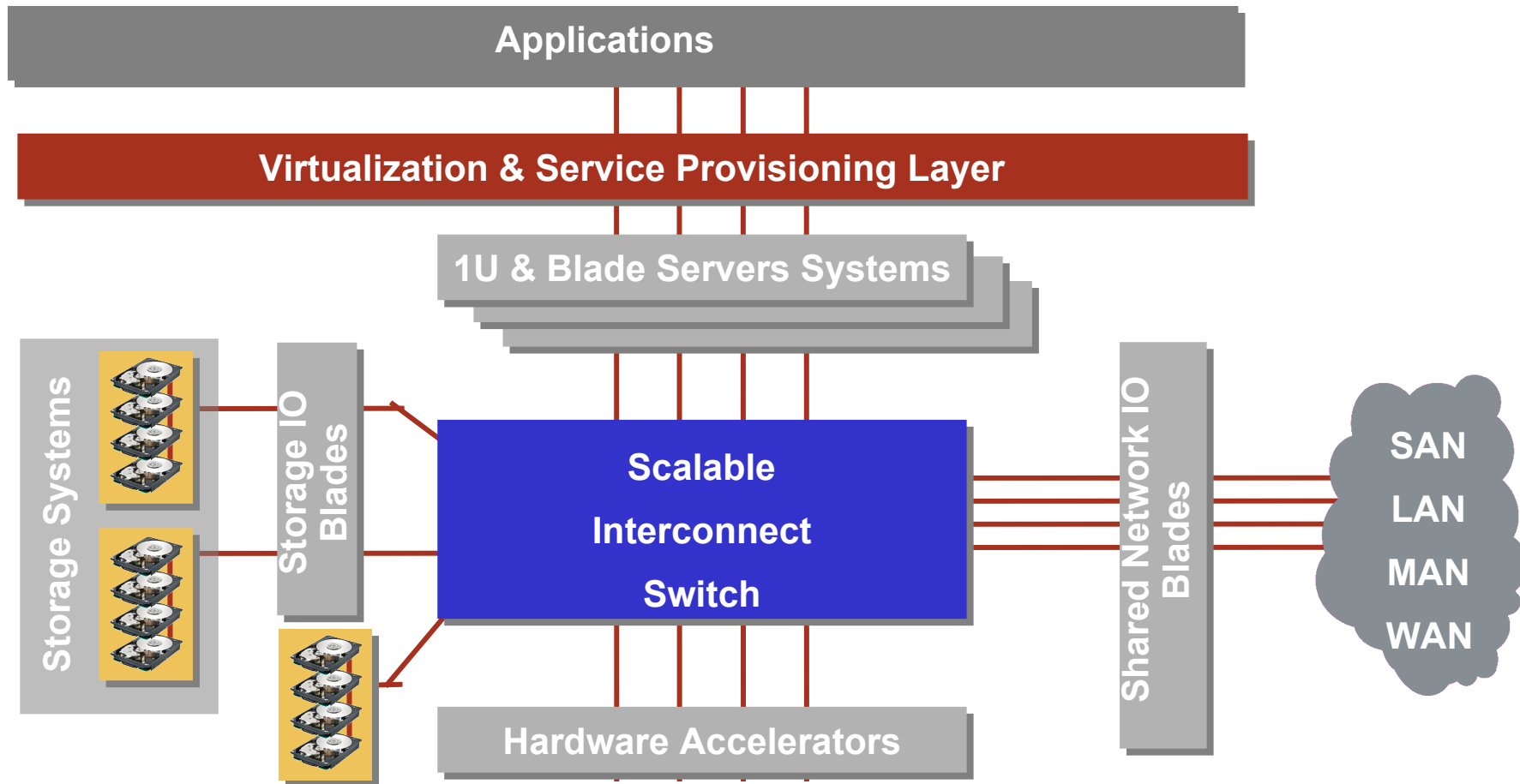
Insert a shared PCIe Fabric and add Multiprotocol



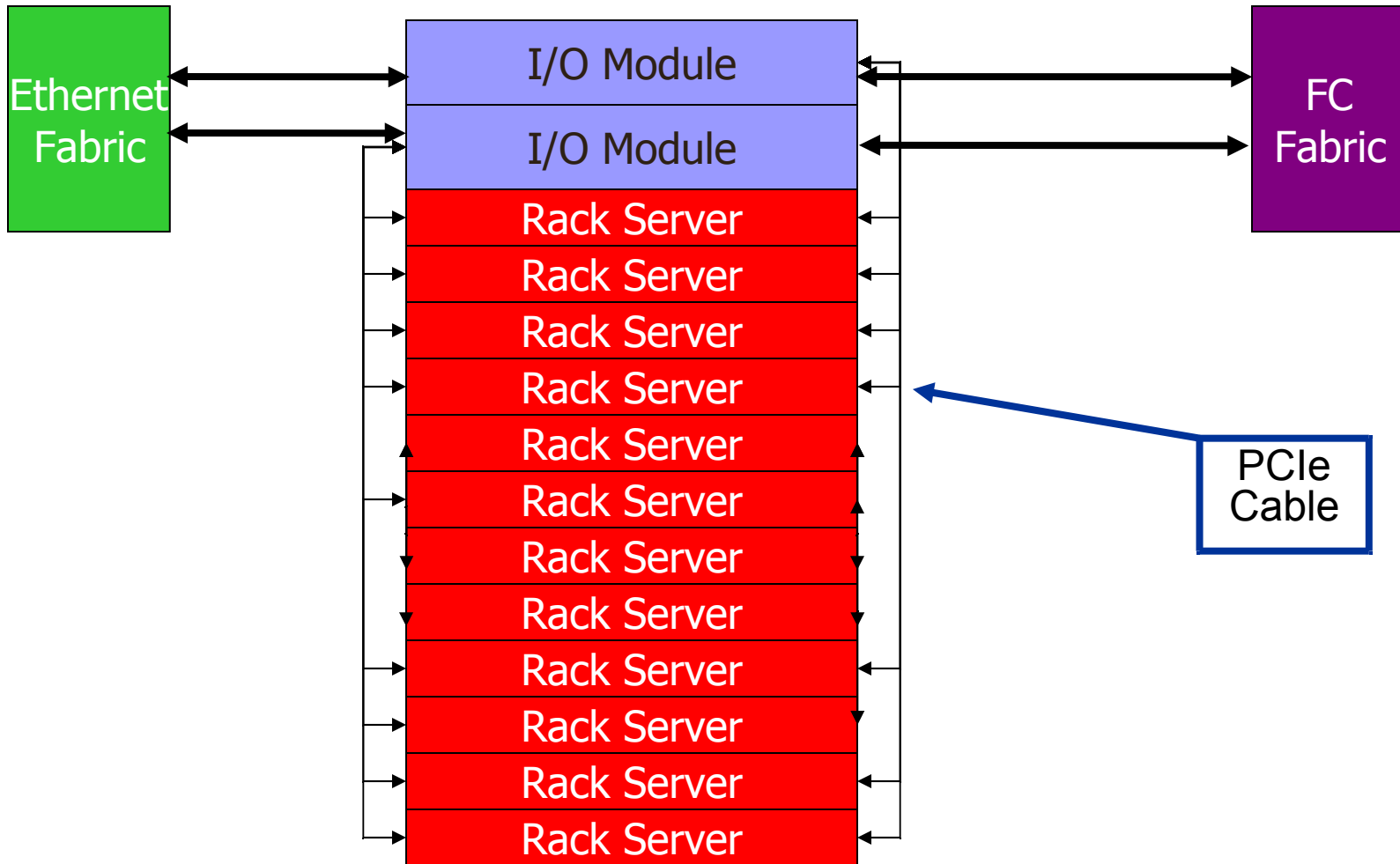
In a Blade Chassis



Mutli Root Virtualization in Blades **SNIA**



In a Rack



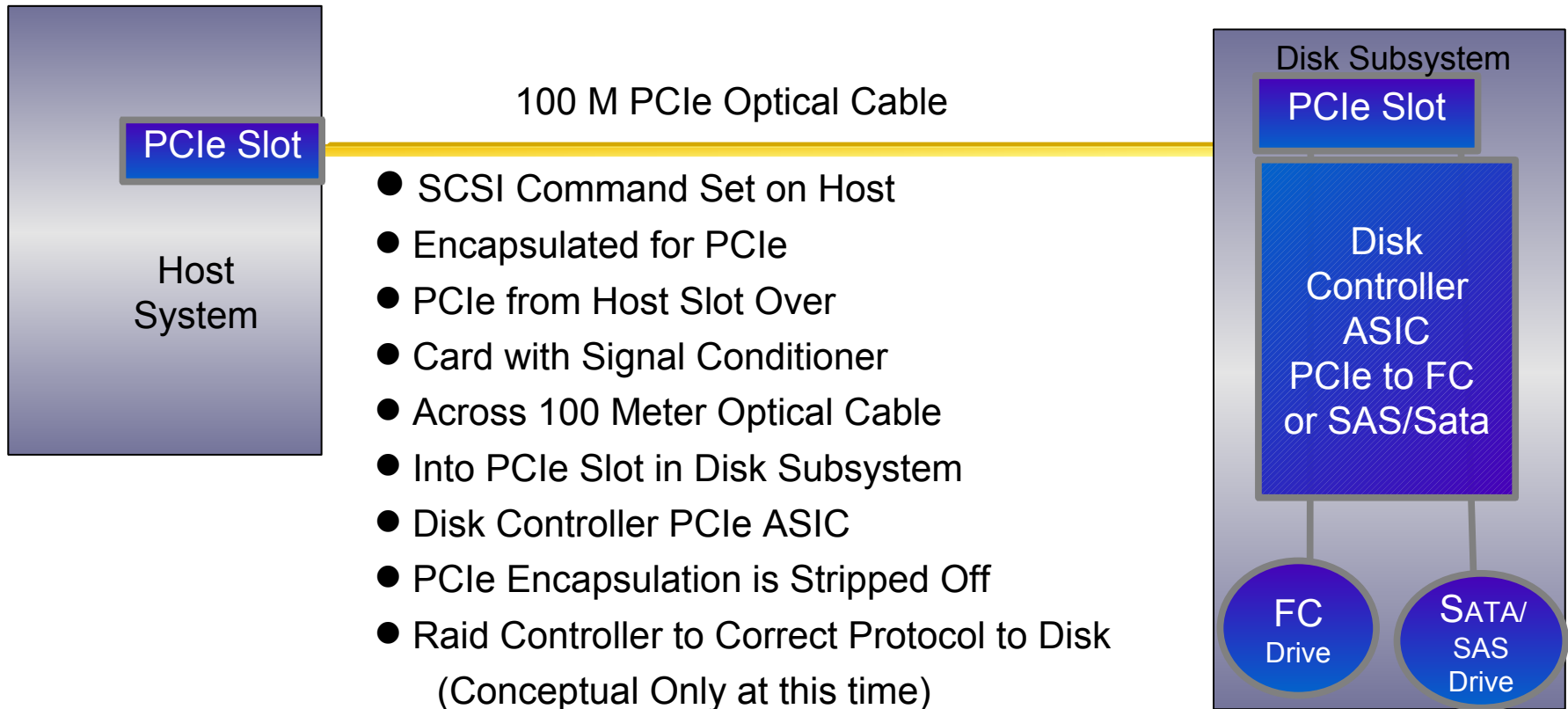
Roll Out of IOV

- **Blade Chassis are First to Roll Out SR IOV**
 - Limited IO Slots
 - Space Constraints
 - Discouraged by OS Uniqueness
- **Servers coming in 2010**
 - Especially Ethernet but also Storage (FC & SAS)
- **MR IOV**
 - No Offerings Yet
 - Great for Blades Sharing High Speed/Bandwidth Ports
 - Each OS must work with IOV Management Layer

Impact / Benefit to Storage

- **PCI Express provides**
 - Full Bandwidth Dual Ported 4 & 8 Gb FC
 - Full Bandwidth for QDR and EDR IB
 - Full Bandwidth SAS 1.0 & 2.0
 - Legacy Support via PCI-X
- **IOV takes it one step further**
 - Ability for System Images to Share IO across OS Images
 - Backplane for Bladed Environments
- **Extension of PCIe**
 - Possible PCIe attached storage devices

Future Storage Attach Model



PCIe 3.0 Root Complexes

- Integrated into the **CPU** (no longer a standalone northbridge)
- Multiple **X8s** from each **Socket** (2-4)
- **RDMA** or **Ethernet Tunneling CPU to CPU** in same chassis

Network Adapters

- **40 GbE Controllers (FCoE, iSCSI, NAS)**
- **Quad 10 GbE Copper & Optical**
- **Dual Ported EDR IB**

Storage Adapters

- SAS 3.0, 8 & 16 ports
- 16 GbE FC HBAs
- Converged Network Adapters (CNA) at 40 GbE

Storage

- SSS (Solid State Storage)

PCIe cards

Trays of Flash DIMMS

2.5" and 3.5" form factors

- 2.5" and 3.5" SAS 10 K RPM drives capacity 500 GB to 2 TB
- 2.5" and 3.5" SATA 2.0 drives capacity 500 GB to over 4 TB
- Tray interfaces 16 Gb FC, SAS 3.0, EDR IB

New Data Center of the 2020

System Based IO is Ethernet

Embedded in the CPU

One type of IO in the Center

Enhanced Loss-less Ethernet Fabric

40 Gb or 100 Gb Ethernet

Normal Ethernet

Low Latency Ethernet

iSCSI

NAS

FCoE

New Storage Interface

Storage All Attached to Ethernet Fabric

Large SSS devices

Large Capacity 2.5"/3.5" SAS and SATA drives

New Storage Interface

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.

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

- Please send any questions or comments on this presentation to SNIA:
tracknetworking@snia.org

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

SNIA Education Committee

Alex Nicolson
John Fehrer
Joe White