PCI Express Impact on Storage Architectures and Future Data Centers

Ron Emerick, Oracle Corporation

SNIA Legal Notice

PCI Express Impact on Storage Architectures and Future Data Centers

◆ PCI Express Gen2 and Gen3, IO Virtualization, FcoE, SSD 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
- Implications and Impacts of FCoE, SSD and IO to storage Connectivity
- IO Virtualization connectivity possibilities in the Data Center (via PCIe)

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, 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 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
  - ATA/PATA to SATA
  - SCSI to SAS
  (UWDIS may finally be gone)
- Move 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
  - ◆ 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

PCI Express Device A

Signal

Wire

Link

Lane
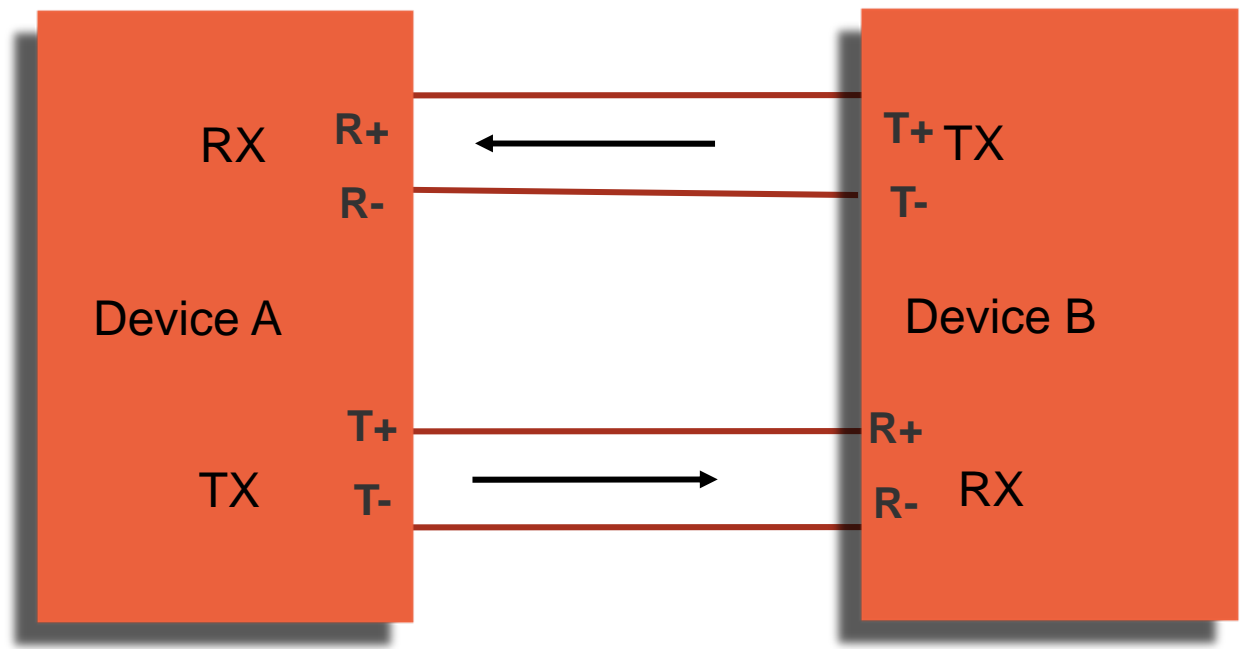
PCI Express Device B
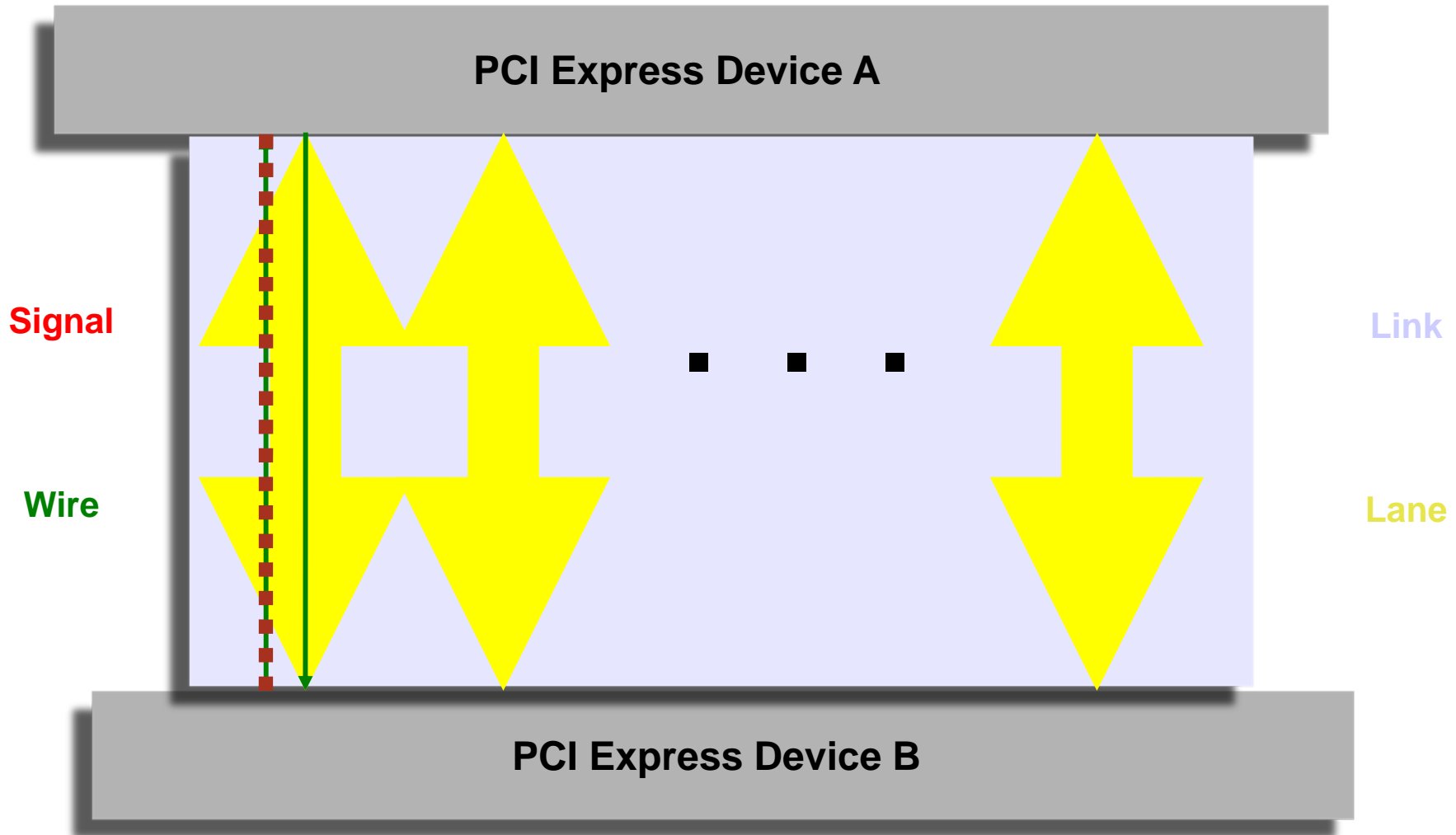
PCIe What's A Lane

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

Lane

Device A

P O R T

P O R T

Device B

Link
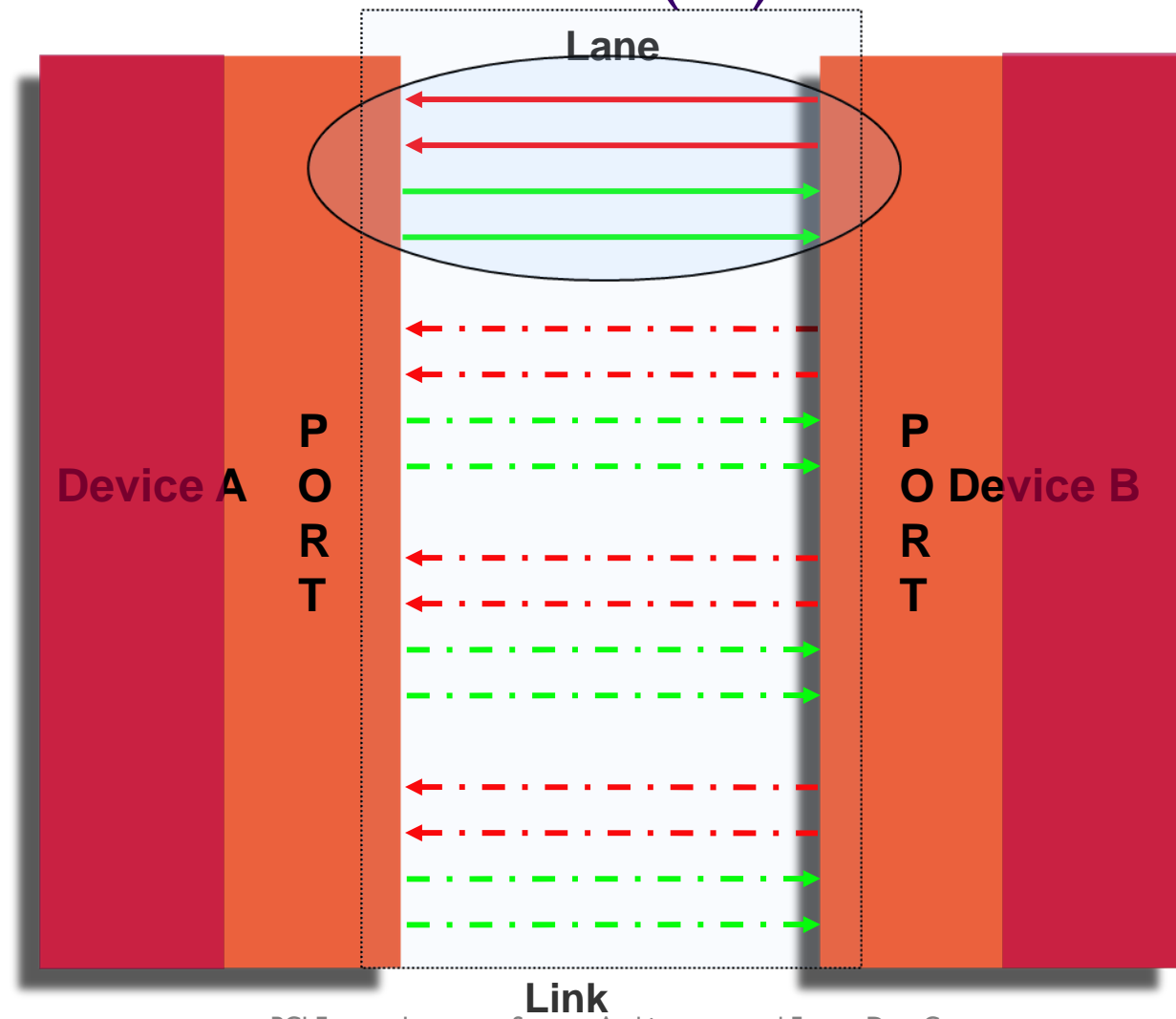
Transaction Types

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/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**

PCI-X vs PCIe Throughput

How does PCI-X compare to PCI Express?
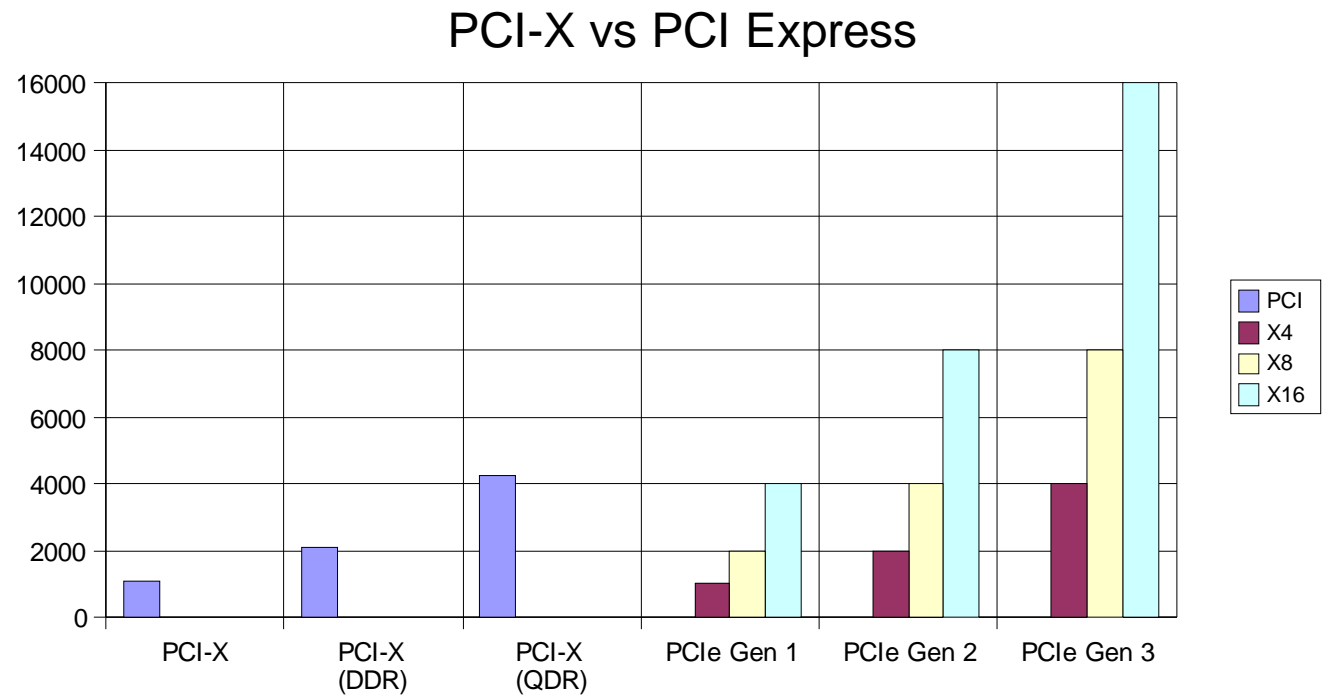
PCI-X vs PCI Express

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

IO Bandwidth Needs

PCI Express Bandwidth
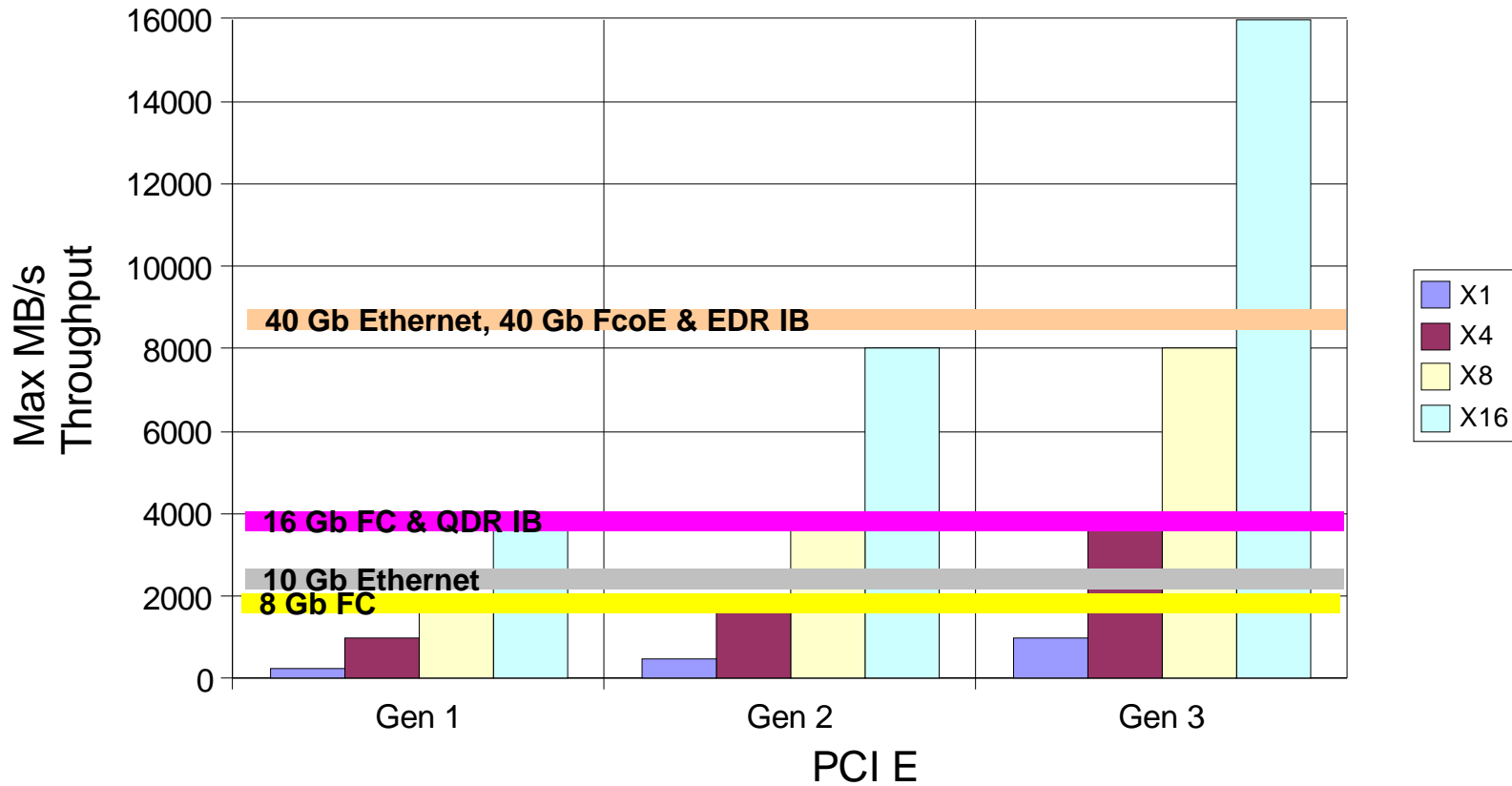
Sample PCI Express Topology

Figure 1-2: Example Topology
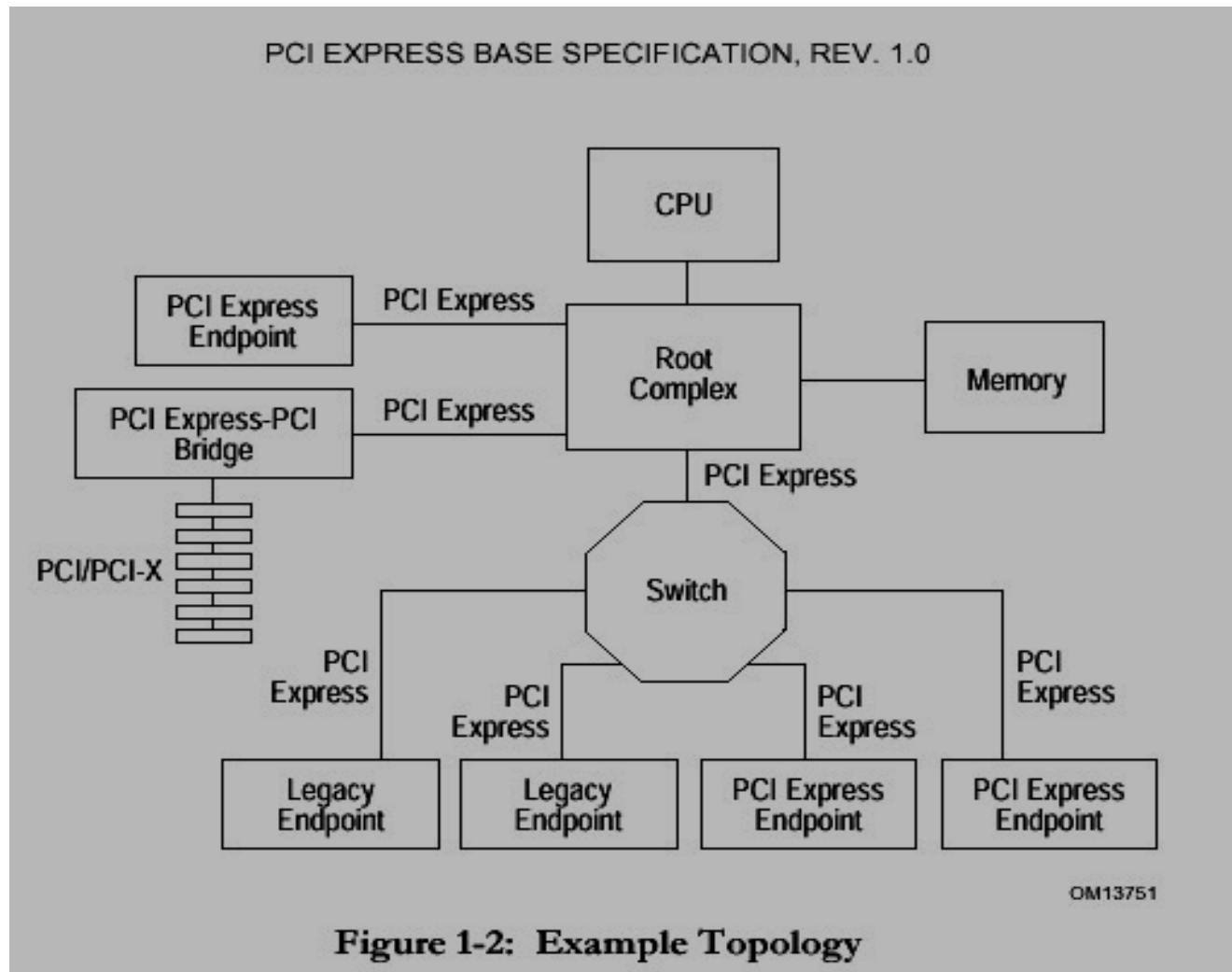
Benefits of PCI Express

> **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

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**
- **Allows SAS routing across PCIe Connector**
- **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**

Cover

Card-Edge Connector

Module Board

EMI Gasket

Module Carrier

Button

Auxiliary Venting

Ejection Lever

Primary Venting
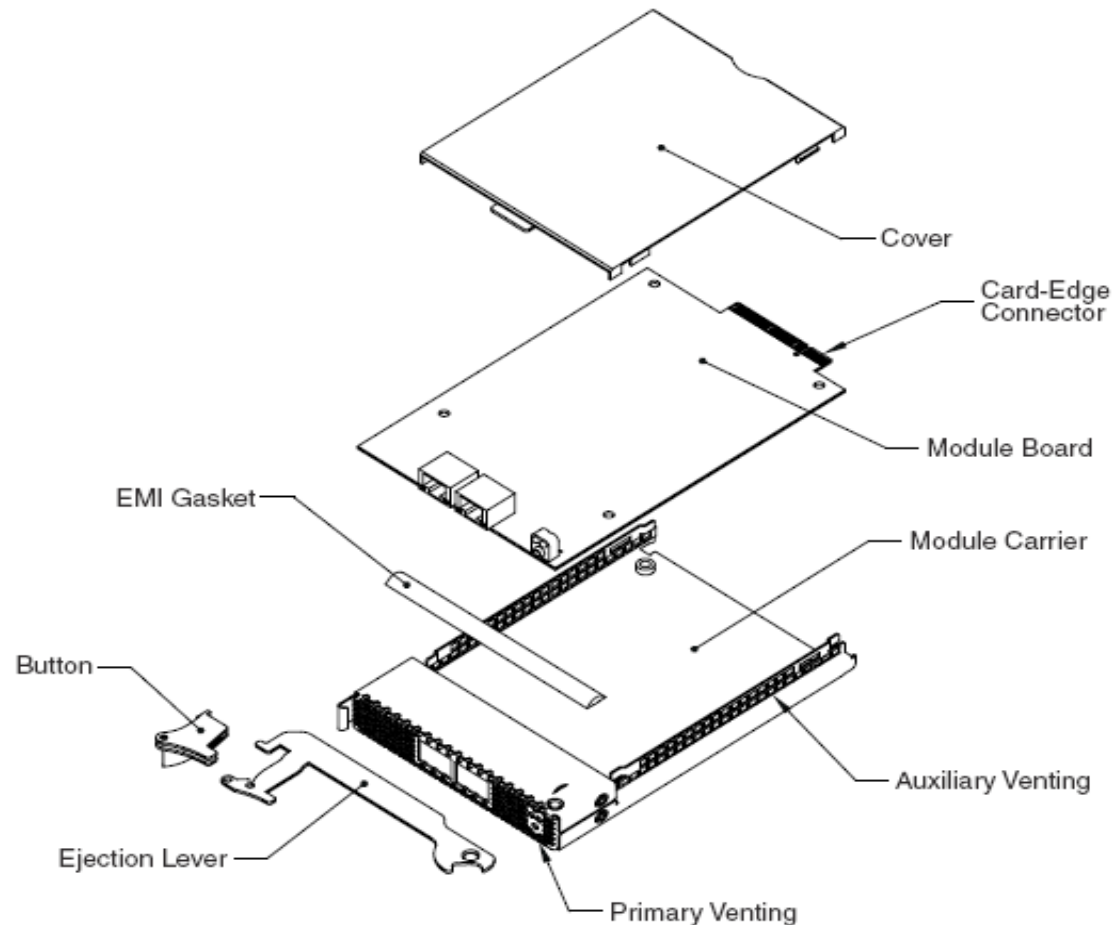
PCI Express In Industry

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, 4/8 Gb FC, SAS, IB
- x16 High Performance Graphics @ 150 W and more

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), scrambling only
  - Gen 3.0 Base Spec at 0.7, 0.9 and 1.0 later this year
- External expansion
  - Copper cable and connector specified, Optical cable available
- Geneseo enhancements to PCIe 2.0 in 3.0
  - Standard for co-processors, accelerators
  - Encryption, visualization, mathematical modelling
- PCIe IO Virtualization (SR/ MR IOV)
  - Architecture allows shared bandwidth

Evolving System Architectures

- **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

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 1ru Trays of Flash

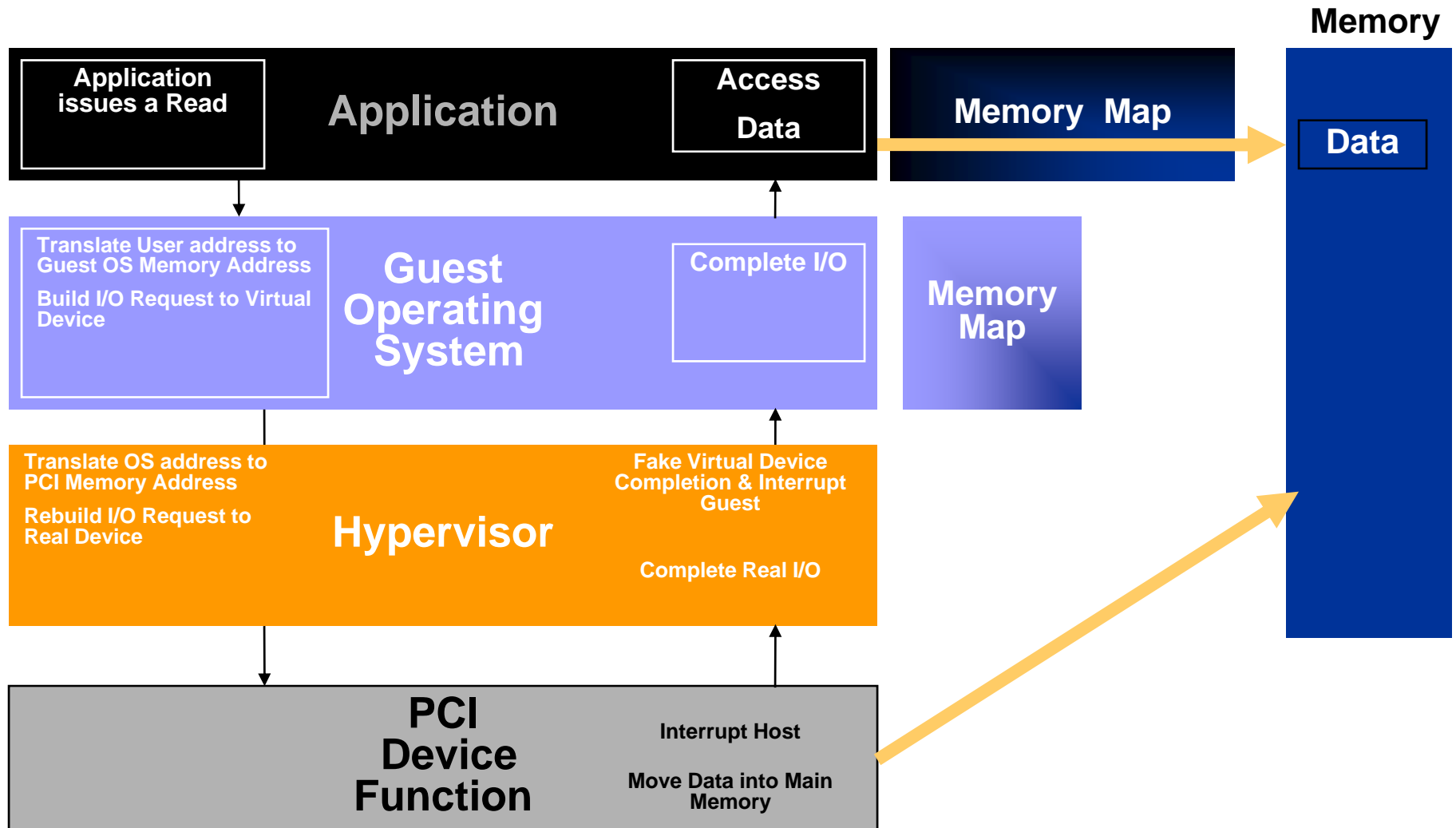# Share the IO Components

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

Memory

Application

Application issues a Read

Access Data

Memory Map

Data

Guest Operating System

Translate User address to Guest OS Memory Address

Build I/O Request to Virtual Device

Complete I/O

Memory Map

Hypervisor

Translate OS address to PCI Memory Address

Rebuild I/O Request to Real Device

Fake Virtual Device Completion & Interrupt Guest

Complete Real I/O

PCI Device Function

Interrupt Host

Move Data into Main Memory
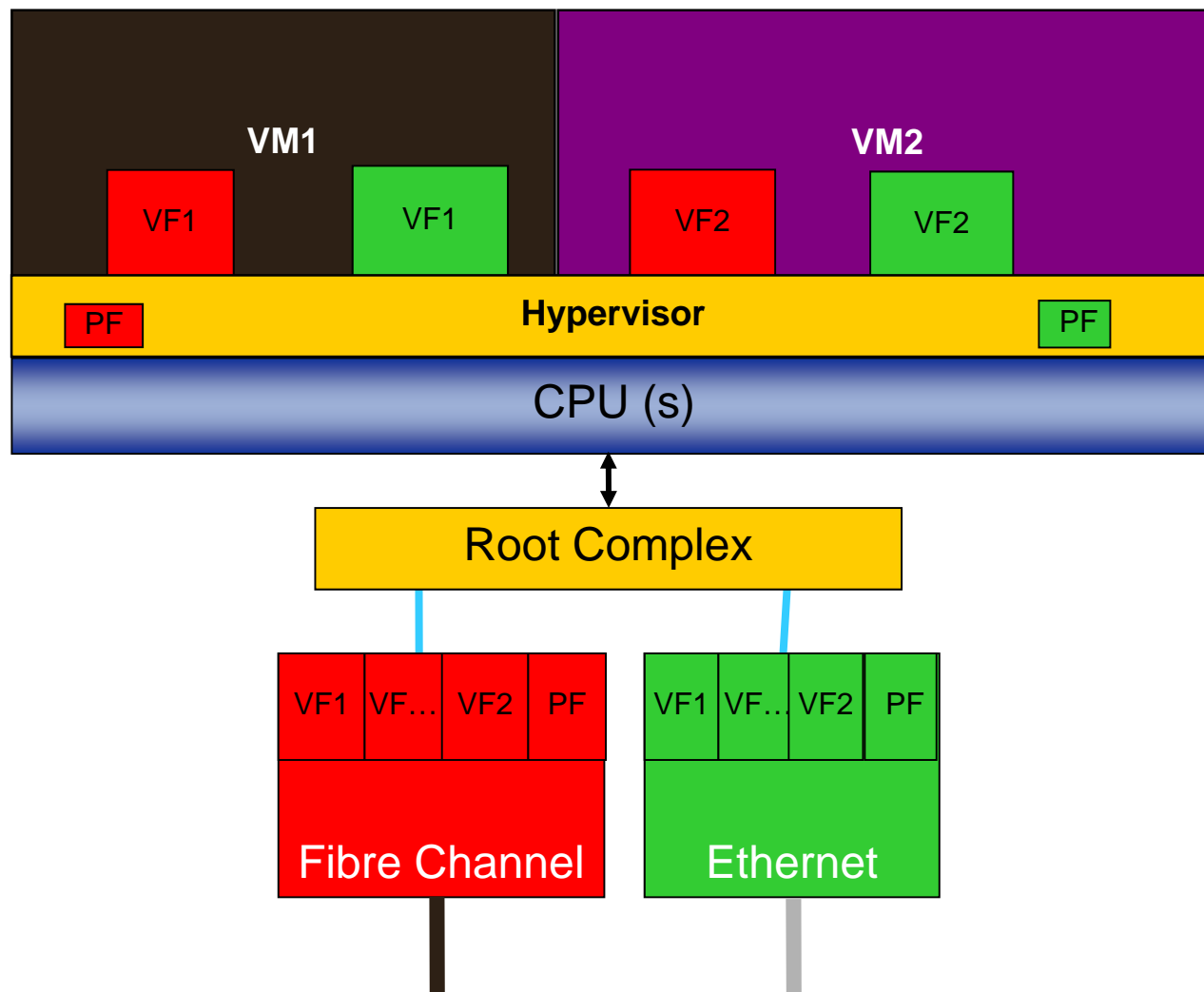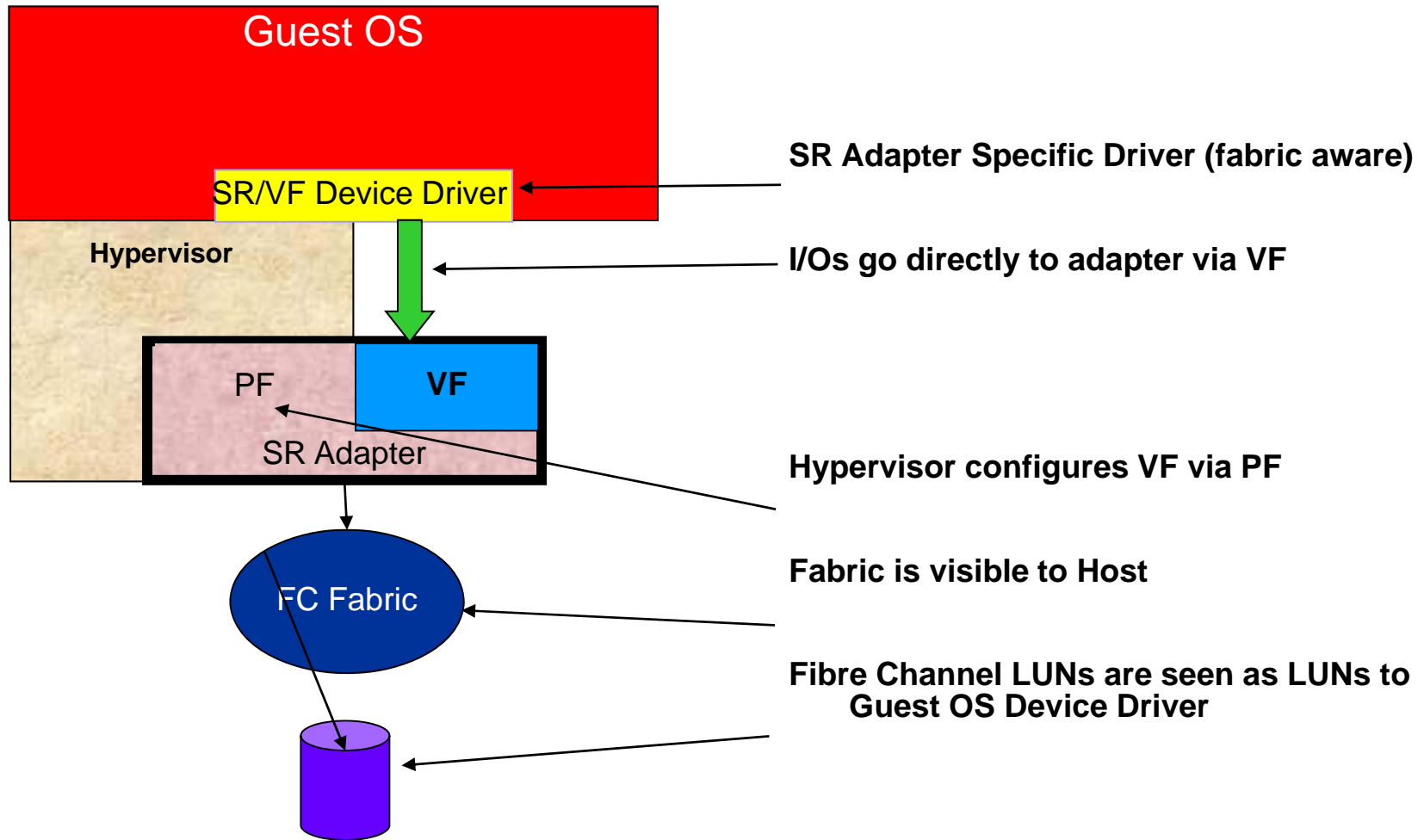
Single Root IOV

- 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

PCI-SIG Single Root

Fibre Channel & SR Virtualization

Guest OS

SR/VF Device Driver

Hypervisor

PF   **VF**

SR Adapter

FC Fabric

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
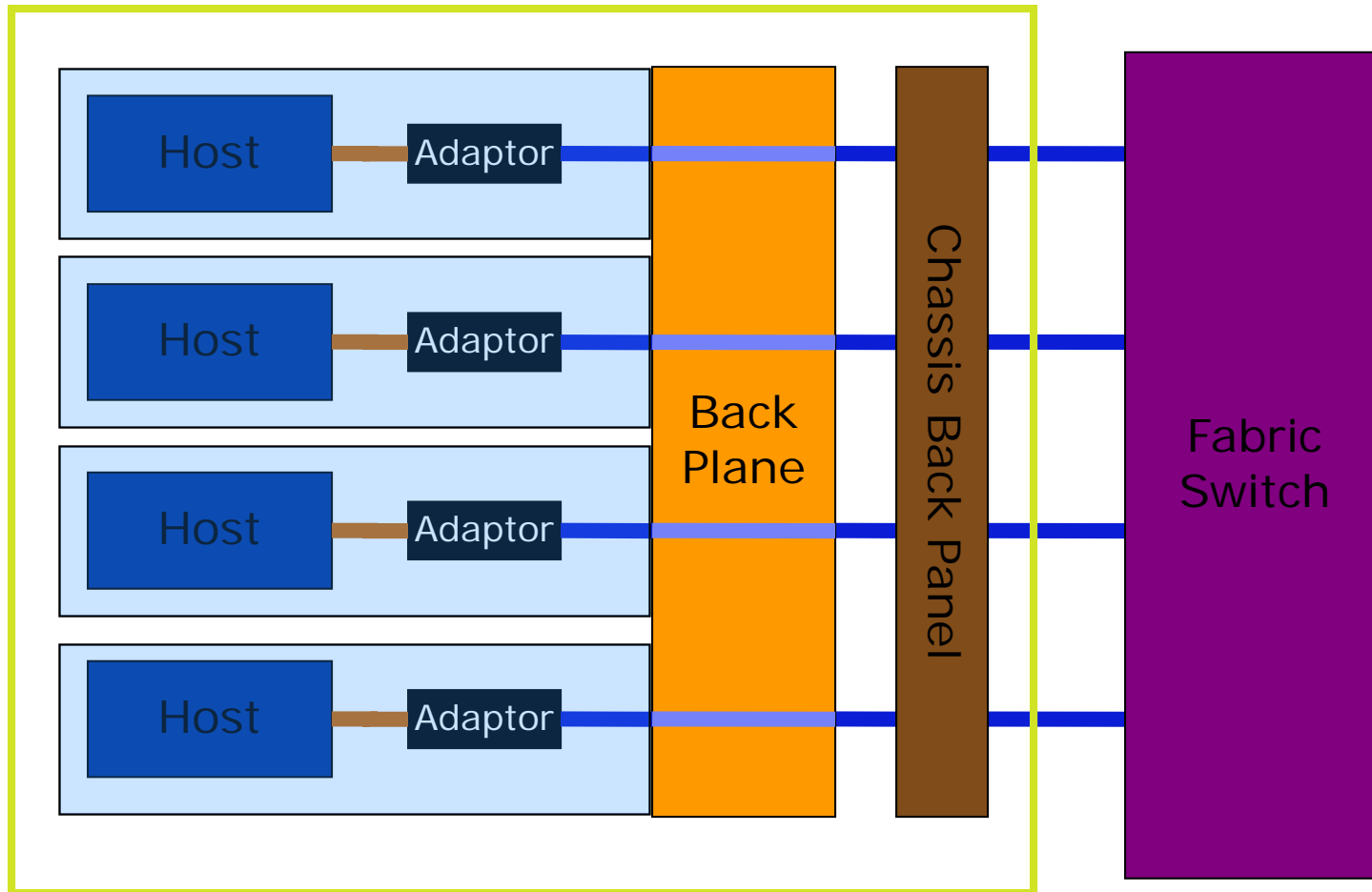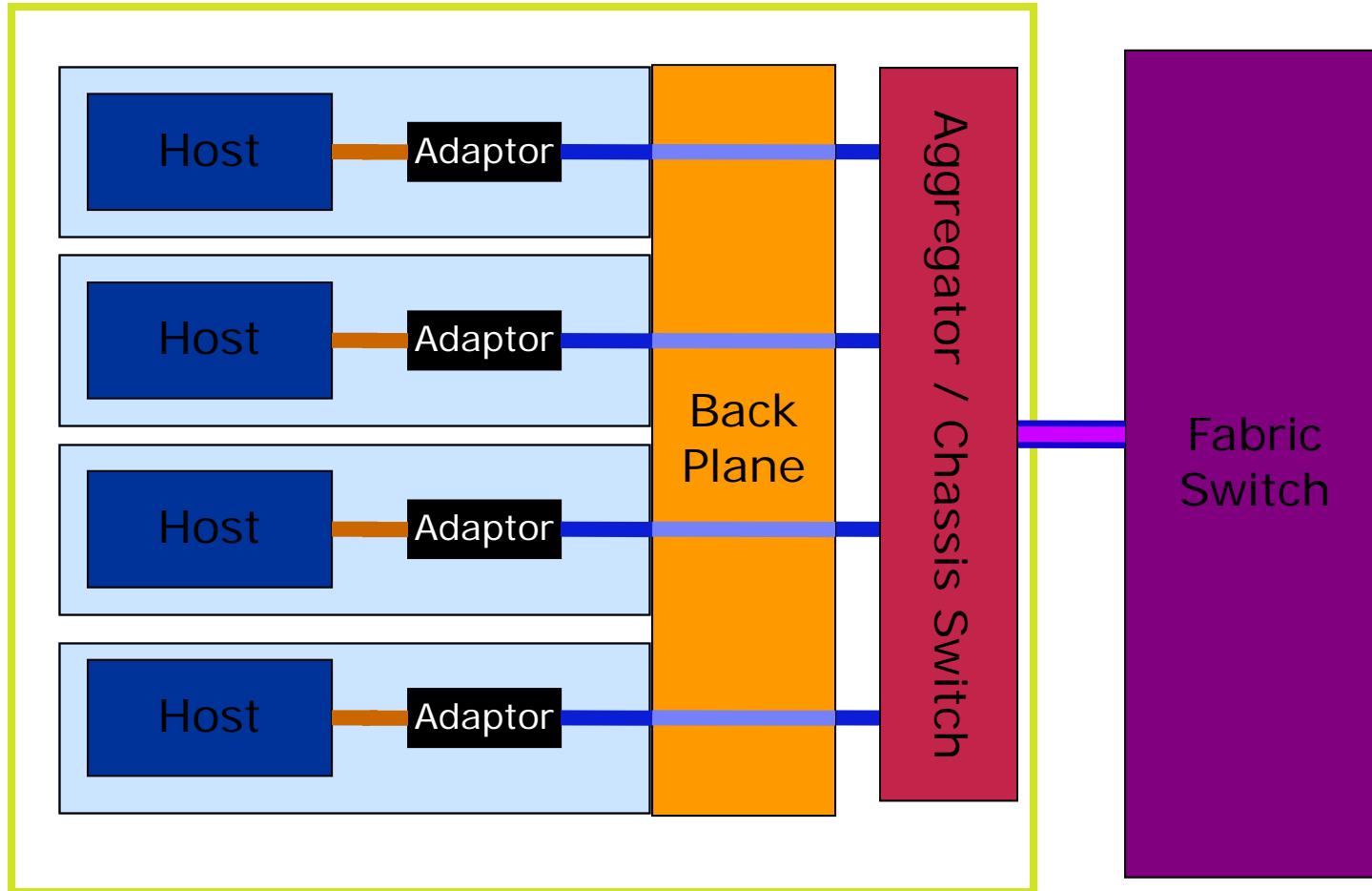
MultiRoot IOV

Virtualizing an IO adapter for multiple Hosts

It Started With Pass Thru

Chassis Switching to Aggregation

Host — Adaptor

Host — Adaptor

Host — Adaptor

Host — Adaptor

Back Plane

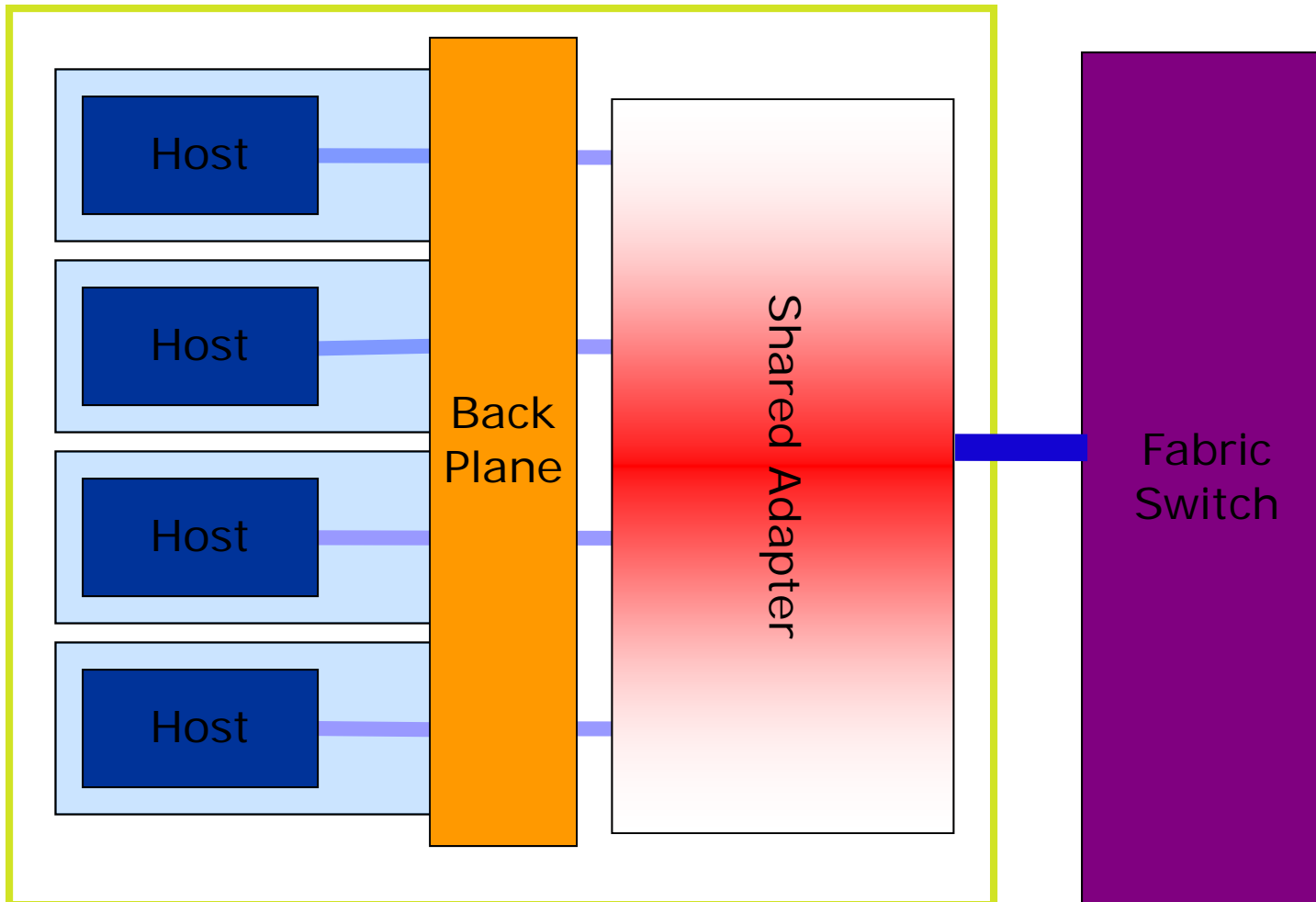Aggregator / Chassis Switch

Fabric Switch

Stretch the PCIe Bus Across the Backplane
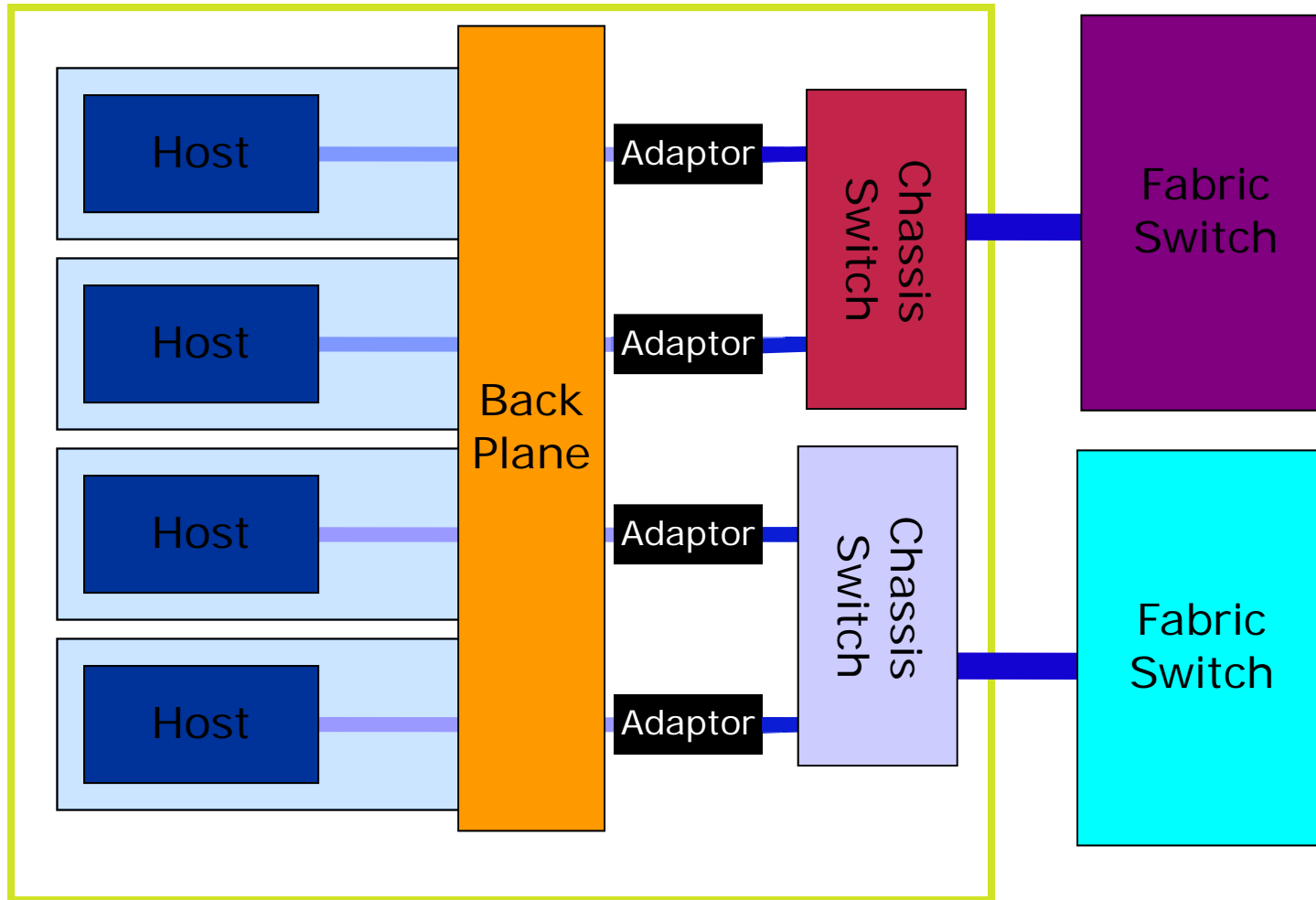
Merge the Adapter and Aggregator

Insert a shared PCIe Fabric and add Multiprotocol
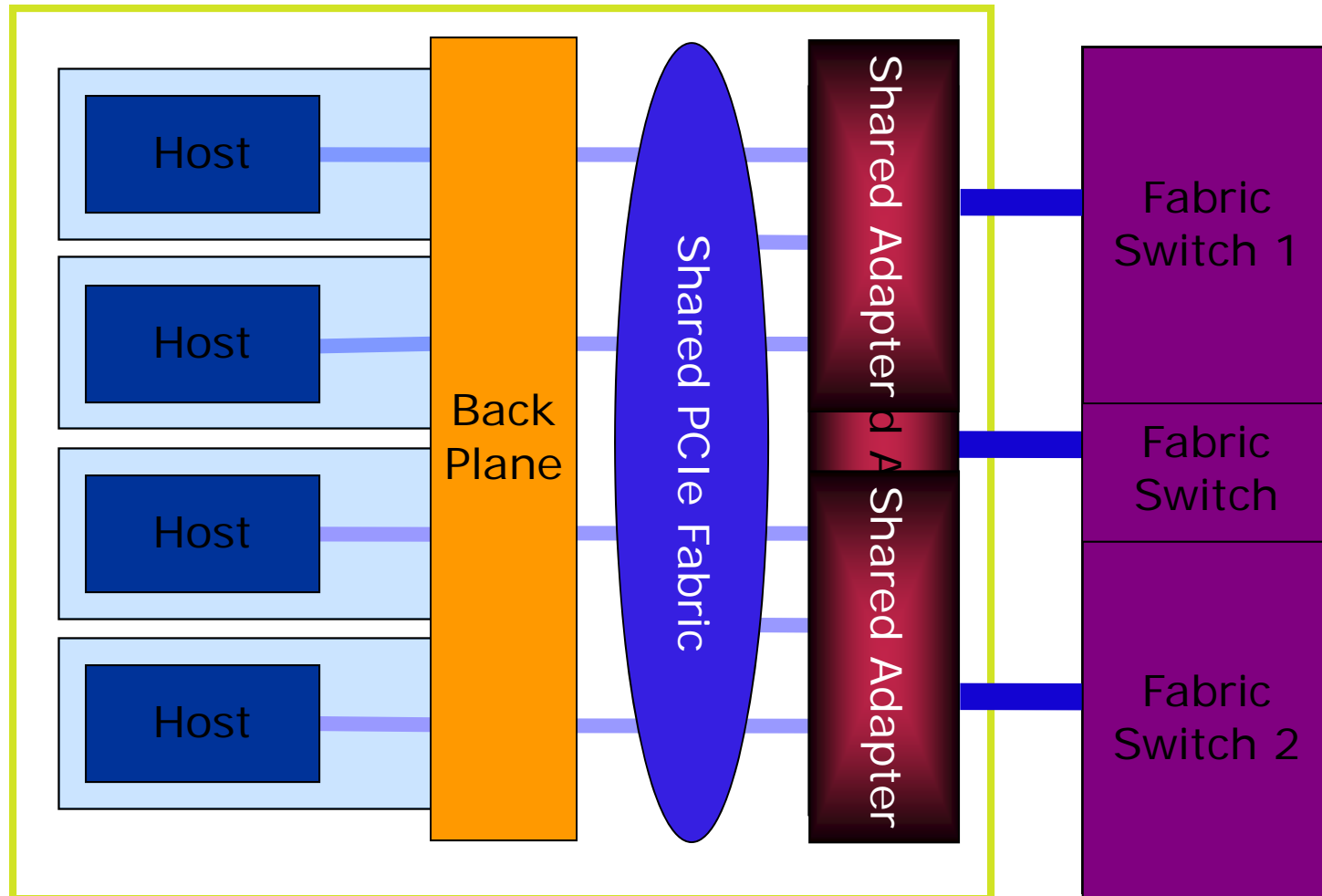
In a Blade Chassis
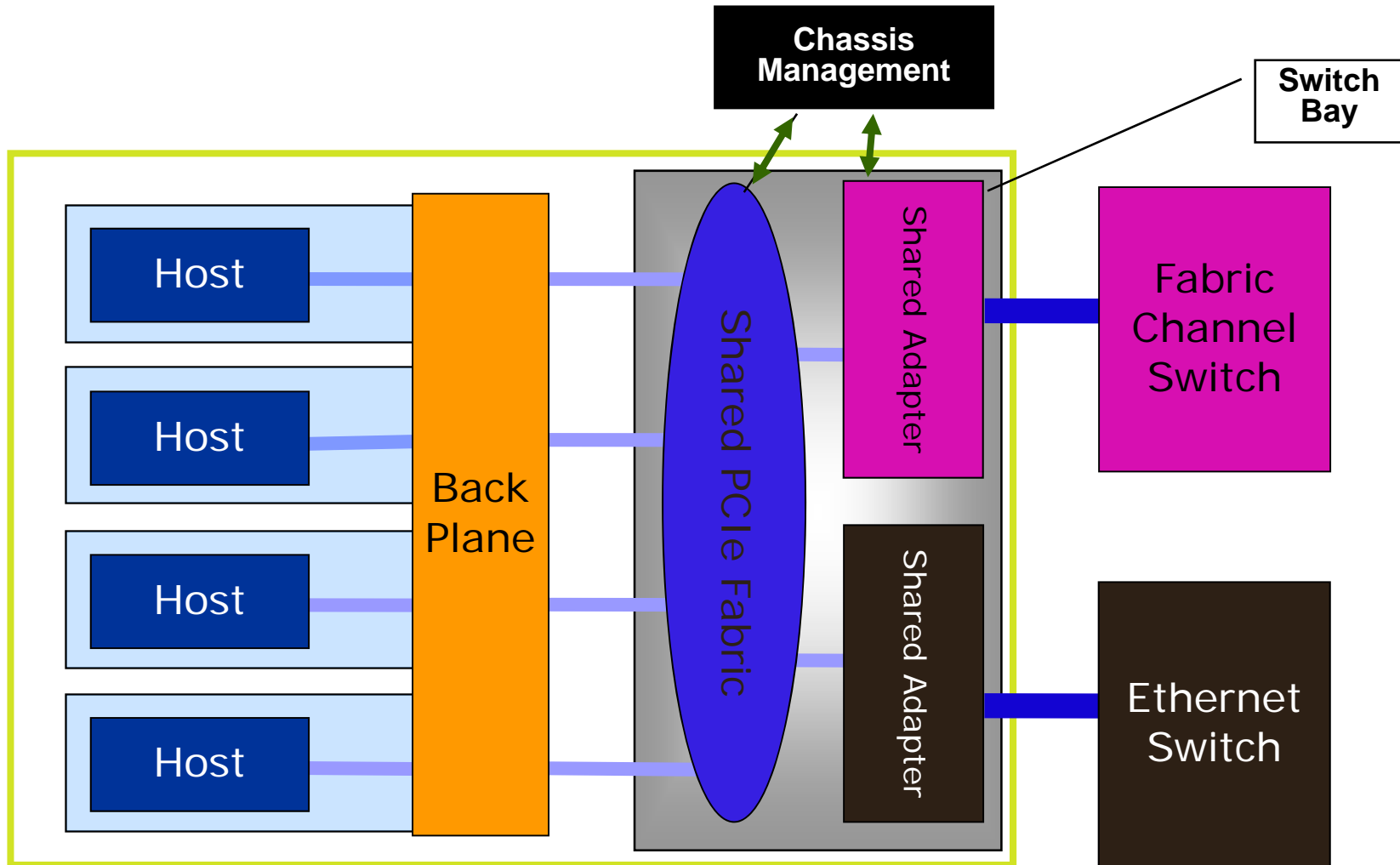
Mutli Root Virtualization - Blades

Applications

Virtualization & Service Provisioning Layer

1U & Blade Servers Systems

Storage Systems

Storage IO Blades

Scalable Interconnect Switch

Shared Network IO Blades

SAN

LAN

MAN

WAN

Hardware Accelerators

In a Rack

Ethernet Fabric

I/O Module

I/O Module

FC Fabric

Rack Server

Rack Server

Rack Server

Rack Server

Rack Server

Rack Server

Rack Server

Rack Server

Rack Server

Rack Server

Rack Server

Rack Server

PCIe Cable

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**
  - Yes this year!
  - 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
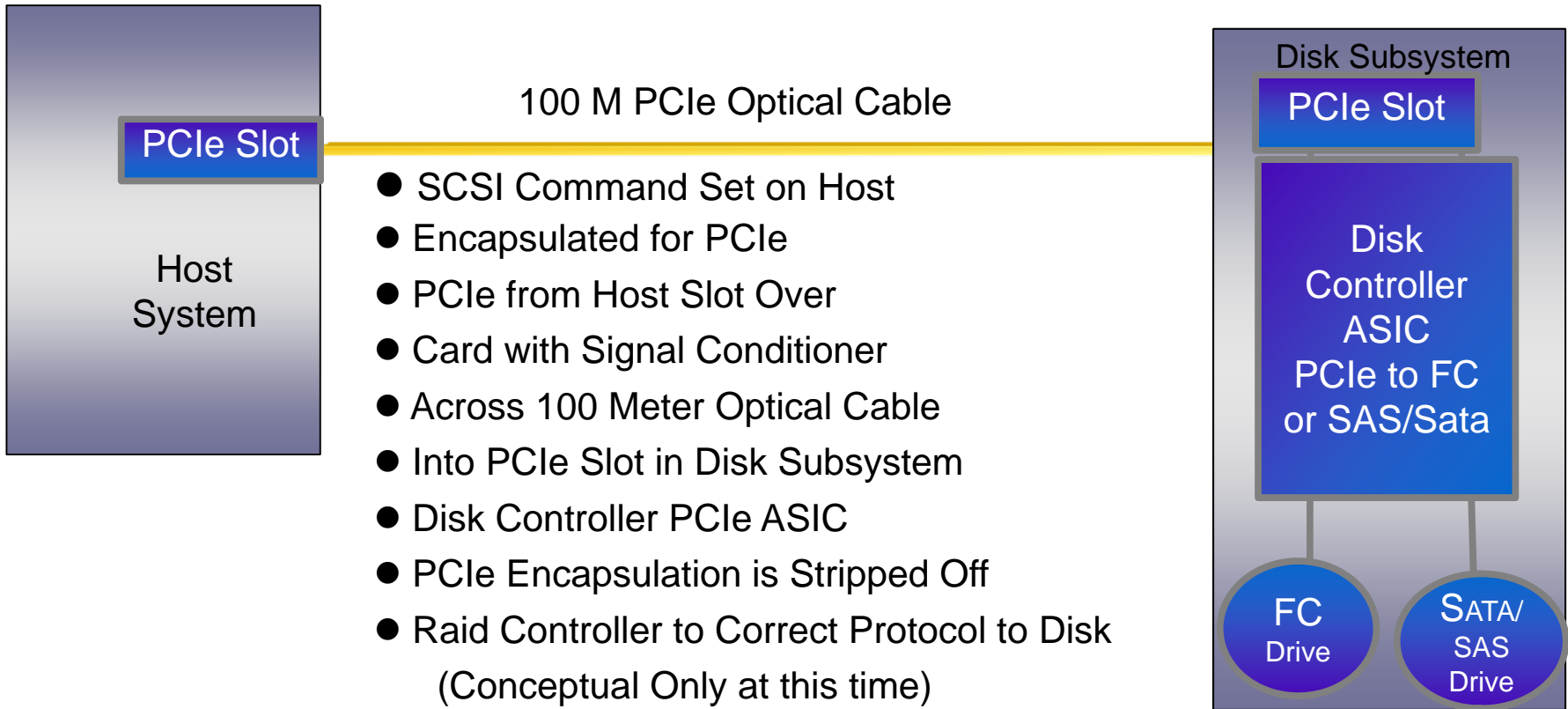   - Legacy Support via PCI-X
- **IOV takes it one step further**
    - Ability for System Images to Share IO across OS Images
    - Backplane for Bladed Environments
- **Extension of PCIe**
    - Possible PCIe attached storage devices

Future Storage Attach Model

Host System

PCIe Slot

100 M PCIe Optical Cable

- 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)

Disk Subsystem

PCIe Slot

Disk Controller ASIC PCIe to FC or SAS/Sata

FC Drive

S_{ATA}/ SAS Drive

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

Q&A / Feedback

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