SNIA

Education

PCI Express Impact on Storage Architectures

Ron Emerick, Sun Microsystems

SNIA Legal Notice

- 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.

Abstract

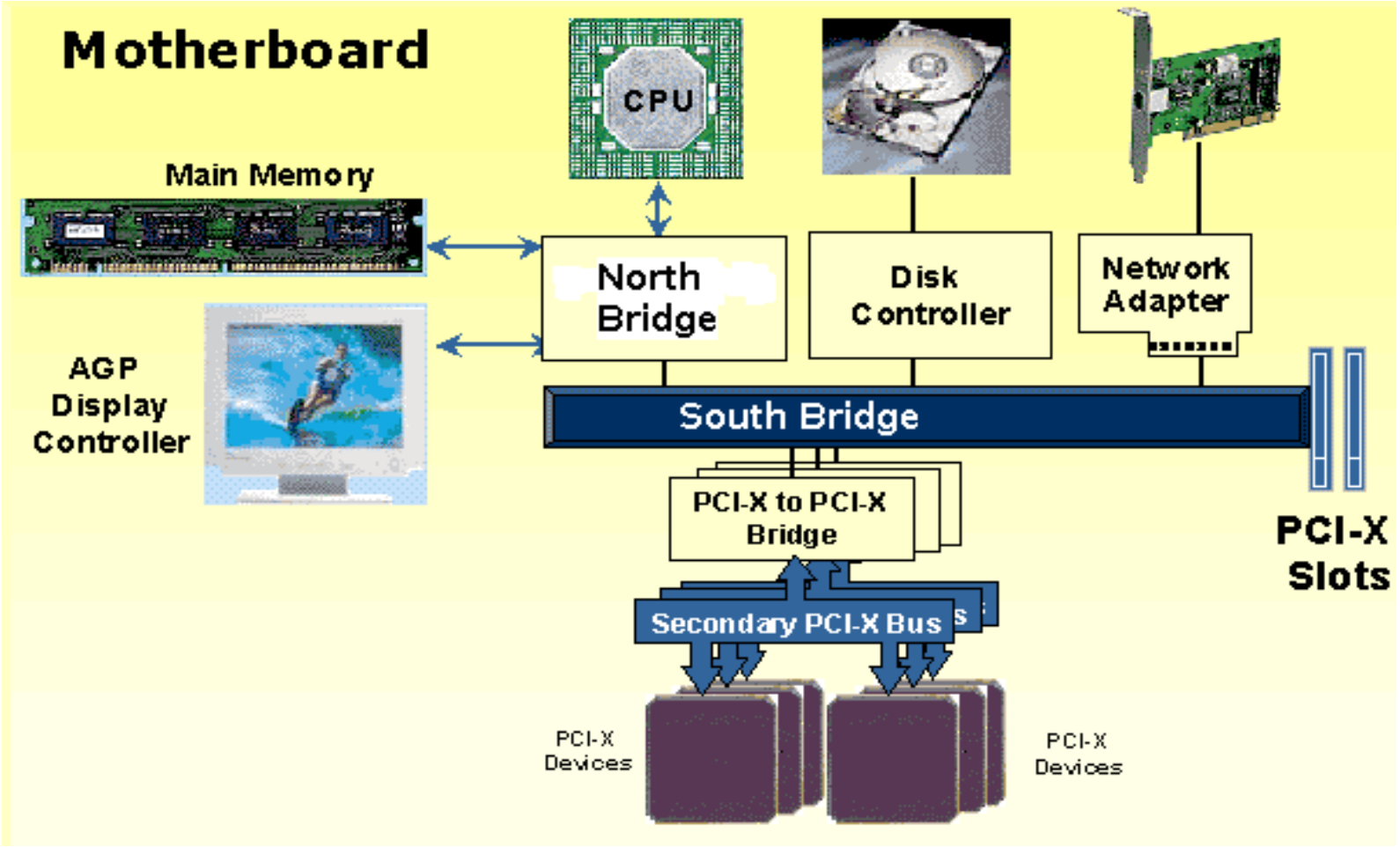
PCI Express Impact on Storage Architecture and Capabilities

- PCI Express Gen2 and Gen3, IO Virtualization, FCoE, SSD are here or coming soon. This session describes PCI Express, Single Root and Multi Root IO Virtualization and discusses the implications on FCoE, SSD and impacts of all these changes on storage connectivity, storage transfer rates. This tutorial will provide the attendee with:
  - Basic knowledge of PCI Express Architecture, PCI Express Roadmap, System Root Complexes and IO Virtualization
  - Expected Industry Roll Out of latest IO Technologies and required Root Complex capabilities
  - Implications on FCoE, SSD and IO to Storage Connectivity
  - Anticipated Impacts of these Technologies on Storage Environments
  - IO Virtualization connectivity possibilities in the Data Center (via PCI Express)

Agenda

- IO Architectures
  - PCI Changing to PCI Express
  - 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

Typical PCI Implementation



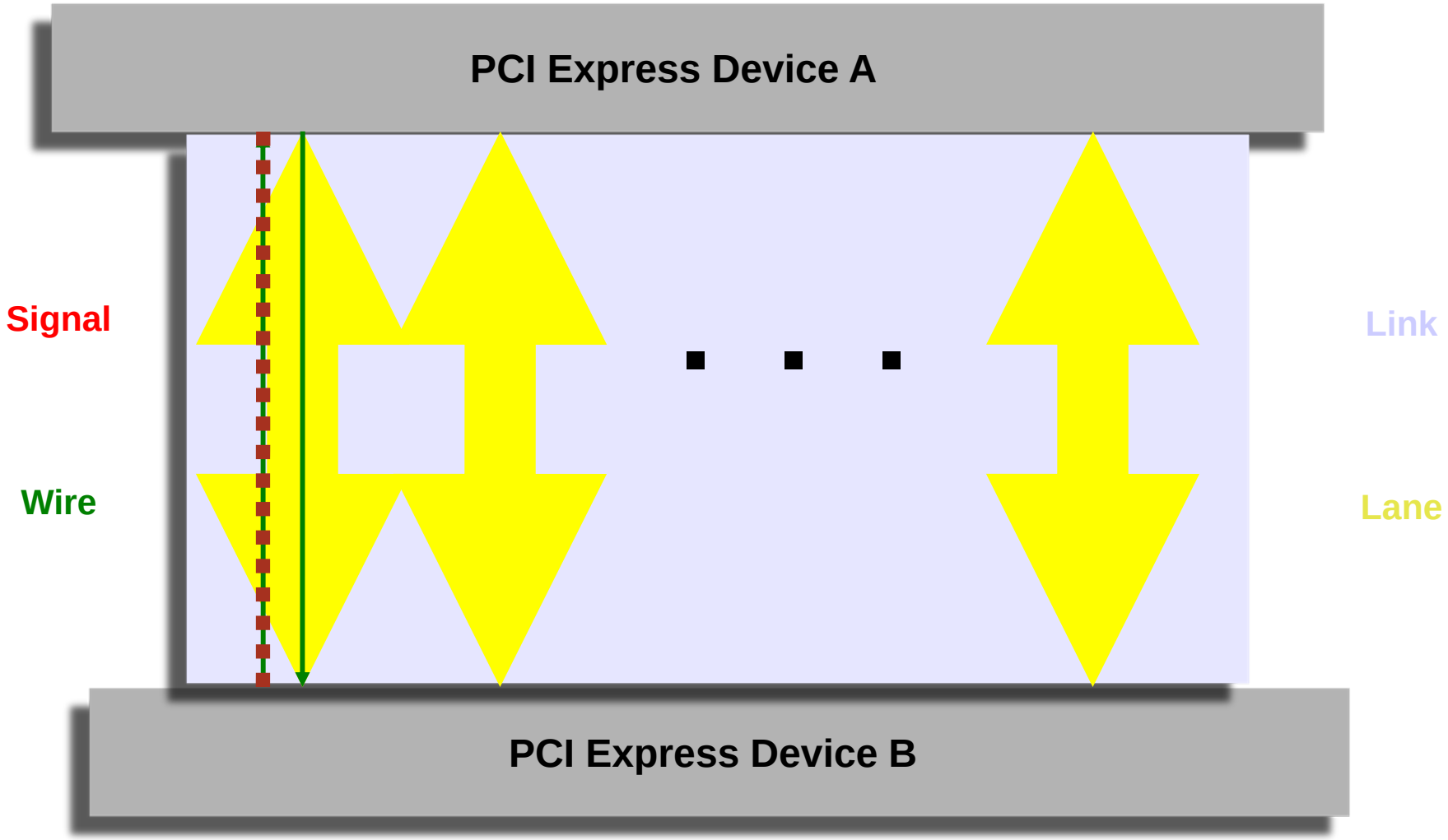
Changing I/O Architecture

- 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

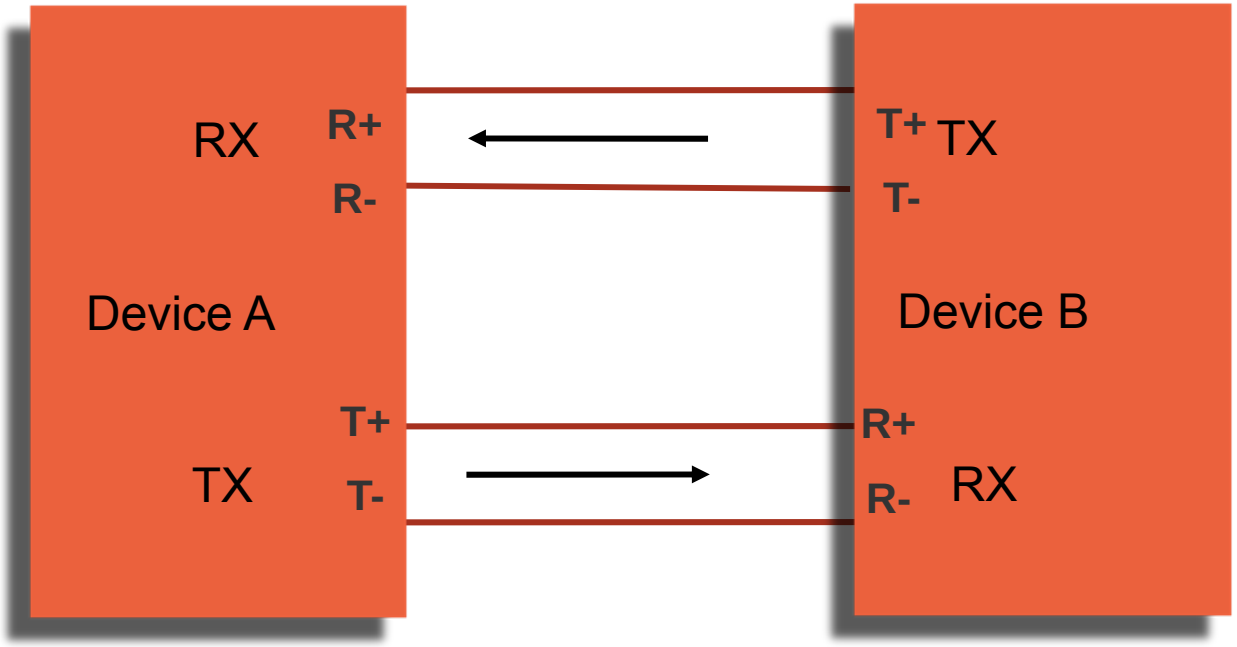
- PCI Express Architecture is a high performance, IO interconnect for peripherals in computing/ communication platforms
- Evolved from PCI and PCI-X™ Architectures
  - Yet PCI Express architecture is significantly different from its predecessors PCI and PCI-X
- PCI Express is a serial point- to- point 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 Terminology



PCIe What's A Lane

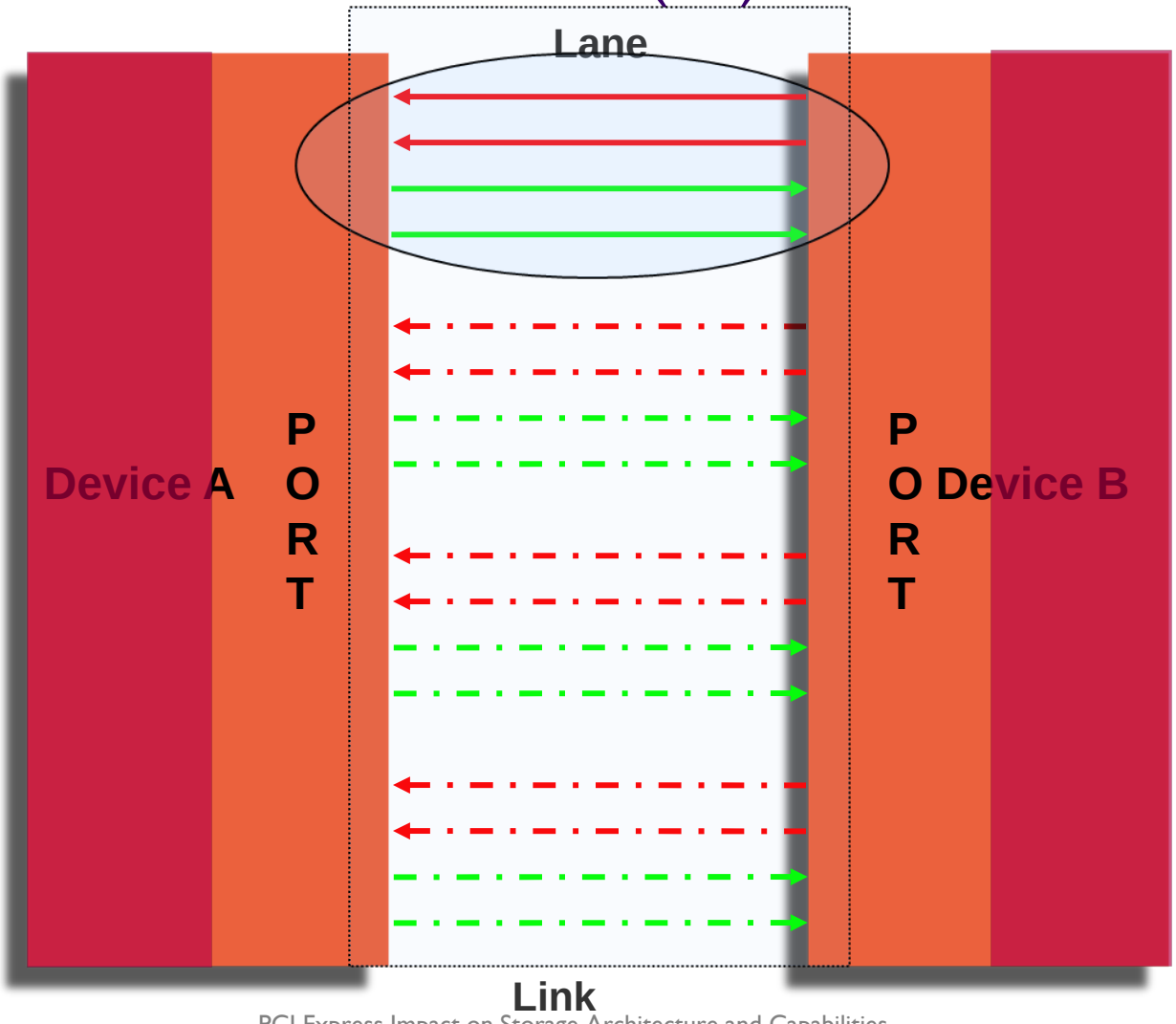
Point to Point Connection Between Two PCIe Devices



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

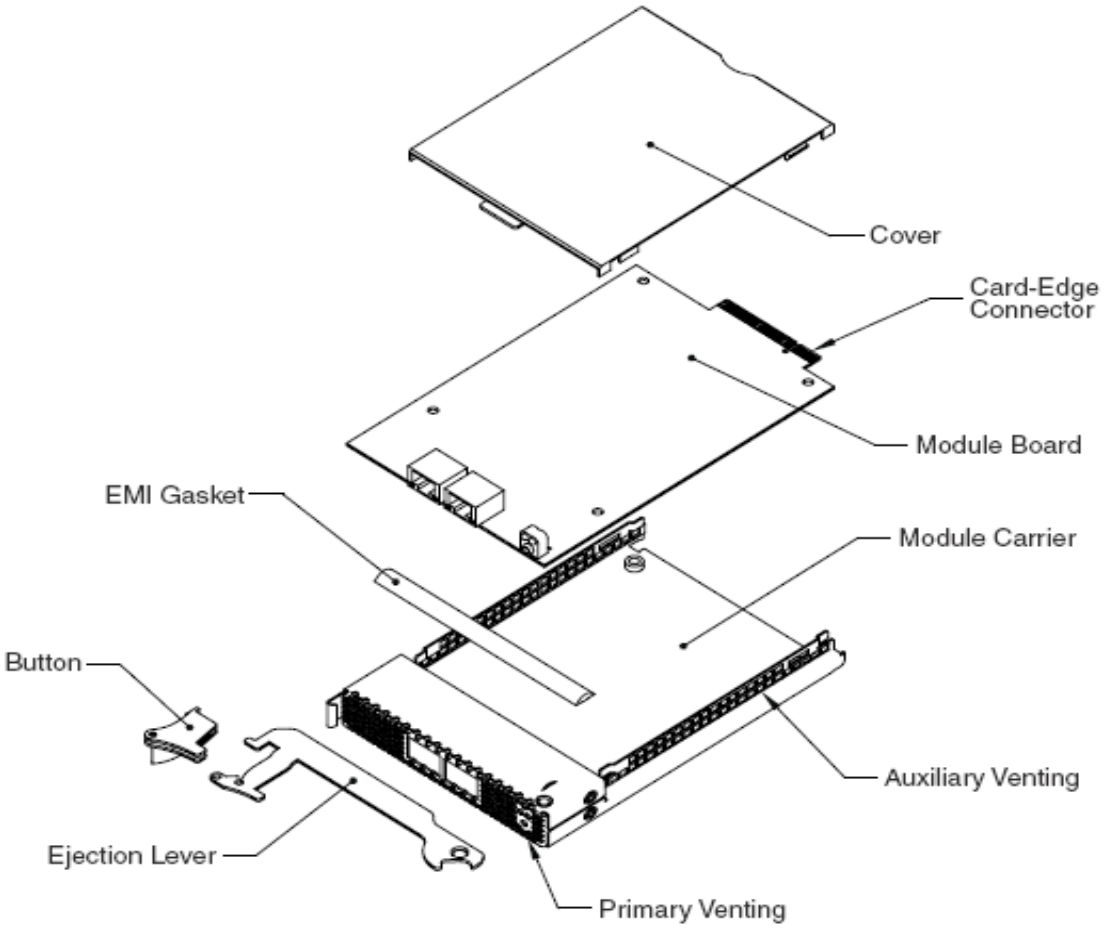


PCI Express Overview

- 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

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



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)

Education SNIA

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.
  - Supports Message Signal Interrupt similar to PCI-X
  - Encodes legacy PCI Interrupt signals (INT) within Message Transactions

PCI Express Throughput

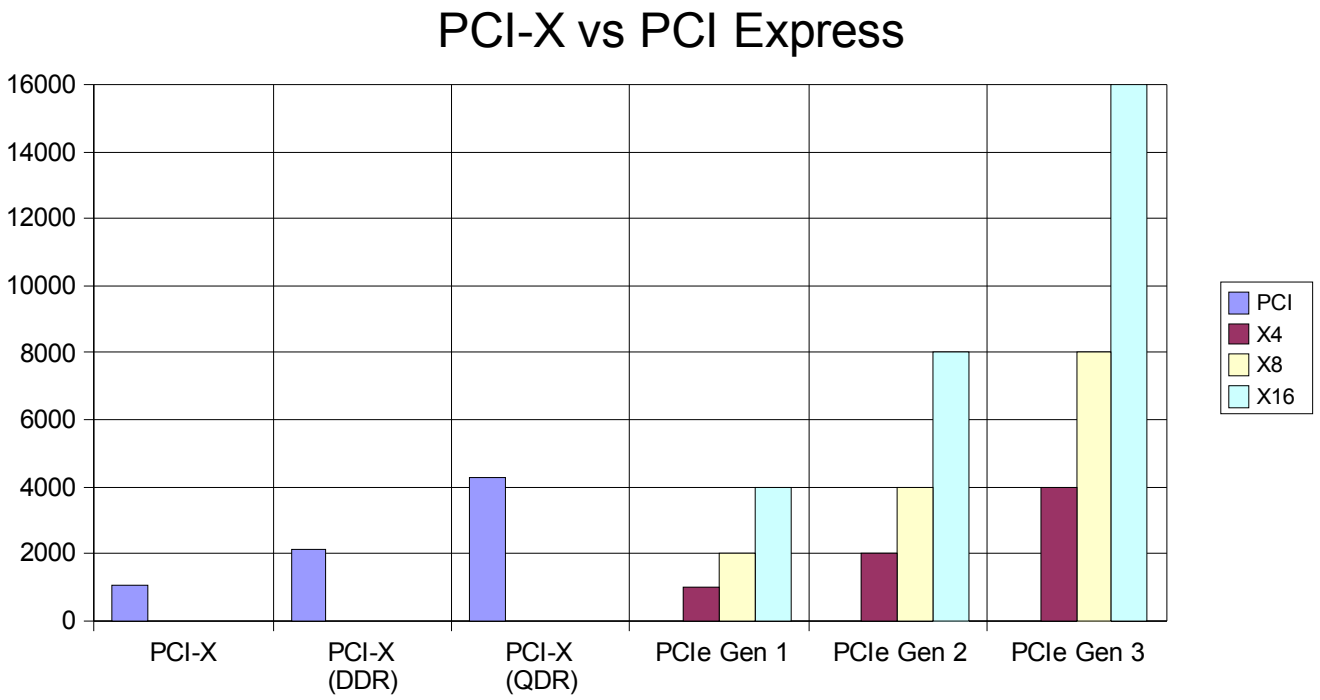
Link Width		X1	X2	X4	X8	X16	X32
Aggregate BW (Gbytes/s)	Gen1 (2004)	0.5	1	2	4	8	16
	Gen2 (2007)	1	N/A	4	8	16	32
	Gen3 (2010)	2	N/A	8	16	32	64

- 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

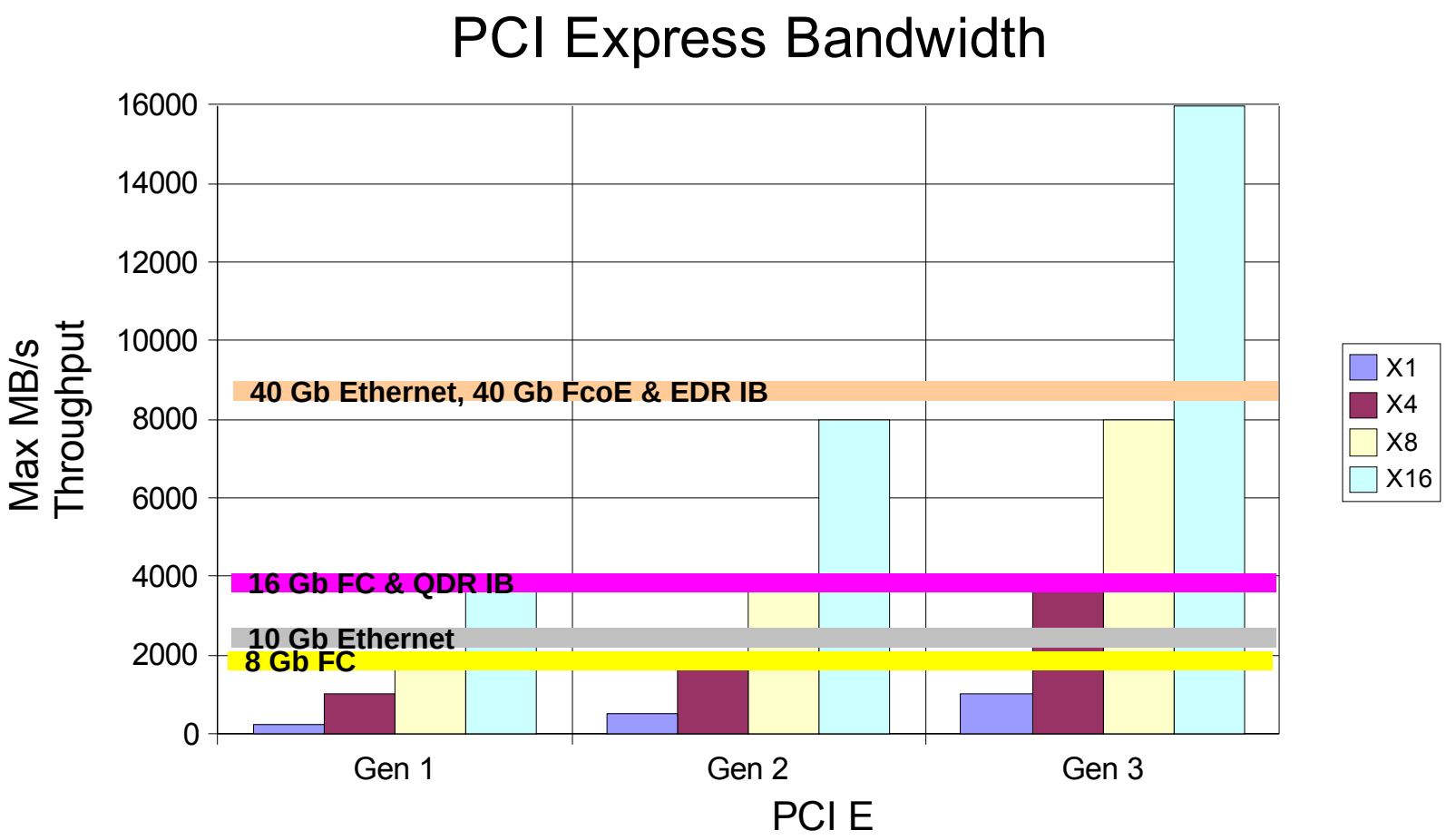
PCI-X vs PCI Express Throughput

How does PCI-X compare to PCI Express?

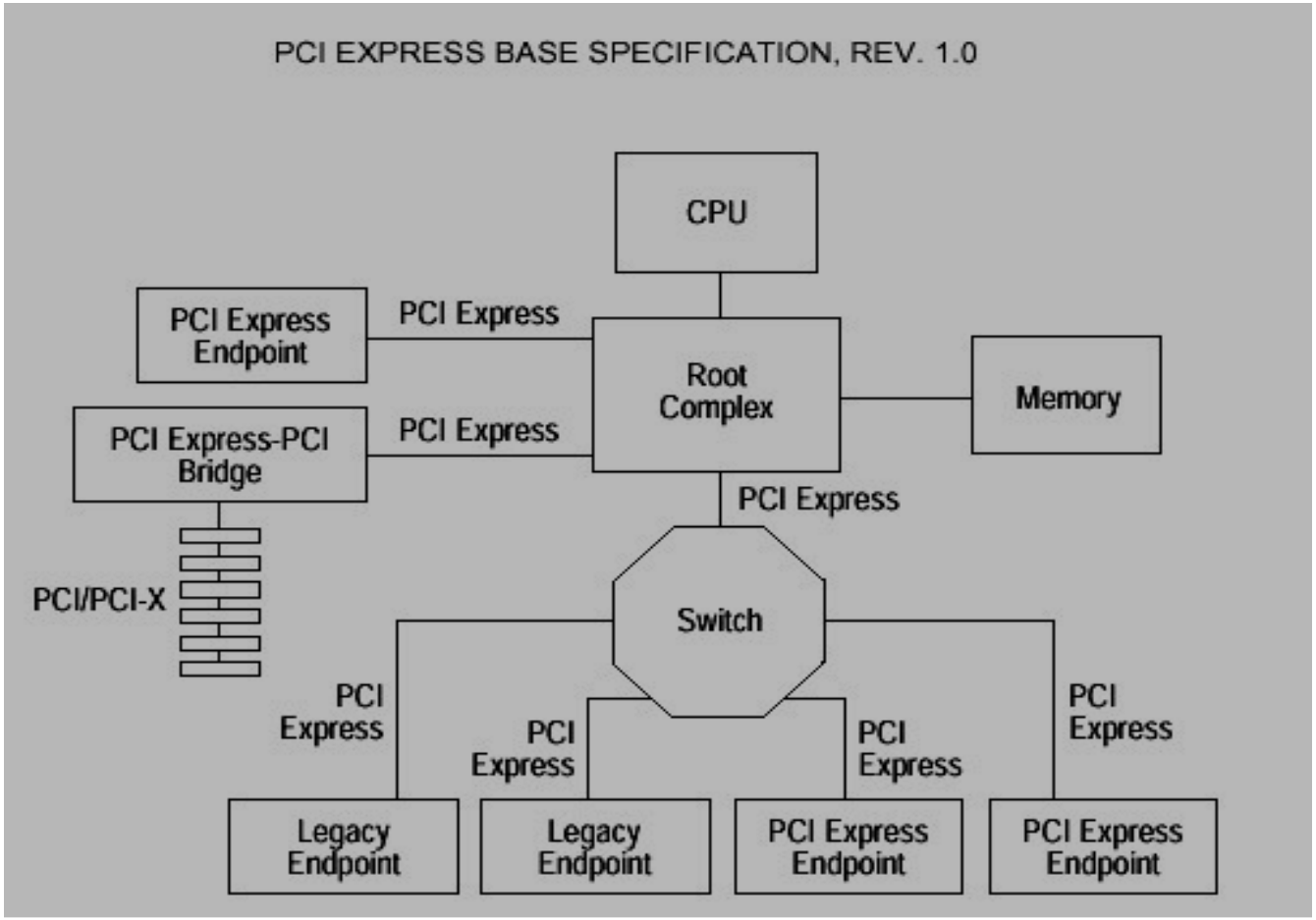


- PCI-X QDR maxs out at 4263 MB/ s per leaf
- PCIe x16 Gen1 maxs out at 4000 MB/ s

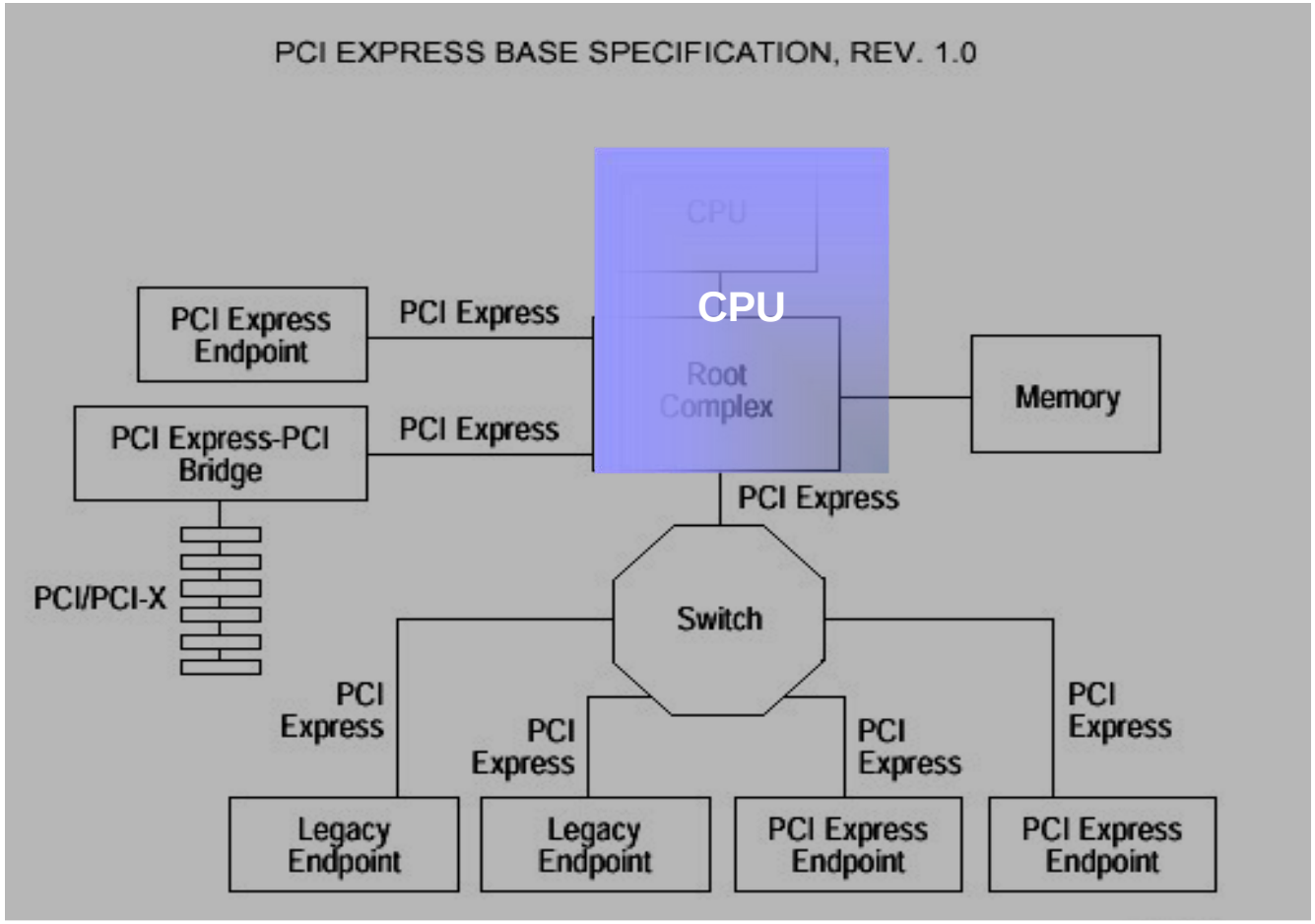
IO Bandwidth Needs



Sample PCI Express Topology



Sample PCI Express 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 Gen2 Switches and Gen1-to-PCI-X Bridges
  - Signal Conditioners, Extend PCIe 5M via Copper, 30M via Optical

PCI Express In Industry

- Gen 1 First Release of Slots in 2005
- Gen 2 Shipped Q4 CY2007
  - First (x16 Gen2) slots shipped in Q4 CY2007
  - Adoption is slower than Gen1
- Systems Currently Shipping
  - Desktops with multiple x16 connectors
  - Servers with multiple x4 and x8 connectors
- Cards Available
  - Gen 1 x4, x8 cards - 10 GbE, Dual/Quad GbE, 4/8 Gb FC, SAS, IB
  - Gen 2 x4, x8 cards – Dual 10 GbE, 8 Gb FC, QDR IB, SAS 2 (soon), QGE, Multi-protocols cards, 2D Graphics, FCoE (soon)
  - All the above are or will be available in EM (Express Module) form factor

Recent PCI Express Changes

- Power increase for Graphics Cards to 300 Watts
- Lanes can be grouped
  - 1x, 4x, 8x, 16x and 32x supported, x2 no longer supported
  - Must support all groupings lower than your width
- 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)
- External expansion
  - Copper connector and connector specified
- 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

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 from DDR to QDR / EDR
    - 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

Existing Serial Interfaces

- Established external transport mechanisms exist
  - Fibre channel - 4 Gb, 8Gb
    - Storage area network standard
  - Ethernet – 10Gb, 40Gb
    - iSCSI/TCP/IP/Ethernet provide a network based solution to SANs
  - InfiniBand - DDR, QDR, EDR
    - Choice for high speed process to processor links
    - Supports wide and fast data channels
  - SAS 1.0, 2.0 (3 Gb, 6 Gb)
    - Serial version of SCSI offers low cost solution
- No need to add to these yet another solution
  - PCI Express is not intended to replace these
  - But backplane IO must support these bandwidths

10G Ethernet or InfiniBand?

10G Ethernet

+ Provides a network based solution to SANs
- Requires Ethernet Switch in Box
- QoS mechanisms required
- Direct interface to root complex
- Low overhead stack
+ Fiber Channel over Ethernet

Check out SNIA Tutorials:

FcoE: Fiber Channel Over Ethernet Fabric Consolidation with InfiniBand

Ethernet Enhancements for Storage: Deploying FCoE

InfiniBand

+ Higher speed capability (4X QDR – 40 Gb/s, EDR – 80 Gb/s)
+ Established stack
+ Choice for high speed, low latency processor to processor links
- Requires IB Fabric
- Requires all protocols to move to IB based via TCAs

Backplane Serial Interfaces

- Inside the box solution today
  - Mid-plane or center-plane communication
  - Possible solutions are
    - 10 Gb Ethernet
    - InfiniBand
    - PCI Express
- Will it be 10GbE, InfiniBand, PCIe IOV ?
- Root In the Host Architectures are All PCIe, SR/MR IOV works with Root Complex,

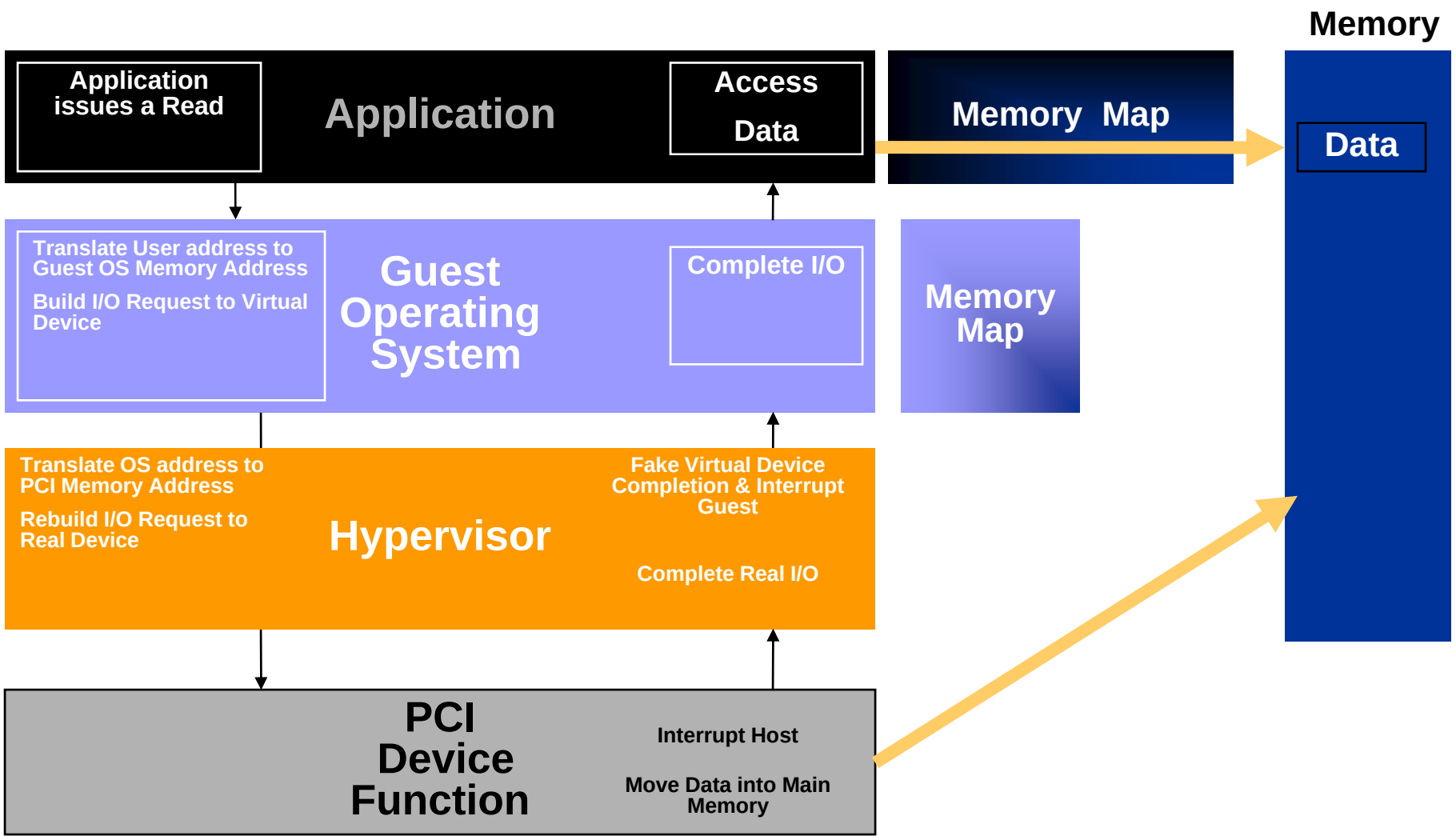
Or PCI Express

- Requires IO Virtualization
  - SR – Single Root
  - MR – Multi Root
- Based Upon PCI SIG Standards

Single Root IOV

Better IO Virtualization for Virtual Machines

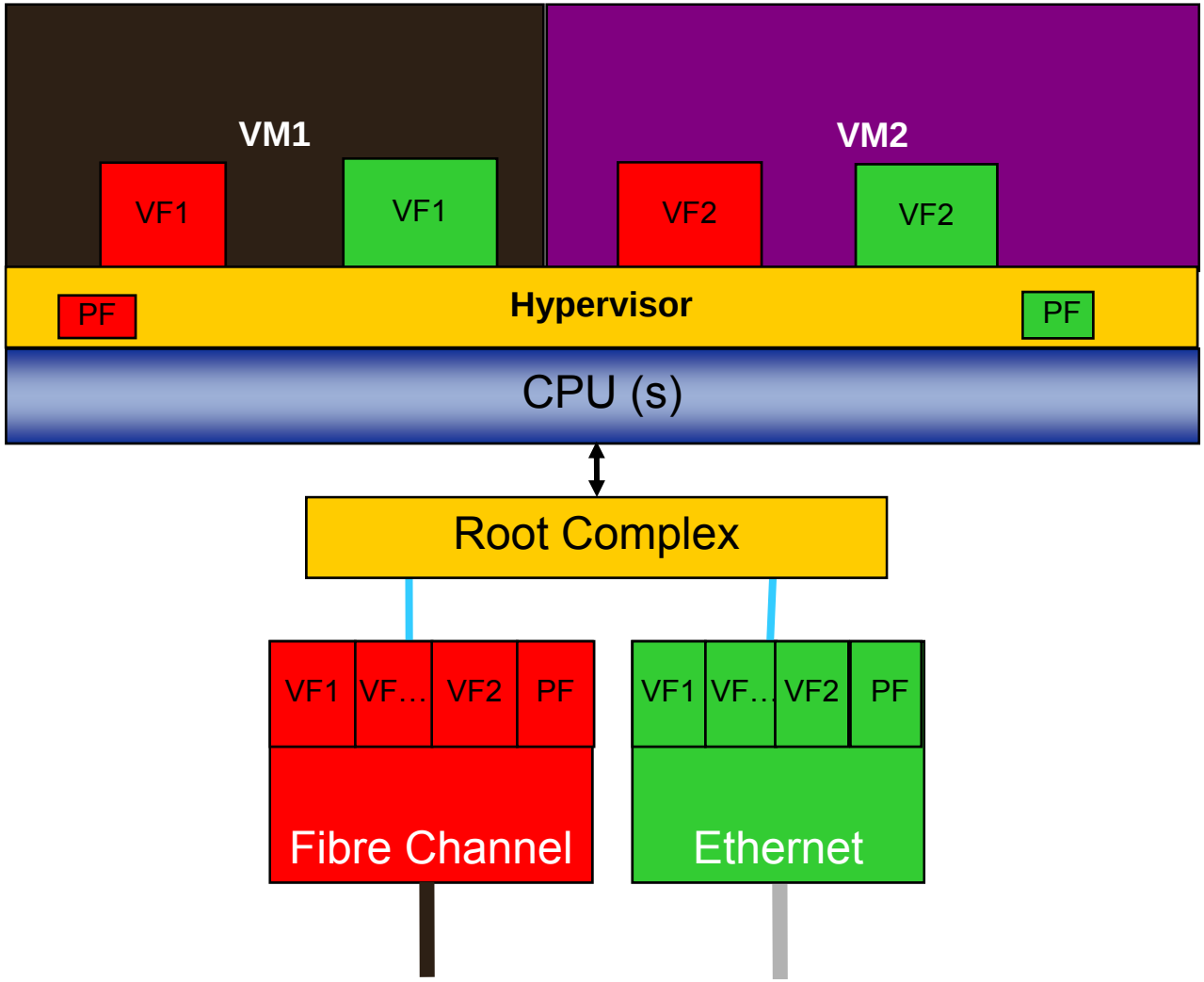
System I/O with a Hypervisor



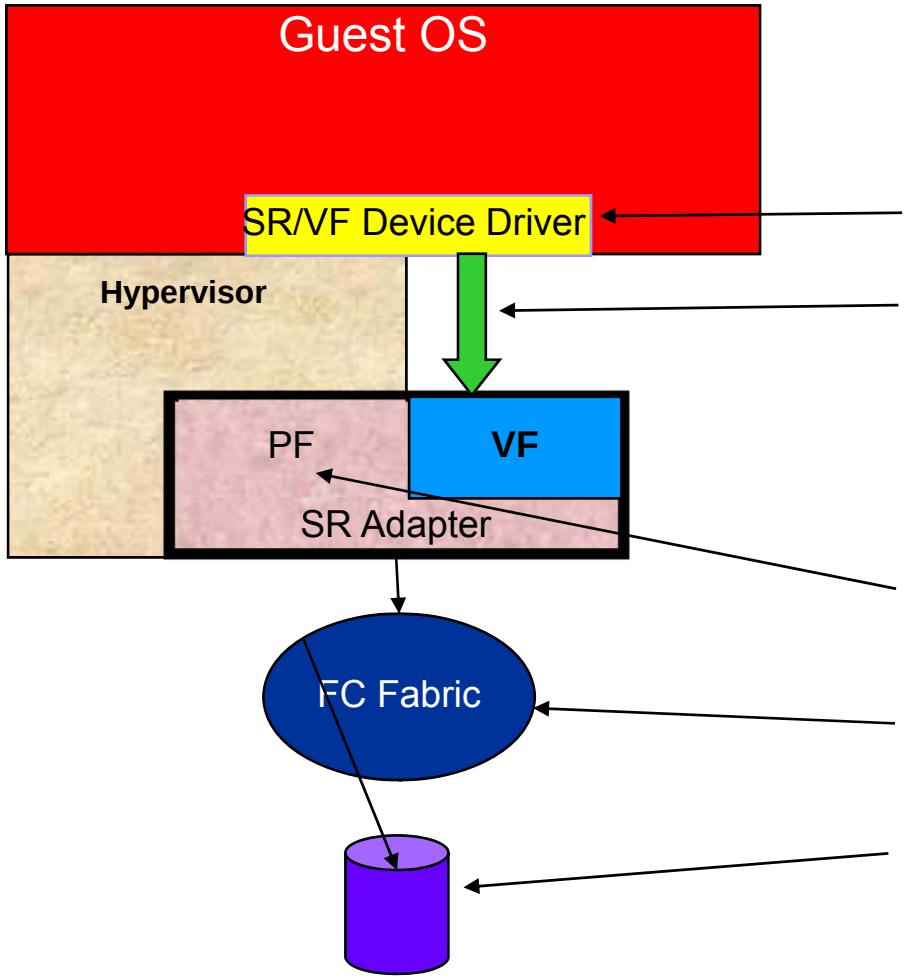
Single Root IOV

- 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 and SR Virtualization



SR Adapter Specific Driver (fabric aware)

I/Os go directly to adapter via VF

Hypervisor configures VF via PF

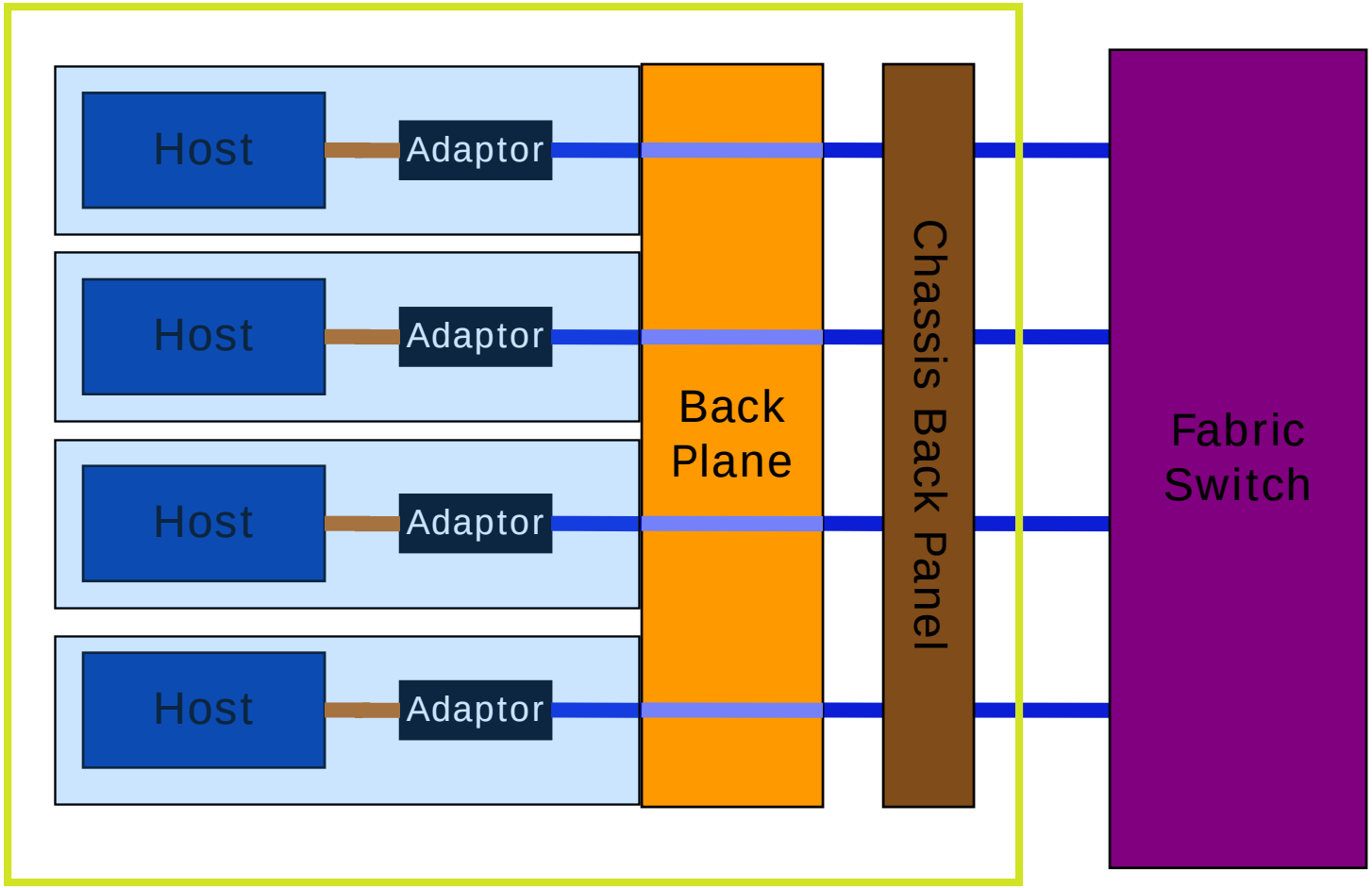
Fabric is visible to Host

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

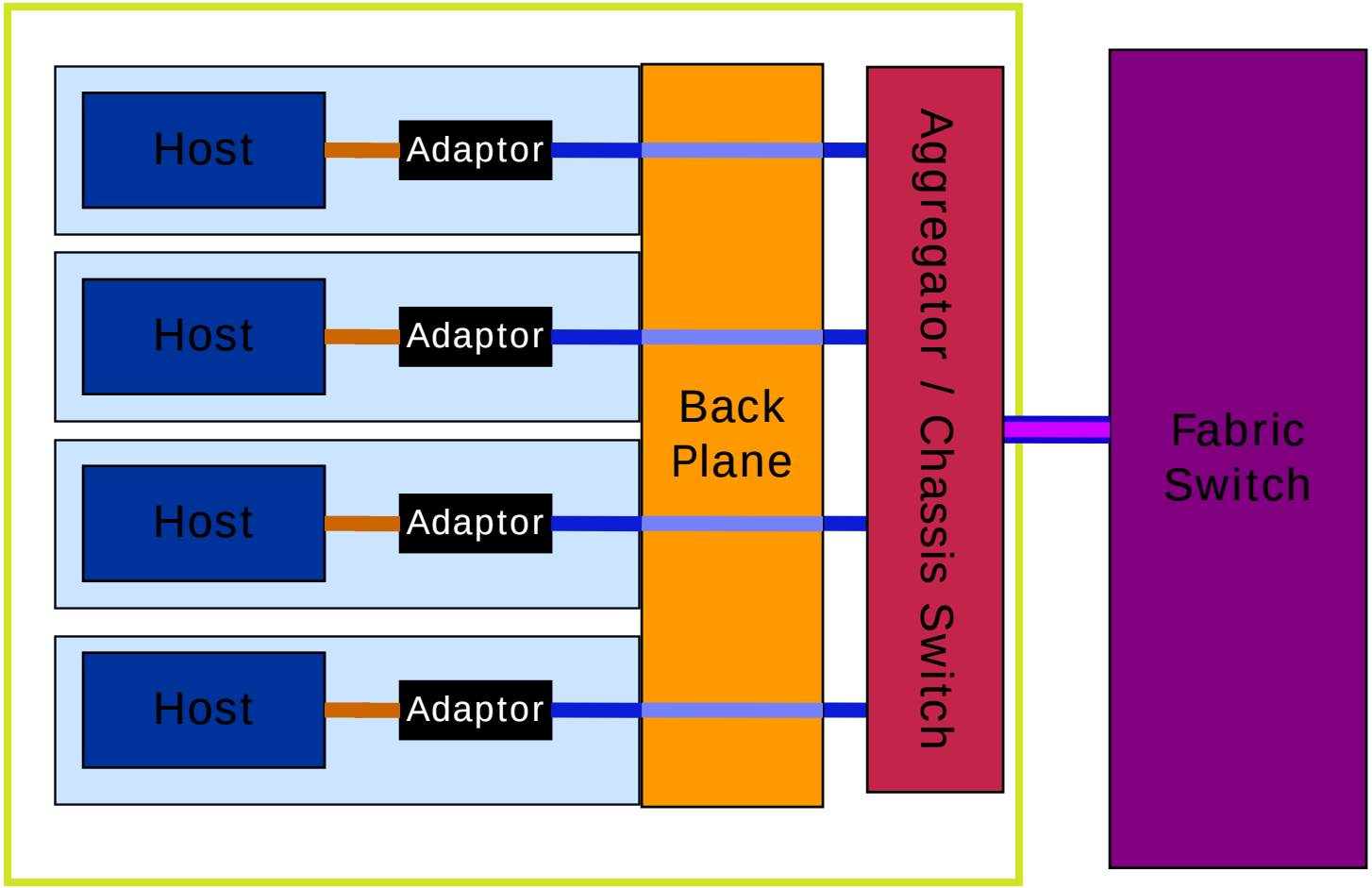
MultiRoot IOV

Virtualizing an IO adapter for multiple Hosts

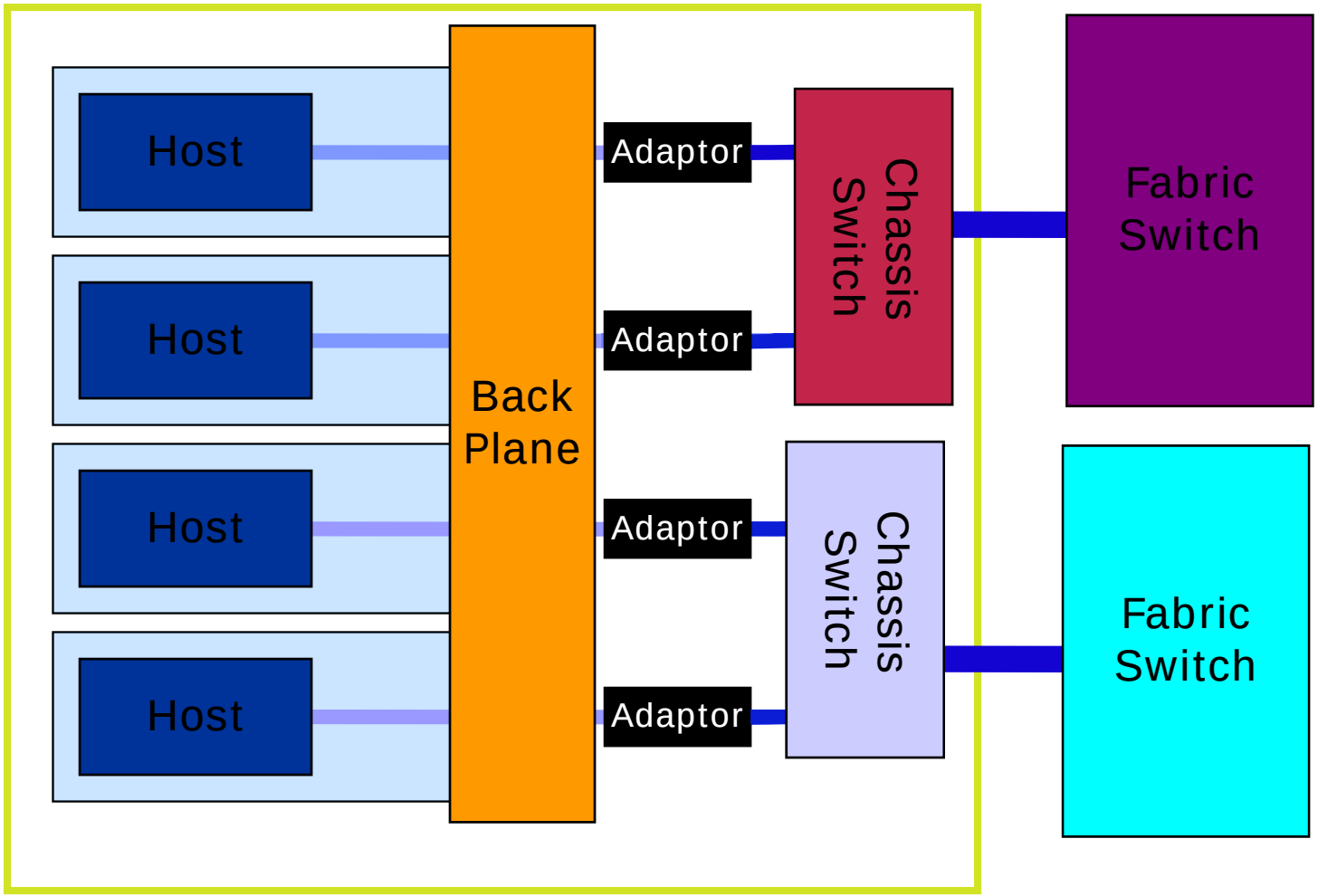
It Started With Pass Thru



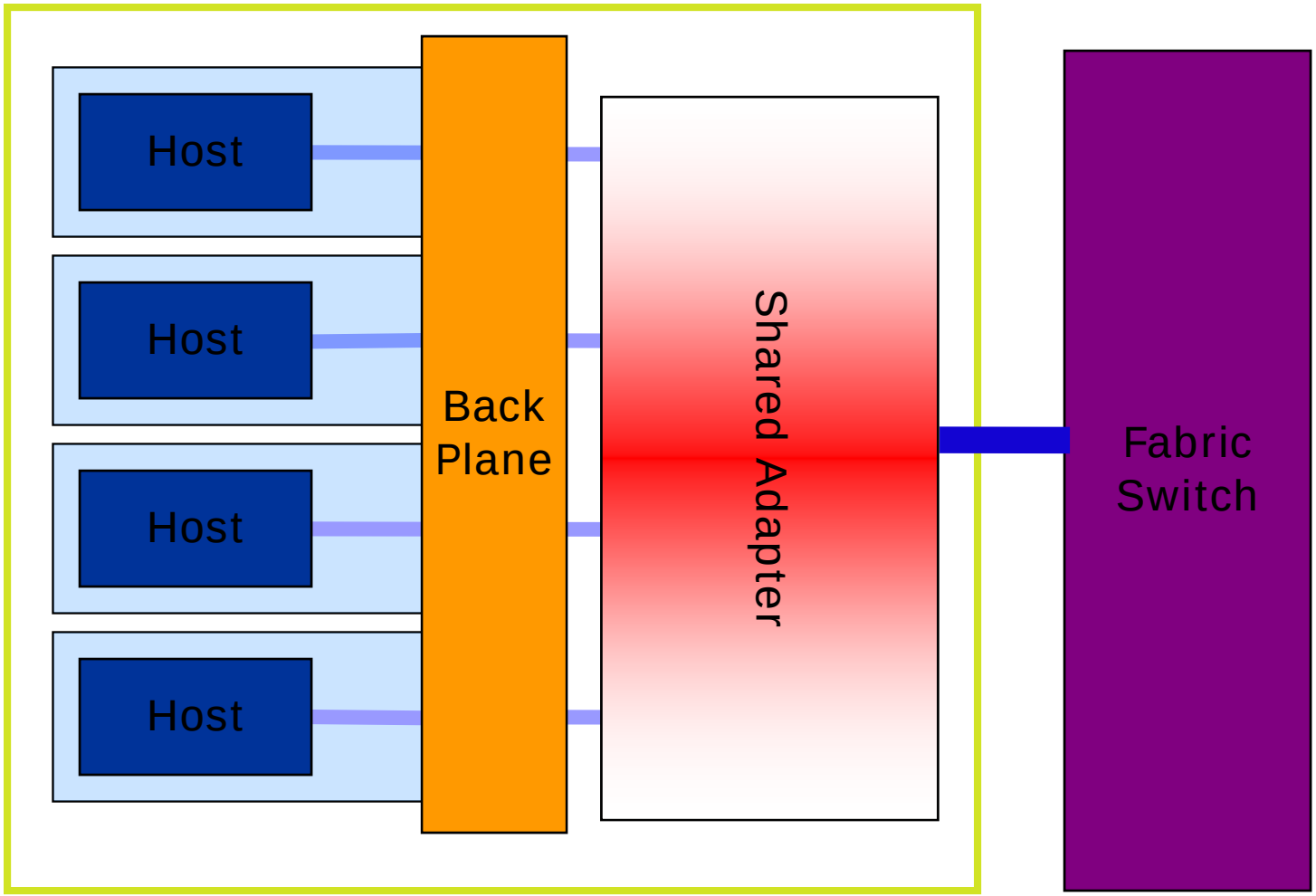
Chassis Switching to Aggregation



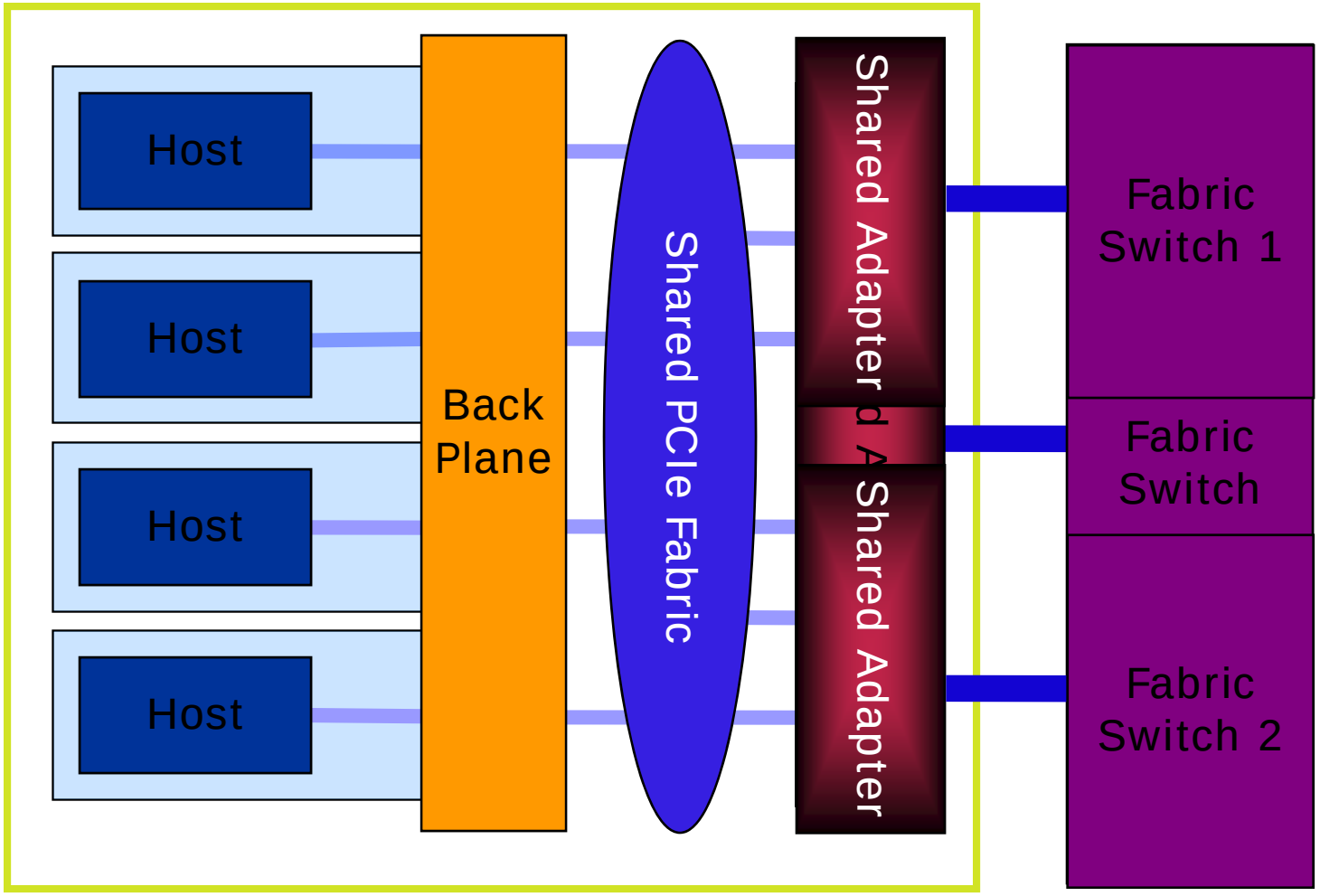
Stretch the PCIe Bus Across the Backplane



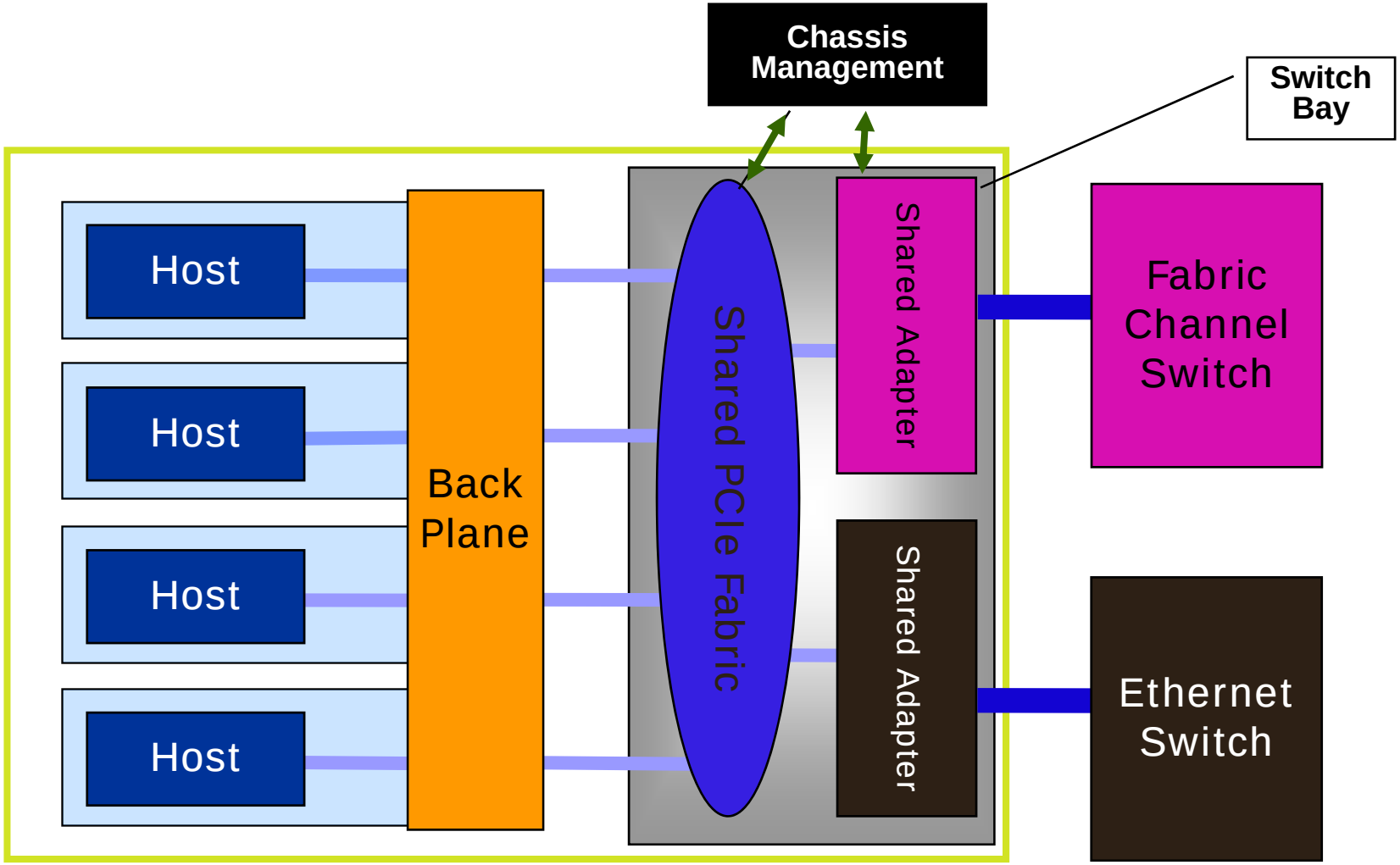
Merge the Adapter and Aggregator



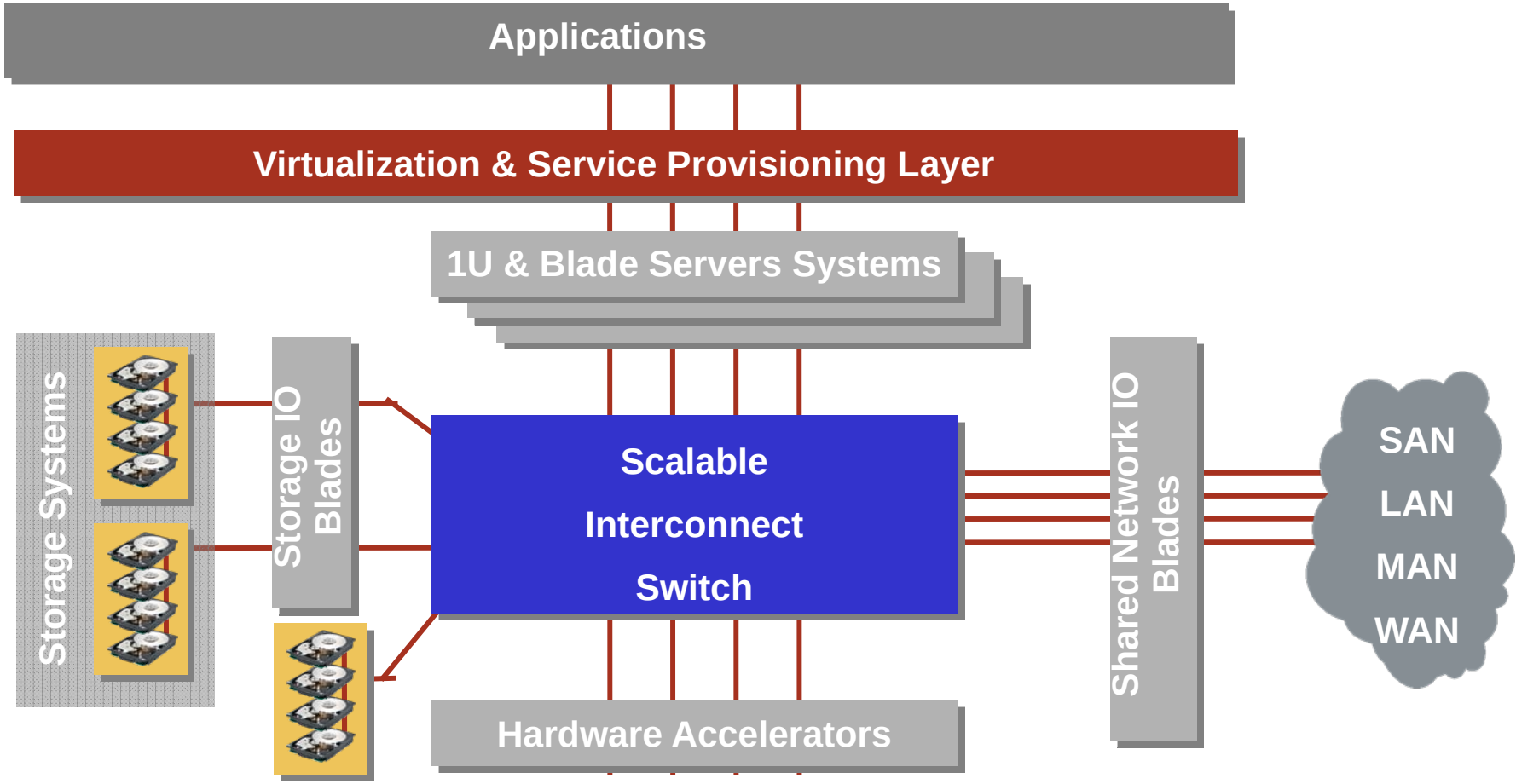
Insert a shared PCIe Fabric and add Multiprotocol



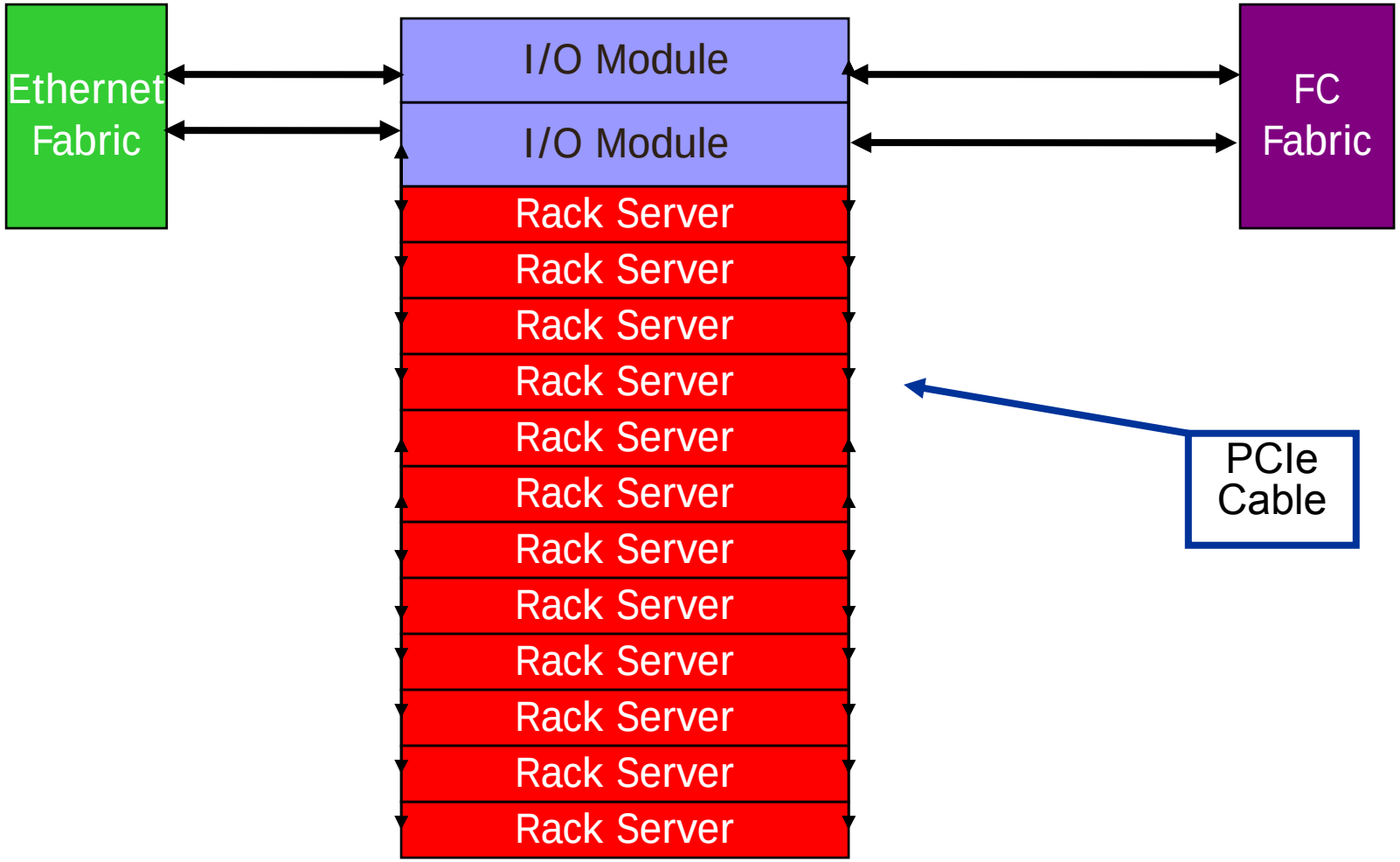
In a Blade Chassis



Mutli Root Virtualization in Blades



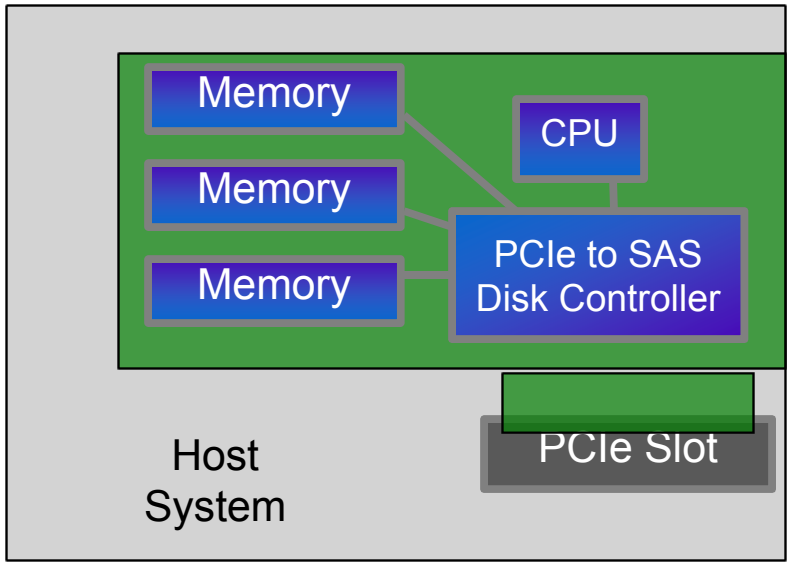
In a Rack



Impact / Benefit to Storage

- PCI Express provides
  - Full Bandwidth Dual Ported 4 & 8 Gb FC
  - Full Bandwidth Dual Ported 10 GbE/iSCSI/FCoE
  - Full Bandwidth for QDR and EDR IB
  - Full Bandwidth SAS 1.0 & 2.0
  - Legacy Support via PCI-X
  - Access to SSS via PCIe
- 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

PCIe and SSS Device on Card



- SCSI Command Set on Host
- JBOD or Raid SAS Disk Controller
- RDMA Over PCIe
  - Card CPU to System Memory
  - System CPU to SSS Device
- PCIe Encapsulation is Stripped Off

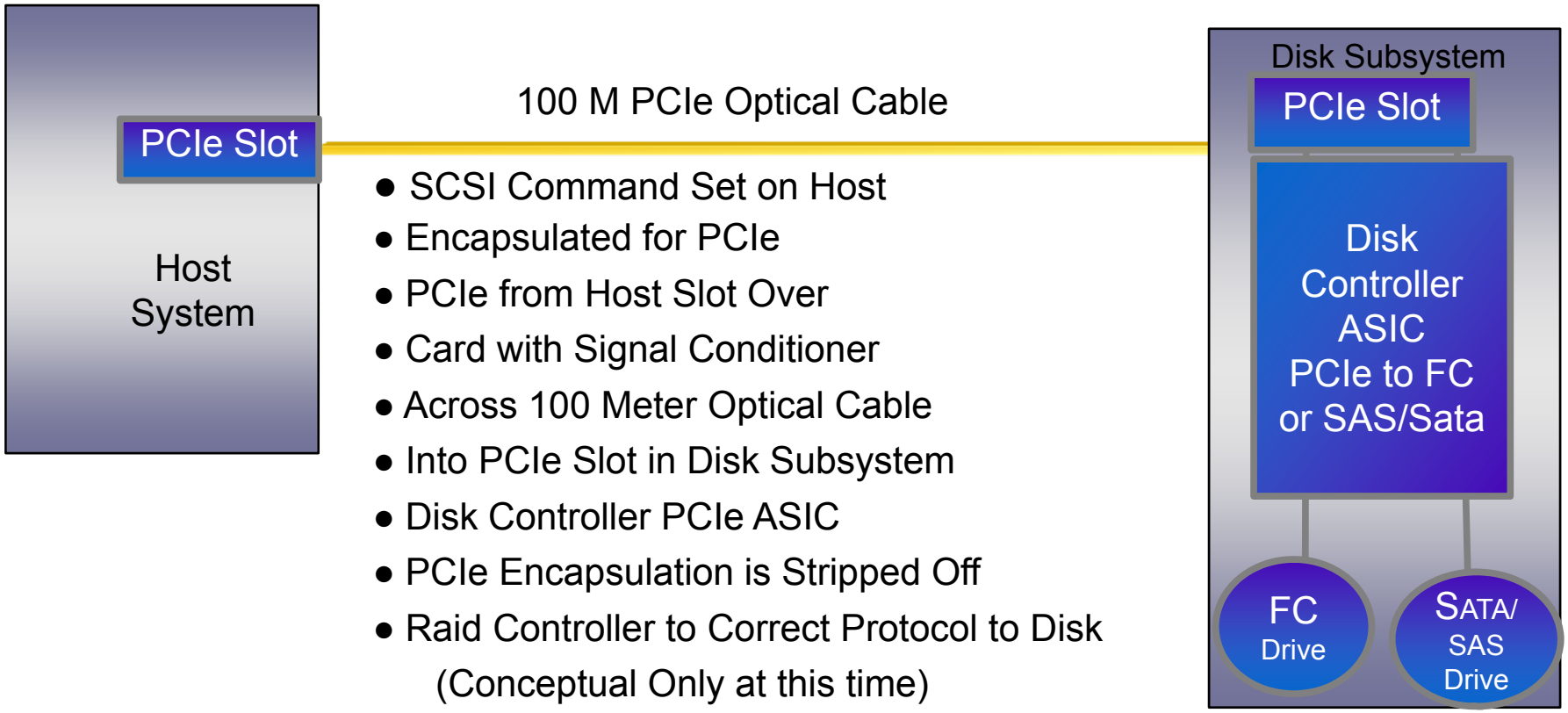
Available soon, from multiple vendors

Check out SNIA Tutorials:

Solid State Storage in a Hard Disk Package

Solid State Storage Reliability and Data Integrity

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

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

Howard Goldstein
Alex Nicolson
Rob Peglar
Joe White