



Education

# PCI Express and Storage

*Ron Emerick, Sun Microsystems*

# SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA.
- Member companies and individuals may use this material in presentations and literature under the following conditions:
  - ◆ Any slide or slides used must be reproduced without modification
  - ◆ The SNIA must be acknowledged as 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.

## PCI Express and Storage

- ◆ System IO Architectures are changing to PCI Express, 10 GbE and InfiniBand. As multi-root IO Virtualization is being defined, shared IO infrastructures are on the horizon. This session discusses the impacts of all these changes on storage connectivity, storage transfer rates, as well as the implications to Storage Industry and Data Center Infrastructures. This tutorial will provide the attendee with:
  - ◆ Basic knowledge of PCI Express and System Root Complexes and IO Virtualization.
  - ◆ Anticipated Impacts (benefits and exposures) of these Technologies on Storage Environments.
  - ◆ IO Virtualization connectivity possibilities provided by PCI Express.

## ➤ Impact of Server Architectures on I/O

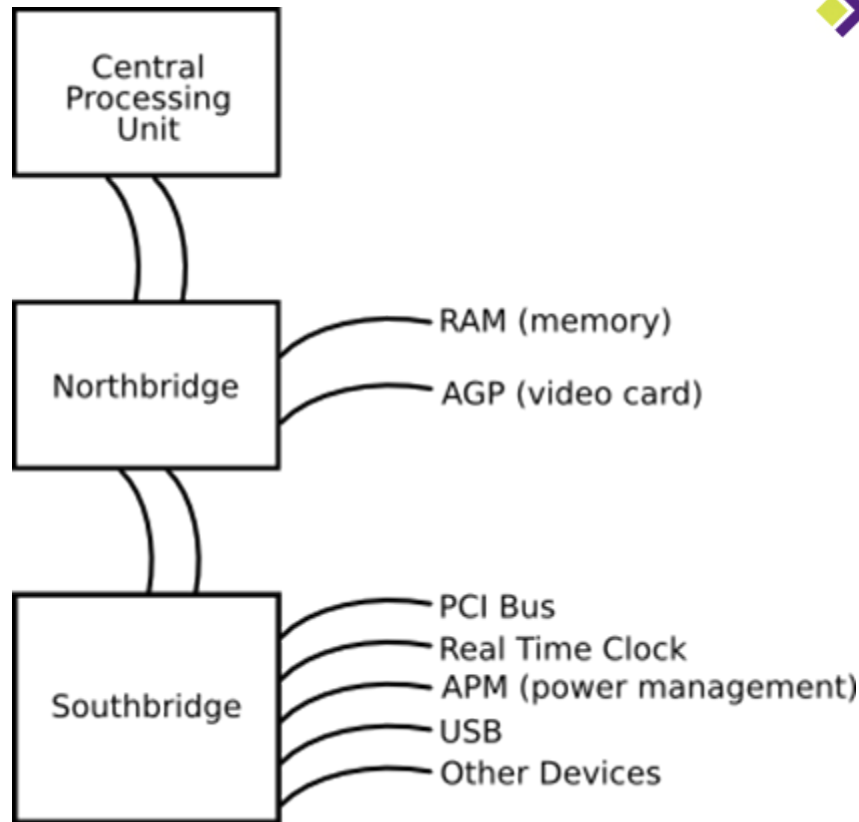
- ◆ Evolving PCI Architectures
- ◆ Current need to change
- ◆ New PCI Express based architectures
  - Review of PCI Express

## ➤ I/O Evolving Beyond the Motherboard

- ◆ Serial Interfaces
  - GbE & 10 GbE
  - InfiniBand
  - PCIe IO Virtualization
- ◆ Review of InfiniBand
- ◆ Review of PCI Express IO Virtualization
- ◆ Impact of PCI Express on Storage

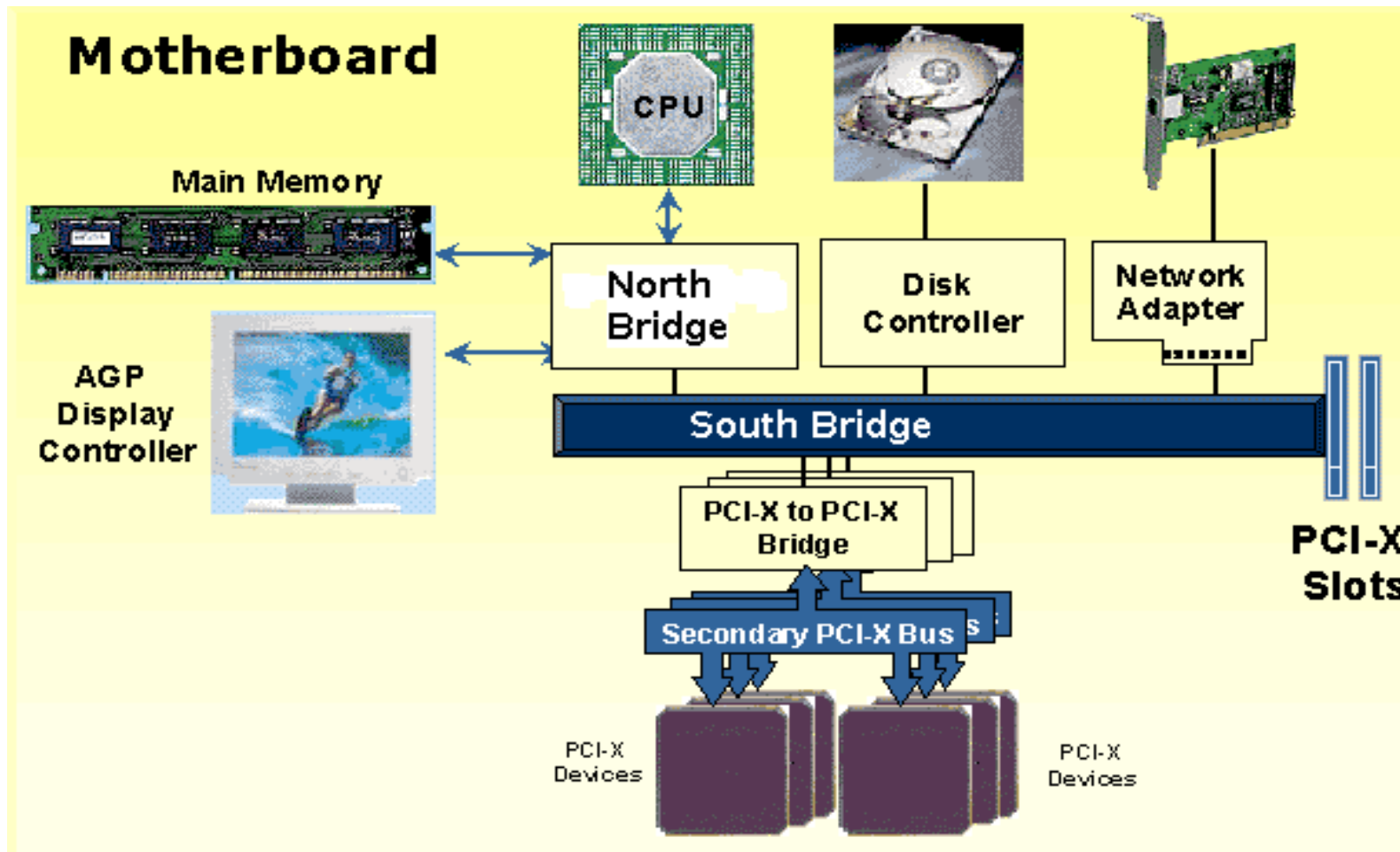
# Yesterdays Architecture

## ➤ IO Architecture



- ◆ North Bridge (Root Complex)
  - The portion of a computer chipset that connects between the CPU and the major interfaces on the computer including memory, AGP port and South Bridge.
- ◆ South Bridge
  - Connects legacy I/O, USB and the PCI Bus

# Typical PCI Implementation



# Changing I/O Architecture

- PCI provides a solution to connect processor to I/O
  - ◆ 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
  - ◆ SCSI to SAS
- Move parallel PCI to serial PCI Express

# PCIe Physical Connection

## ➤ Higher Throughput on Fewer Pins

- ◆ X1 Link consists of 1 Lane or 1 Differential Pair in Each Direction
  - Uses 4 signal Pins
- ◆ X16 Link or 16 Lanes consists of 16 Signal Pairs
  - Uses 64 signal Pins (62.5 MB/pin/sec)
- ◆ PCIX 533 Mhz
  - Uses 150 Pins (28.4 MB/pin/sec)

## ➤ PCIe - Point-to-Point Interconnect

- ◆ Switches provide expansion
- ◆ PCIe-to-PCIX Bridges provide legacy support

# PCI Express Overview(1 of 2)

- Uses PCI constructs
  - ◆ Same Memory, I/O 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
  - ◆ I/O can exchange data
  - ◆ System boards can exchange data

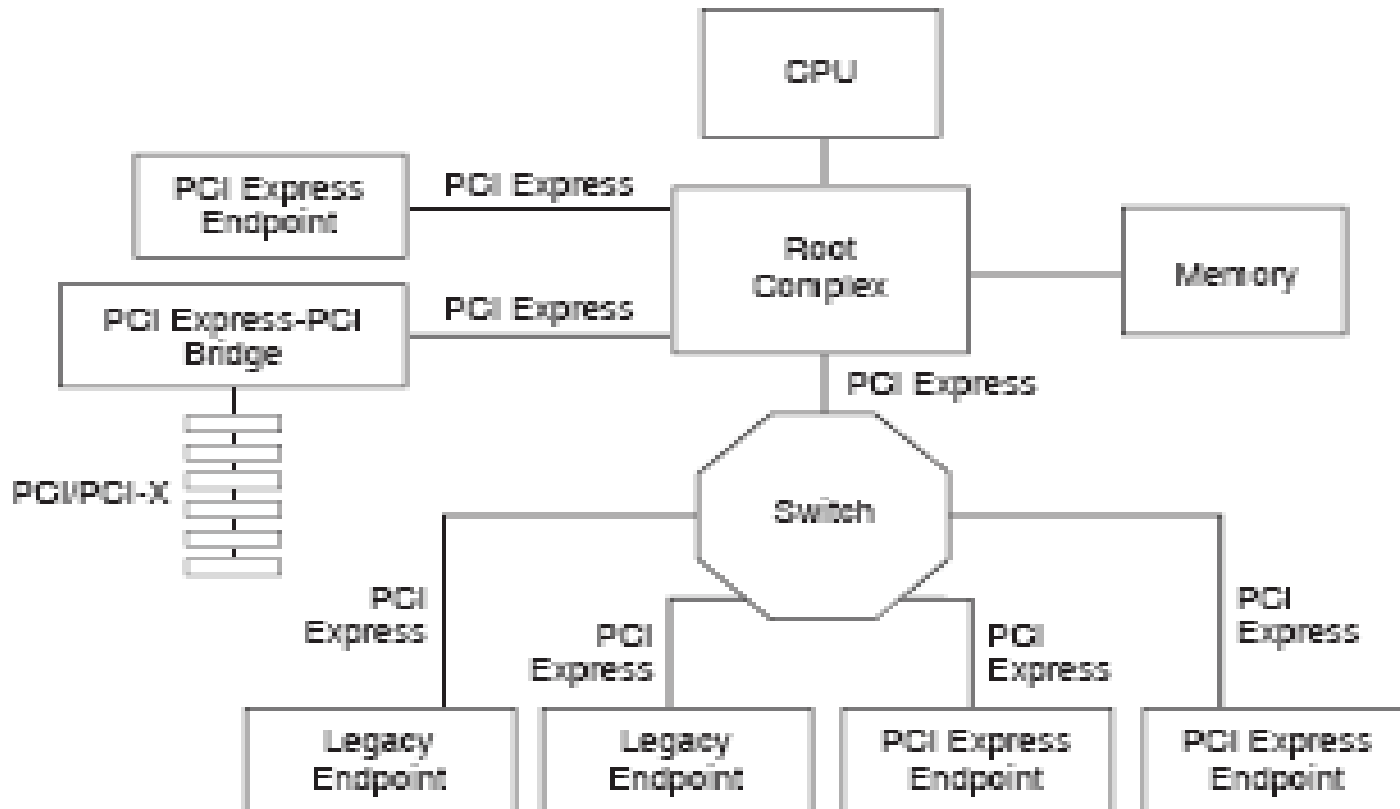
# PCI Express Overview(2 of 2)

- Receive and Transmit lanes
- Lane data rate
  - ◆ Currently 2.5Gbits/sec (SDR) with 8/10bit encoding
- Lanes can be grouped
  - ◆ 1x, 2x, 4x, 8x, 16x and 32x supported
- External expansion
  - ◆ Optical Cable and connector specified in 1.1 spec

# Recent PCI Express Changes

- Power increase for Graphics Cards to 300Watts
- Lanes can be grouped
  - ◆ 1x, 4x, 8x, 16x and 32x supported
  - ◆ Must support all groupings lower than your width
  - ◆ 2x no longer supported
- Performance roadmap
  - ◆ Gen 2.0 Doubled to 5Gbits/sec (DDR) with 8/10bit encoding
  - ◆ Gen 3.0 Doubles again to 8Gbits/sec (no 8b/10b)
- External expansion
  - ◆ Copper connector and connector specified
- Geneseo enhancements to PCIe 2.0 (IDF 2006)
  - ◆ Standard for co-processors, accelerators
    - > Encryption, visualization, mathematical modeling

# Sample PCIe Topology



OM13751

**Figure 1-2: Example Topology**

# Comparison of Freq & Slots

Bus Type	Clock Frequency	Peak Bandwidth	Number of Card Slots per Bus
<b>PCI</b>	33 Mhz	266 MB/s @ 64 bit	4-5
<b>PCI</b>	66 Mhz	533 MB/s @ 64 bit	1-2
<b>PCI-X</b>	133 Mhz	1066 MB/s @ 64 bit	1-2
<b>PCI-X (DDR)</b>	266 Mhz	2131 MB/s @ 64 bit	1
<b>PCI-X (QDR)</b>	533 Mhz	4263 MB/s @ 64 bit	1

<b>PCI-e x1</b>	1 lane @ 2500 Mhz	250 MB/s FD (500 MB/s) 2500 Mbits/s @ 1 bit (8/10 bit encoding)	Point to Point Switched
<b>PCI-e x8</b>	8 lanes @ 2500 Mhz	2000 MB/s FD (4000 MB/s) 2500 Mbits/s/lane @ 1 bit (8/10 bit encoding)	Point to Point Switched
<b>PCI-e x16</b>	16 lanes @ 2500 Mhz	4000 MB/s FD (8000 MB/s) 2500 Mbits/s/lane @ 1 bit (8/10 bit encoding)	Point to Point Switched

# Comparison of Freq & Slots

Architecture Type	Clock Frequency	Peak Bandwidth	Number of Card Slots per Bus
<b>PCI-X</b>	133 Mhz	1066 MB/s @ 64 bit	1-2
<b>PCI-X (DDR)</b>	266 Mhz	2131 MB/s @ 64 bit	1
<b>PCI-X (QDR)</b>	533 Mhz	4263 MB/s @ 64 bit	1

<b>PCIe 1.1 – x8 Gen 1</b>	8 lanes @ 2500 Mhz	2000 MB/s FD (4000 MB/s) 2500 Mbits/s/lane @ 1 bit (8/10 bit encoding)	Point to Point Switched
<b>PCIe 2.0 – x8 Gen 2</b>	8 lanes @ 2500 Mhz	4000 MB/s FD (8000 MB/s) 5000 Mbits/s/lane @ 1 bit (8/10 bit encoding)	Point to Point Switched

# Comparison of Freq & Slots

Architecture Type	Clock Frequency	Peak Bandwidth	Number of Card Slots per Bus
<b>PCI-X</b>	133 Mhz	1066 MB/s @ 64 bit	1-2
<b>PCI-X (DDR)</b>	266 Mhz	2131 MB/s @ 64 bit	1
<b>PCI-X (QDR)</b>	533 Mhz	4263 MB/s @ 64 bit	1

<b>PCIe 1.1 – x8 Gen 1</b>	8 lanes @ 2500 Mhz	2000 MB/s FD (4000 MB/s) 2500 Mbits/s/lane @ 1 bit (8/10 bit encoding)	Point to Point Switched
<b>PCIe 2.0 – x8 Gen 2</b>	8 lanes @ 2500 Mhz	4000 MB/s FD (8000 MB/s) 5000 Mbits/s/lane @ 1 bit (8/10 bit encoding)	Point to Point Switched
<b>PCIe 3.0 – x8 Gen 3</b>	8 lanes @ 8000 Mhz	8000 MB/s FD (16000 MB/s) 8000 Mbits/s/lane @ 1 bit (no encoding)	Point to Point Switched

# Throughput Requirements

Adapter Type	Expected Throughput	Theoretical Maximum	Released	Data Center Rollout
Quad Gb Ethernet	3.6+ Gb/s	424 MB/s	Now	Now
10 GbE Port	9+ Gb/s FD	1275 MB/s	Now	2006
10 Gb iSCSI w/ TOE	9+ Gb/s FD	1275 MB/s	2006-7	2007-8
InfiniBand 4X-SDR	7 Gb/s	8 Gb/s	Now	2005-6 (clustering)
InfiniBand 4X-DDR	15 Gb/s	16 Gb/s	Now	Now
SAS 2.0 Quad Port	21+ Gb/s	24 Gb/s	2007	2009
4 Gb FC Port	400 MB/s (800 MB/s FD)	425 MB/s (850 MB/s FD)	2005	2006
8 Gb FC Port	800 MB/s (1600 MB/s FD)	850 MB/s (1700 MB/s FD)	2008	2009

## ➤ Maximum requirements met using PCI Express

- ◆ PCIe x8 (Gen 1), provides 2000 MB/s      Full Duplex (4000 MB/s)
- ◆ PCIe x8 (Gen 2), provides 4000 MB/s      Full Duplex (8000 MB/s)
- ◆ PCIe x8 (Gen 3), provides 8000 MB/s      Full Duplex (16000 MB/s)

# Throughput Requirements

Adapter Type	Expected Throughput	Theoretical Maximum	Released	Data Center Rollout
Quad Gb Ethernet	3.6+ Gb/s	424 MB/s	Now	Now
10 GbE Port	9+ Gb/s FD	1275 MB/s	Now	2006
10 Gb iSCSI w/ TOE	9+ Gb/s FD	1275 MB/s	2006-7	2007-8
InfiniBand 4X-SDR	7 Gb/s	8 Gb/s	Now	2005-6 (clustering)
InfiniBand 4X-DDR	15 Gb/s	16 Gb/s	Now	Now
SAS 2.0 Quad Port	21+ Gb/s	24 Gb/s	2007	2009
4 Gb FC Port	400 MB/s (800 MB/s FD)	425 MB/s (850 MB/s FD)	2005	2006
8 Gb FC Port	800 MB/s (1600 MB/s FD)	850 MB/s (1700 MB/s FD)	2008	2009

## ➤ Maximum requirements met using PCI Express

- ◆ PCIe x8 (Gen 1), provides 2000 MB/s      Full Duplex (4000 MB/s)
- ◆ PCIe x8 (Gen 2), provides 4000 MB/s      Full Duplex (8000 MB/s)

# Benefits of PCI Express

- Lane expansion to match need
  - ◆ x1 Low Cost Simple Connector
  - ◆ x4 or x8 PCIe Adapter Cards
  - ◆ x16 High Performance Graphics
  
- 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

# PCI Express 1.x In Industry

## ➤ First Release of Slots in 2005

- ◆ x16 High Performance Graphics
- ◆ x1 Low Cost Simple Connector
- ◆ Desktop Systems

## ➤ Systems Currently Shipping

- ◆ Desktops with multiple x16 connectors
- ◆ Servers with multiple x8 connectors

## ➤ Cards Available

- ◆ x4, x8 cards 10 GbE, Dual/Quad GbE, 4 Gb FC, SAS, IB
- ◆ X16 High Performance Graphics @ 150 W

# PCI Express 1.x In Industry

## ➤ Infrastructure Built

- ◆ PCIe Switches
  - > Multi-ported
  - > Root complex based
- ◆ PCIe-to-PCIX Bridges
  - > Provides access for Legacy Devices

## ➤ PCIe 1.0 Components

- ◆ Switches
- ◆ Bridges
- ◆ Endpoints

# PCI Express Roadmap

## ➤ PCIe 1.0a/1.1

- ◆ Shipping now

## ➤ PCIe 2.0

- ◆ Spec is Approved
- ◆ First slots will ship in late 2007
- ◆ PCIe 2.0 cards backward compatible

## ➤ PCIe 3.0

- ◆ Doubles effective bandwidth again

## ➤ PCIe IOV

- ◆ Architecture allows shared bandwidth

# Evolving I/O Architectures

- Processor speed increase slowing
  - ◆ Multi core processors increasing
  - ◆ Requires multi root complex architectures
  
- Requires high speed interface for interconnect
  - ◆ Minimum 10G data rates
  - ◆ Must support backplane distances
    - > Bladed systems
    - > Single box clustered processors
  - ◆ Need backplane reach, cost effective interface to I/O
  
- Options
  - ◆ Use an existing I/O interface like 10G/Infiniband
  - ◆ Enhance PCI Express

# Drivers for New IO Architectures

- Interface speeds are increasing
  - ◆ Ethernet moving from GbE to 10G and FC from dual 4 Gb to dual 8 Gb
    - > Single applications struggle to supply enough bandwidth to fill these links
  - ◆ Processor speed growth is being replaced by multiple core processors
    - > Requires applications to share these links
  - ◆ PCI Express IOV would allow this to happen
  
- High Availability Increasing in Importance
  - ◆ Requires duplicated processors, IO modules and interconnect
  - ◆ Use of shared virtual I/O simplifies and reduces costs and power
    - > Shared I/O support N+1 redundancy for I/O, 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

# Existing Serial Interfaces

- Established external transport mechanisms exist
  - ◆ Fibre channel
    - > Storage area network standard
  - ◆ 10G Ethernet
    - > Provides a network based solution to SANs
  - ◆ InfiniBand
    - > Choice for high speed process to processor links
    - > Supports wide and fast data channels
  - ◆ SAS
    - > 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 I/O must support these bandwidths

## ➤ Inside the box no clear solution today

### ◆ 10G Ethernet

- Provides a network based solution to SANs

### ◆ InfiniBand

- Choice for scalable, high speed, low latency processor to processor links
- Supports wide and fast data channels

### ◆ PCI Express

- Requires multi root complex support

## ➤ Will it be 10GbE, InfiniBand or PCIe IOV?

# 10G Ethernet or InfiniBand?

## ➤ 10G Ethernet

- ◆ May seem like the obvious choice, but currently lacks
  - QoS mechanisms required
  - Direct interface to root complex
  - Low overhead stack
  - Fiber Channel over Ethernet



## ➤ InfiniBand

- ◆ Has established stack
- ◆ Has higher speed capability
  - SDR, DDR and QDR data rates
  - 4x to 12X widths
  - 500 MB/s to 6 GB/s



# PCI SIG Developing PCIe IOV

## ▶ PCI Express IOV Sig

- ◆ Initial specification 2007
- ◆ Architecture allows shared bandwidth
- ◆ Complementary to PCI Express
  - > Independent of 1.1, 2.0 or 3.0
- ◆ Single Root IOV
  - > One Root Complex
  - > Multiple System Images
- ◆ Multi Root IOV
  - > Multiple Root Complexes
  - > Multiple System Images

# What is PCI Express IOV

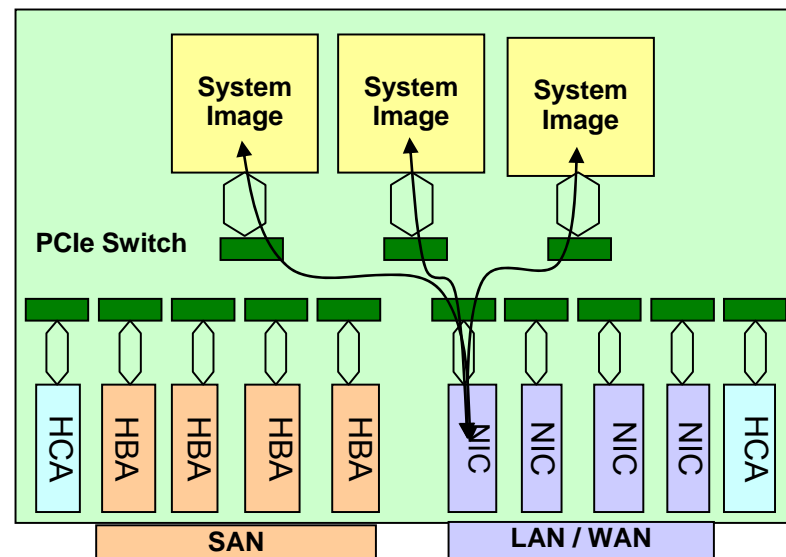
## ➤ Virtualization Technology Overview

- ◆ System Virtualization allows the sharing of physical resources across different System Images (SI)
  - > Processors
  - > Memory
  - > IO components (PCIe Devices)
- ◆ PCIe IO Virtualization
  - > Load balance across SIs
  - > Provides bandwidth management
- ◆ Using the IO components
  - > Each device has multiple Physical Functions (PF)
    - Physical Functions have full PCI configuration space
    - Read/Write
    - Reset
  - > Each PF can have multiple Virtual Functions (VF)
    - One or more VF
    - SI only know about VF's
  - > Requires an IOV Manager
    - Keeps a VF from one SI to affect other SIs

# Single Root IOV

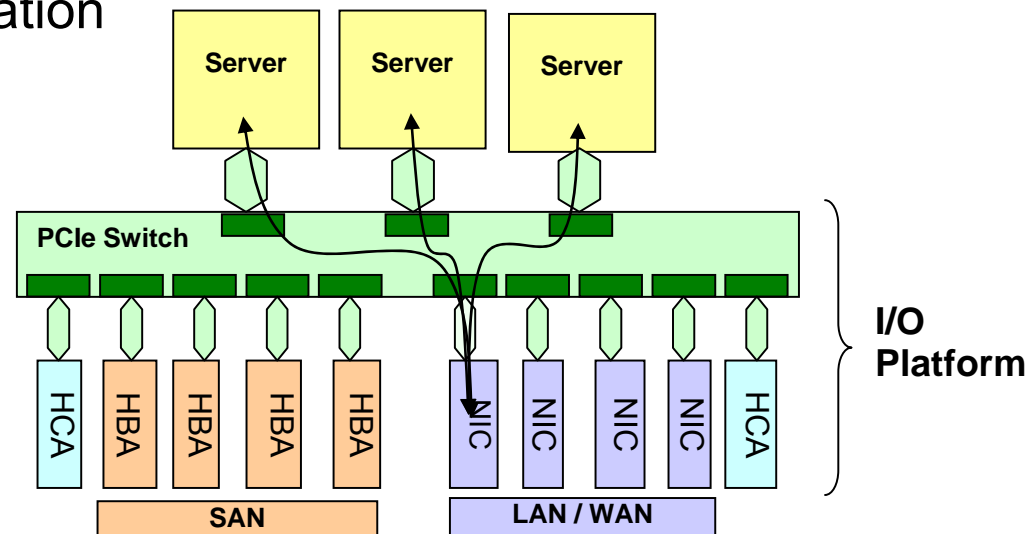
## ➤ Single root complex with Multiple OSES

- ◆ Multiple independent OS images (Linux, Unix, Windows)
- ◆ IO devices are shared by the OS Images
- ◆ Virtualization occurs in the Control Domain and I/O card
- ◆ Device drivers remain almost unchanged
  - Requires additional management
- ◆ Allows bandwidth allocation
- ◆ Supports hot plug
- ◆ Supports remote configuration

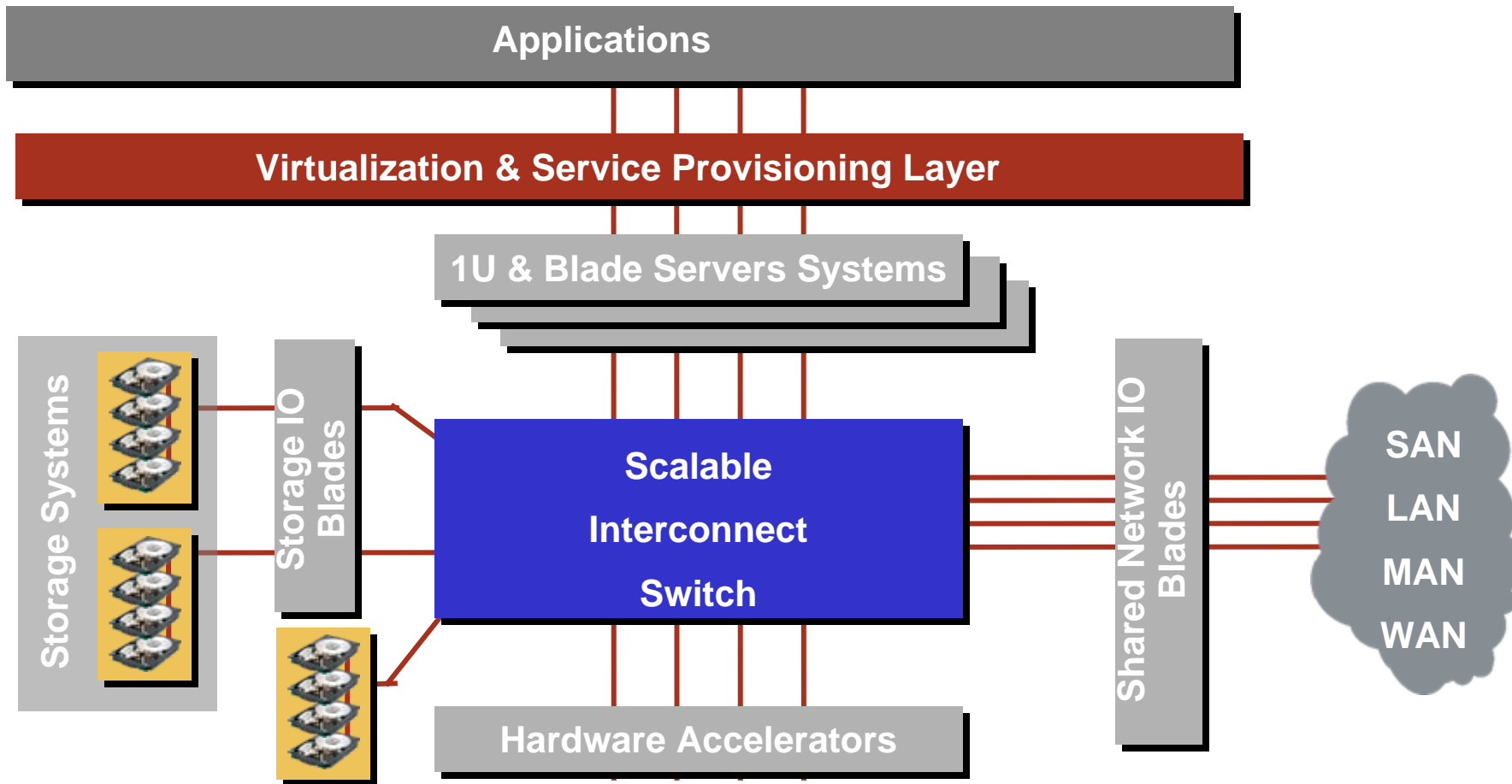


# Multi-Root PCI Express IOV

- ▶ **Multiple root complexes share a single I/O**
  - ◆ Virtualization occurs in the PCIe switch and I/O card
  - ◆ Device drivers remain almost unchanged
    - ▶ Requires additional management
  - ◆ Allows bandwidth allocation
  - ◆ Supports hot plug
  - ◆ Supports remote configuration

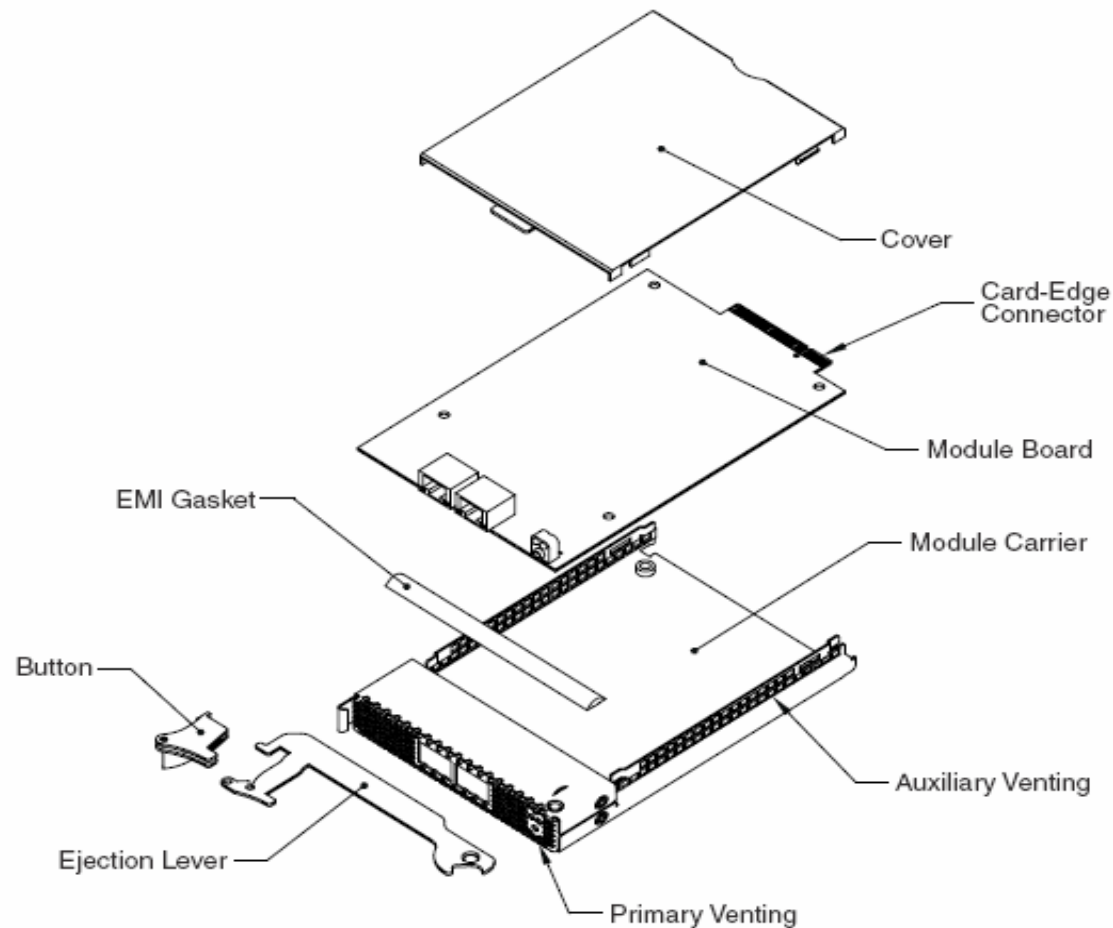


# Mutli Root Virtualization in Blades



# Express Module (EM)

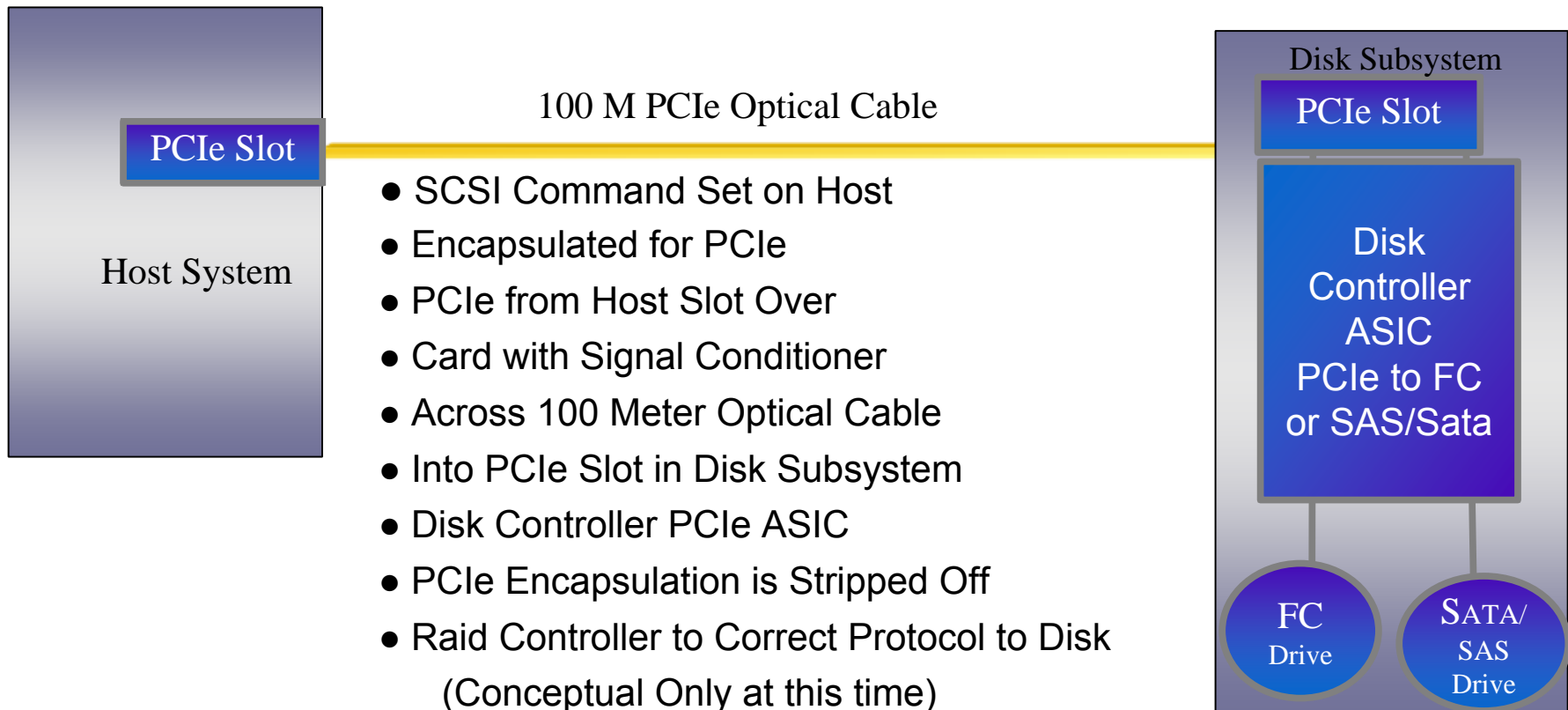
- ◆ Developed by the PCI-SIG (formally Server IO Modules)
  - ◆ Fully compatible with latest PCI Express specification
  - ◆ Designed to support future generations of PCI Express
- ◆ Adds the necessary hot plug hardware and software
- ◆ Commodity pricing model using standard PCI Express silicon and ½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



# Impact / Benefit to Storage

- PCI Express provides
  - ◆ Full Bandwidth Dual Ported 4 Gb FC
  - ◆ Full Bandwidth SAS
  - ◆ Legacy Support via PCIX
- 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



**PCI** – Peripheral Component Interconnect. An open, versatile I/O 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 I/O system to the CPU and memory.

**HBA** – Host Bus Adapter.

**IOV** – IO Virtualization

Single root complex IOV – Sharing an I/O resource between multiple operating systems on a HW Domain

Multi root complex IOV – Sharing an I/O resource between multiple operating systems on multiple HW Domains

- Please send any questions or comments on this presentation to SNIA: [tracknetworking@snia.org](mailto:tracknetworking@snia.org)

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

*SNIA Education Committee*

**Rob Peglar  
Howard Goldstein  
Carl Hensler  
Paul Millard  
Ron Emerick**

# Appendix

# InfiniBand HW Architecture

## ➤ Physical Layer (SDR)

- ◆ 2.5 Gb/s signaling rate
- ◆ Multiple Link Widths
  - > 1X, 4X, 12X (2.5, 10, 30 Gb/s)

## ➤ Distance

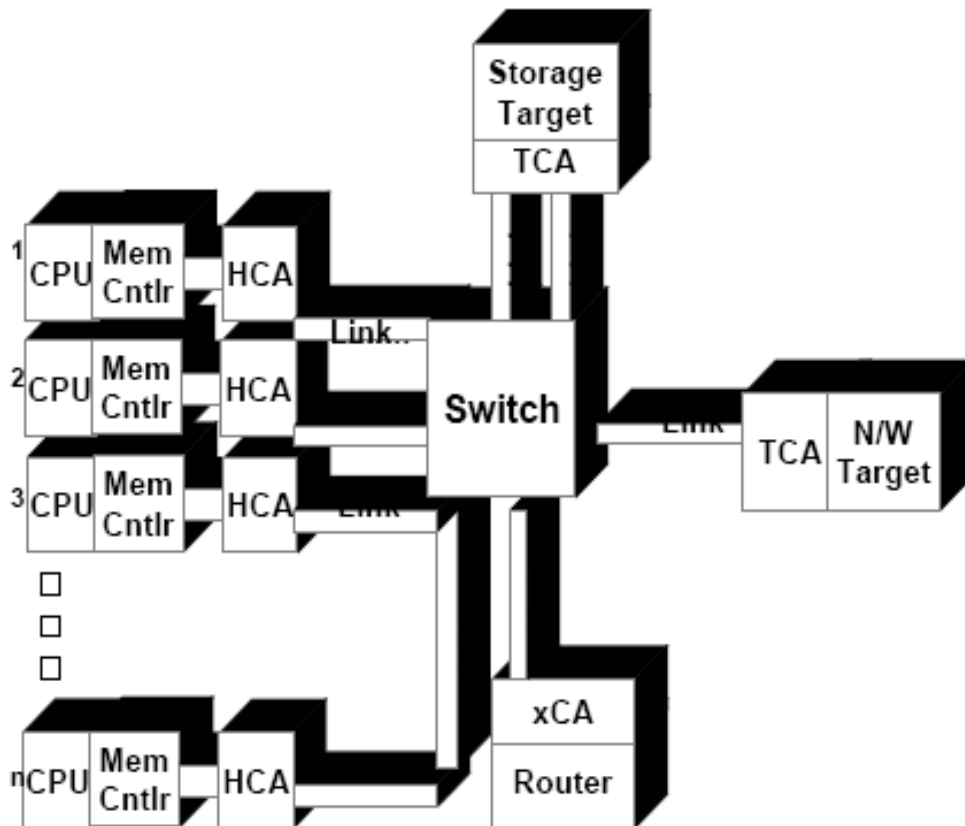
- ◆ Copper 15 M 4x SDR, 10 M 12x SDR
- ◆ Optical 150 M (12X)

## ➤ Topology

- ◆ Connection Orientated, Switched Fabric
- ◆ 64K Nodes per Sub-net
- ◆ Multiple subnets bridged w/router
  - > IPV6 addressing x-subnet

## DDR - Doubles the Signaling Rate

# InfiniBand Started As ...



## Host Channel Adaptor: HCA

- Connects a processor or CPU to the I/O fabric

## Target Channel Adaptor: TCA

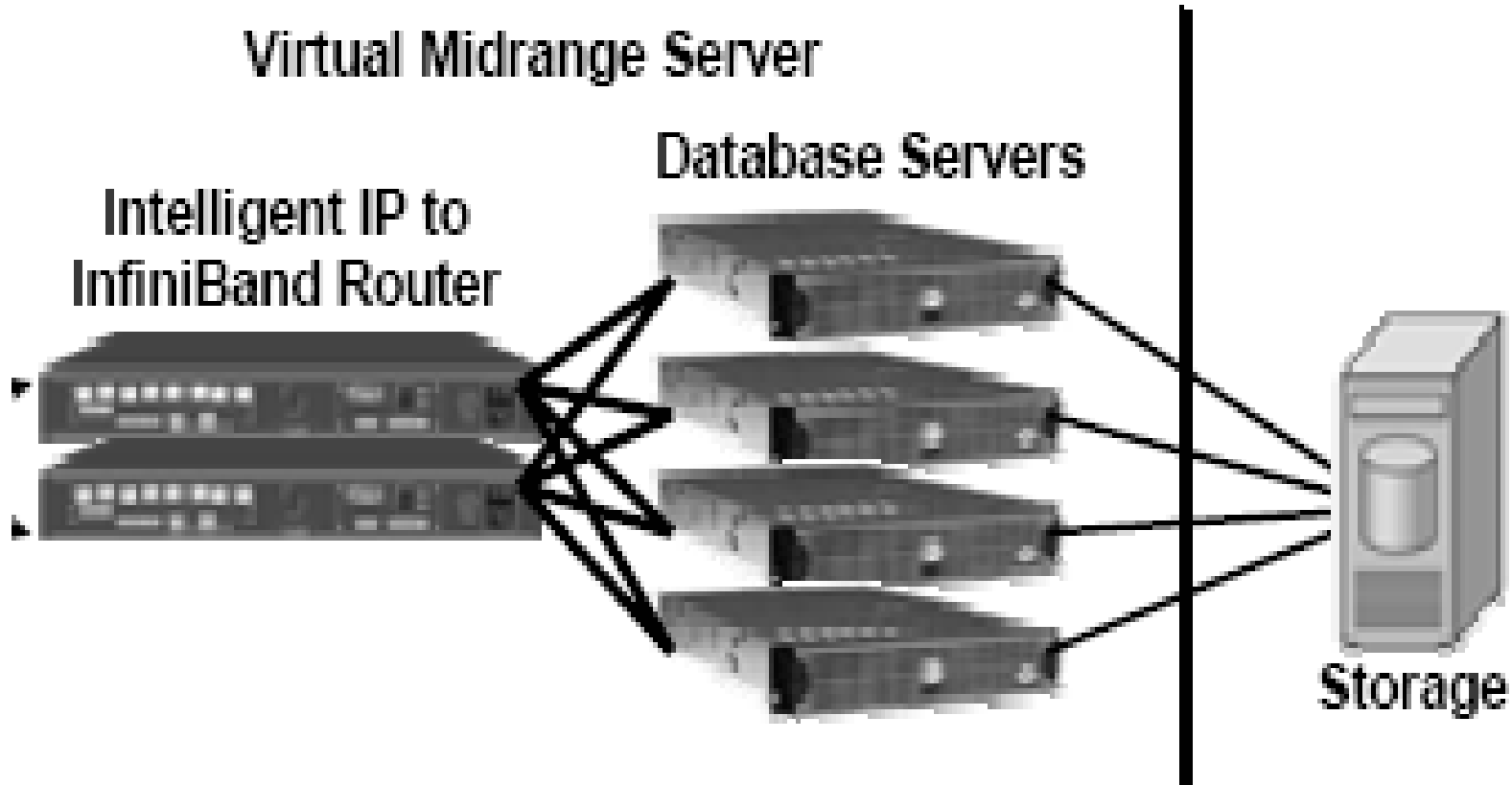
- Connects I/O controllers e.g. Fibre Channel, Ethernet and SCSI to InfiniBand

Switch: Building block to create a subnet

## Links Speed (per second)

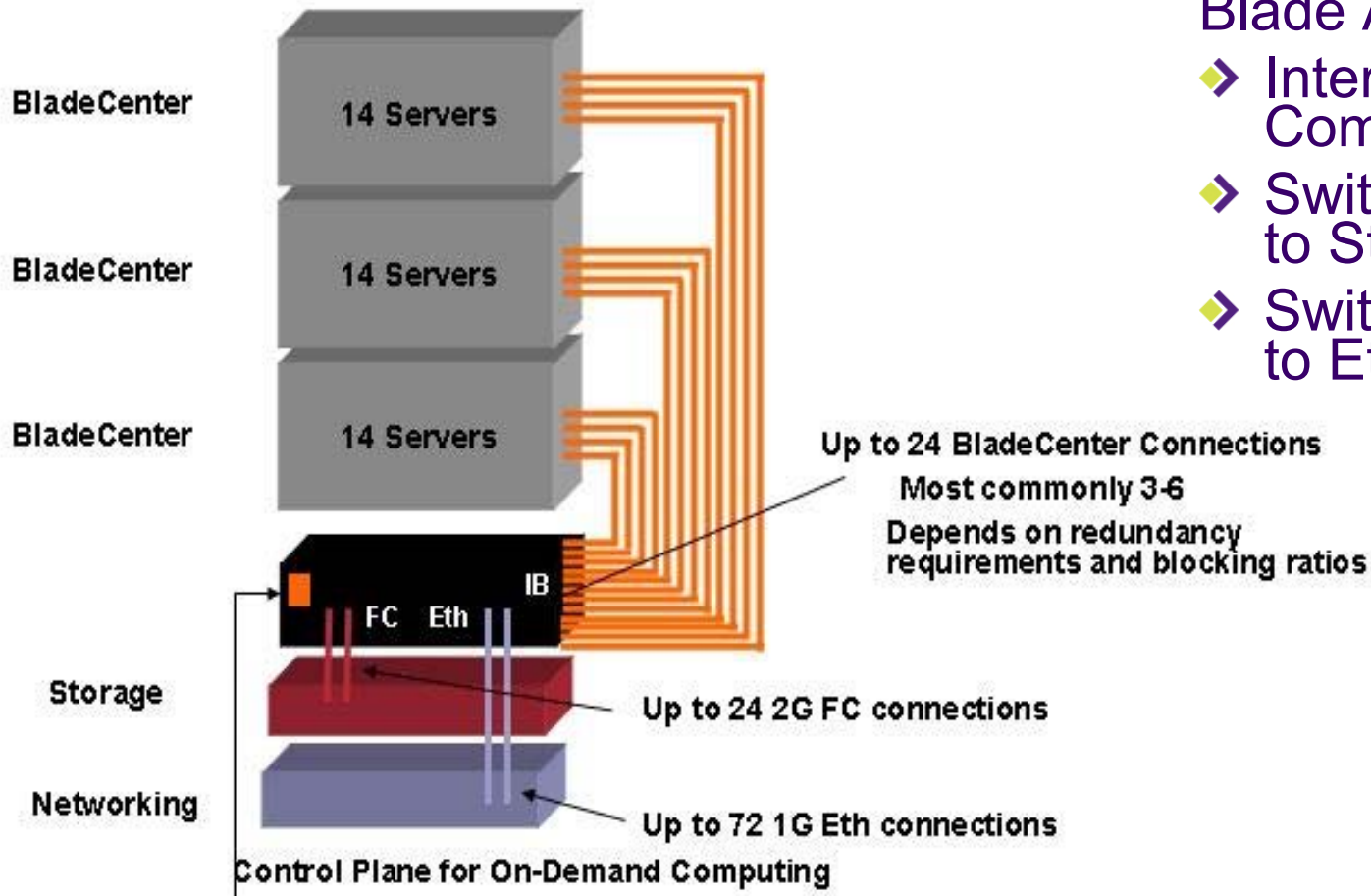
2.5 GB, 10GB, 30GB

# InfiniBand As Rolling Out



- Application Clustering
- Inter-Tier & Inter-Processing Communications
- Gateways, Routers, Bridges & Switches

# InfiniBand Possibilities



## Blade Applications

- Inter-Tier Communications
- Switched Gateway to Storage
- Switched Gateway to Ethernet

## ➤ Implemented

- ◆ IB HCAs for Clustering
- ◆ IB Switches
- ◆ IB to Ethernet Gateways
- ◆ IB to FC Gateways
- ◆ IB Attached Storage

## ➤ Still Under Consideration

- ◆ Additional IB Gateways
- ◆ IB as Blade Backplanes