



Education

Comparing Server I/O Consolidation Solutions: iSCSI, InfiniBand and FCoE

Gilles Chekroun

Errol Roberts

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.

➤ Comparing Server I/O Consolidation: iSCSI, Infiniband and FCoE

This tutorial gives an introduction to Server I/O consolidation, having one network interface technology (Standard Ethernet, Data Center Ethernet, InfiniBand), to support IP applications **and block level storage** (iSCSI, FCoE and SRP/iSER) applications. The benefits for the end user are discussed: less cabling, power and cooling. For these 3 solutions, iSCSI, Infiniband and FCoE, we compare features like Infrastructure / Cabling, Protocol Stack, Performance, Operating System drivers and support, Management Tools, Security and best design practices.

Agenda

- Definition of Server I/O Consolidation
- Why Server I/O Consolidation
- Introducing the 3 solutions
 - iSCSI
 - InfiniBand
 - FCoE
- Differentiators
- Conclusion

Definition of Server I/O Consolidation

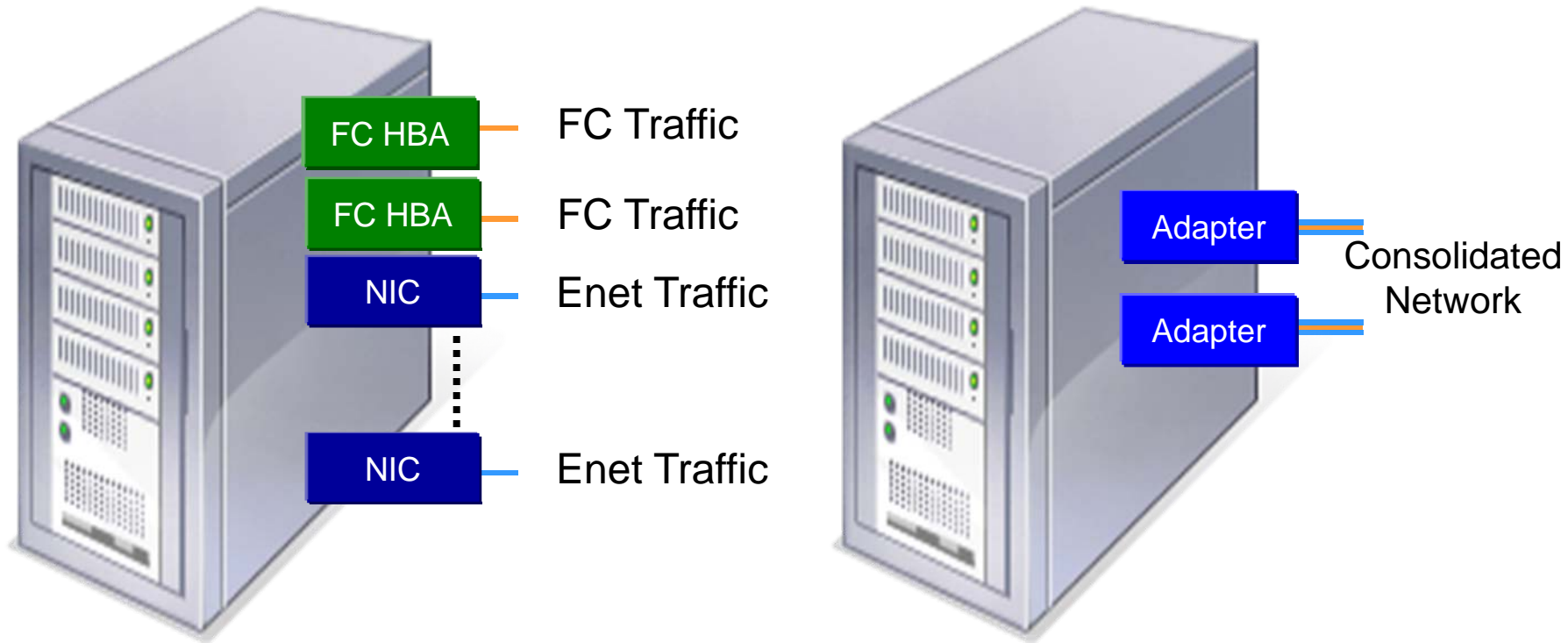
What is Server I/O Consolidation

- IT Organizations operate multiple parallel networks
 - ◆ IP Applications (including NFS, NAS,...) over a Ethernet network *)
 - ◆ SAN over a Fibre Channel network
 - ◆ HPC/IPC over an InfiniBand network **)
- Server I/O consolidation combines the various traffic types onto a single interface and single cable
- Server I/O consolidation is the first phase for a Unified Fabric (single network)

*) In this presentation we cover only **Block Level Storage** solutions, not **File Level (NAS, NFS,..)**

***) For the remaining part, we don't cover **HPC**; for lowest latency requirements, **InfiniBand** is the best and most appropriate technology.

I/O Consolidation Benefits



- ◆ **Adapter:**
 - › **NIC for Ethernet/IP or HCA for InfiniBand or Converged Network Adaptor (CNA) for FCoE**
- ◆ **Customer Benefit:**
 - › **Fewer NIC's, HBA's and cables, lower CapEx, OpEx (power, cooling)**

Why Server I/O Consolidation ?

The drive for I/O Consolidation

- Multicore – Multisocket CPUs
- Server Virtualization software (Hypervisor)
- High demand for I/O bandwidth
- Reductions in cables, power and cooling, therefore reducing OpEx/CapEx
- Limited number of interfaces for Blade Servers
- Consolidated Input into Unified Fabric

- Virtual networks growing faster and larger than physical
 - ◆ Network admins are getting involved in virtual interface deployments
 - ◆ Network access layer needs to evolve to support consolidation and mobility
- Multi-core Computing driving Virtualization & new networking needs
 - ◆ Driving SAN attach rates higher (10%→40%→Growing)
 - ◆ Driving users to plan now for 10GE server interfaces
- Virtualization enables the promise of blades
 - ◆ 10GE and FC are highest growth technologies within blades
 - ◆ Virtualization and Consolidated I/O removes blade limitation

10GbE Drivers in the Datacenter



Multi-Core CPU architectures allowing bigger and multiple workloads on the same machine



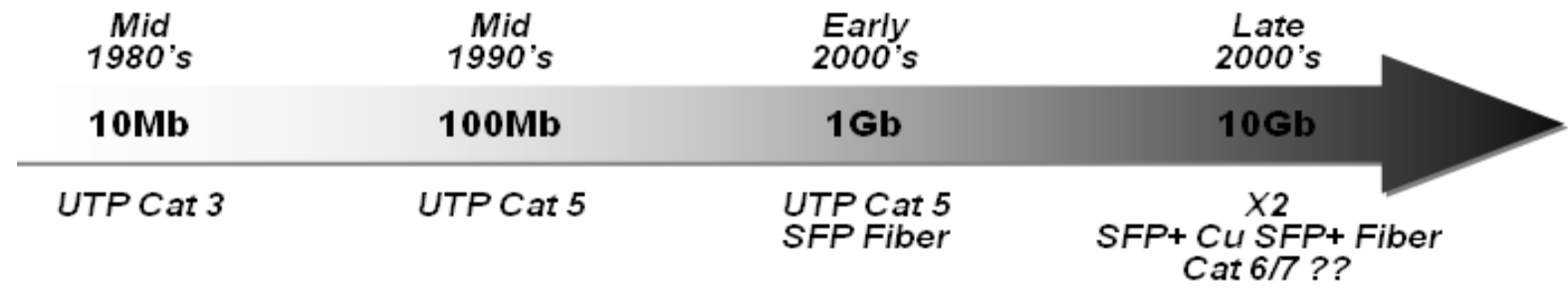
Server virtualization driving the need for more bandwidth per server due to server consolidation



Growing need for network storage driving the demand for higher network bandwidth to the server

Multi-Core CPUs and Server Virtualization driving the demand for higher bandwidth network connections

Evolution of Ethernet Physical Media



Technology	Cable	Distance	Power (each side)	Transceiver Latency (link)
SFP+ CU Copper	Twinax	10m	~0.1W	~0.1µs
SFP+ USR ultra short reach	MM OM2 MM OM3	10m 100m	1W	~0
SFP+ SR short reach	MM 62.5µm MM 50µm	82m 300m	1W	~0
10GBASE-T	Cat6 Cat6a/7 Cat6a/7	55m 100m 30m	~8W ~8W ~4W	2.5µs 2.5µs 1.5µs

Therefore . . .

**10 Gigabit Ethernet is the
enabler
for I/O Consolidation
in next generation Data Centers**

Introducing the three solutions

Server I/O Consolidation Solutions

➤ iSCSI

- ◆ LAN: Based on Ethernet and TCP/IP
- ◆ SAN: Encapsulates SCSI in TCP/IP

➤ InfiniBand

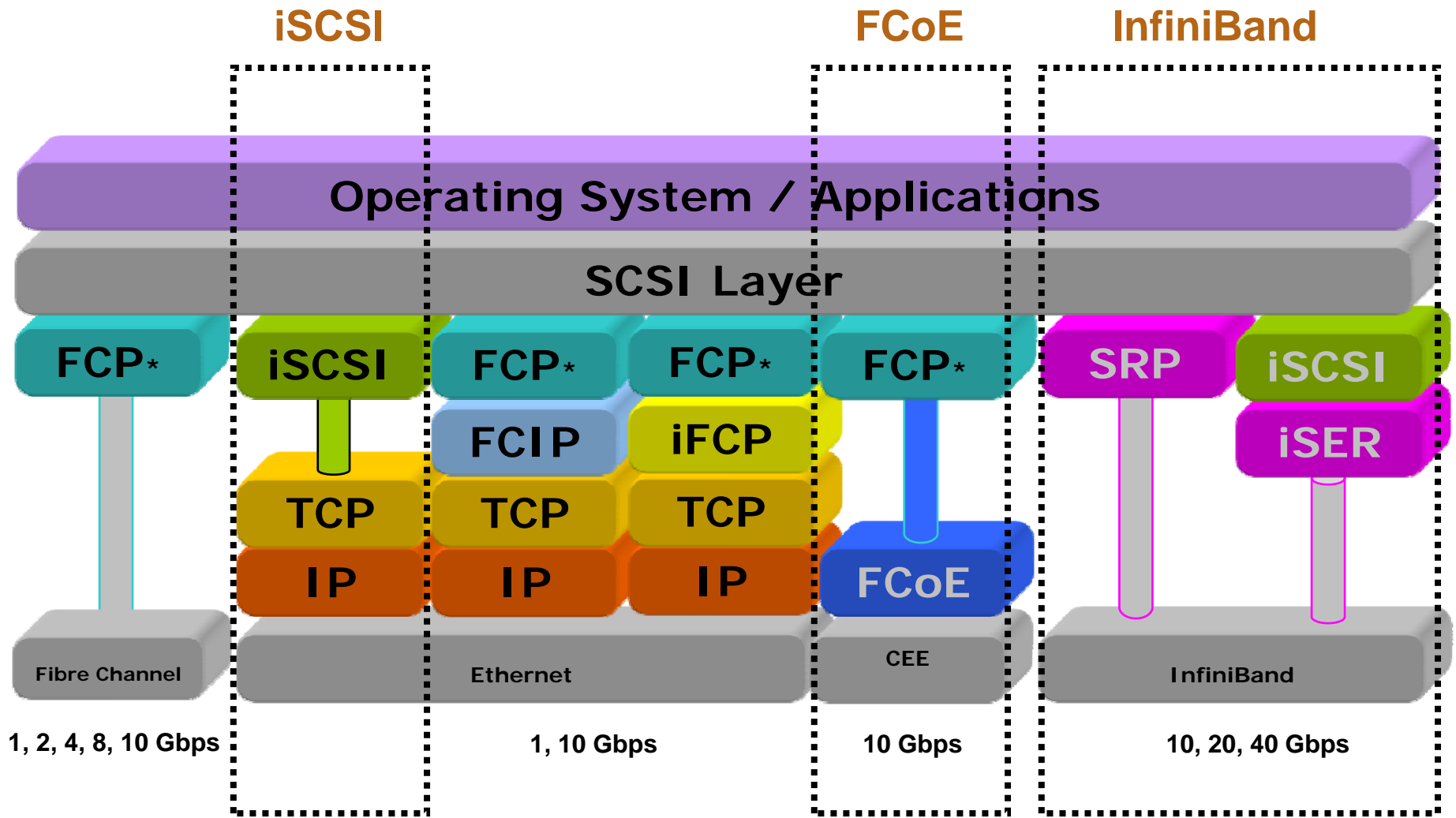
- ◆ LAN: Transports IP over InfiniBand (IPoIB); Socket Direct Protocol (SDP) between IB attached servers
- ◆ SAN: Transports SCSI over Remote DMA protocol (SRP) or iSCSI Extensions for RDMA (iSER)
- ◆ HPC/IPC: Message Passing Interface (MPI) over InfiniBand network

➤ FCoE

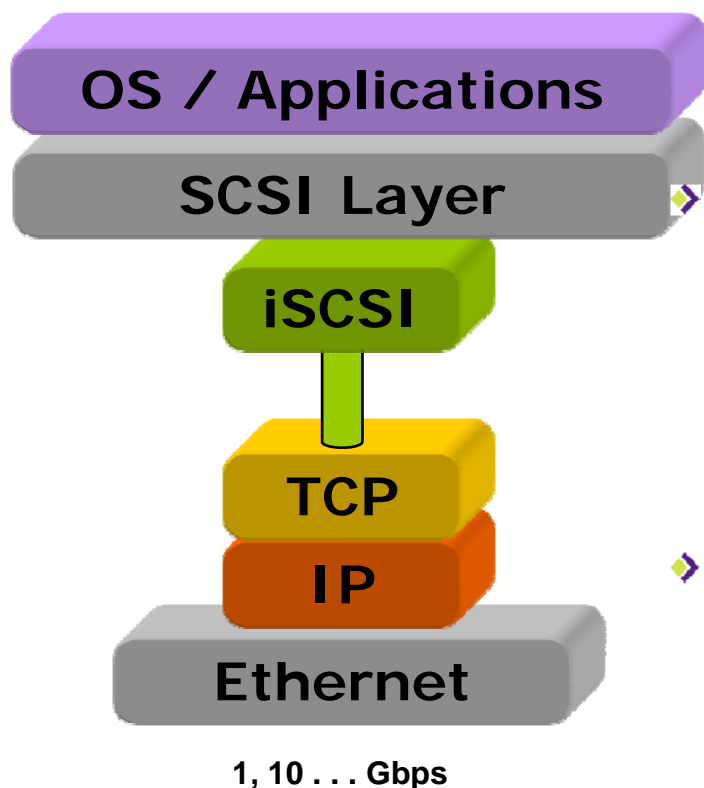
- ◆ LAN: Based on Ethernet (CEE *) and TCP/IP
- ◆ SAN: Maps and transports Fibre Channel over CEE *

* CEE (Converged Enhanced Ethernet) is an architectural collection of Ethernet extensions designed to improve Ethernet networking and management in the Data Center; also called DCB (Data Center Bridging) or DCE (Data Center Ethernet)

Encapsulation technologies



* Includes FC Layer



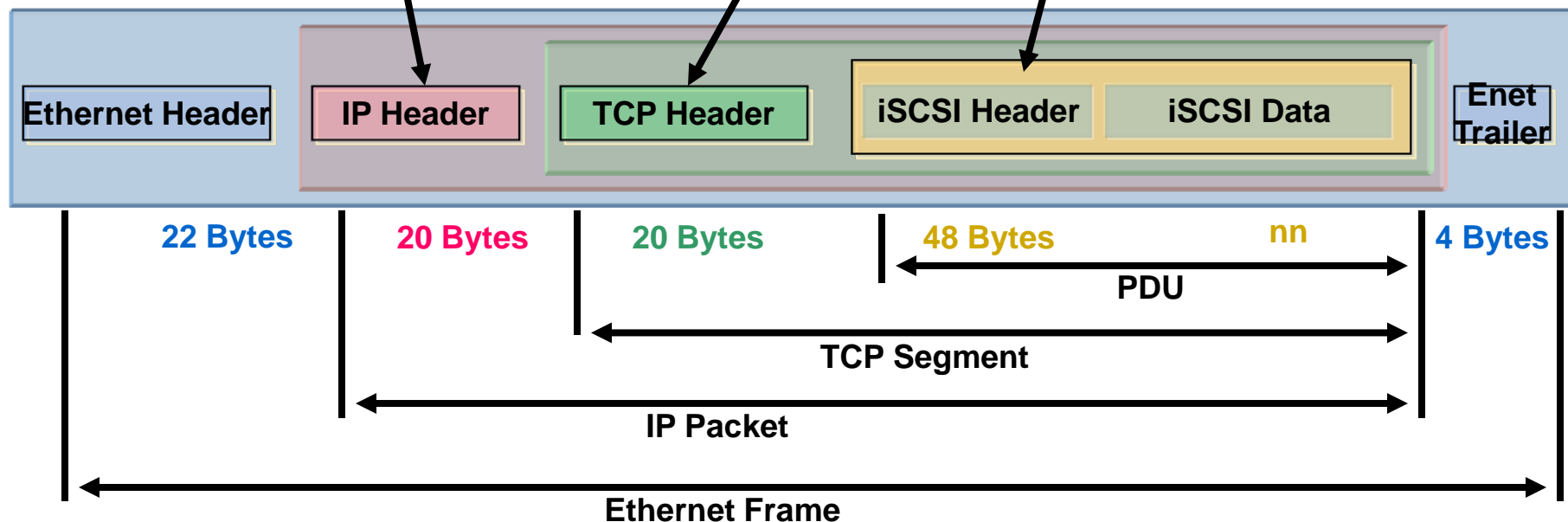
- **A SCSI transport protocol that operates over TCP**
 - ◆ Encapsulates **SCSI CDBs** (operational commands: e.g. read or write) and data into **TCP/IP** byte-streams
 - ◆ Allows **iSCSI Initiators** to access **IP-based iSCSI targets** (either natively or via **iSCSI-to-FC gateway**)
- **Standards status – IETF standard**
 - ◆ **RFC 3720 on iSCSI**
 - ◆ **Collection of RFCs describing iSCSI**
 - **RFC 3347—iSCSI Requirements**
 - **RFC 3721—iSCSI Naming and Discover**
 - **RFC 3723—iSCSI Security**
- **Broad industry support**
 - ◆ **Operating System vendors support their iSCSI drivers**
 - ◆ **Gateway (Routers, Bridges) and Native iSCSI storage arrays**

iSCSI Messages

Contains routing information so that the message can find its way through the network

Provides information necessary to guarantee delivery

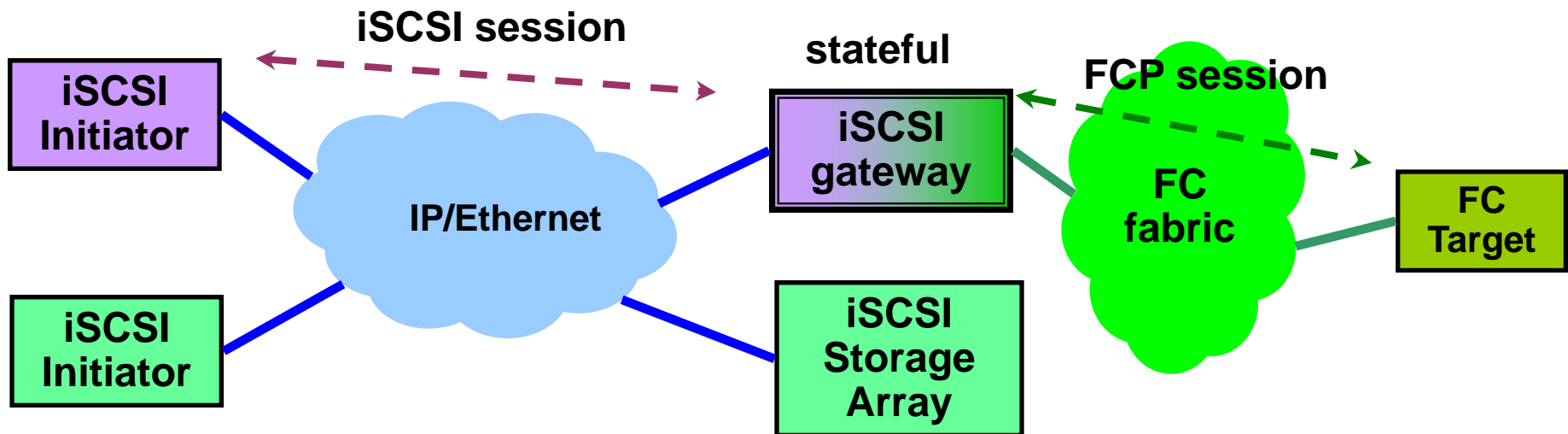
Explains how to extract SCSI commands and data



iSCSI Topology

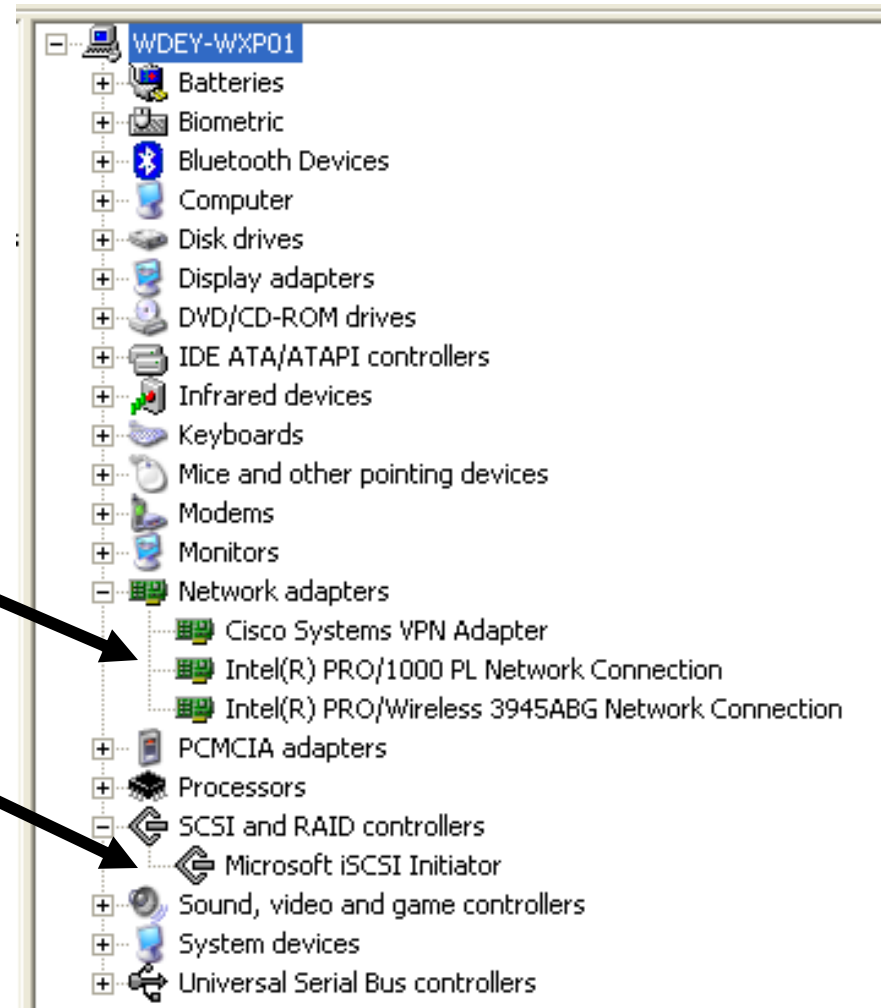
➤ Allows I/O consolidation

- ◆ iSCSI is proposed today as an I/O consolidation option
- ◆ Native (iSCSI Storage Array) and Gateway solutions

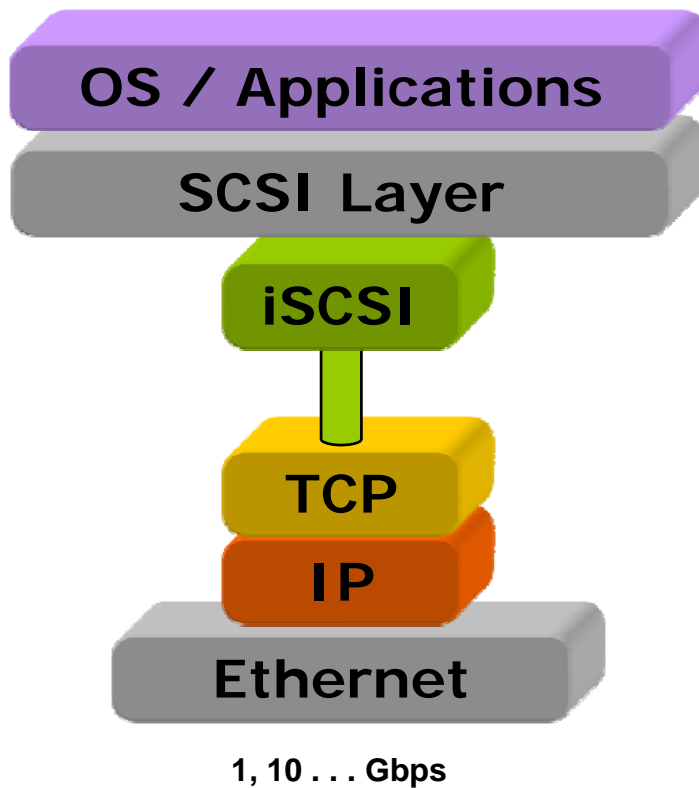


View from Operating System

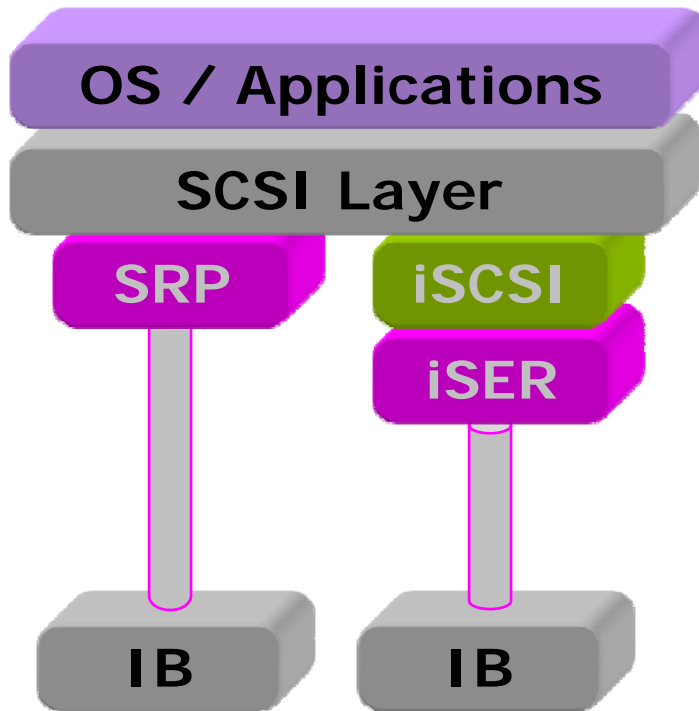
- Operating System sees:
 - ◆ 1 Gigabit Ethernet adapter
 - ◆ iSCSI Initiator



iSCSI based I/O Consolidation



- **Overhead of TCP/IP Protocol**
- **TCP for recovery of lost packets**
- **It's SCSI not FC**
- **LAN/Metro/WAN (Routable)**
- **Security of IP protocols (IPsec)**
- **Stateful gateway (iSCSI <-> FCP)**
- **Mainly IG Initiator (Server)**
- **10G for iSCSI Target recommended**
- **Can use existing Ethernet switching infrastructure**
- **Offload Engine (TOE) suggested (virtualized environment support ?)**
- **QoS or separate VLAN for storage traffic suggested**
- **New Management Tools**
- **Might require different Multipath Software**
- **iSCSI Boot Support**



10, 20, 40 Gbps (4X SDR/DDR/QDR)

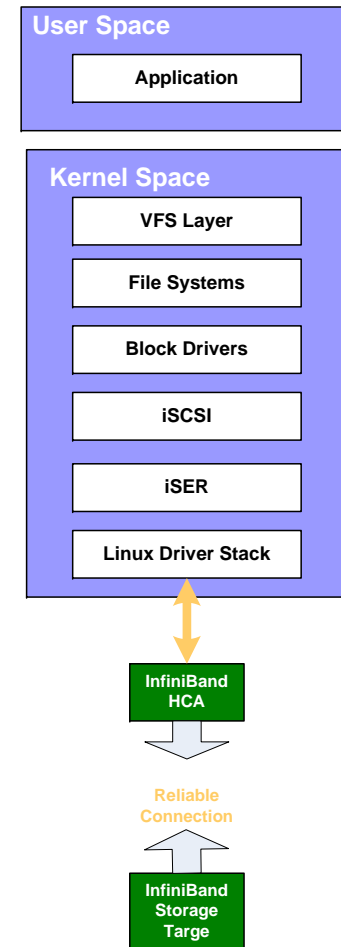
- **Standards-based interconnect**
 - ◆ <http://www.infinibandta.org>
- **Channelized, connection-based interconnect optimized for high performance computing**
- **Supports server and storage attachments**
- **Bandwidth Capabilities (SDR/DDR/QDR)**
 - ◆ 4x—10/20/40 Gbps: 8/16/32 Gbps actual data rate
 - ◆ 12x—30/60/120 Gbps: 24/48/96 Gbps actual data rate
- **Built-in RDMA as core capability for inter-CPU communication**

InfiniBand: SCSI RDMA Protocol (SRP)

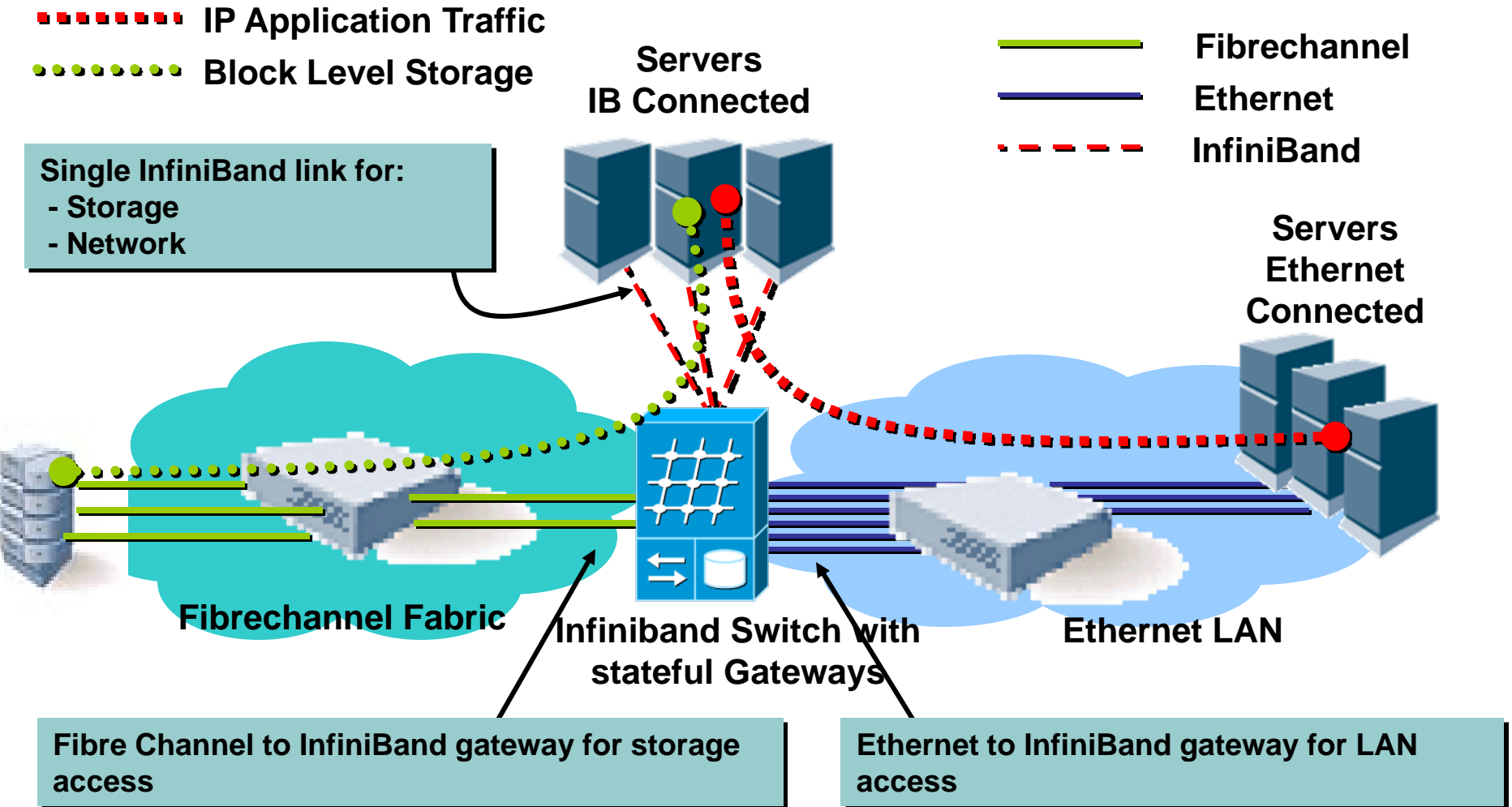
- SCSI Semantics over RDMA fabric
- Provides High Performance block-level storage access
- Not IB specific - Standard specified by T10
<http://www.t10.org>
- Host drivers tie into standard SCSI I/F in kernel
- Storage appears as SCSI disks to local host
- Can be used for end-to-end IB storage (No FC)
- Can be used for SAN Boot over InfiniBand
- Usually in combination with a Gateway to native Fibre Channel targets

InfiniBand: iSCSI Extensions for RDMA (iSER)

- IETF Standard
- Enables iSCSI to take advantage of RDMA.
- Mainly offloads the data path
- Leverages iSCSI management and discovery architecture
- Simplifies iSCSI protocol details such as data integrity management and error recovery
- Not IB Specific
- Needs a iSER Target to work end to end



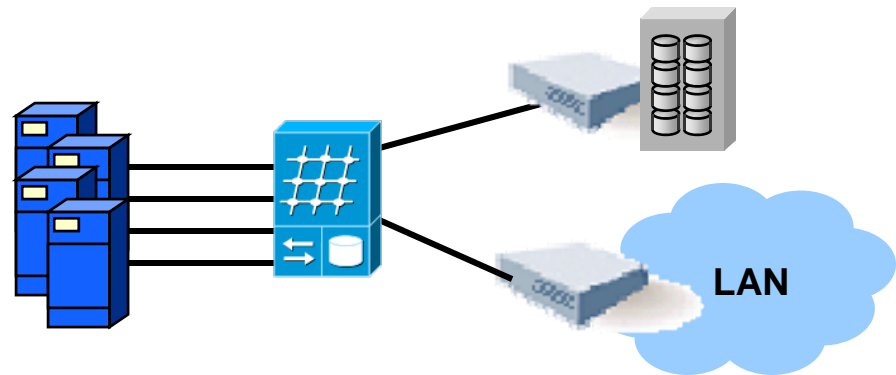
InfiniBand Gateway Topology: Gateways for Network and Storage



Physical vs. Logical view

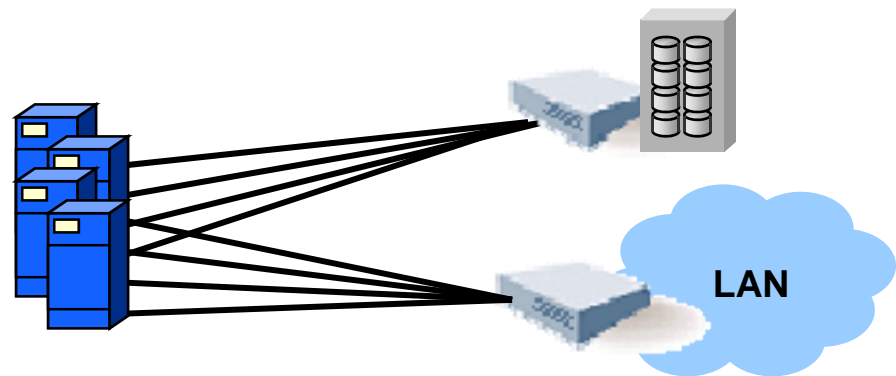
Physical View

- Servers connected via IB
- SAN attached via FC
- LAN attached via Gigabit Ethernet

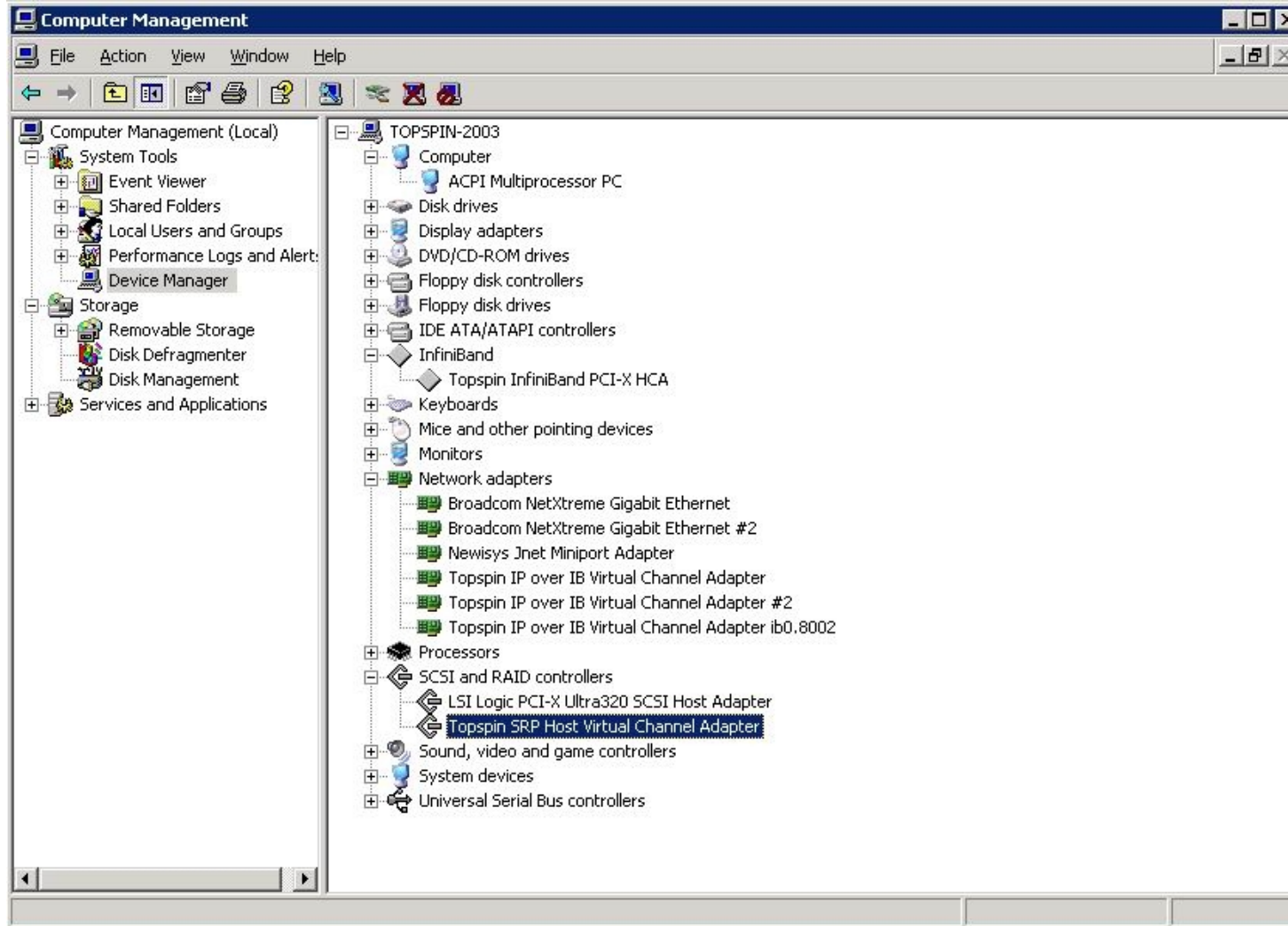


Logical View

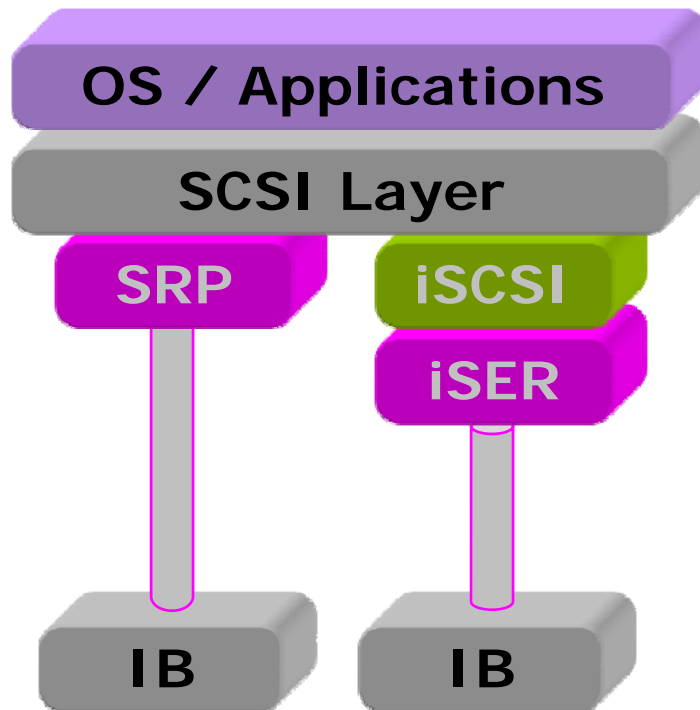
- Hosts present PWWN on SAN
- Hosts present IP address on VLAN



View from Operating System

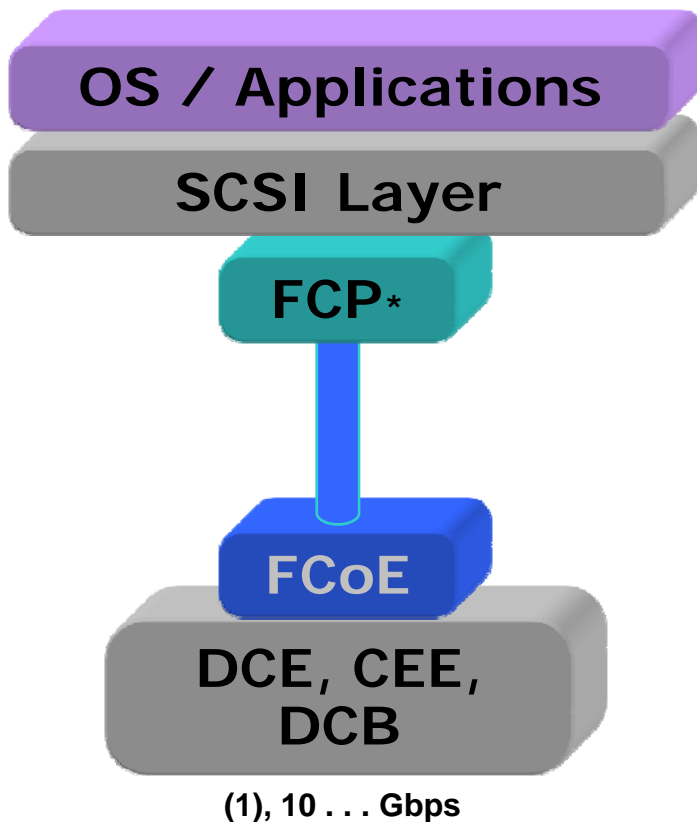


InfiniBand based I/O Consolidation



10, 20, 40 Gbps (4X SDR/DDR/QDR)

- Requires new Eco system (HCA, cabling, switches)
- Mostly copper cabling, limited distance but Fiber is available
- Datacenter protocol
- New driver (SRP)
- Stateful Gateway from SRP to FCP (unless native IB attached disk array)
- RDMA capability of HCA used
- Low CPU overhead
- Payload is SCSI not FC
- Concept of Virtual links and QoS in InfiniBand
- Boot Support

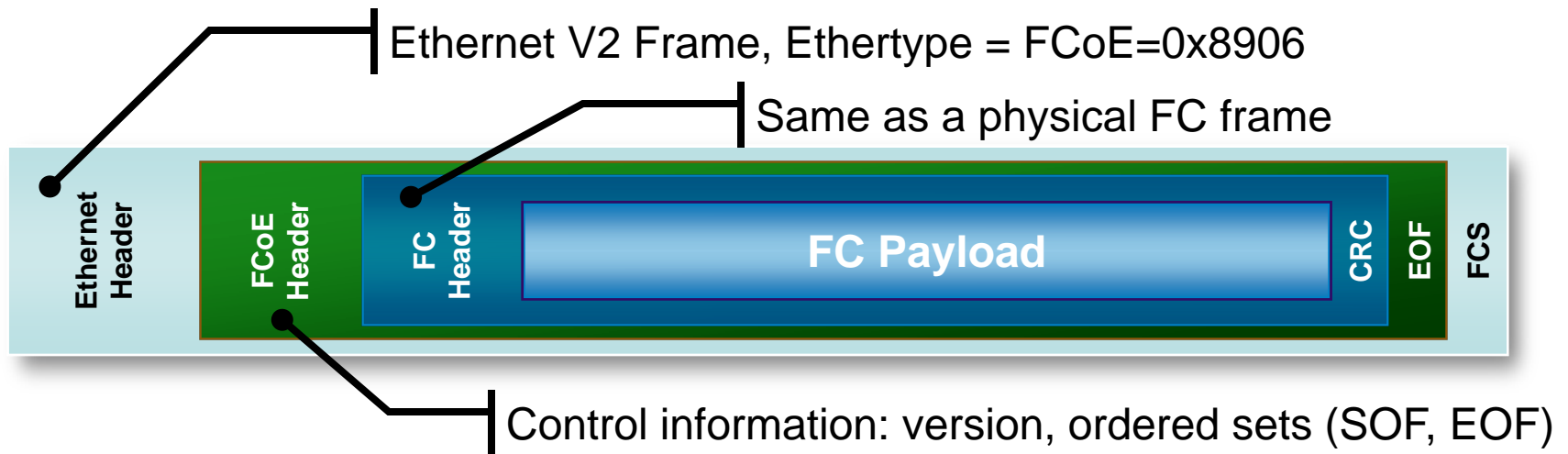


- From a Fibre Channel standpoint it's Fibre Channel encapsulated in Ethernet
- From an Ethernet standpoint it's just another ULP (Upper Layer Protocol)
- FCoE is an extension of Fibre Channel onto a Lossless (Data Center) Ethernet fabric
- FCoE is managed like FC at initiator, target, and switch level, completely based on the FC model
 - ◆ Same host-to-switch and switch-to-switch behavior of FC
 - ◆ in order frame delivery or FSPF load balancing
 - ◆ WWNs, FC-IDs, hard/soft zoning, DNS, RSCN
- Standards Work in T11, IEEE and IETF not yet final

* Includes FC Layer

FCoE Enablers

- 10Gbps Ethernet
- Lossless Ethernet
 - ◆ Matches the B2B credits used in Fibre Channel to provide a lossless service
- Ethernet mini jumbo frames (2180 Bytes)
 - ◆ Max FC frame payload = 2112 bytes



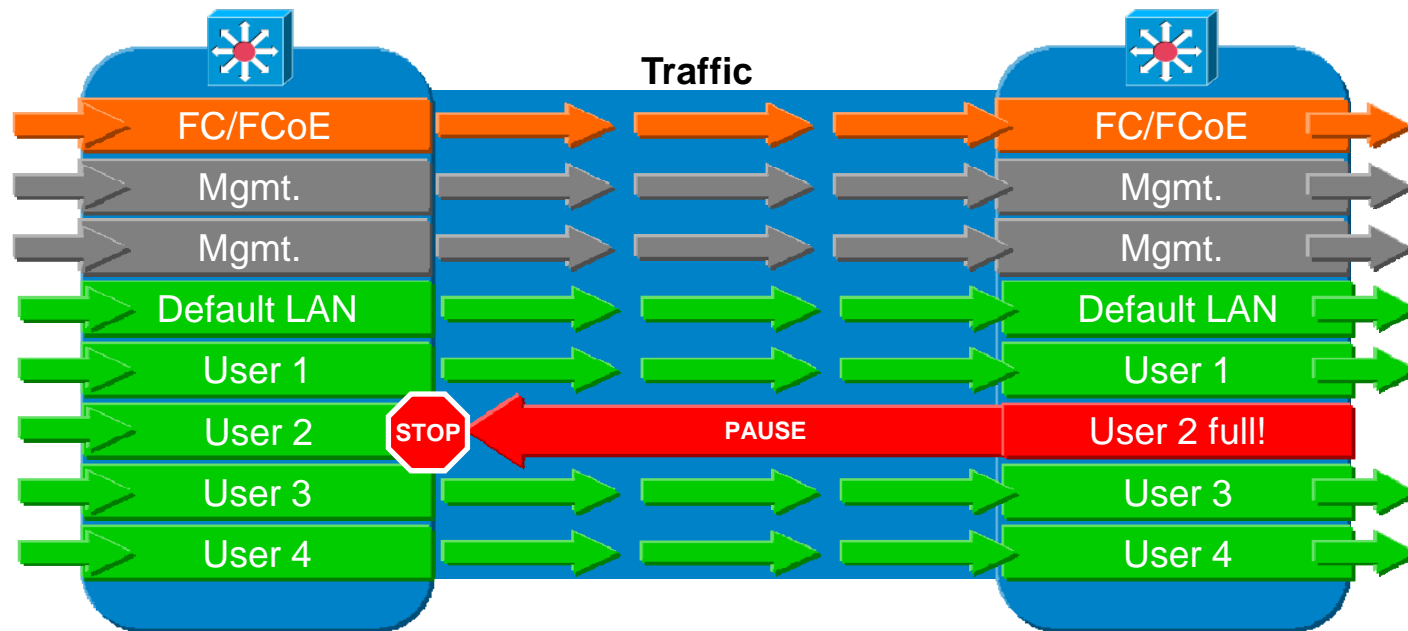
Ethernet Enhancements

- Enhanced Ethernet for Data Center Applications
- Transport of FCoE
- Enabling Technology for I/O Consolidation and Unified Fabric

Feature / Standard	Benefit
Priority Flow Control (PFC) IEEE 802.1Qbb	Enable multiple traffic types to share a common Ethernet link without interfering with each other
Bandwidth Management IEEE 802.1Qaz	Enable consistent management of QoS at the network level by providing consistent scheduling
Congestion Management IEEE 802.1Qau	End-to-end congestion management for L2 network
Data Center Bridging Exchange Protocol (DCBX)	Management protocol for enhanced Ethernet capabilities
L2 Multipath for Unicast and Multicast	Increase bandwidth, multiple active paths. No spanning tree

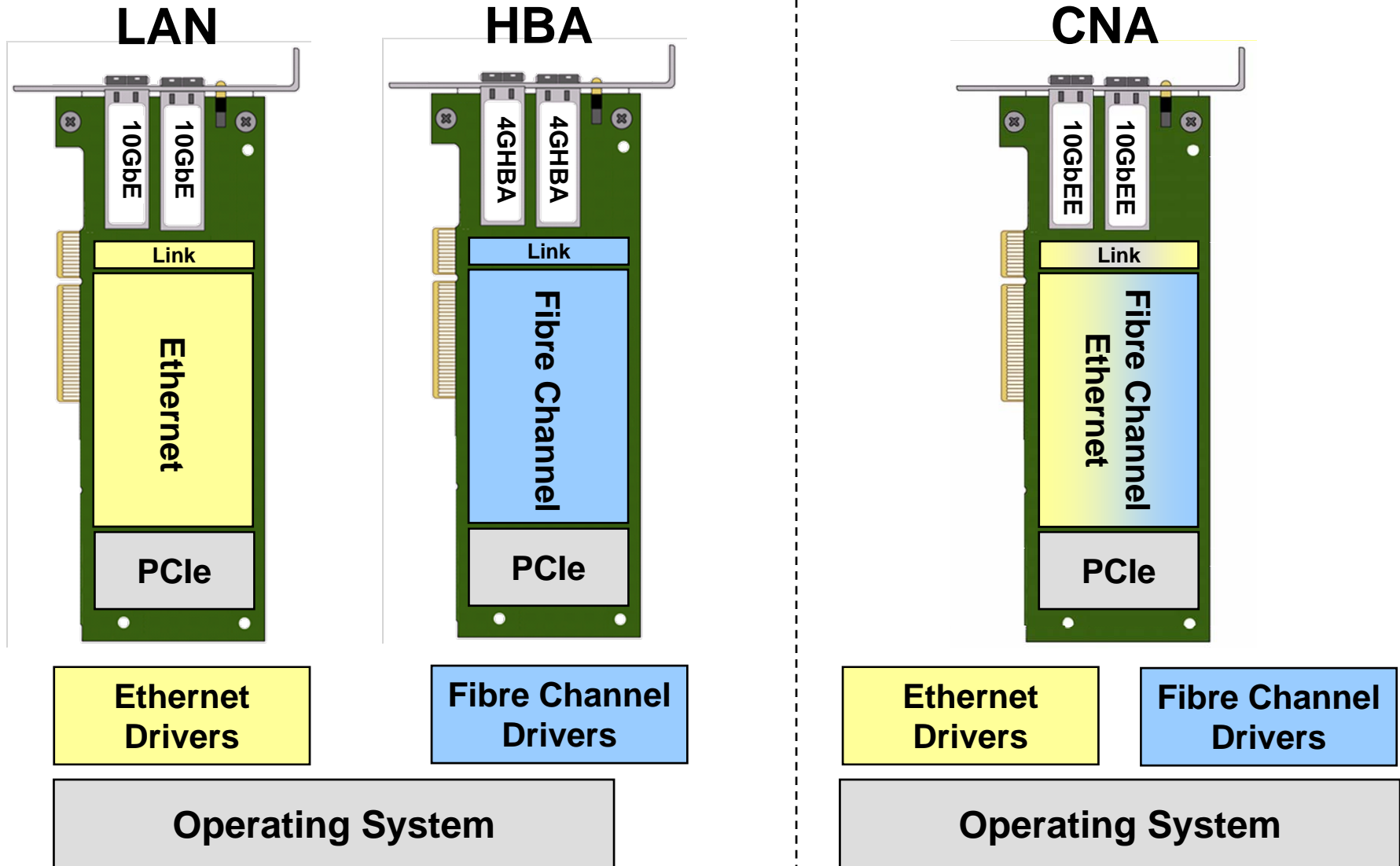
*) T11 BB 5 group has only required that Ethernet switches have standard Pause (Link Pause), and baby Jumbo frame capability; which means no I/O consolidation support.

FCoE Enabler: Priority Flow Control



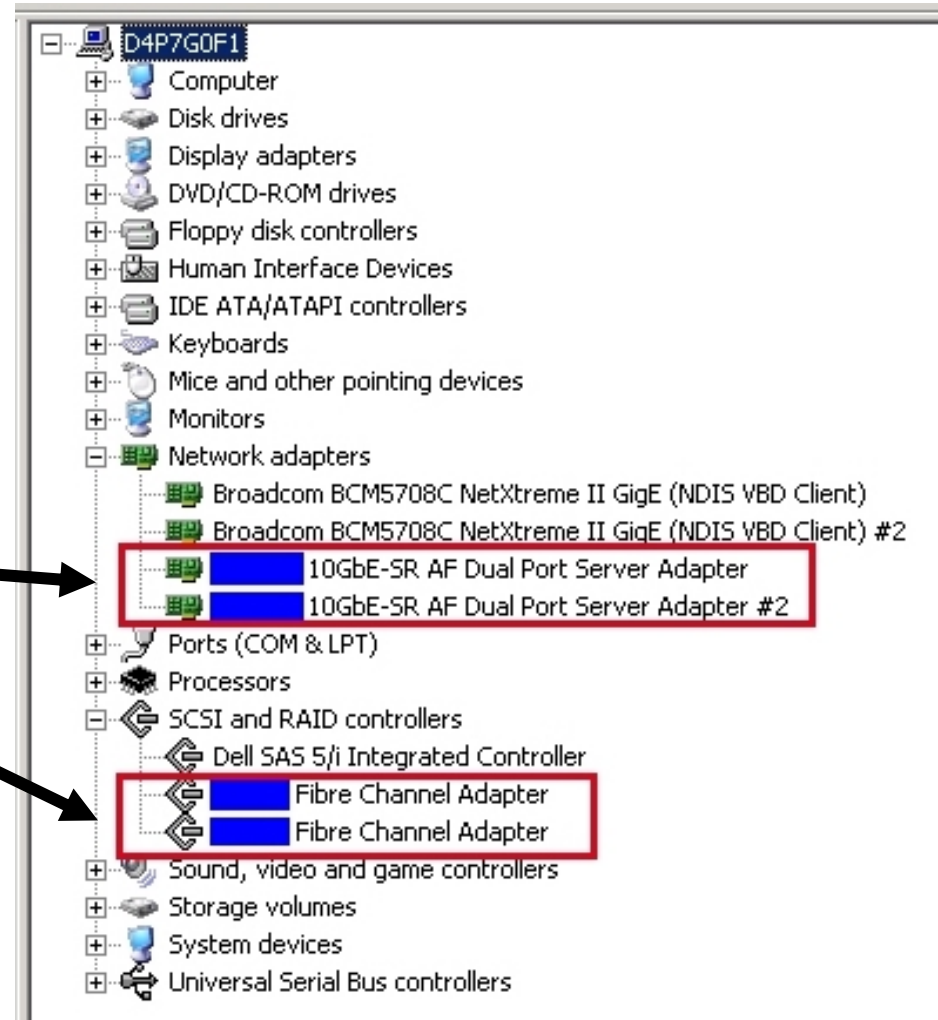
- ◆ Enables lossless Fabrics for each class of service
- ◆ **PAUSE** sent per virtual lane when buffers limit exceeded
- ◆ Network resources are partitioned between VL's (E.g. input buffer and output queue)
- ◆ The switch behavior is negotiable per VL
- ◆ InfiniBand uses a similar mechanism for multiplexing multiple data streams over a single physical link

FCoE Enabler: Converged Network Adapter (CNA)

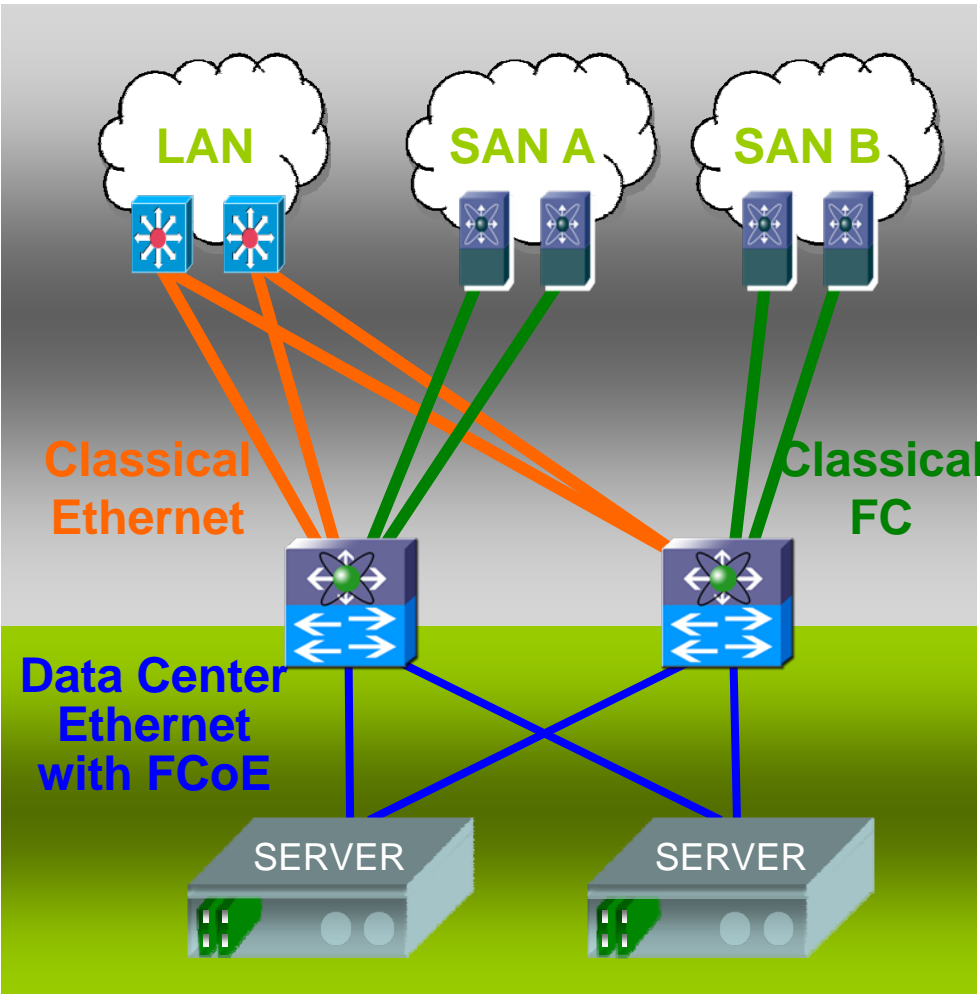


View from Operating System

- Standard drivers
- Same management
- Operating System sees:
 - ◆ Dual port 10 Gigabit Ethernet adapter
 - ◆ Dual Port 4 Gbps Fibre Channel HBAs



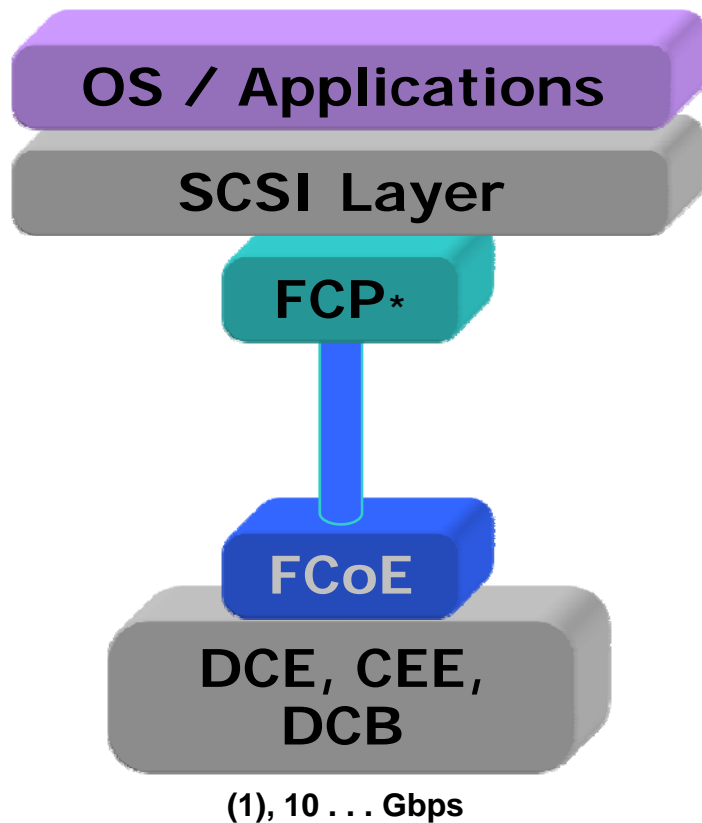
FCoE I/O Consolidation Topology



➤ FCoE Target:

- ◆ **Dramatic reduction in adapters, switch ports and cabling**
 - 4 cables to 2 cables per server
- ◆ **Seamless connection to the installed base of existing SANs and LANs**
- ◆ **High performance frame mappers vs. gateway bottlenecks**
- ◆ **Effective sharing of high bandwidth links**
- ◆ **Consolidated network infrastructure**
 - **Faster infrastructure provisioning**
- ◆ **Lower TCO**

FCoE based I/O Consolidation



- FCP layer untouched
- Requires Baby Jumbo Frames (2180 Bytes)
- Non IP routable Datacenter protocol
- Datacenter wide VLAN's
- Same management tools as for Fibre Channel
- Same drivers as for Fibre Channel HBA's
- Same Multipathing software
- Simplified certifications with storage subsystem vendors
- Requires lossless (10G) Ethernet switching fabric
- May require new host adaptors (unless FCoE software stack)
- Boot Support

* Includes FC Layer

Differentiators

Storage Part of I/O Consolidation

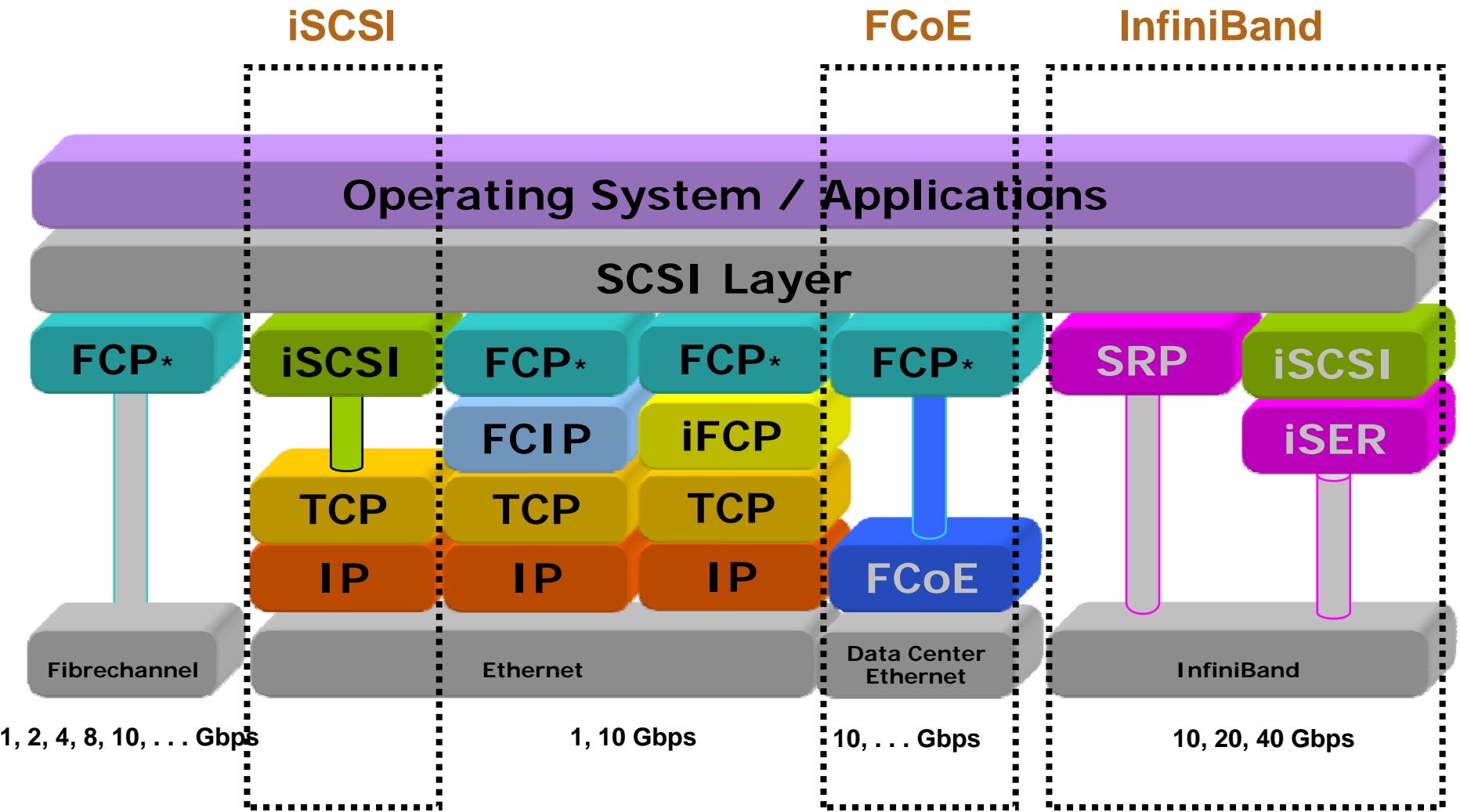
	iSCSI	FCoE	IB-SRP
Payload	SCSI	Fibre Channel	SCSI
Transport	TCP/IP	Data Center Ethernet	InfiniBand
Scope	LAN/MAN/WAN	Datacenter	Datacenter
Bandwidth/Performance	Low/Medium	High	High
CPU Overhead	High	Low	Low
Gateway Overhead	High	Low	High
FC Security Model	No	Yes	No
FC Software on Host	No	Yes	No
FC Management Model	No	Yes	No
Initiator Implementation	Yes	Yes	Yes
Target Implementation	Yes	Yes	Yes
IP Routable	Yes	No	N/A

Storage Part of I/O Consolidation

	iSCSI	FCoE	IB-SRP
Virtual Lanes	No	Yes	Yes
Congestion Control	TCP	Priority Flow Control	Credit based
Gateway Functionality	stateful	stateless	stateful
Connection Oriented	Yes	No	Yes
Access Control	IP/VLAN	VLAN / VF	Partitions
RDMA primitives	defined	defined	defined
Latency	Medium	Low	Very low
Adapter	NIC	CNA	HCA

Conclusion

Encapsulation Technologies



* Includes FC Layer

Conclusion

- Server I/O Consolidation is driven by high I/O bandwidth demand
- I/O Bandwidth demand is driven by Multicore / Socket Server and Virtualization
- TCP/IP (iSCSI), Data Center Ethernet (FCoE) and InfiniBand (SRP, iSER) are generic transport protocols allowing Server I/O Consolidation
- Server I/O Consolidation is the first phase, consolidating input into a Unified Fabric

Thank You !



Check out **SNIA Tutorial**:

- **Fibre Channel Technologies: Current and Future**
- **iSCSI tutorial / IP storage protocols**
- **Fabric Consolidation with InfiniBand**
- **FCoE: Fibre Channel over Ethernet**

- 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

**Gilles Chekroun
Errol Roberts
Walter Dey**

**Howard Goldstein
Bill Lulofs
Dror Goldenberg**

**Marco Di Benedetto
Carlos Pereira
James Long**

<http://www.fcoe.com/>

<http://www.t11.org/fcoe>

http://www.fibrechannel.org/OVERVIEW/FCIA_SNW_FCoE_WP_Final.pdf

http://www.fibrechannel.org/OVERVIEW/FCIA_SNW_FCoE_flyer_Final.pdf

<http://www.fibrechannel.org/FCoE.html>