Comparing Server I/O Consolidation Solutions

with an eye on Storage Networks

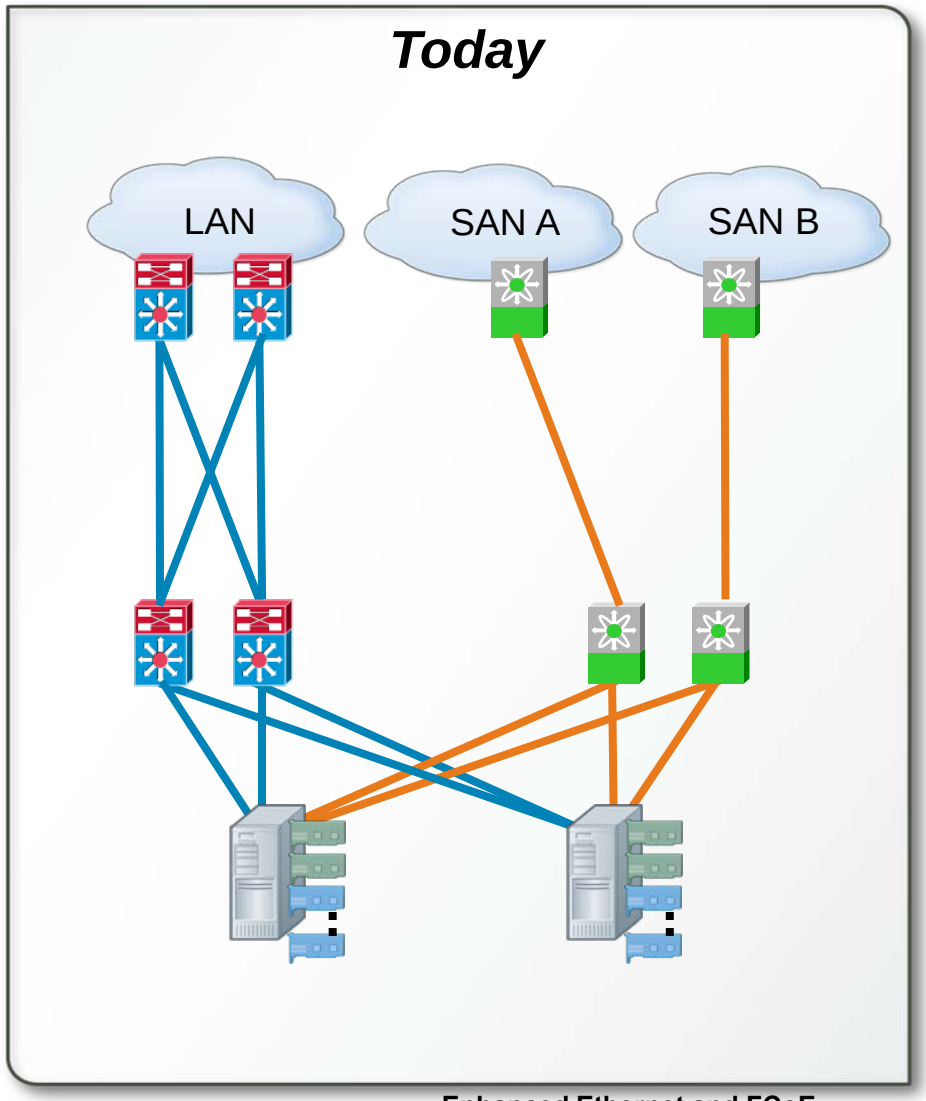
NSC'08 Storage track
Oct 14-15th, 2008 NSC, Linköping, Sweden

Bjørn R. Martinussen
Consulting System Engineer
Cisco Systems Europe
brm@cisco.com

Server I/O Consolidation

In the industry as a whole...

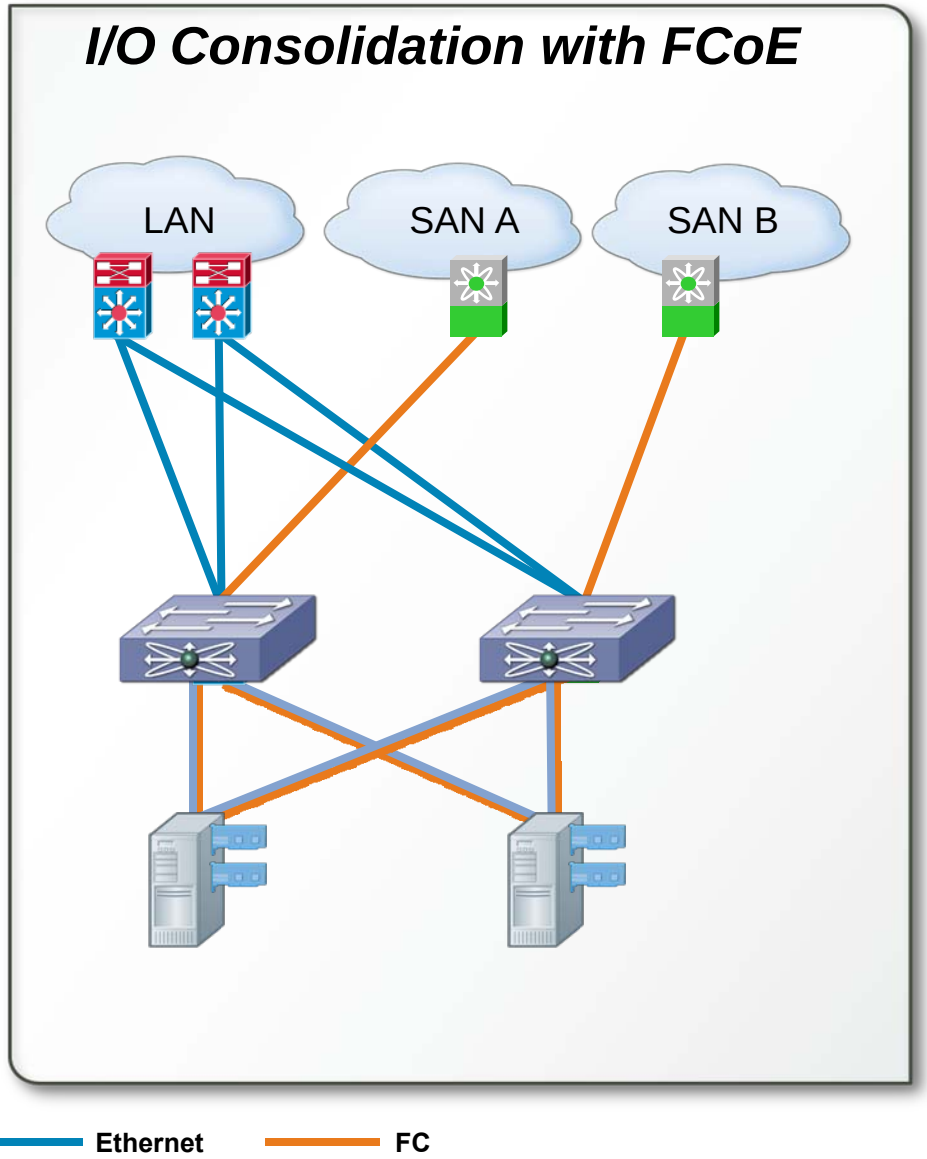
Server Networking Today



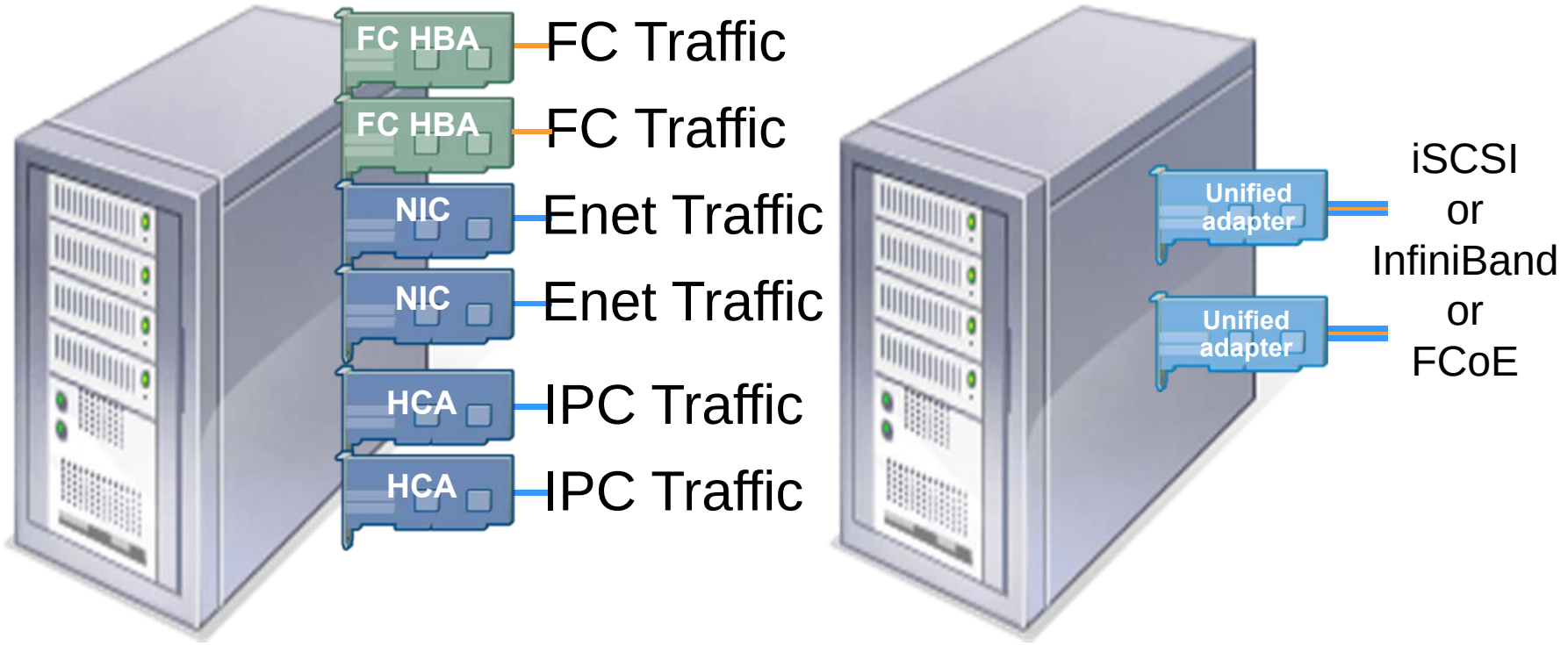
- **Today:**
  - **Parallel LAN/SAN Infrastructure**
  - **Inefficient use of Network Infrastructure**
  - **5+ connections per server – higher adapter and cabling costs**
    - **Adds downstream port costs;**
      cap-ex and op-ex
    - **Each connection adds additional points of failure in the fabric**
  - **Longer lead time for server provisioning**
  - **Multiple fault domains – complex diagnostics**
  - **Management complexity**

Server Networking Future

- **I/O consolidation**
  - **Reduction of server adapters**
  - **Simplification of access layer & cabling**
  - **Gateway free implementation – fits in installed base of existing LAN and SAN**
  - **L2 Multipathing Access – Distribution**
  - **Lower TCO**
  - **Fewer Cables**
  - **Investment Protection (LANs and SANs)**
  - **Consistent Operational Model**



Server Networking and I/O Consolidation



Adaptor: NIC for Ethernet/IP, HCA for InfiniBand, Converged Network Adaptor (CNA) for FCoE

Customer Benefit: Fewer NIC's, HBA's and cables, lower CapEx, OpEx (power, cooling)

Summary of Server I/O Consolidation Solutions

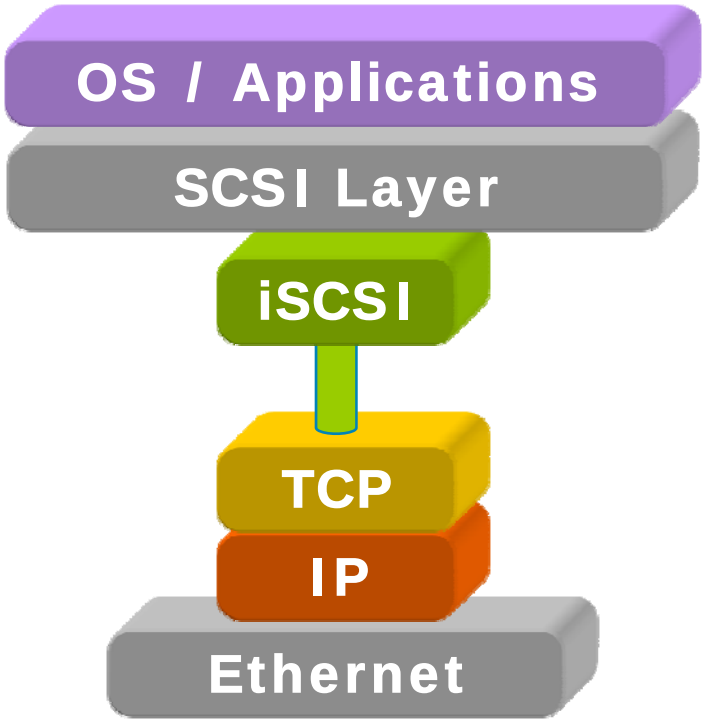
Server I/O Consolidation Solutions *)

- **iSCSI (could argue NAS technologies for files)**
  - **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 (Data Center Ethernet) and TCP/IP**
  - **SAN: Maps and transports Fibre Channel over Data Center Ethernet (lossless Ethernet) **)**

***) Past: Fibrechannel: SCSI and IP over FC; Future: IB over Ethernet (D. Goldenberg, Mellanox)**

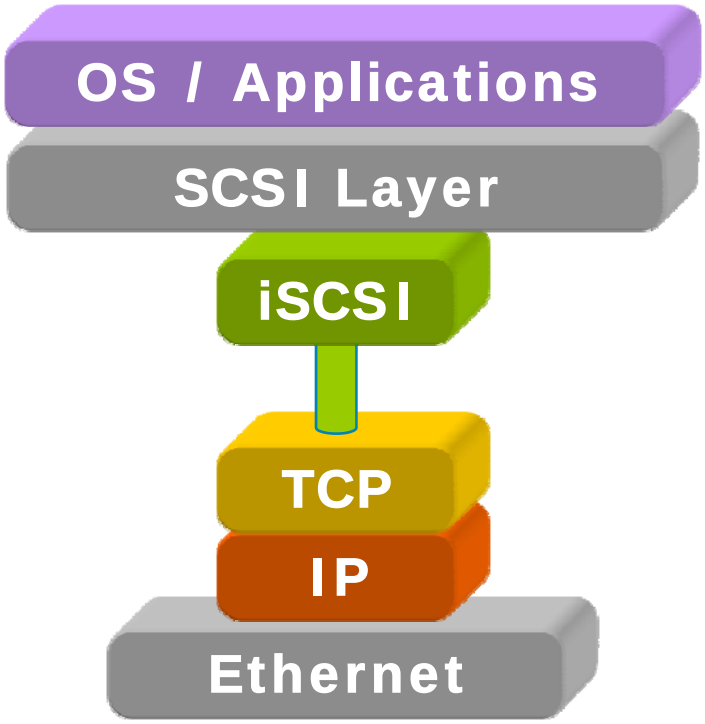
****) Data Center Ethernet is an architectural collection of Ethernet extensions designed to improve Ethernet networking and management in the Data Center; also called CEE (Converged Enhanced Ethernet), DCB (Data Center Bridging), DCE**

iSCSI



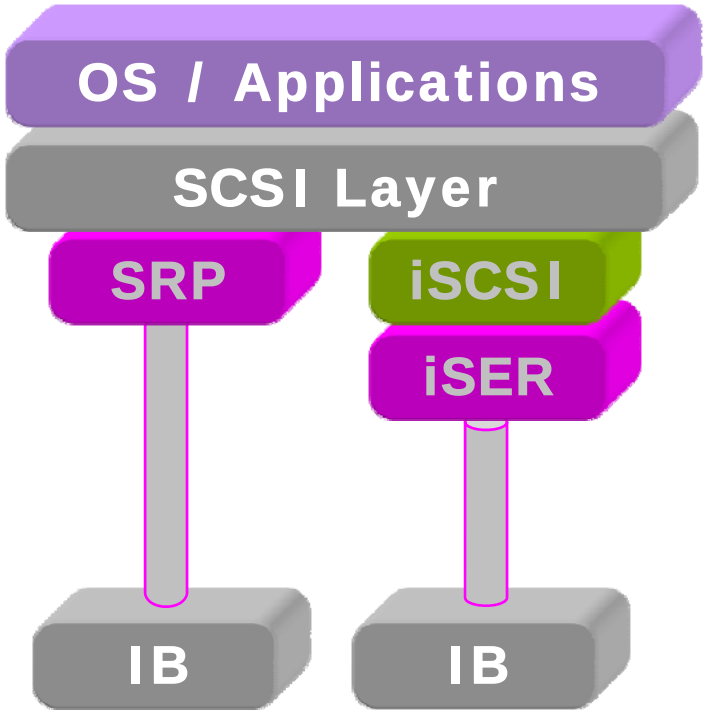
- A SCSI transport protocol that operates over TCP
  - Encapsulates SCSI CDBs (operational commands: e.g. read or write) and data into TCP/IP byte-streams (defined by SAM-2—SCSI Architecture Model 2)
  - Allows iSCSI Initiators to access IP-based iSCSI targets (either natively or via iSCSI-to-FC gateway)
- Standards status
  - 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 based I/O Consolidation



- Overhead of TCP/IP Protocol
- It's SCSI not FC
- LAN/Metro/WAN (Routable)
- Security of IP protocols (IPsec)
- Stateful gateway (iSCSI <-> FCP)
- Mainly 1G 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

InfiniBand



10, 20 Gbps (4X SDR/DDR)

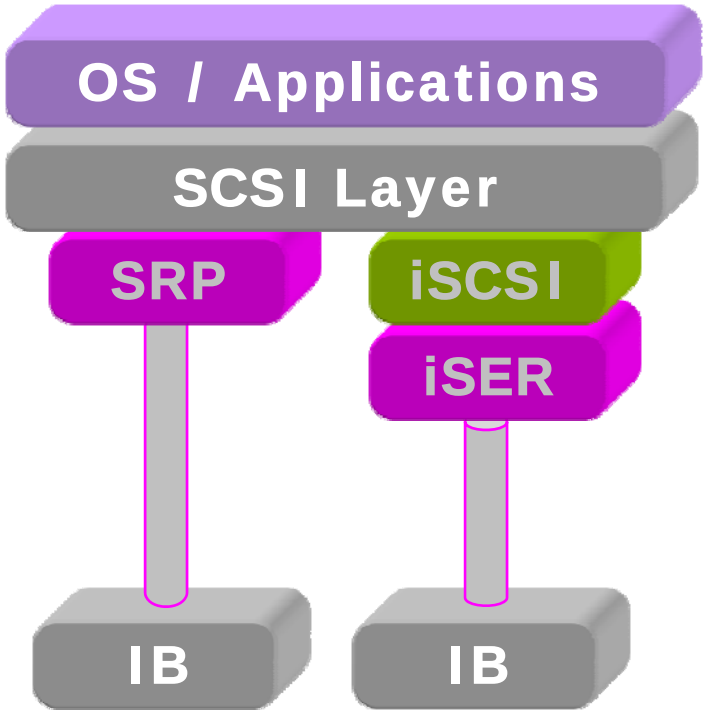
- Standards-based interconnect

  http://www.infinibandta.org
- Channelized, connection-based interconnect optimized for high performance computing
- Supports server and storage attachments
- Bandwidth Capabilities (SDR/DDR)

  4x—10/20 Gbps: 8/16 Gbps actual data rate

  12x—30/60 Gbps: 24/48 Gbps actual data rate
- Built-in RDMA as core capability for inter-CPU communication

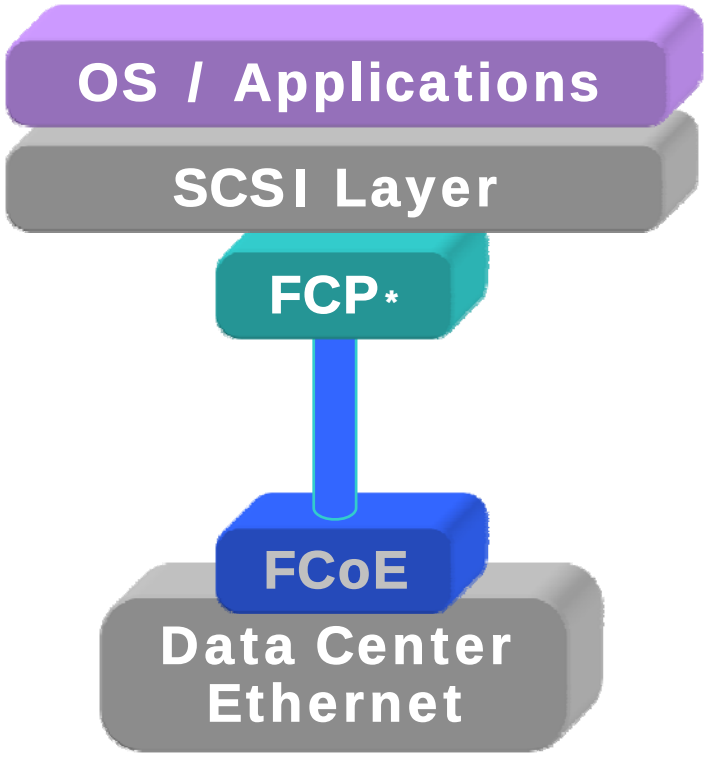
InfiniBand based I/O Consolidation



10, 20 Gbps (4X SDR/DDR)

- Requires new Eco system (HCA, cabling, switches)
- Mostly copper cabling, limited distance
- 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

FCoE

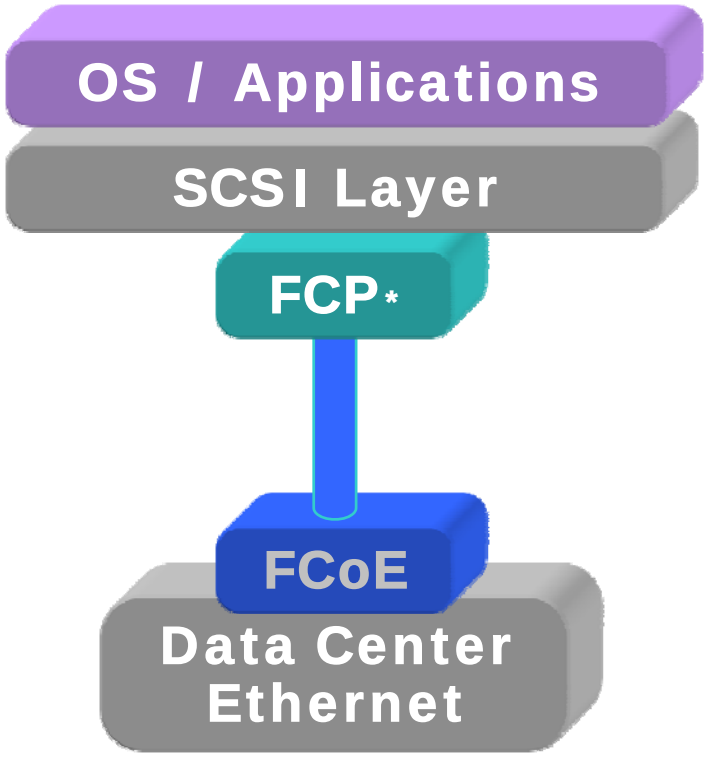


(1), 10 . . . Gbps

* Includes FC Layer

- From a Fibre Channel standpoint it's Fibrechannel 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

FCoE based I/O Consolidation



(1), 10 . . . Gbps

* Includes FC Layer

- FCP layer untouched
- Requires Baby Jumbo Frames (2180 Bytes)
- Nonroutable 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

Enabling Technologies

Three Challenges + One

Can Ethernet be Lossless?

Is a Credit Scheme Required?

Is Lossless Better?

PLUS...

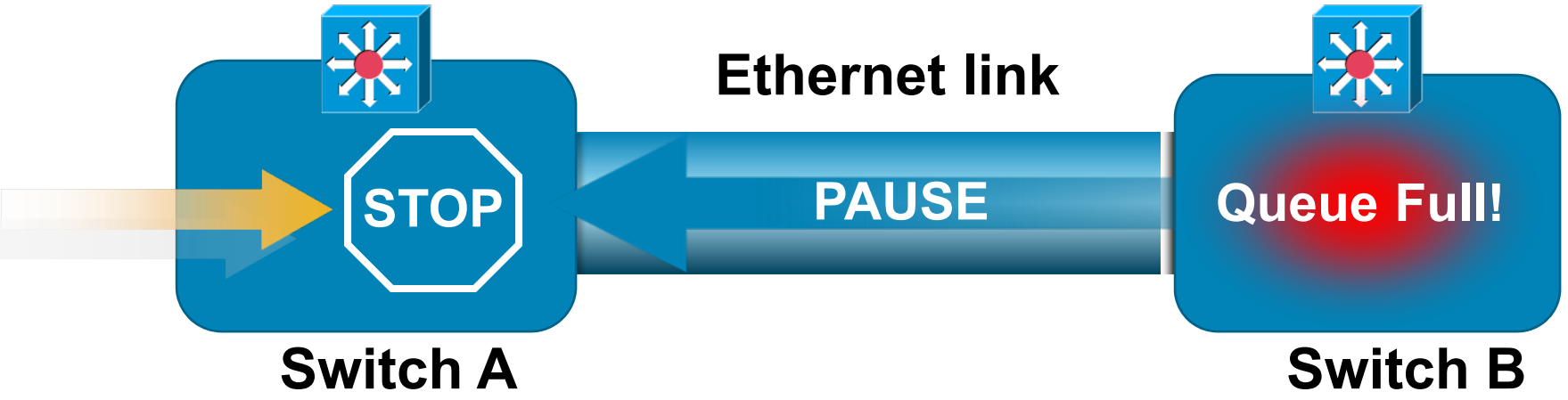
Is Anything Else Required?

Why Are Frames Lost?

Collision	Transmission Error	Congestion
• No longer present in Full Duplex Ethernet	• Very rare in the Data Center	• Most common cause • Congestion is a switch issue, not a link issue – A full duplex IEEE 802.3 link does not lose frames • It must be dealt with in the bridge/switch – By IEEE 802.1

Can Ethernet be Lossless?

- Yes, with Ethernet PAUSE Frame



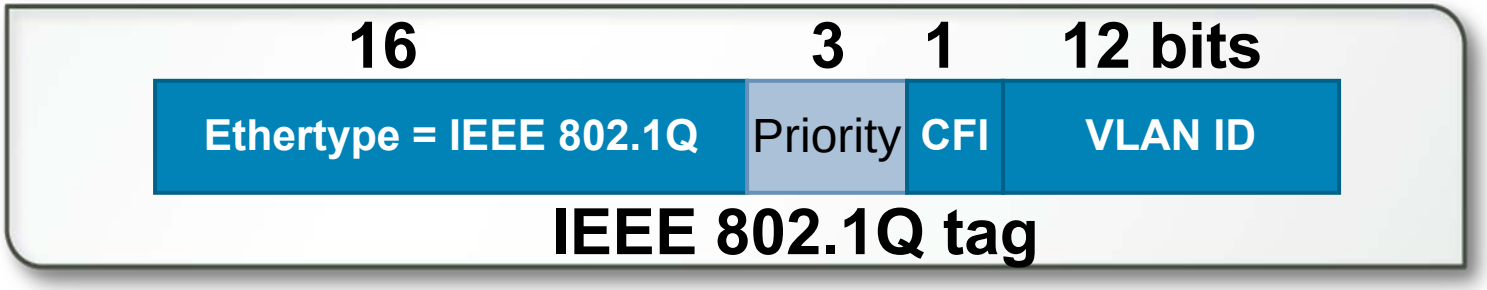
- Defined in IEEE 802.3 – Annex 31B
- The PAUSE operation is used to inhibit transmission of data frames for a specified period of time
- Ethernet PAUSE transforms Ethernet into a lossless fabric

Why is PAUSE not Widely Deployed?

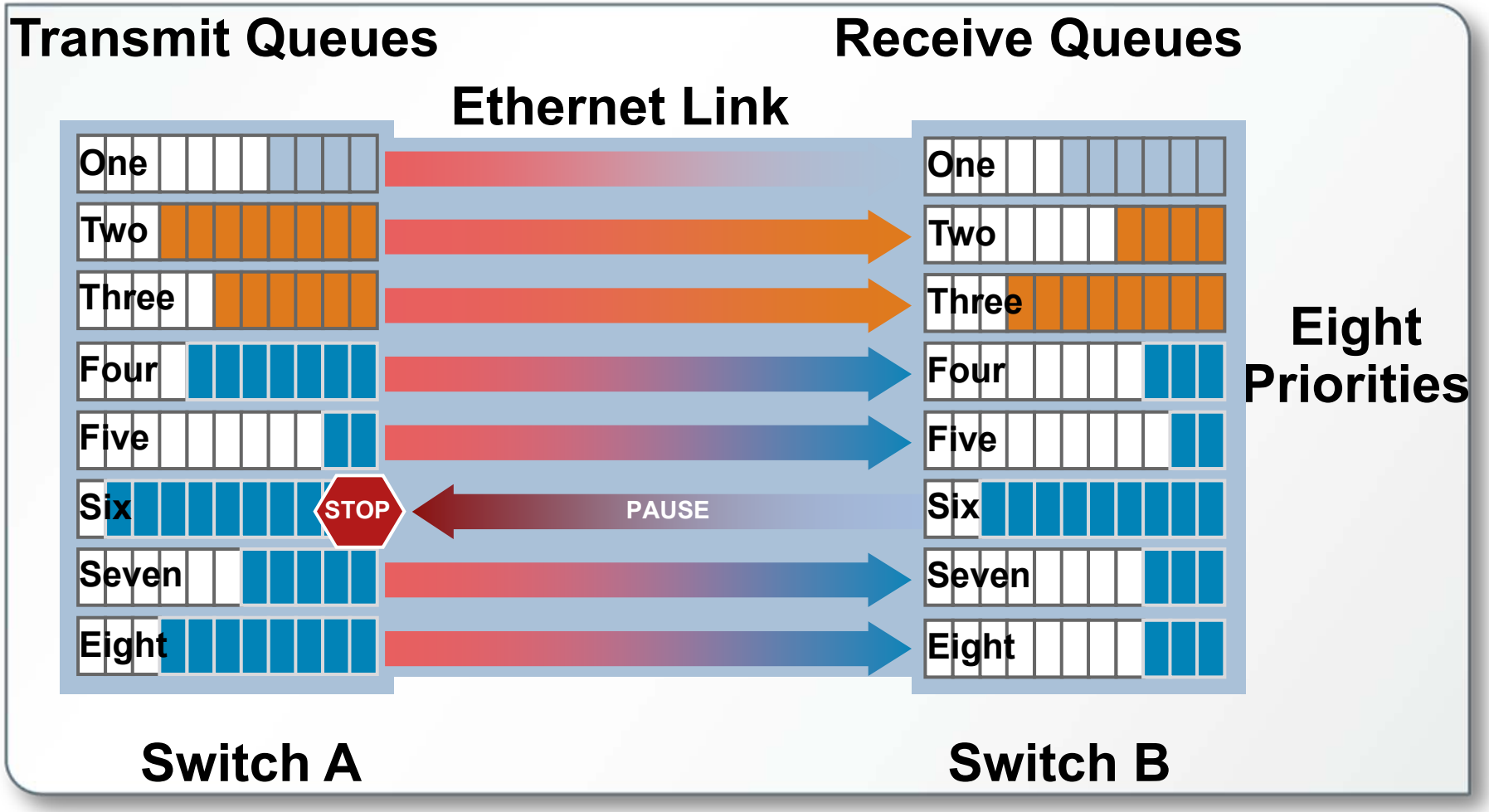
- Inconsistent implementations
  - Standard allows for asymmetric implementations
  - Easy to fix
- PAUSE applies to the whole links
  - Single mechanism for all traffic classes
- This may cause “traffic interference”
  - E.g. Storage traffic paused due to a congestion on IP traffic

Priority Flow Control (PFC)

- aka PPP (Per Priority Pause)
- PFC enables PAUSE functionality per Ethernet priority
  - IEEE 802.1Q defines 8 priorities
  - Traffic classes are mapped to different priorities:
    - no traffic interference
      - IP traffic may be paused while Storage traffic is being forwarded
      - Or, vice versa
  - Requires independent resources per priority (buffers)
- High level of industry support
  - Cisco distributed proposal
  - Standard Track in IEEE 802.1Qb



Priority Flow Control in Action



Is Lossless Better?

Plus

- Frames are not dropped
- FC over lossless Ethernet works well

Neutral

- TCP relies on losses
- We can run it on a priority where we do not enable Pause

Minus

- Congestion Spreading & Head of line blocking

Is Anything Else Required?

Yes

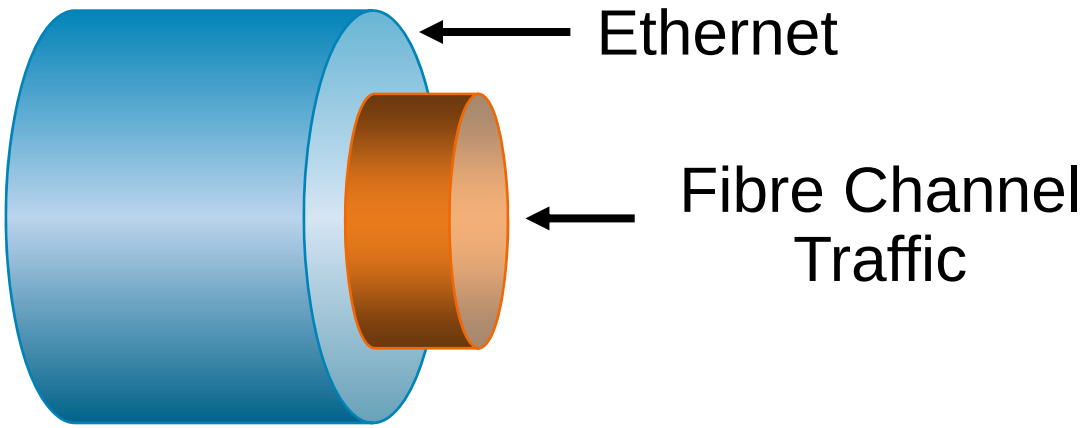
- In order to build a deployable I/O consolidation solution, the following additional components are required:

- Discovery Protocol (DCBX)
- Bandwidth manager
- Congestion Management

FCoE: Fibre Channel over Ethernet

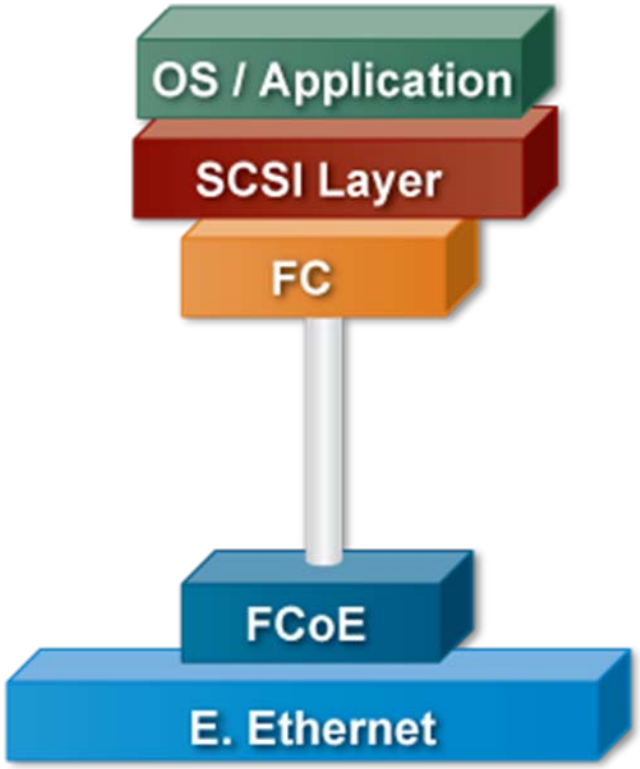
FCoE: FC over Ethernet

- FCoE is I/O consolidation of FC storage traffic over Ethernet
  - FC traffic shares Ethernet links with other traffics
  - Requires a lossless Ethernet fabric



FCoE Protocol Stack

- From a Fibre Channel standpoint it's FC connectivity over a new type of cable called an Ethernet cloud
- From an Ethernet standpoints it's yet another ULP (Upper Layer Protocol) to be transported



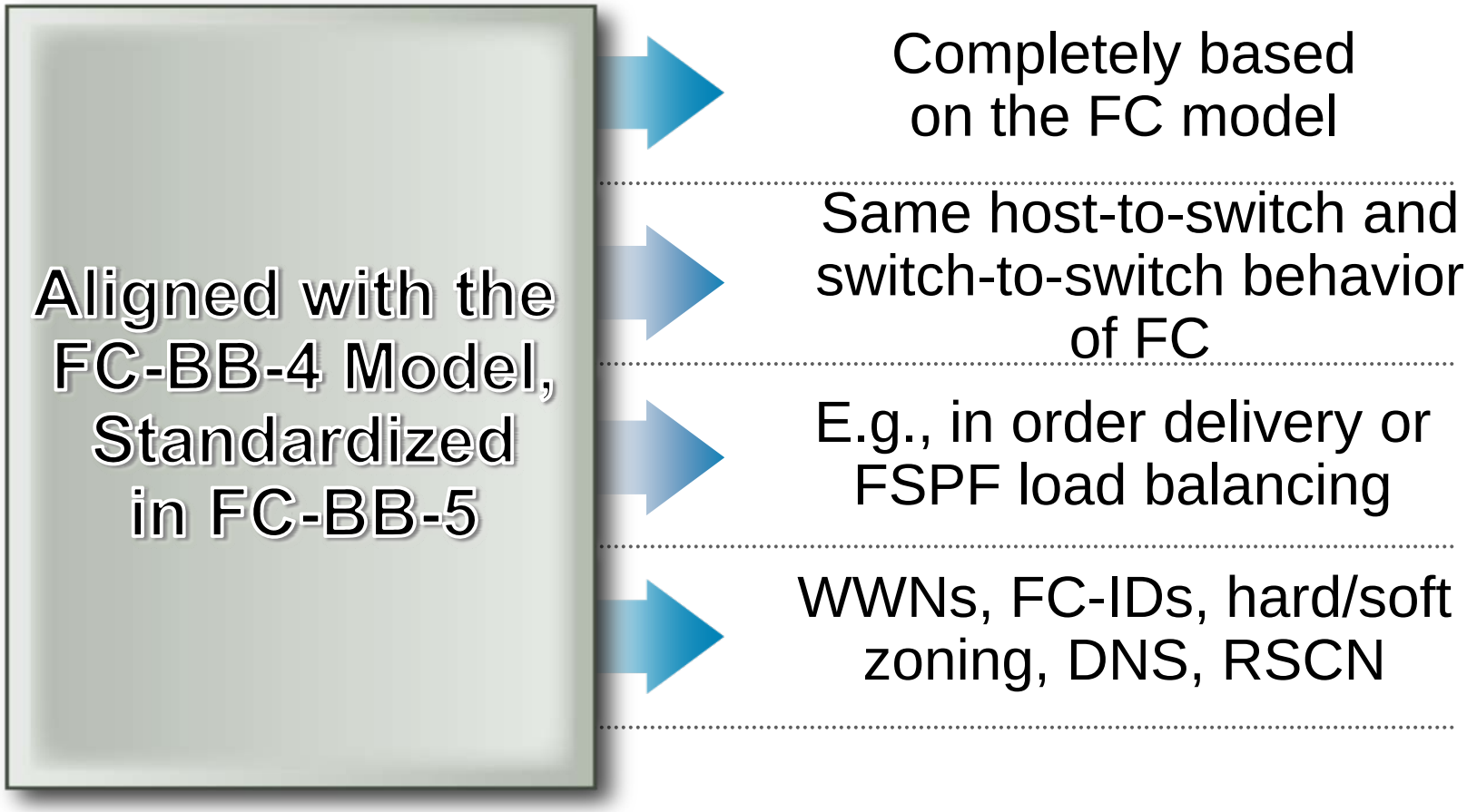
1, 10 . . . Gbps

FCoE Benefits

- FCoE benefits are the same of any I/O consolidation solution
  - Fewer Cables
    - Both block I/O & Ethernet traffic co-exist on same cable
  - Fewer adapters needed
  - Overall less power
- Plus additional advantages of being FC
  - Seamless integration with existing FC SANs
  - No Gateway

FCoE is Fibre Channel

FCoE is Fibre Channel at the host and switch level



Protocol Organization

FCoE is really two different protocols:

FCoE itself

- Is the data plane protocol
- It is used to carry most of the FC frames and all the SCSI traffic

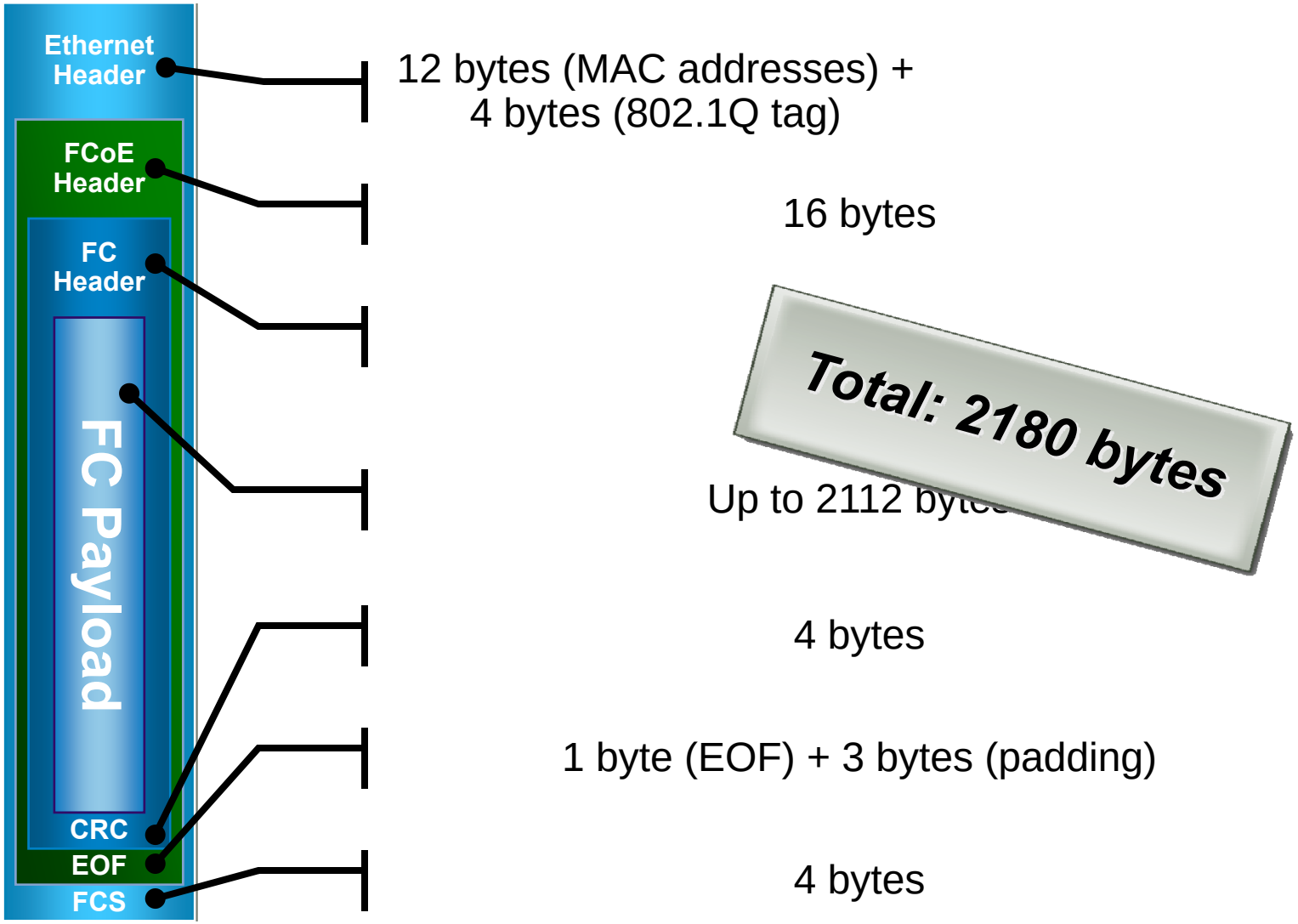
FIP (FCoE Initialization Protocol)

- It is the control plane protocol
- It is used to discover the FC entities connected to an Ethernet cloud
- It is also used to login to and logout from the FC fabric

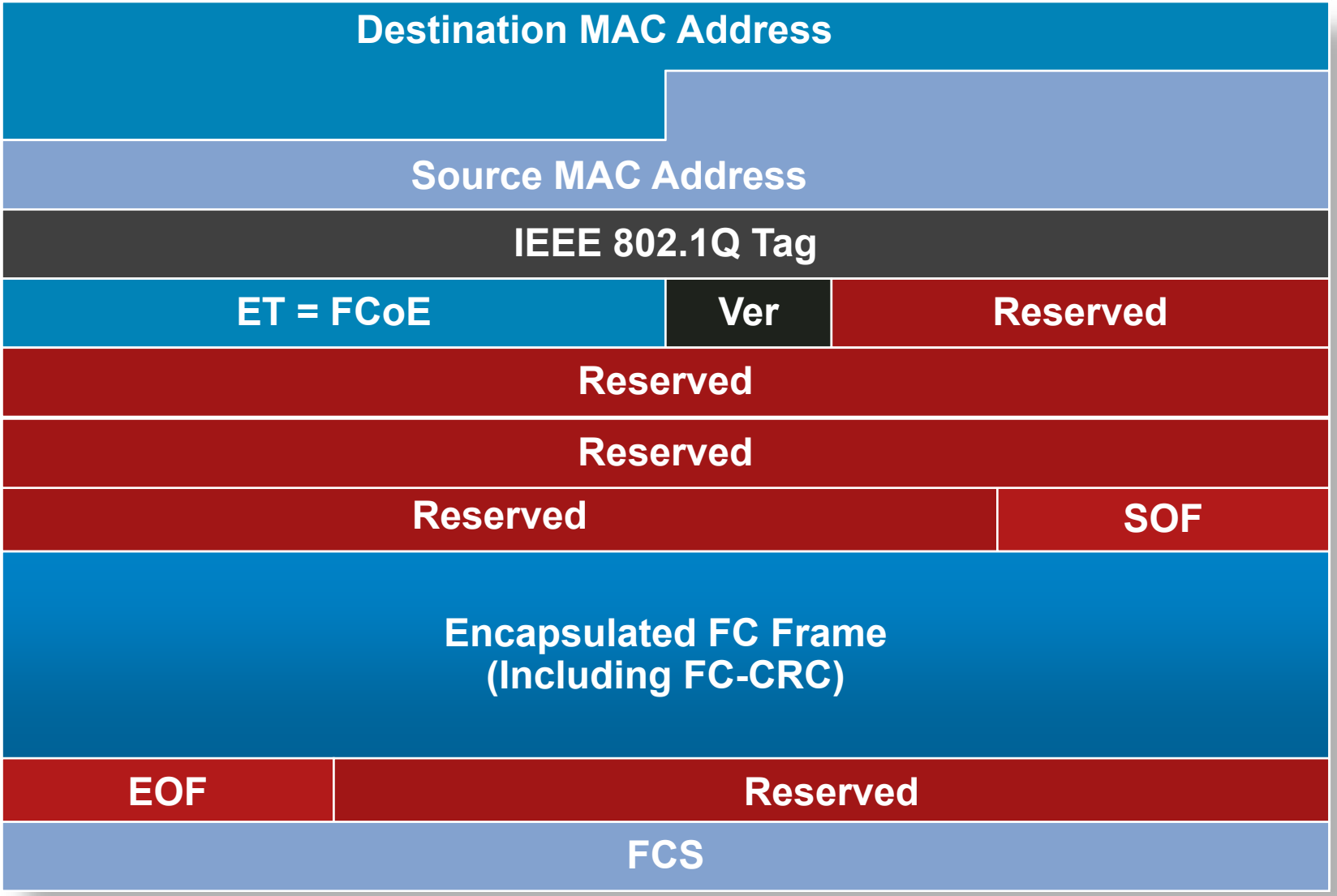
The two protocols have:

- Two different Ethertypes
- Two different frame formats

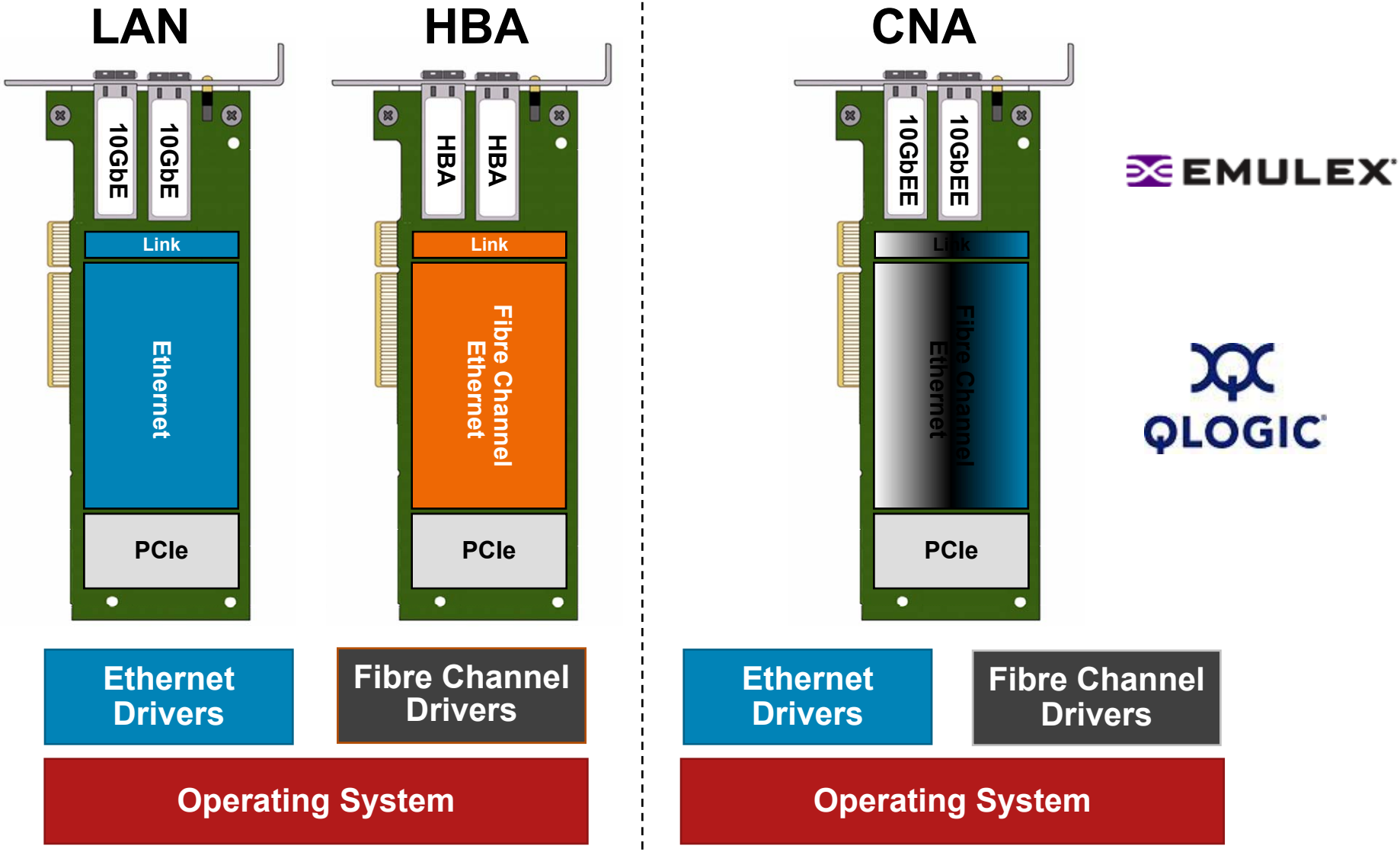
FCoE frame size



FCoE Frame Format



CNA: Converged Network Adapter

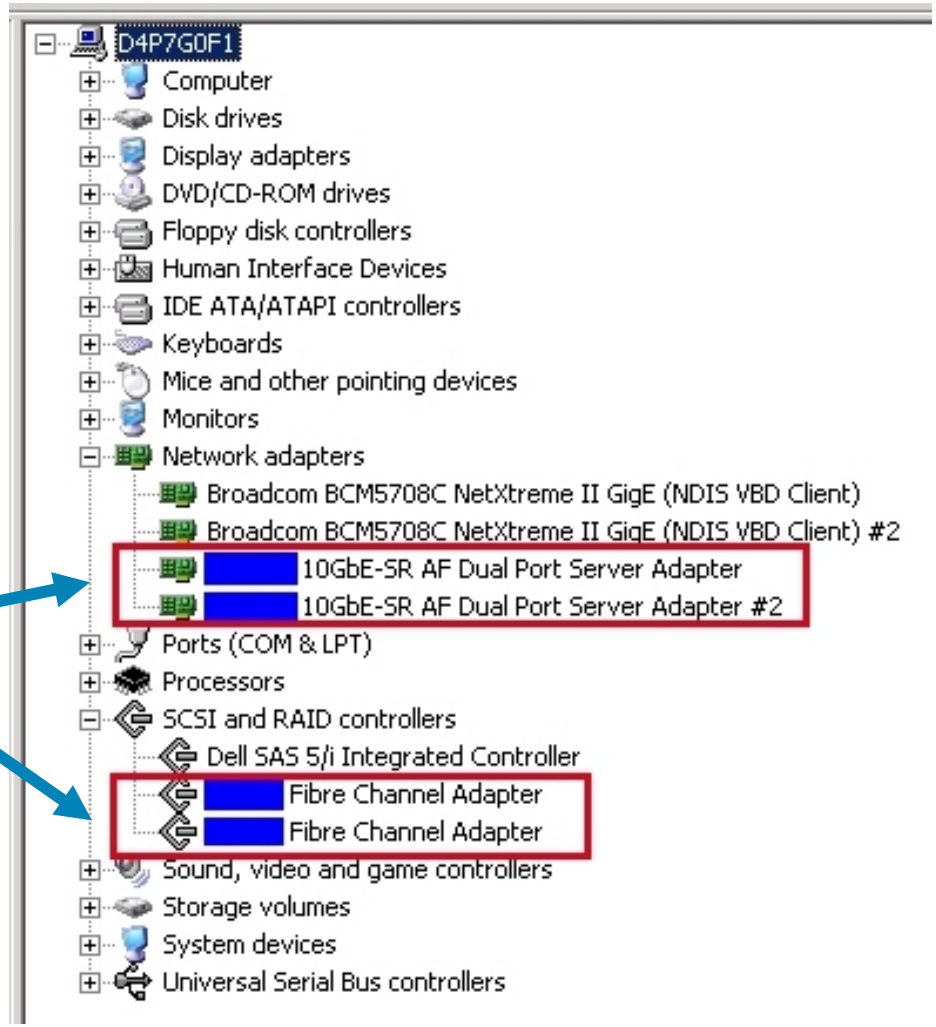


View from Operating System

- Standard drivers
- Same management
- Operating System sees:

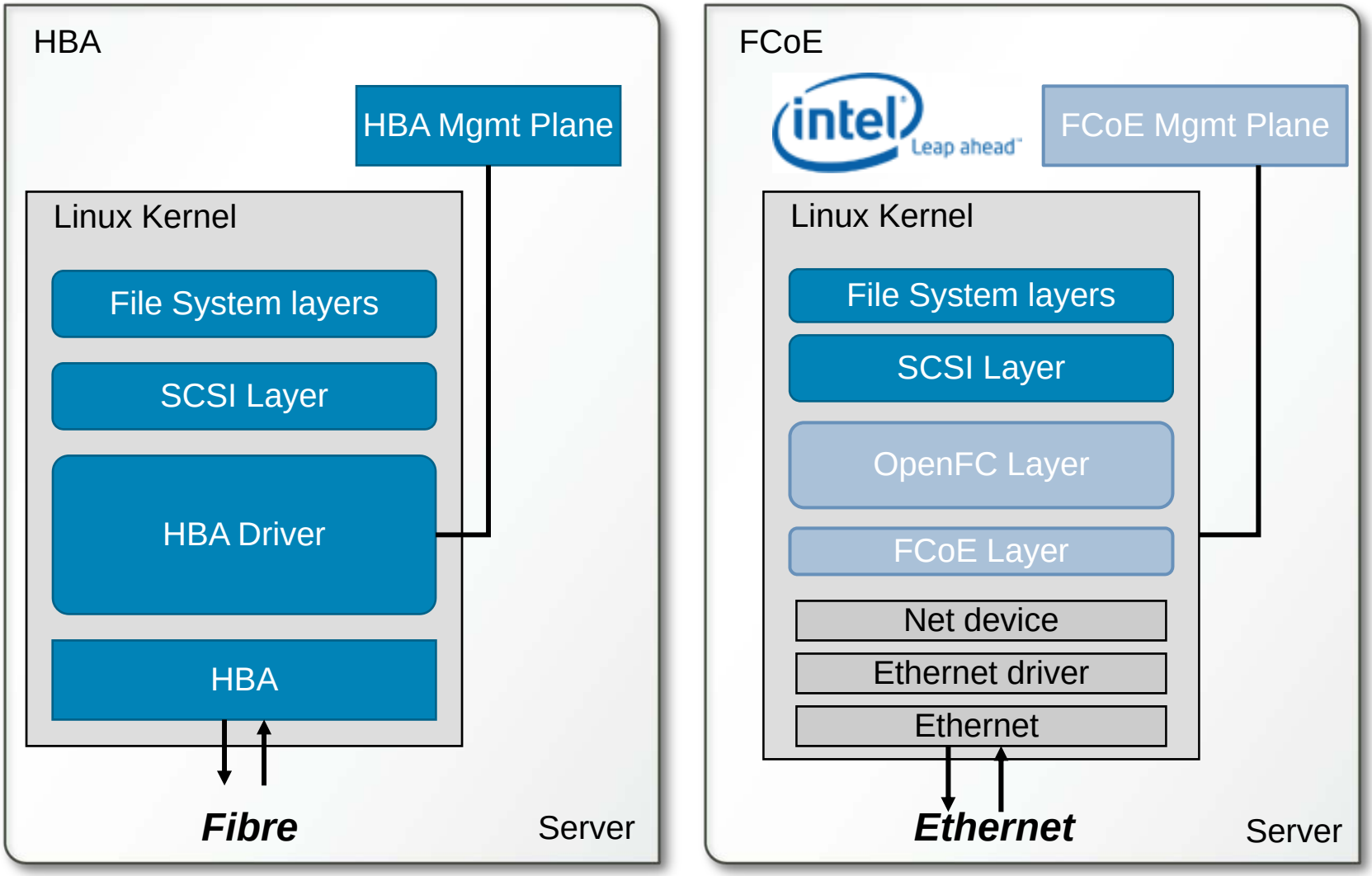
  Dual port 10 Gigabit Ethernet adapter

  Dual Port 4 Gbps Fibre Channel HBAs



Open-FCoE Software

www.open-fcoe.org



FCoE delivers same performance as FC and at 25% lower cost

**intel case study IDFAugust 2008*

Summary

Comparison of Selected Characteristics

	iSCSI	FCoE	IB-SRP
Virtual Lanes	No	Yes (8)	Yes(16)
Congestion Control	TCP	PFC	Credit based
Gateway Functionality	stateful	stateless	stateful
Connection Oriented	Yes	No	Yes
Access Control	IP/VLAN	VLAN / VSAN	Partitions
RDMA primitives	defined	defined	defined
Latency	100s of µs	10s of µs	µs
Adapter	NIC	CNA	HCA