hochschule hof
University of Applied Sciences

UNIVERSIDAD DE MÁLAGA

Overview of advanced storage technologies and storage virtualization

Pablo Pérez Trabado

Dept. of Computer Architecture

University of Malaga (Spain)

Disclaimers

- Some of the figures in the slides are taken from SNIA Education tutorials. To comply with the SNIA conditions of use for these documents, any SNIA tutorial used in the elaboration of these slides has been also provided as a handout

- Many (but not all) figures in the slides have been taken from Internet-available resources, including Wikipedia. Whenever possible by copyright restrictions, the original source has been also provided as a handout along with the slides. Also, whenever possible the original source is quoted. In any case, no ownership claims are made over images in the slides not drawn by the author himself.

hochschule hof
University of Applied Sciences

UNIVERSIDAD DE MÁLAGA

UNIVERSIDAD MÁLAGA

Overview of storage networking technologies

What will we learn?

- The concept of storage networking
- How the Fibre Channel technology provides a storage-specific, high-performance and highly reliable networked interconnection, called Storage Area Network (SAN)
- How a SAN allows both sharing physical storage and data between servers to form a clustered architecture
- High-availability, multipathing and load-balancing
- How SCSI block I/O can be implemented over Ethernet networks using iSCSI or FCoE
- The difference between SAN (block I/O) and NAS (file I/O)
- How a NAS provides an alternative way to implement a storage network

Definition of storage networking

- Term *storage networking* identifies any system in which storage devices are accessed over networked interconnection and transport technologies

- Currently, storage networking can be implemented using one of these options:

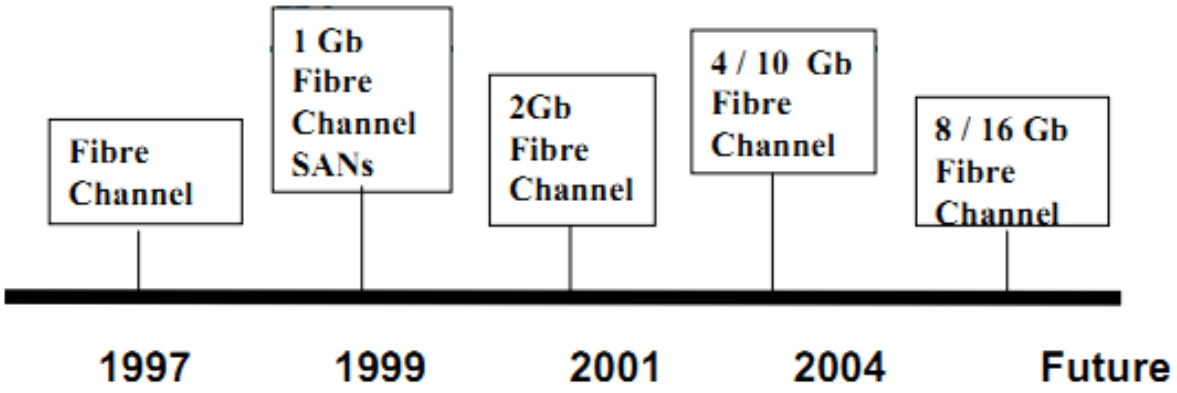
  - SCSI payload transported over storage-specific high-speed network technologies: Fibre Channel

  - SCSI payload transported over standard computer networks technologies: iSCSI, FCoE (Fibre Channel over Ethernet)

  - File-sharing protocols transported over standard computer networks technologies: NAS (Network-Attached Storage)

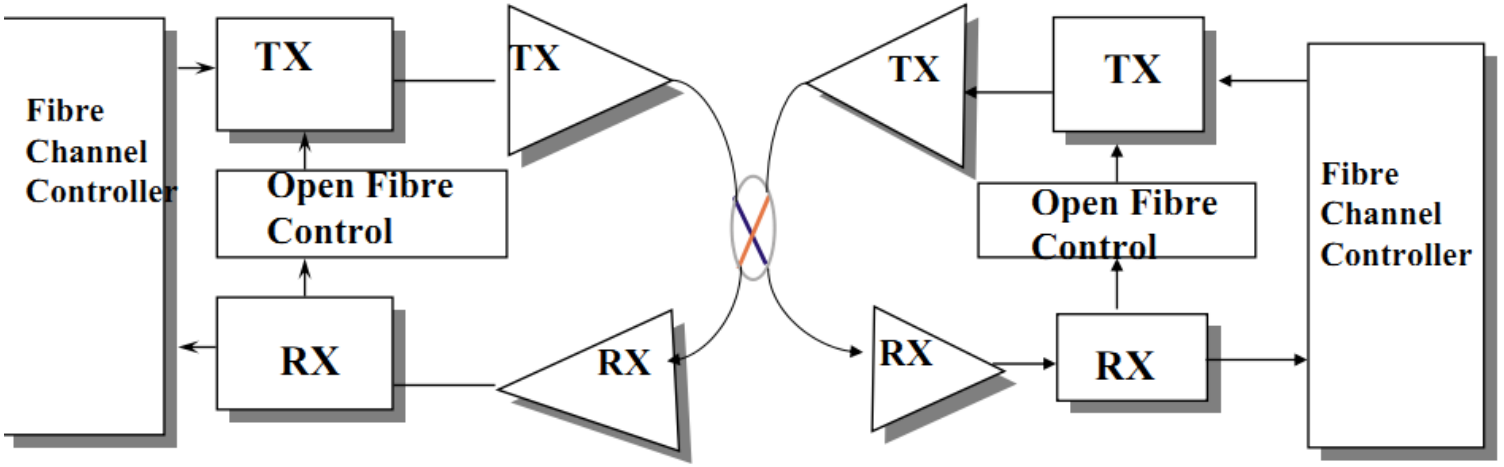
FIBRE CHANNEL

Fibre Channel Technology



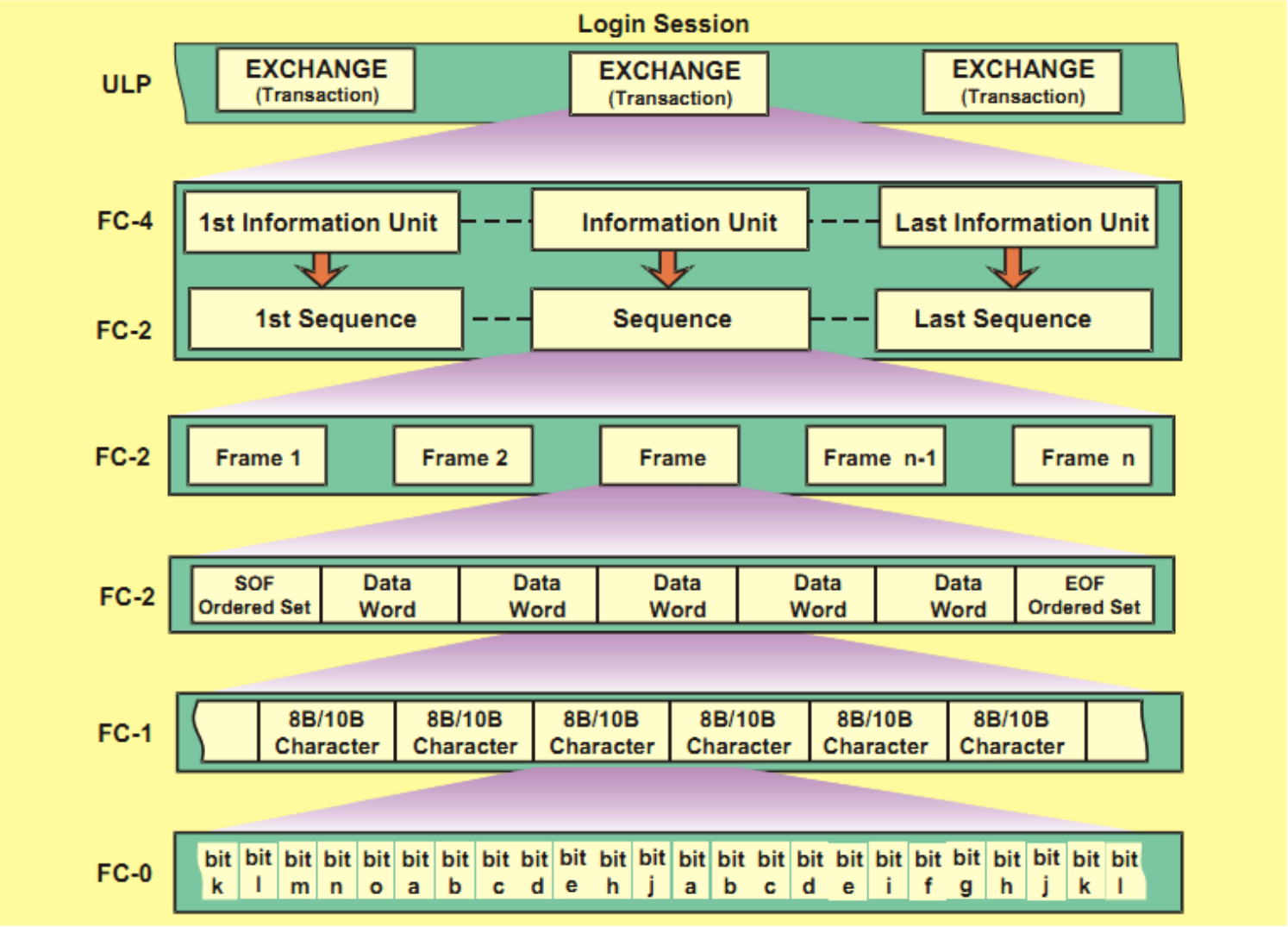
- Fibre Channel (FC) is an open standard for networked serial data transfer, which incorporates:
  - "Channel transport" characteristics of an I/O bus
  - Flexible connectivity and distance reliability of traditional networks
- Current day FC implementations allow reliable operation up to 16 Gbps per link

Fibre Channel Technology

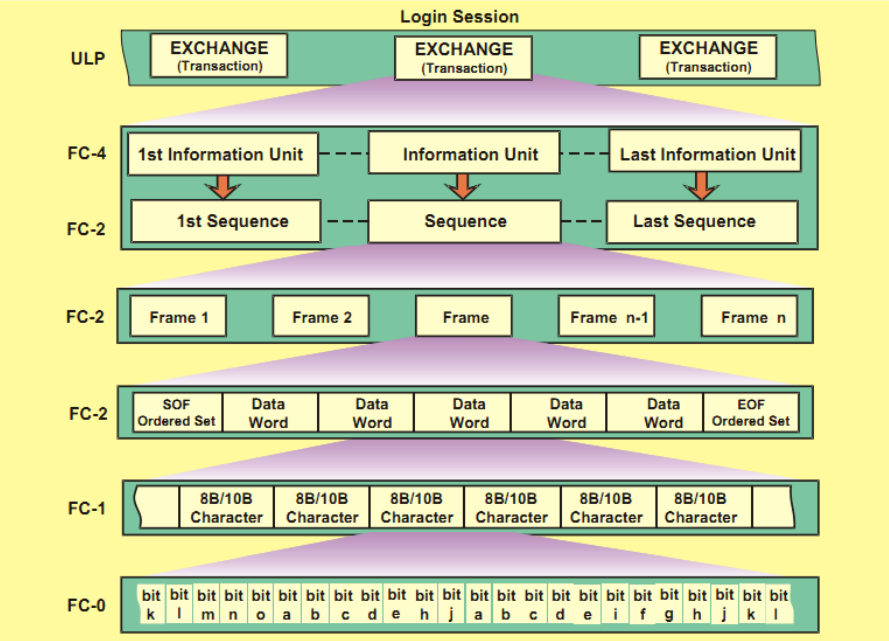


- FC uses a serial, bi-directional link
  - Transmission over optical fiber
  - Transmission over copper, using differential signaling, also standardized
    - No longer used for external interconnects
    - Possibly used by manufacturers for backplanes of FC-to-FC disk arrays
  - Use of optical fiber ensures low BER (Bit Error Rate) and high reliability for long links (tens or a few hundred meters)

FC protocol stack

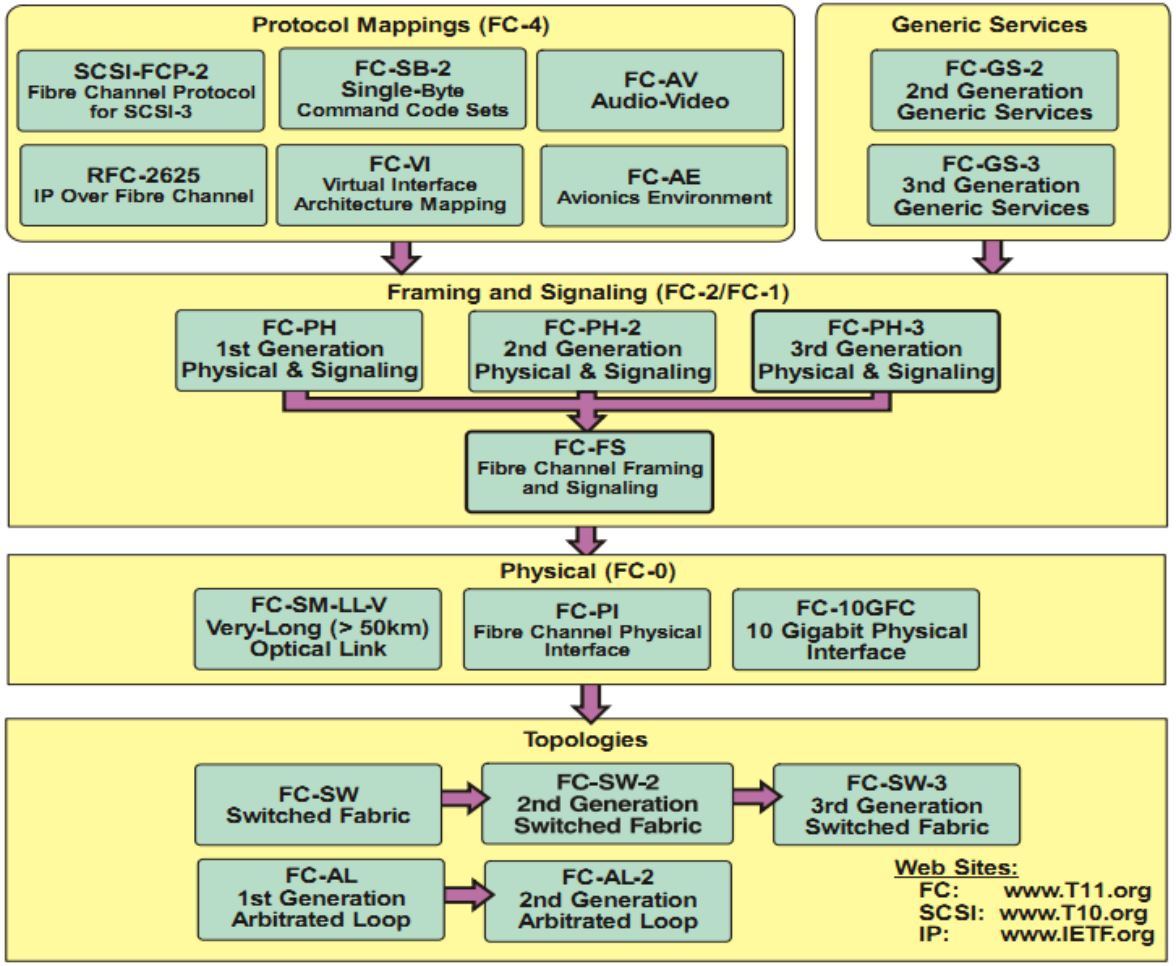


FC protocol stack



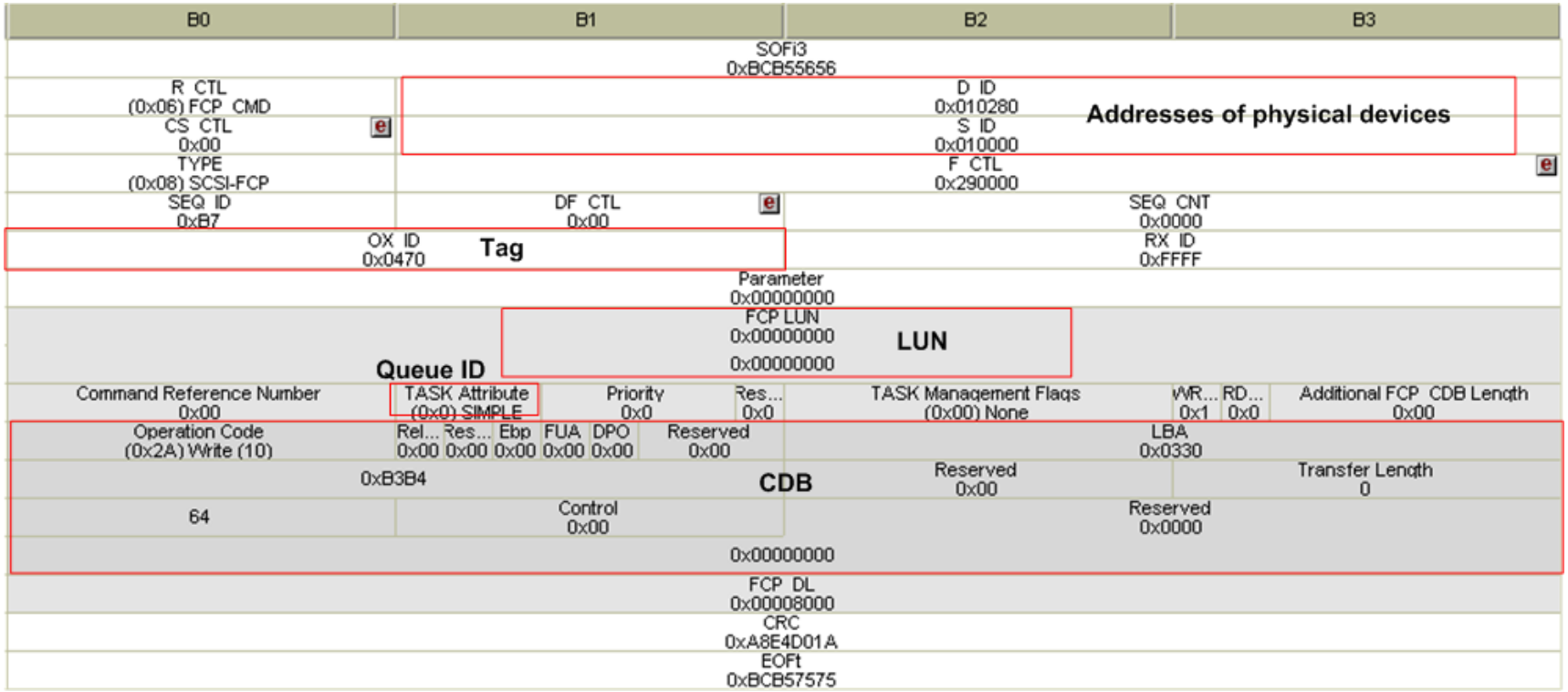
- FC standard is structured as a layered protocol stack
- FC-0 = physical level, signaling
- FC-1 and FC-2 = framing and link-level control
- FC-4 = Transport level. Carries payload edge-to-edge

FC protocol stack



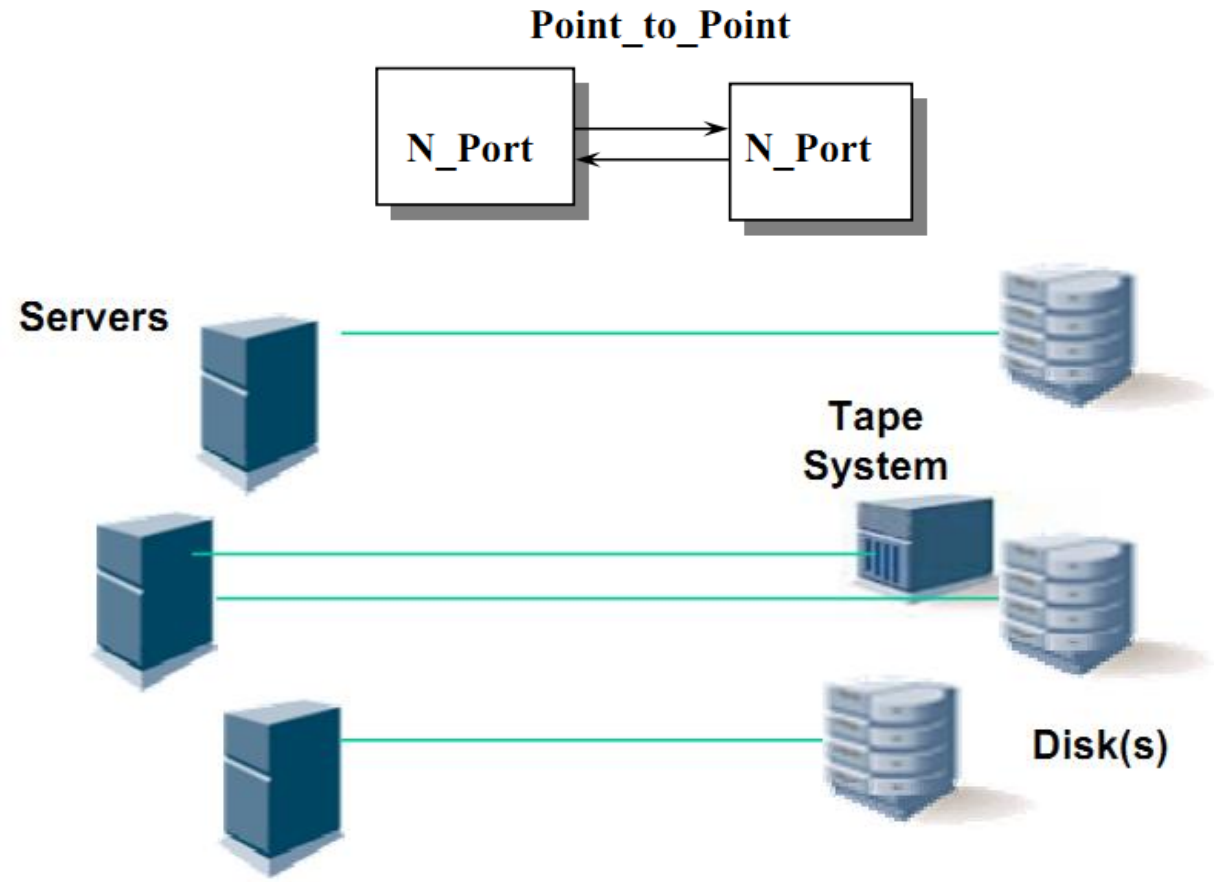
- Typical FC-4 payload is SCSI protocol (FCP)

FC protocol stack



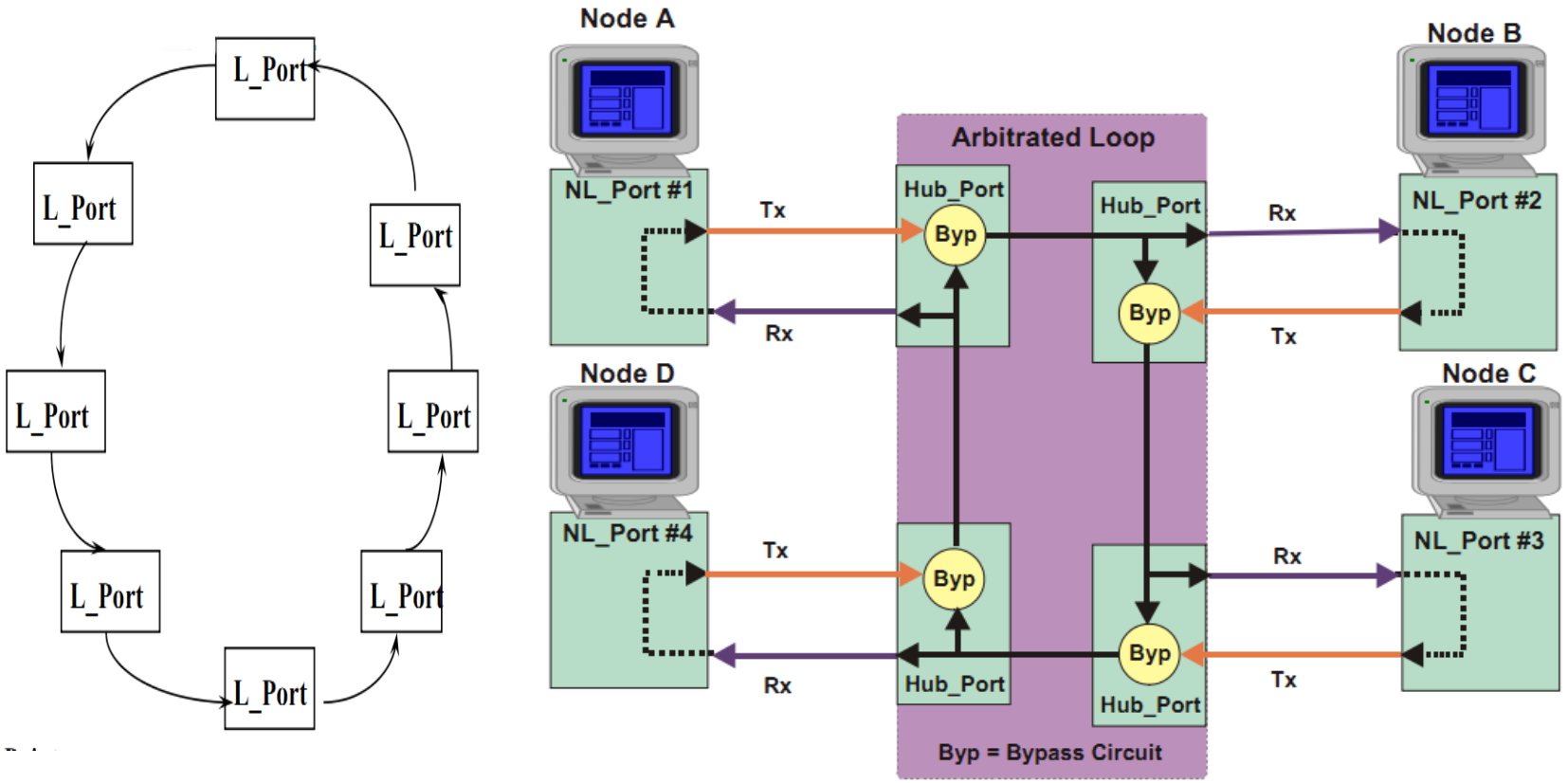
- SCSI protocol commands and responses are encapsulated within payload area of FC frame

Fibre Channel topologies



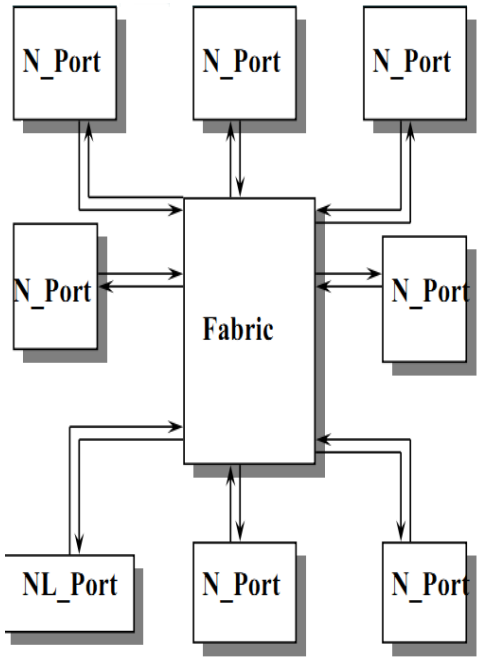
- Simplest topology for FC is DAS (Direct-Attach Storage)
  - Point-to-point interconnection between server and storage

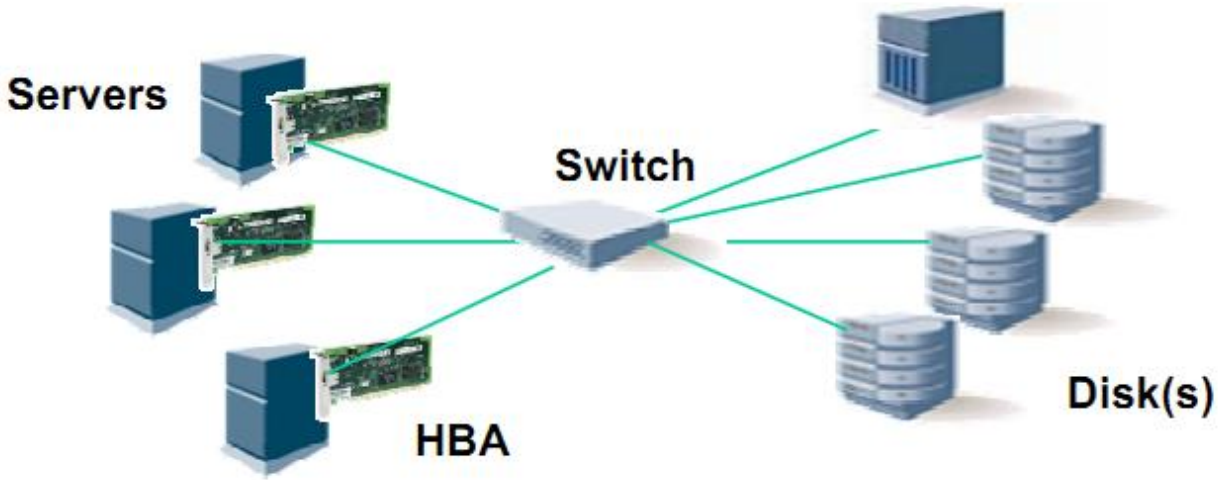
Fibre Channel topologies



- FC supports also arbitrated loop topology
  - Can use hub, but also workable just with cabling
  - Limits overall transfer speed (“bandwidth”) of system

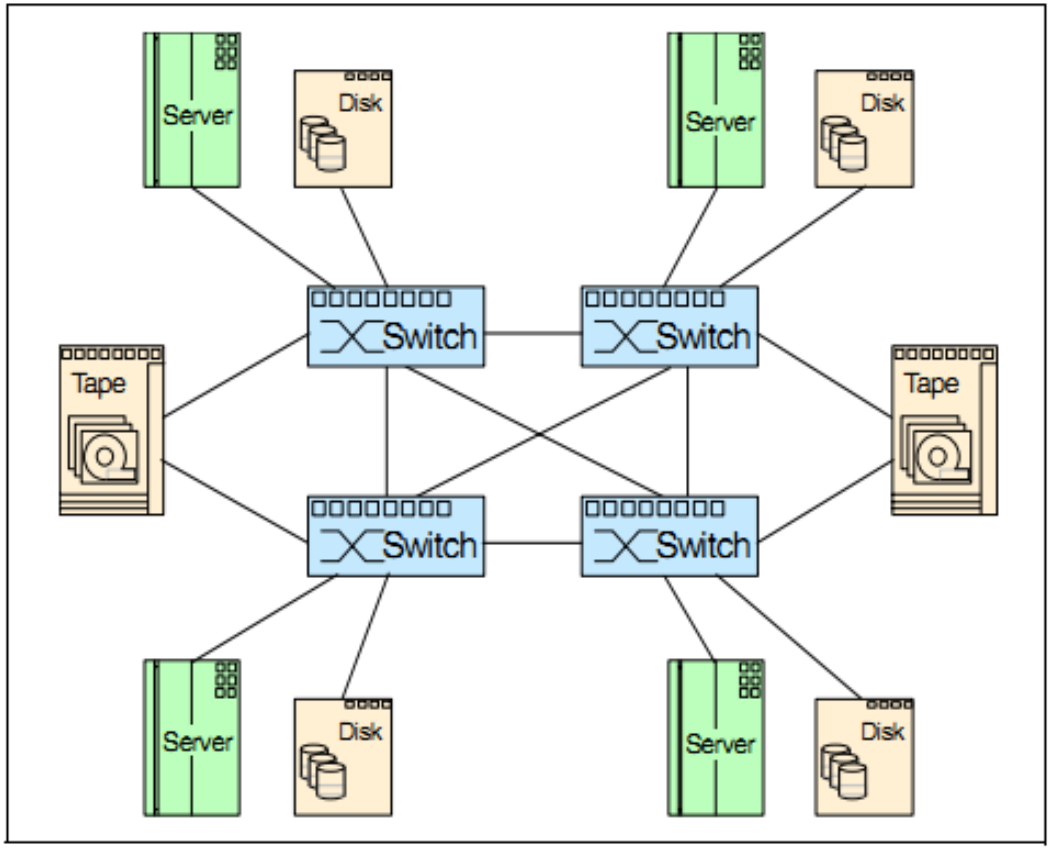
Fibre Channel topologies





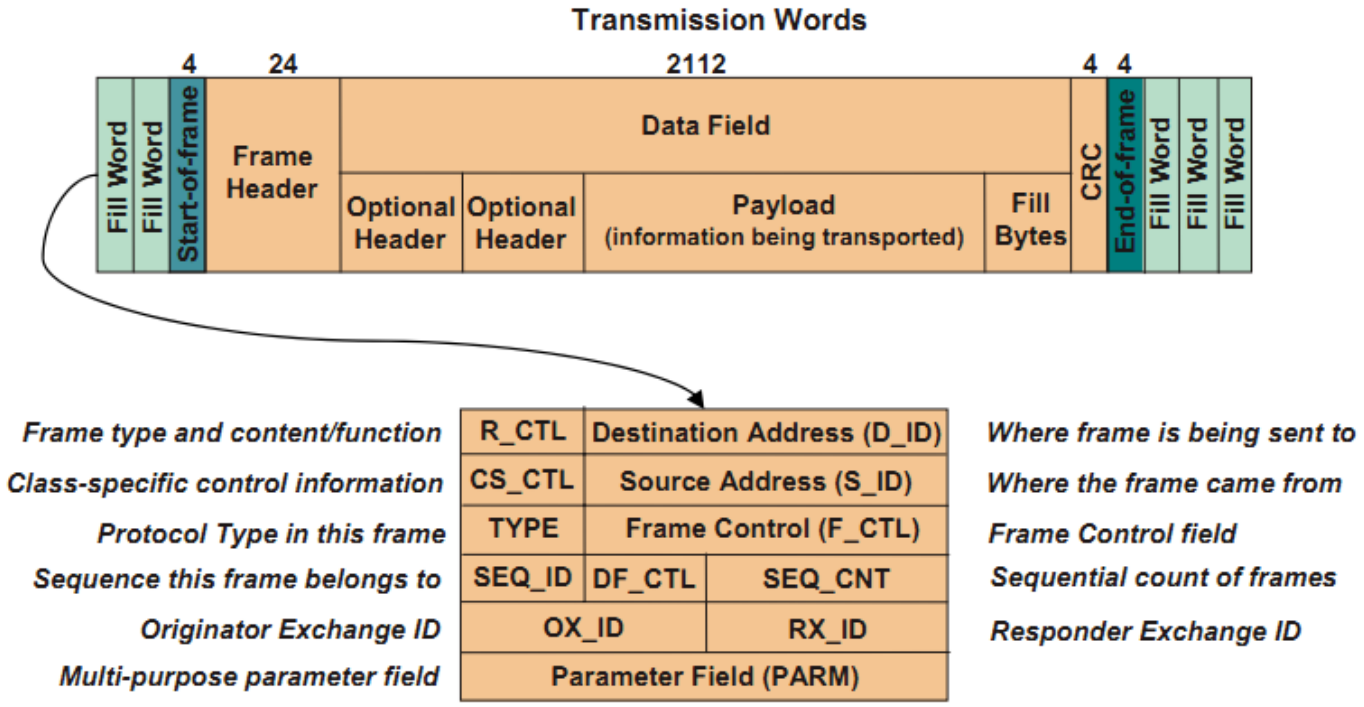
- Most versatile topology is *Fabric*
  - Truly meshed network, using a fabric switch
  - Devices connected by point-to-point links to switch
  - Switch moves data between links
    - While aggregated bandwidth of switch holds, links can work at maximum speed

Fibre Channel topologies



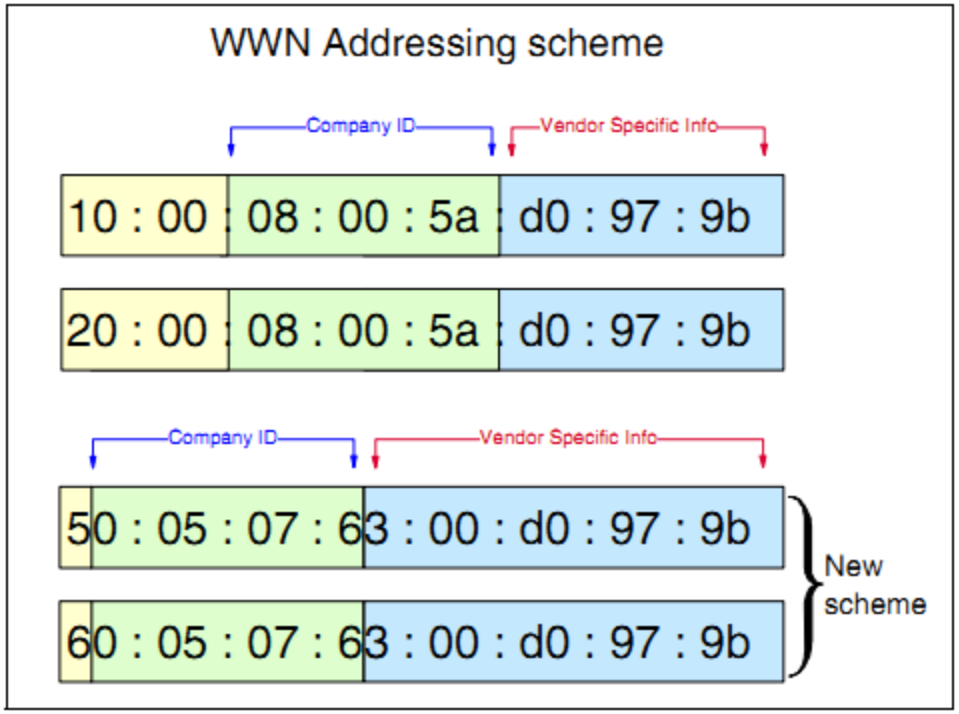
- FC switches can be interconnected to form a complex, switched fabric
  - Known as *meshed topology*

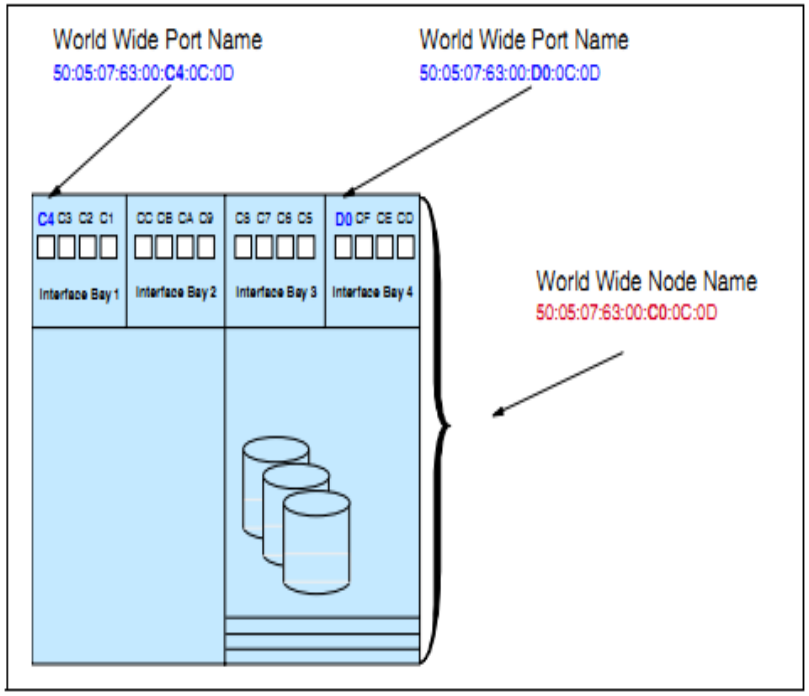
FC addressing



- FC frame requires unambiguous addressing for reliable travel through fabric
  - Frame uses 24-bit address for destination (D_ID) and source (S_ID)
  - Address must uniquely identify each FC port of each FC device

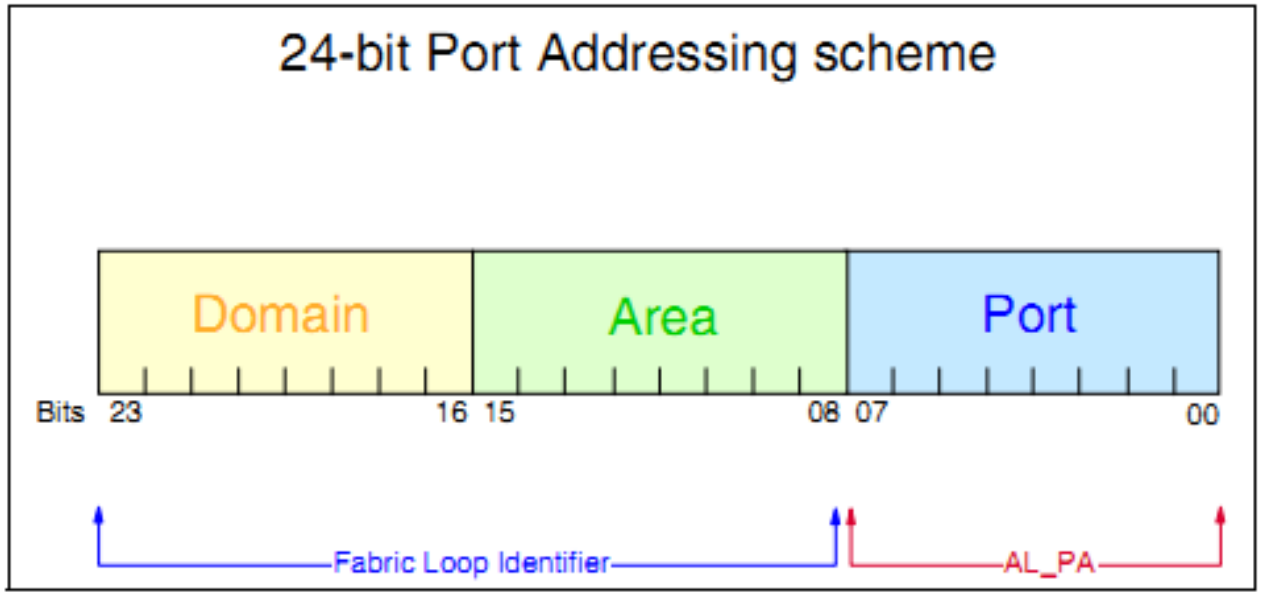
FC Addressing





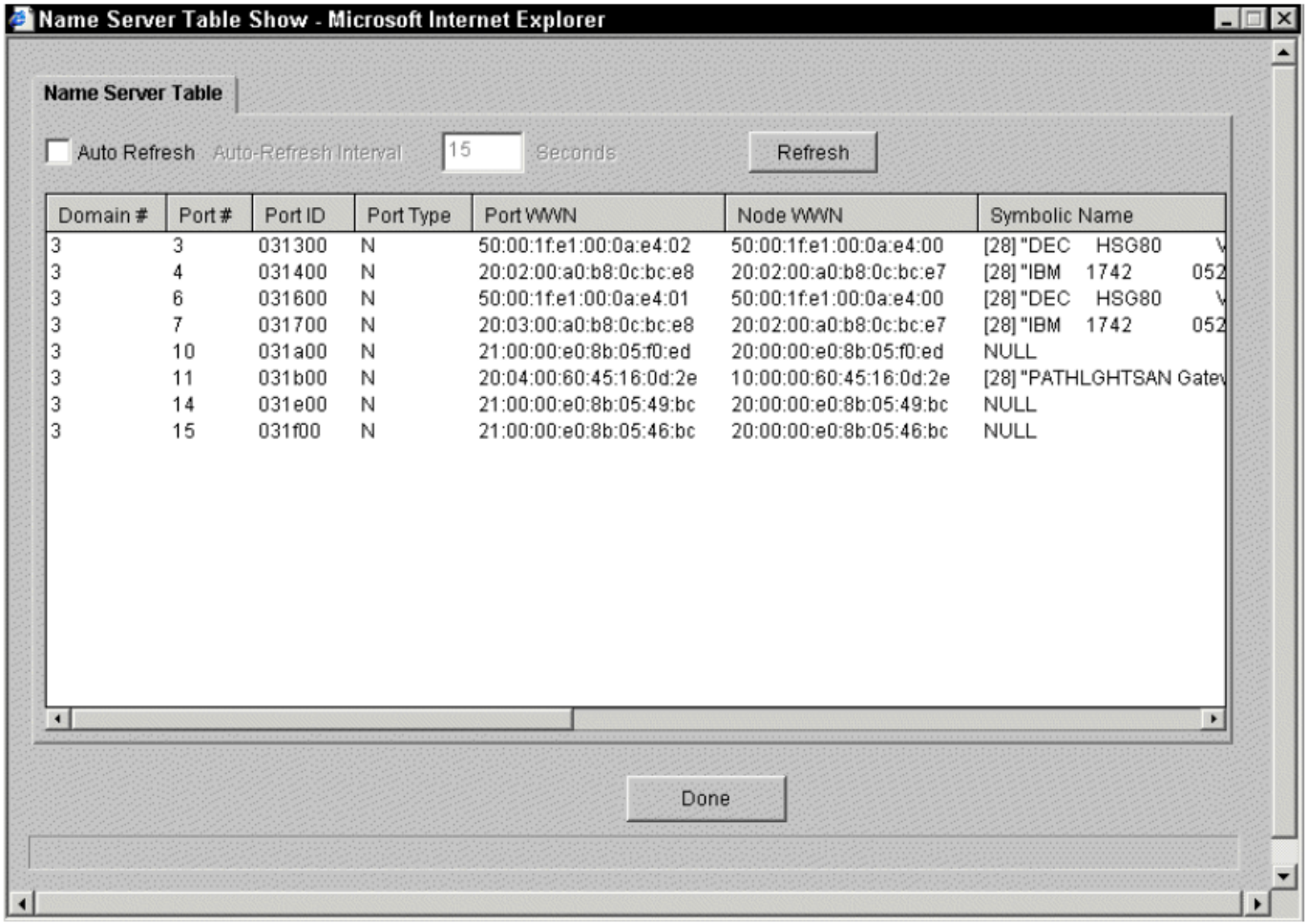
- Each FC device has a world-wide unique identifier, called World Wide Name (WWN)
  - Identifies maker and model, along with individual device
- Device can have also a unique identifier for each port (WWPN)

FC Addressing



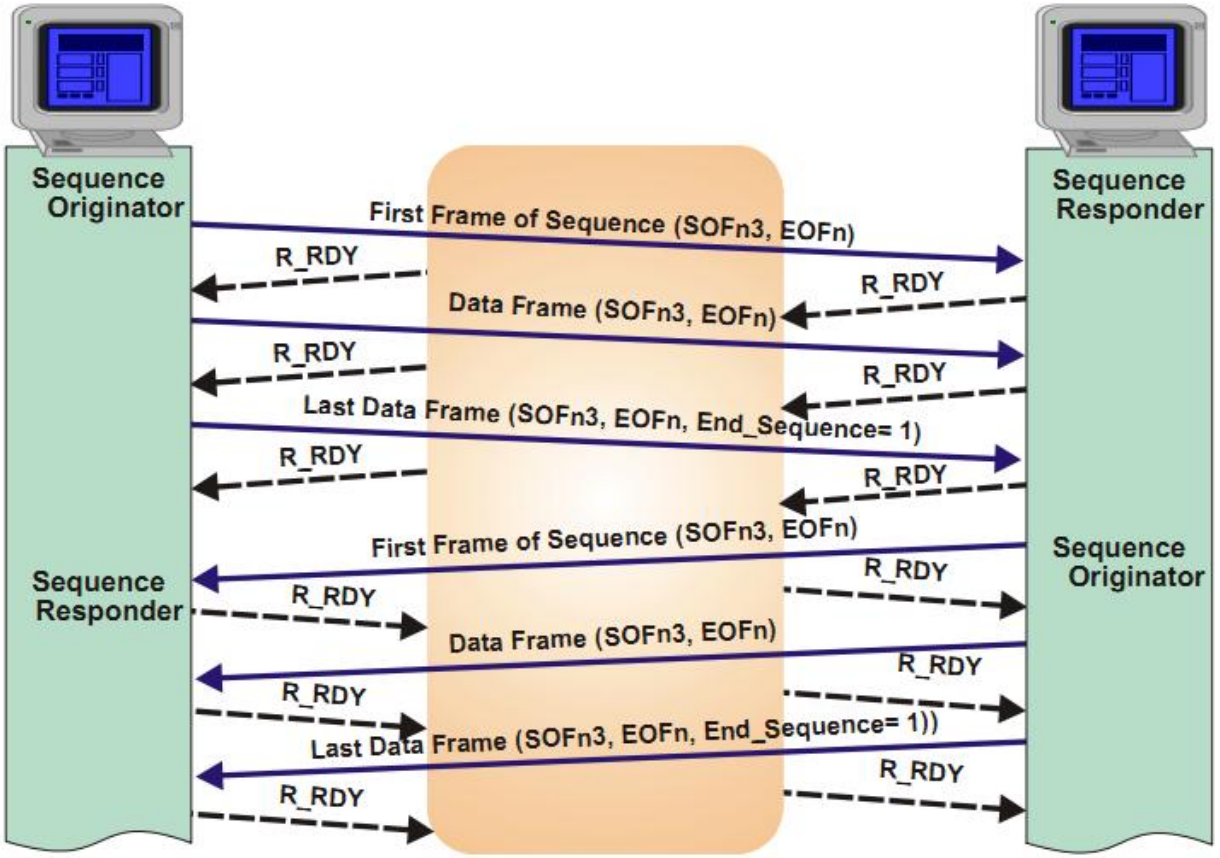
- Switch assigns an unique 24-bit address to each WWN or WWPN that joins the fabric (FLOGI)

FC Addressing



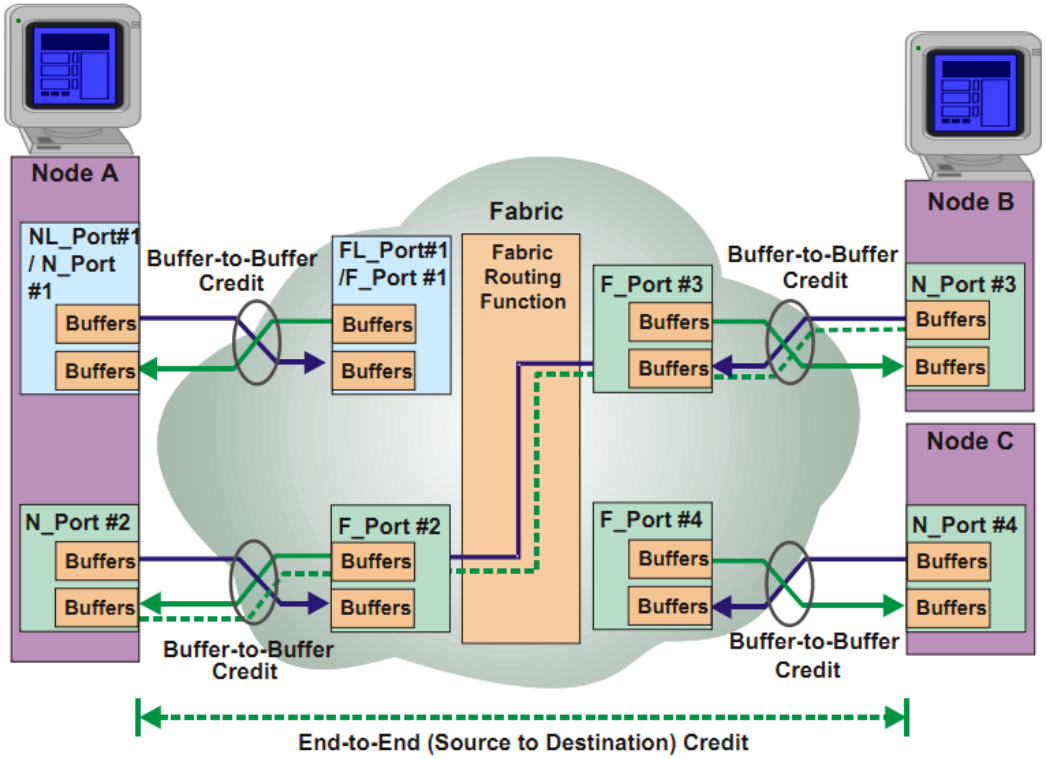
- Switch maintains in RAM translation table between WWN/WWPN and fabric address

FC flow control



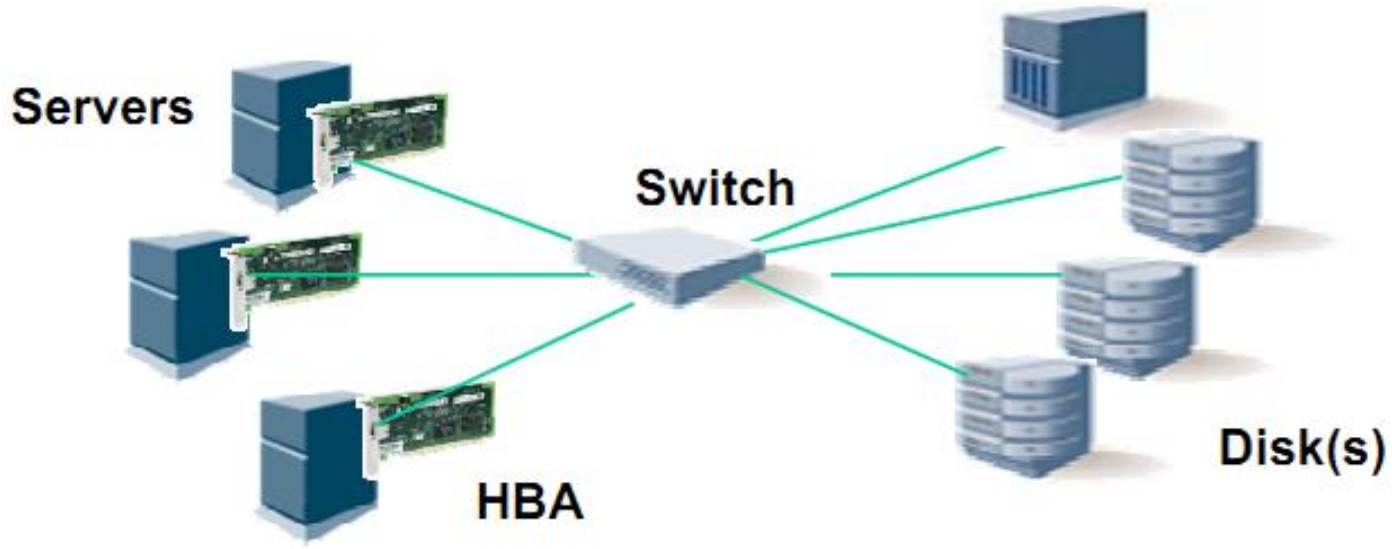
- Usually FC works in *Class3* service (connectionless communication with no end-to-end ACKs)
  - Link layer DO acknowledge every frame
  - Obviously, same kind of flow control is required

FC flow control



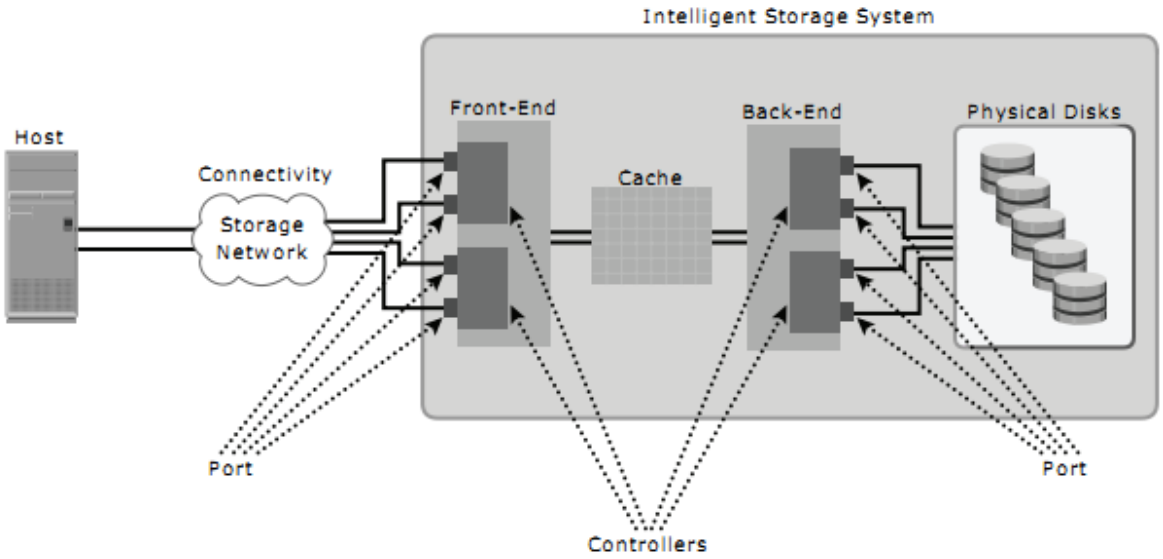
- FC uses link-level flow control, called *Buffer-to-Buffer* credits
  - Essentially, fancy name for a sliding window flow-control protocol
- FC Buffer credits ensure that fabric congestion does not result in packet loss
  - I/O timeouts due to packet loss are nasty

Storage Area Network (SAN)



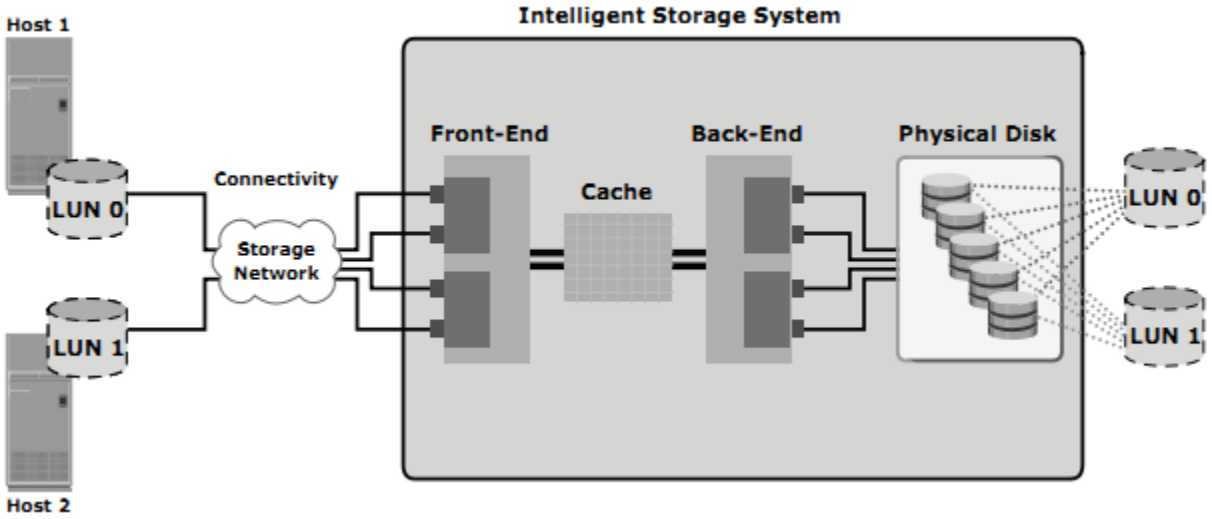
- A Storage Area Network carries data between servers and storage devices through a FC fabric, implemented by one or more FC switches

Storage Area Network (SAN)



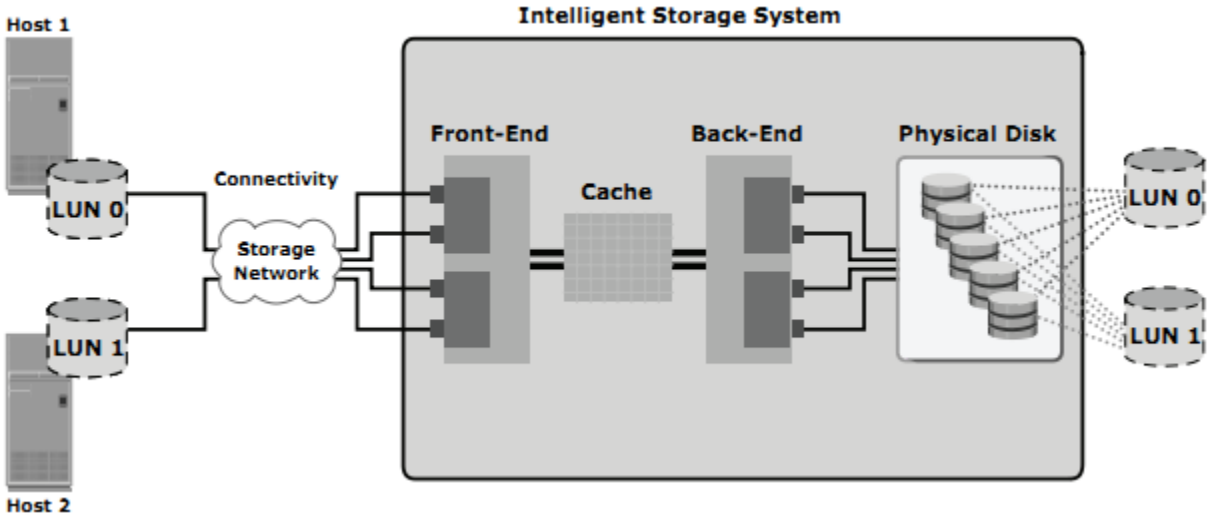
- Front-end controller of smart storage offers FC interface
  - Encapsulates and de-encapsulates SCSI payload onto/from FC frames
  - Allows flexibility for choosing performance/cost of physical storage
  - FC-to-FC: FC front-end, FC backend to physical disks
    - Maximum performance, maximum cost
  - FC-to-SAS: FC front-end, SAS backend: medium performance and cost
  - FC-to-SATA: SATA backend, lower performance and reliability, low cost

Shared physical storage with SANs



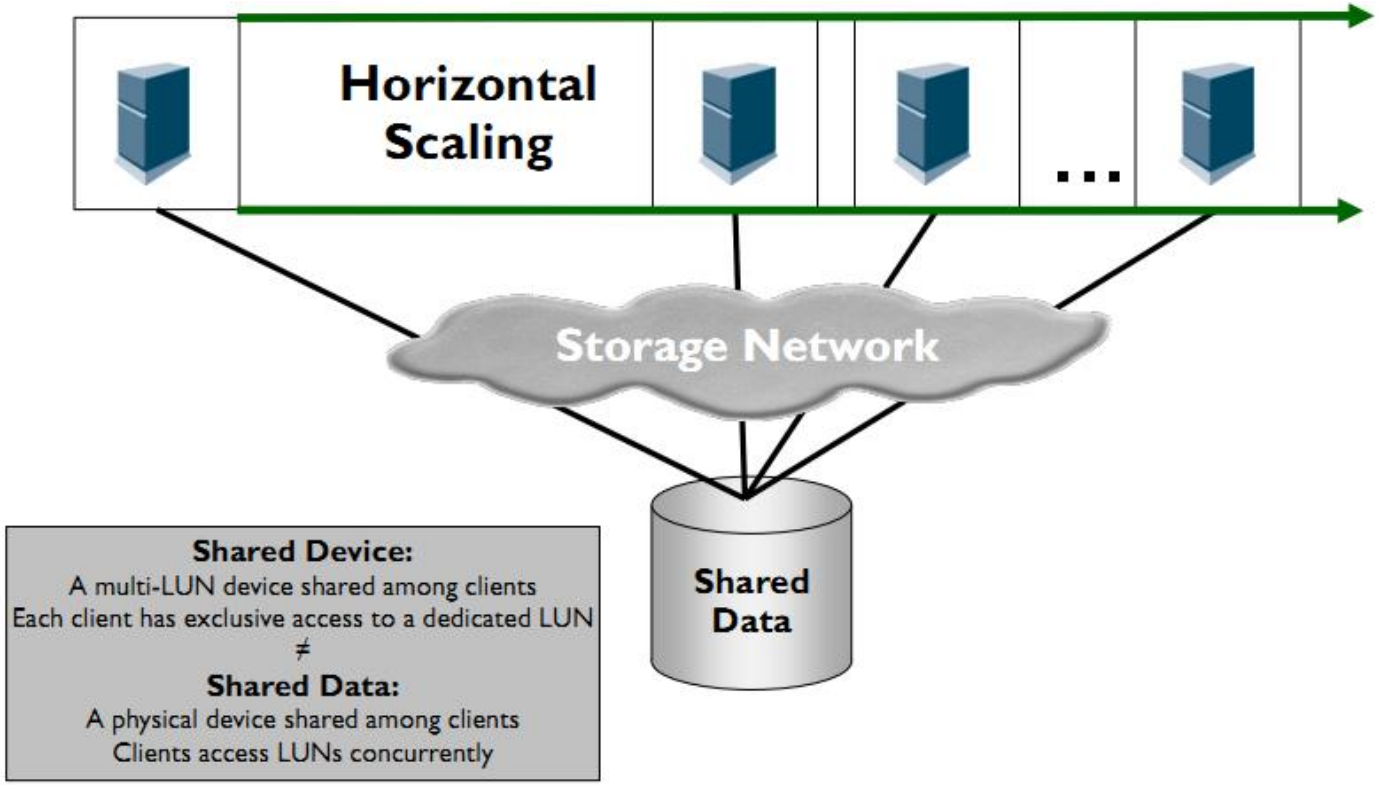
- Front-end controller hides physical disks to servers
  - Storage organized as separate LUNs
- Controller uses unique WWN identification to perform LUN-Mapping
  - Configuration of which server can see which LUN

Shared physical storage with SANs



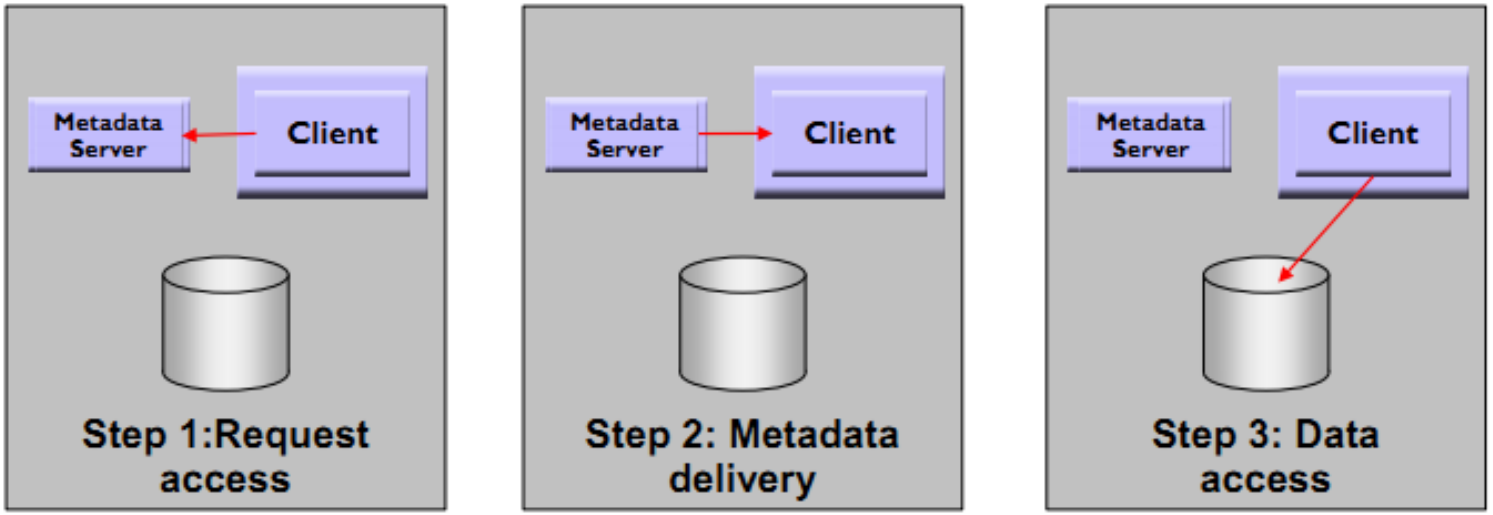
- In simplest configuration, each server will be shown a single LUN
  - No two servers share same LUN
    - Avoids coherency problems in filesystems
  - Reduces costs by centralizing physical storage array

Sharing data with SANs



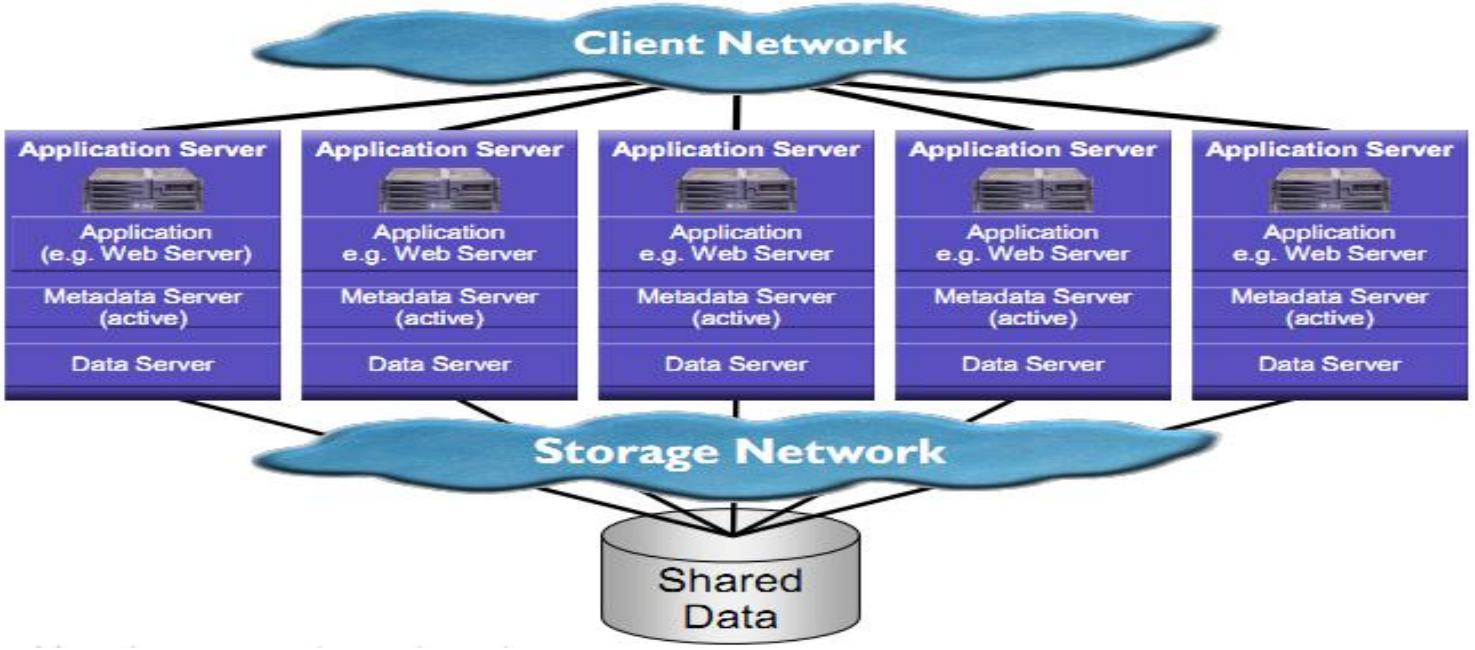
- SANs allow also sharing data between servers
- Front-end controller allows concurrent access to same SCSI LUN from several servers
- Problem: access MUST be coordinated at the filesystem level, or coherency problems will arise soon

Sharing data with SANs



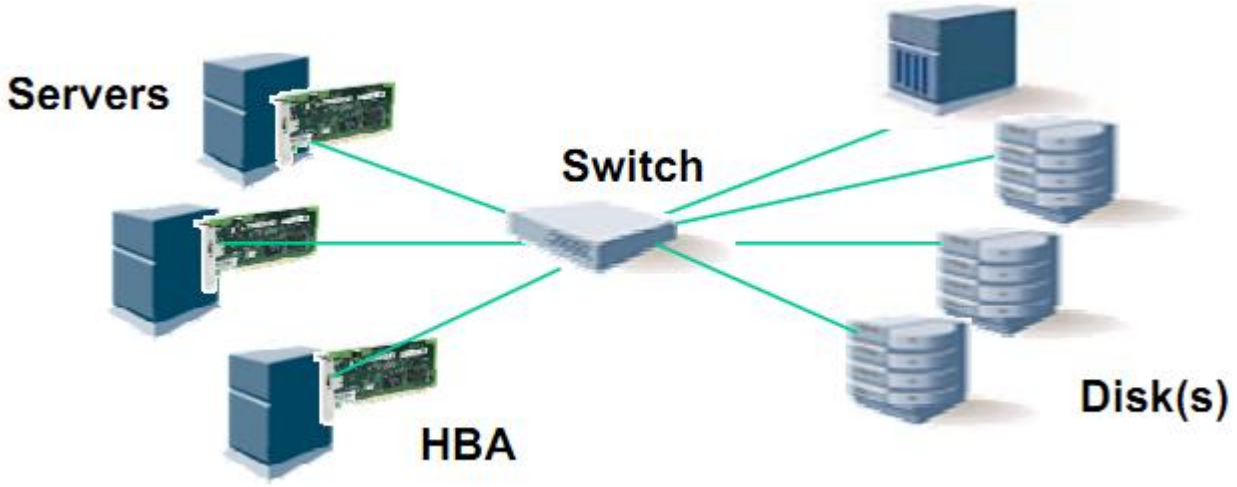
- Solution: clustered file-system
  - Metadata server keeps database with mapping of filesystem contents to SCSI blocks (LBAs)
  - Servers (clients) must request ticket to metadata server before access to data (read or write)
  - Metadata server returns permission and block metadata while maintaining system-wide coherency
  - Client can then do physical access to block device

Sharing data with SANs



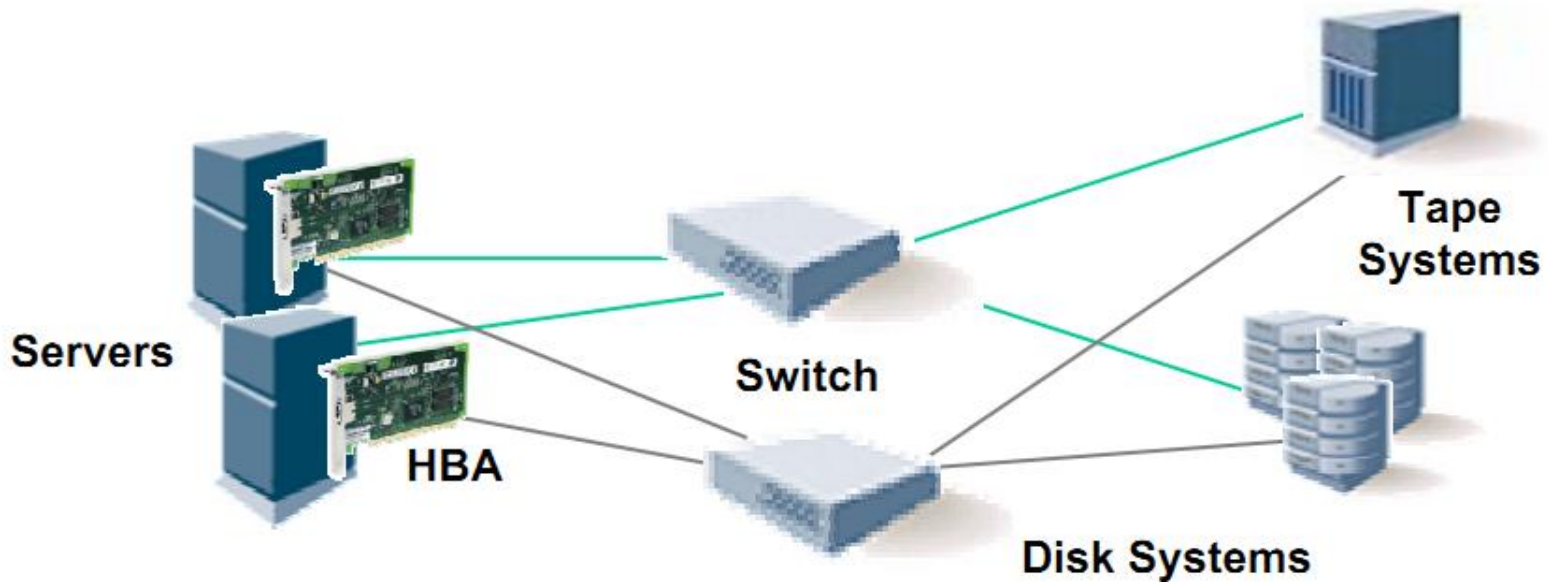
- Metadata server and clients all must have physical, SCSI-level access to shared storage
  - SAN connectivity makes this possible
- In production environment, two metadata servers
  - Active-passive: simpler, no High Availability (downtime to switch server)
  - Active-active: more complex, allows High Availability and load balancing

High Availability and Multipathing



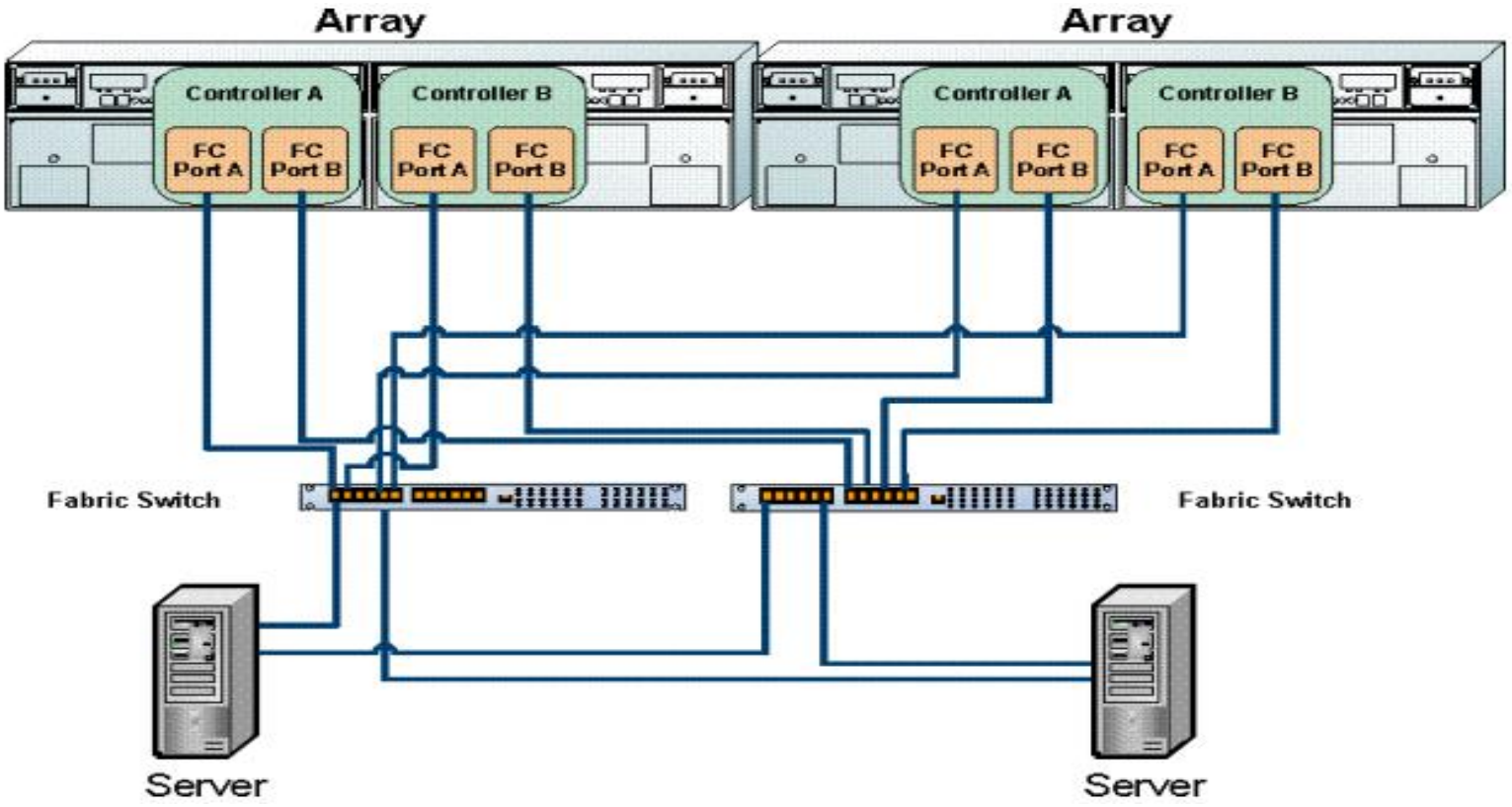
- Problem with SAN of figure: lots of single points of failure
  - Only one FC switch
  - Only one FC link between devices and switch
- Additional problem: congestion if all I/O transactions target the same LUN (link)

High Availability and Multipathing



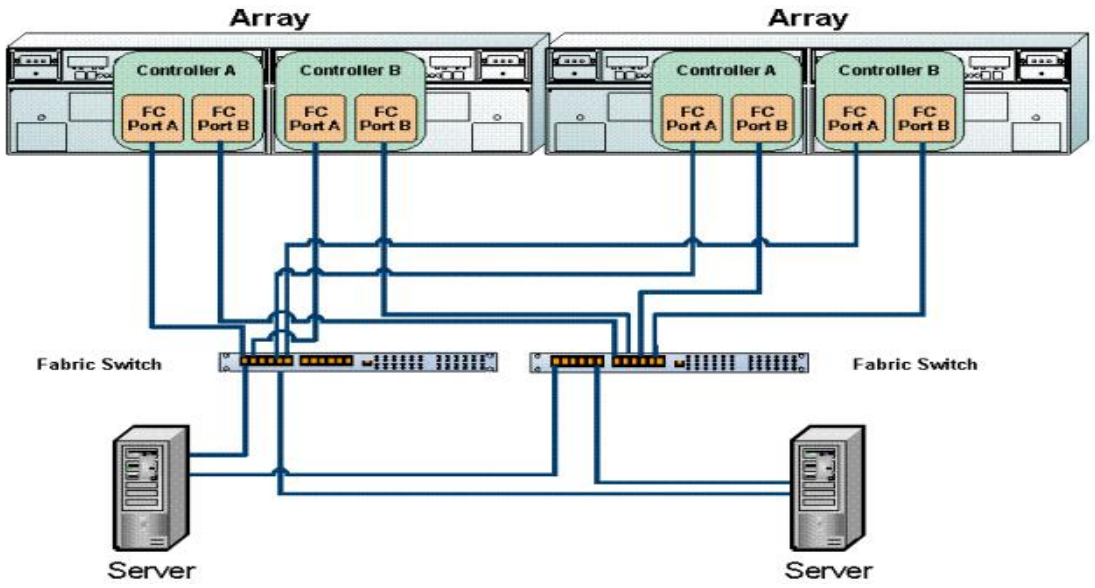
- Reliability of system improves introducing redundancy for all FC components
- System provides failover capability

Multipathing



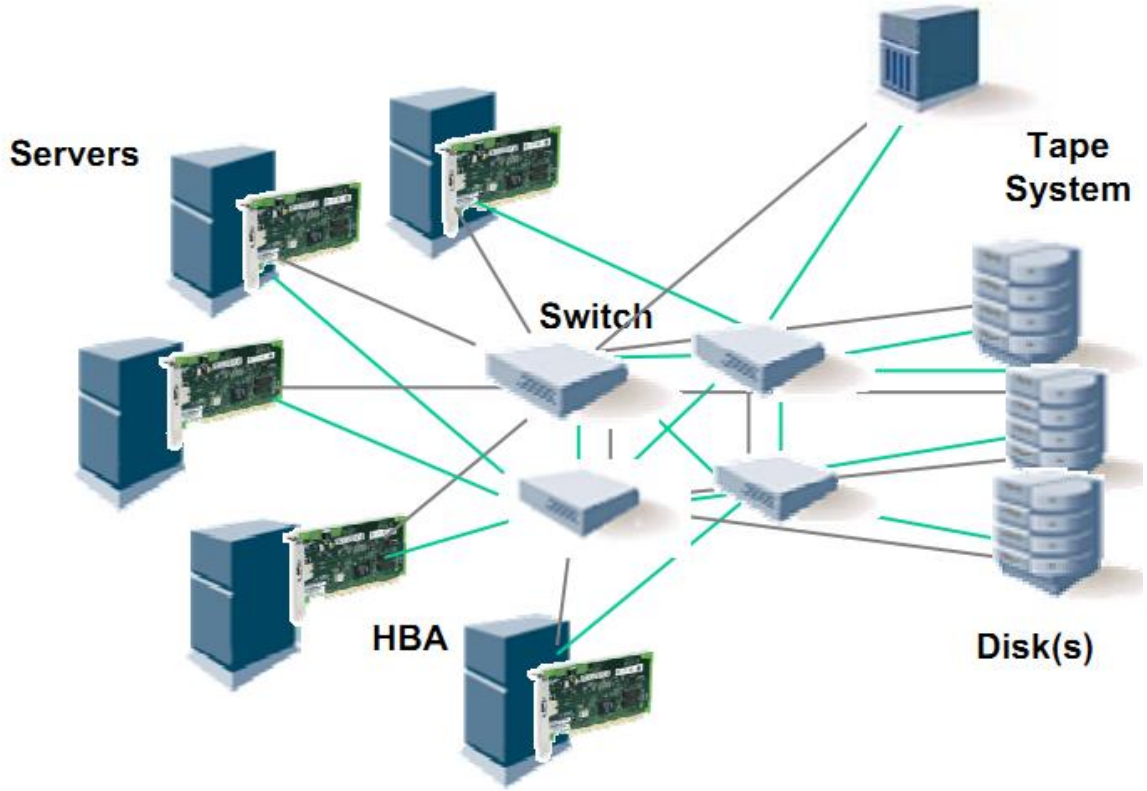
- System provides also multipathing
  - There are at least two different paths through FC fabric between any initiator and any target (LUN)

Multipathing



- Managing multipathing requires multipath-capable FC HBA device driver in servers
  - Driver uses unique SCSI-level identifiers to avoid duplicating LUNs at filesystem level.
  - Manages active-passive or active-active link status
  - If active-active, can also manage load-balancing
    - SCSI traffic split between paths to help reduce congestion

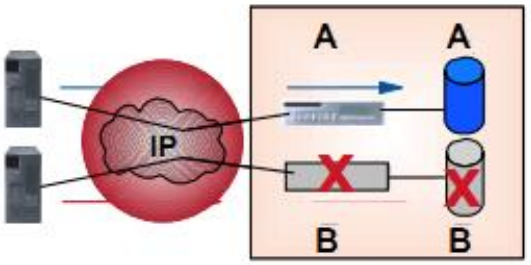
Multipathing and load-balance



- Fabric can be also scaled-out to improve redundancy and load-balancing
  - Avoids running in degraded state due to bandwidth loss if a switch fails
  - Trunking between switches provide necessary cross-over paths

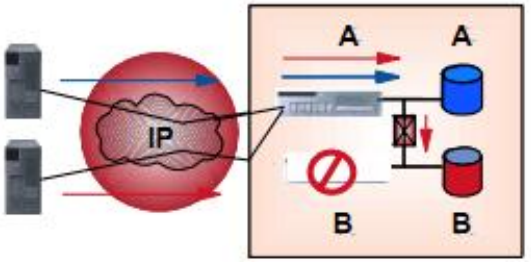
Clustering architectures

- **Shared Null**
  - No failover
  - No clustering
  - No load balancing



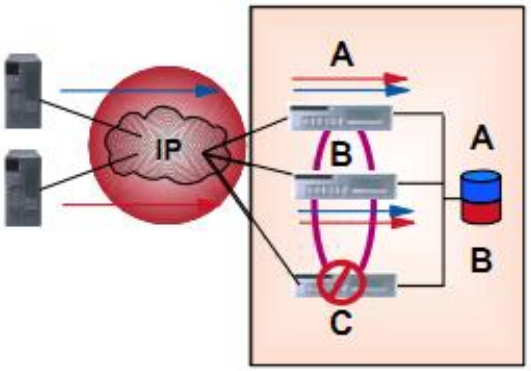
Each node accesses storage separately

- **Shared Nothing**
  - Clustered failover
  - No load balancing



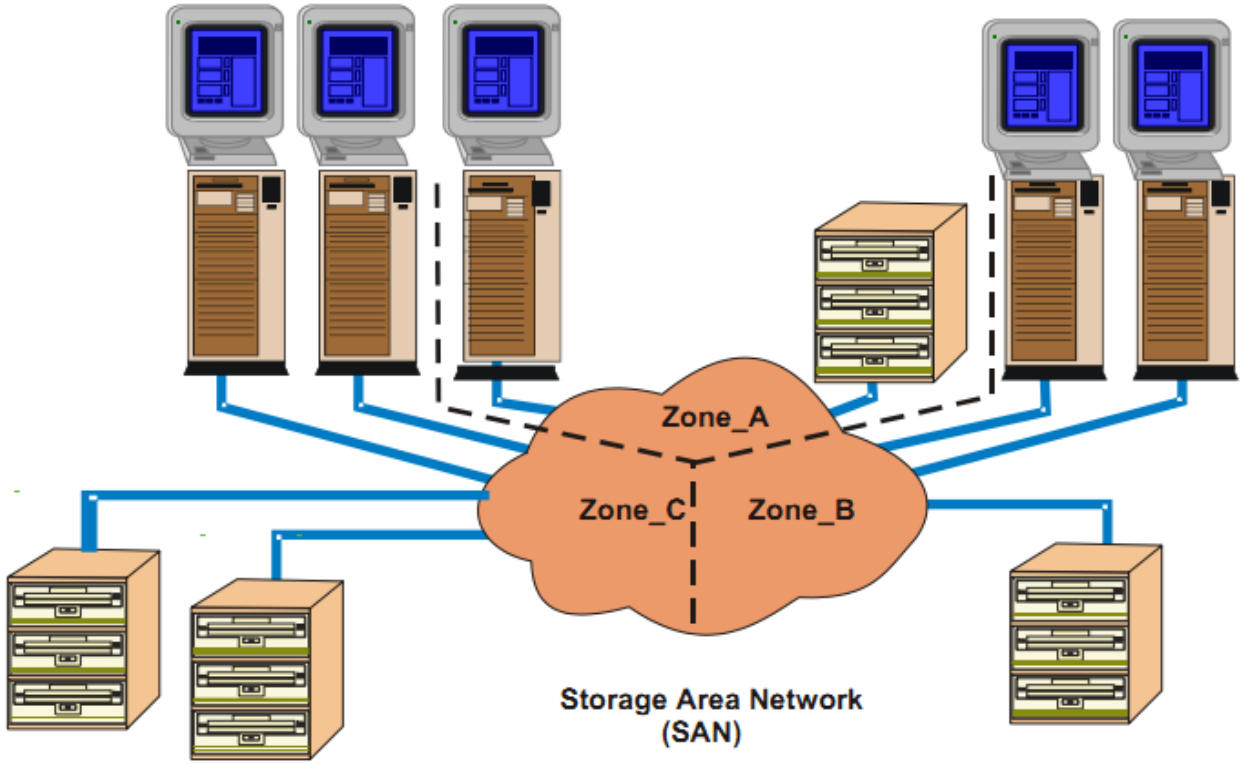
Any node can read or write to any piece of data...
NOT Concurrently

- **Shared Everything**
  - Clustering and immediate failover
  - Load balancing
  - Multinode access to large file systems



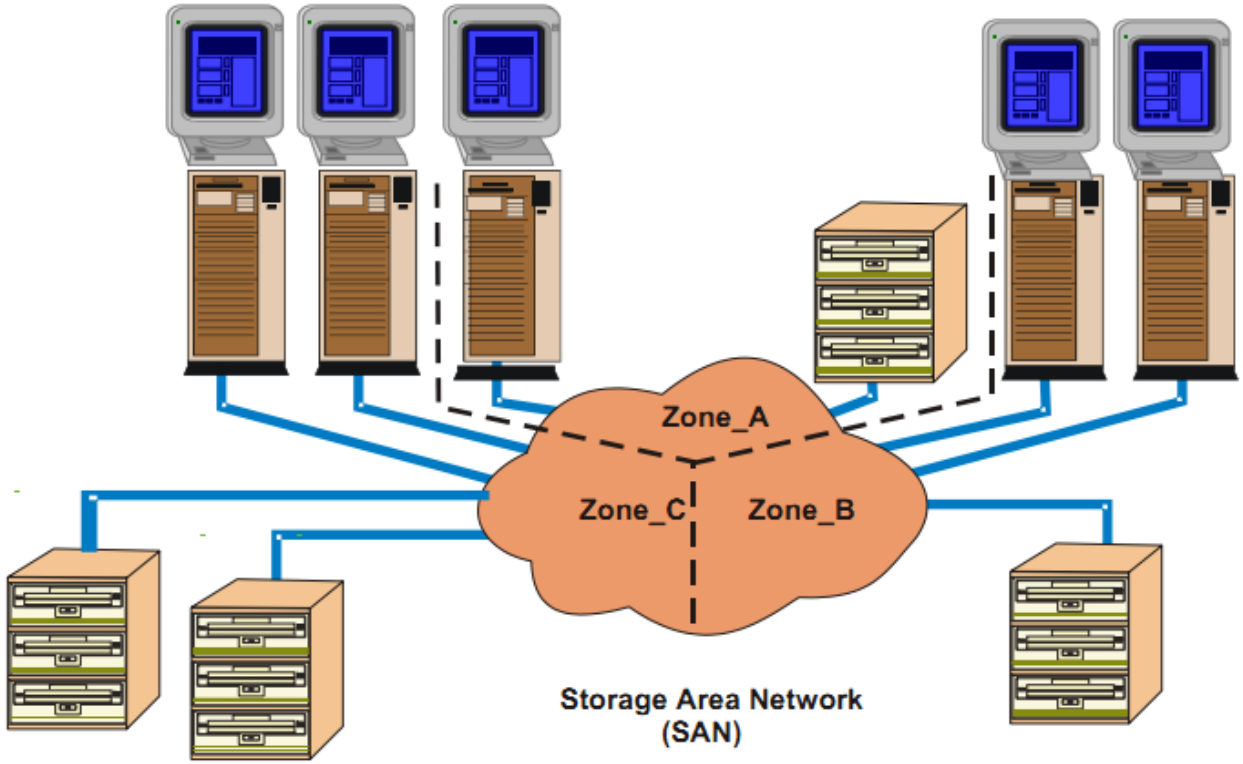
Any node can read or write to any piece of data...
CONCURRENTLY

Zoning



- SAN allows also FC access control = *zoning*
  - Done within the FC switch
  - FC traffic can be allowed or disallowed between given pairs of WWNs
  - Effective way to hide selected targets to selected initiators

Zoning



- Do not confuse with LUN-Mapping
  - Zoning: done at switch, disallows FC traffic between WWNs
  - LUN-Mapping: done at the storage controller, disallows SCSI traffic between a given LUN and a given initiator (identified by its WWN)

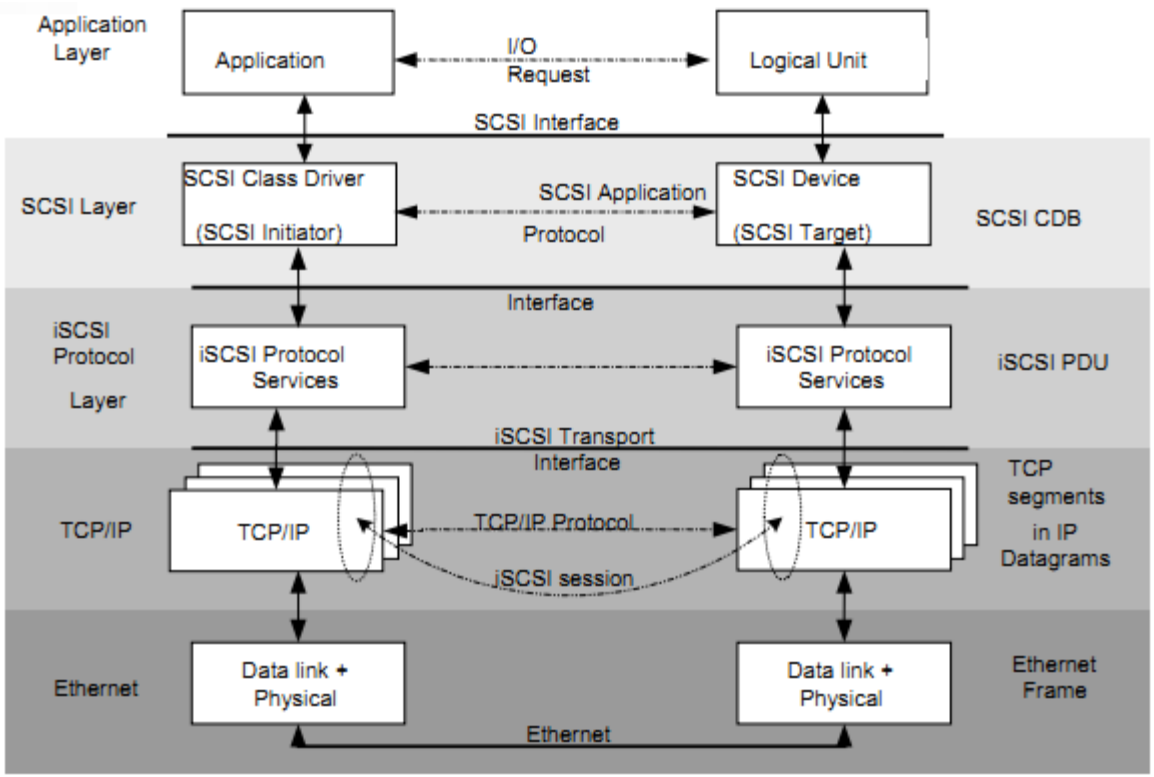
IP Storage Protocol: iSCSI

iSCSI

- iSCSI = Internet SCSI

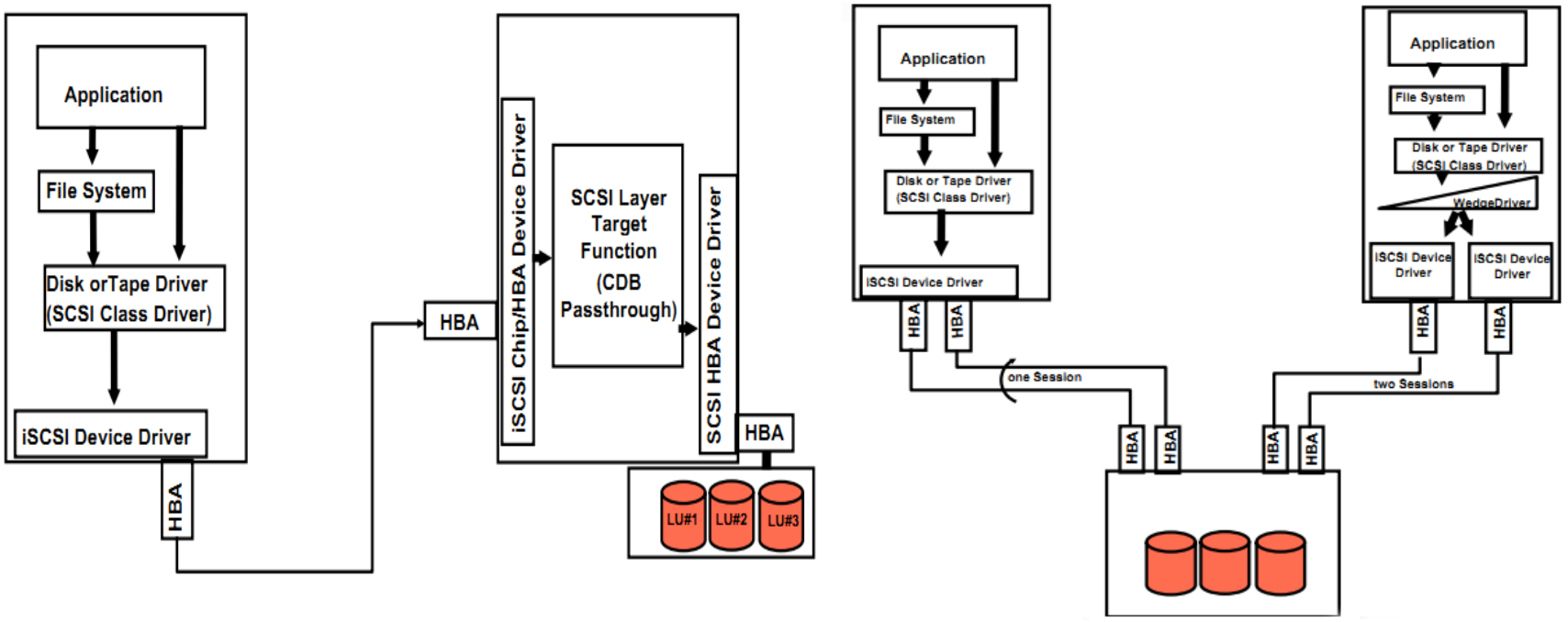
- SCSI payload is encapsulated and transported over TCP/IP network

iSCSI



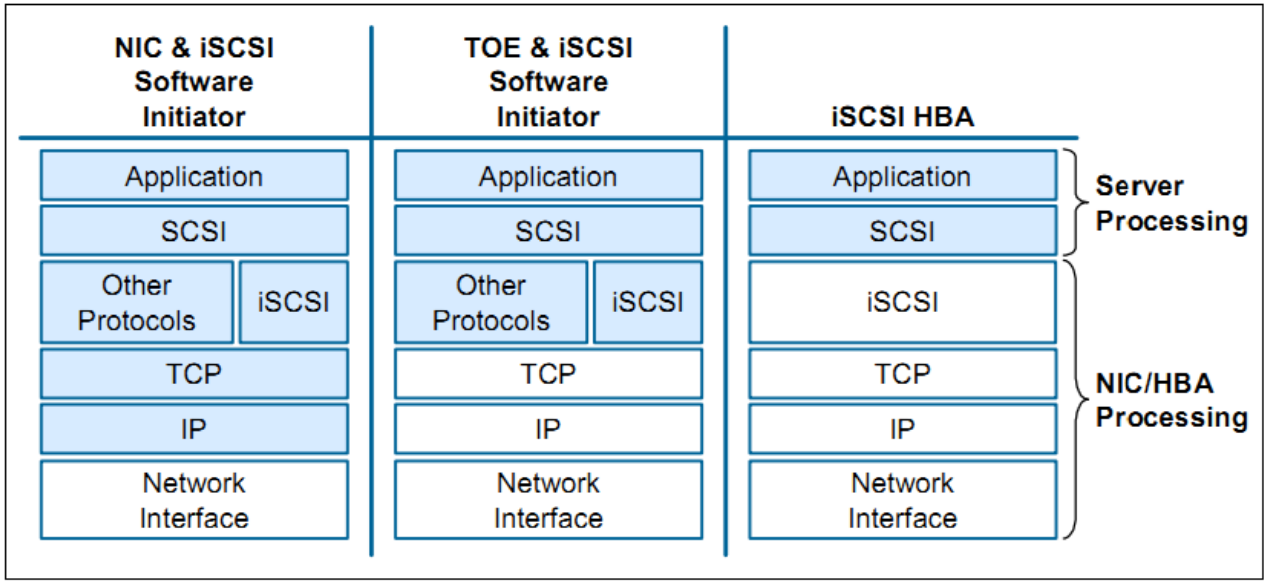
- iSCSI layers a SCSI transport layer and a SCSI protocol layer over the TCP/IP stack
  - TCP provides reliable data transport and delivery
  - IP provides routing between networks
    - Routing over FC is complex and expensive

iSCSI



- iSCSI requires use of iSCSI device driver over network HBA
  - If multiple HBAs, or multiport network HBA, iSCSI allows link aggregation for improved bandwidth or reliability
  - Multipathing also works fine with iSCSI

iSCSI performance

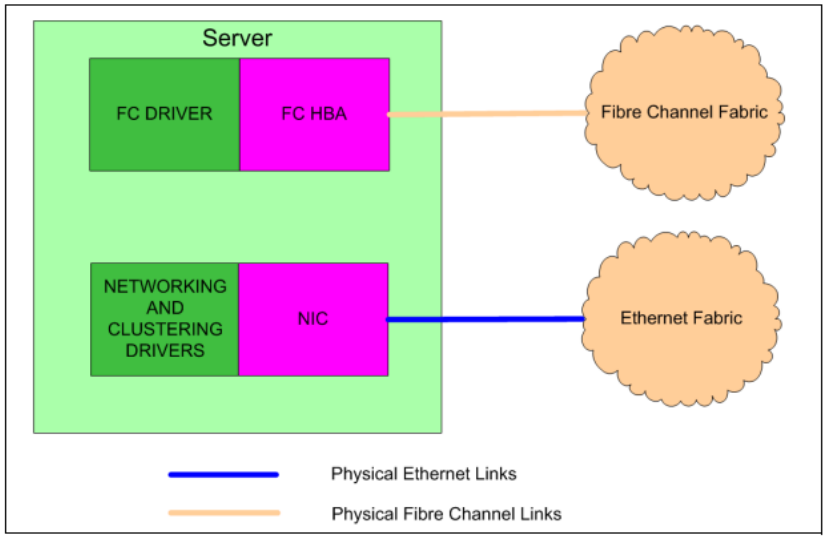


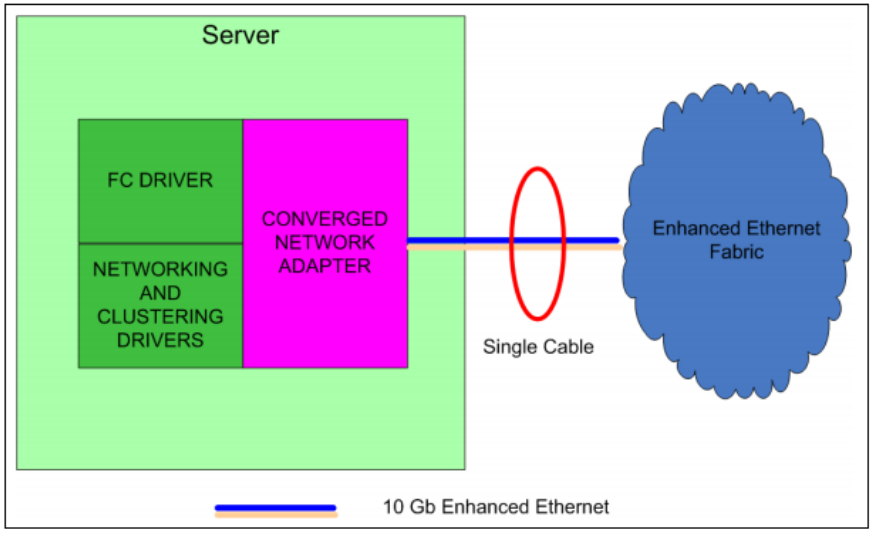
- iSCSI provides a flexible way to give SCSI access between initiators and targets
  - Quite useful for virtualization, as network traffic is much easier to virtualize than FC
- TCP processing overheads must however be carefully watched
  - TCP checksums and iSCSI protocol management can require a noticeable amount of host processing power
  - Hardware processing in iSCSI HBAs improves performance, but at a cost

Converged networking with FCoE

(Fibre Channel over Ethernet)

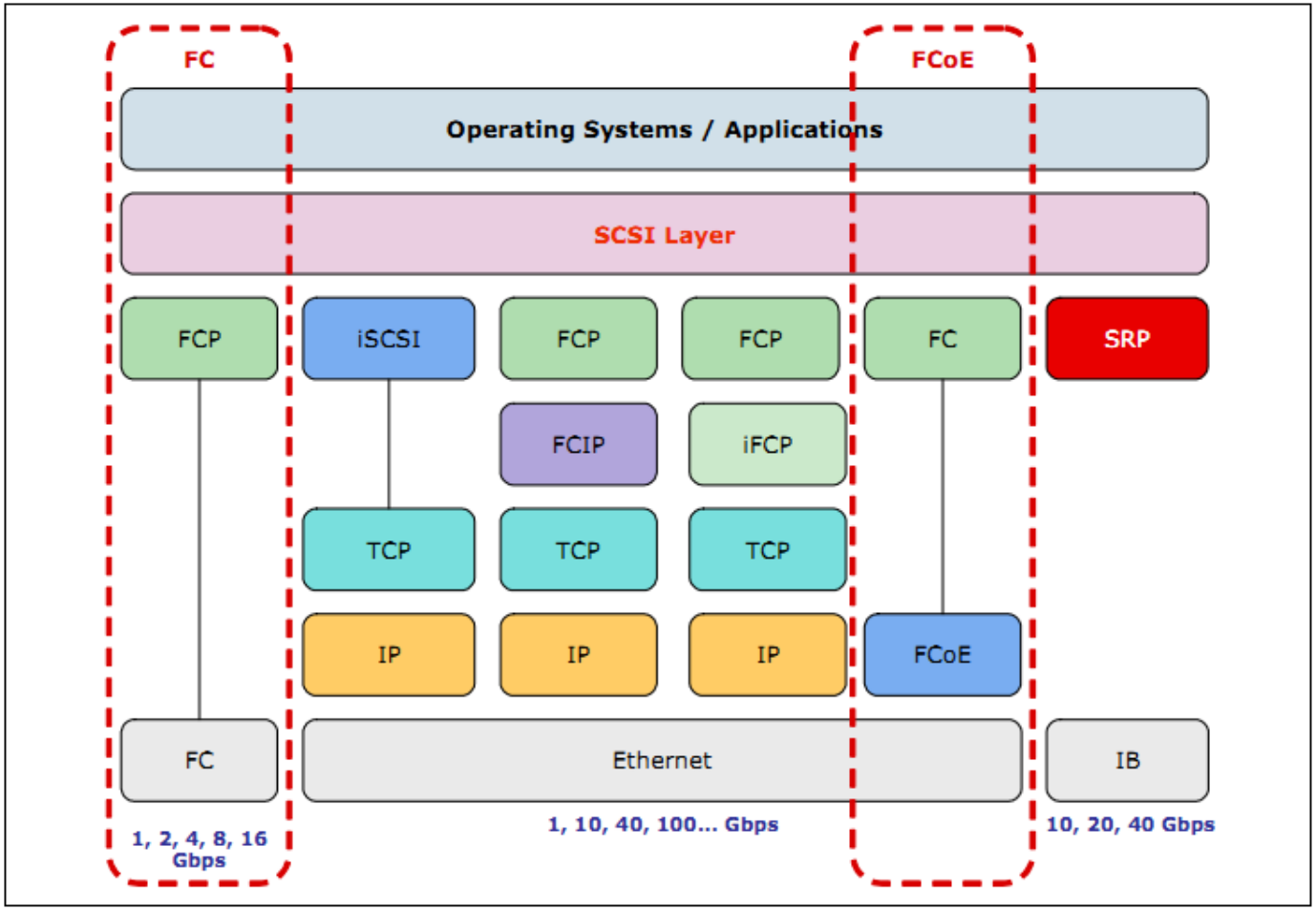
Converged networks





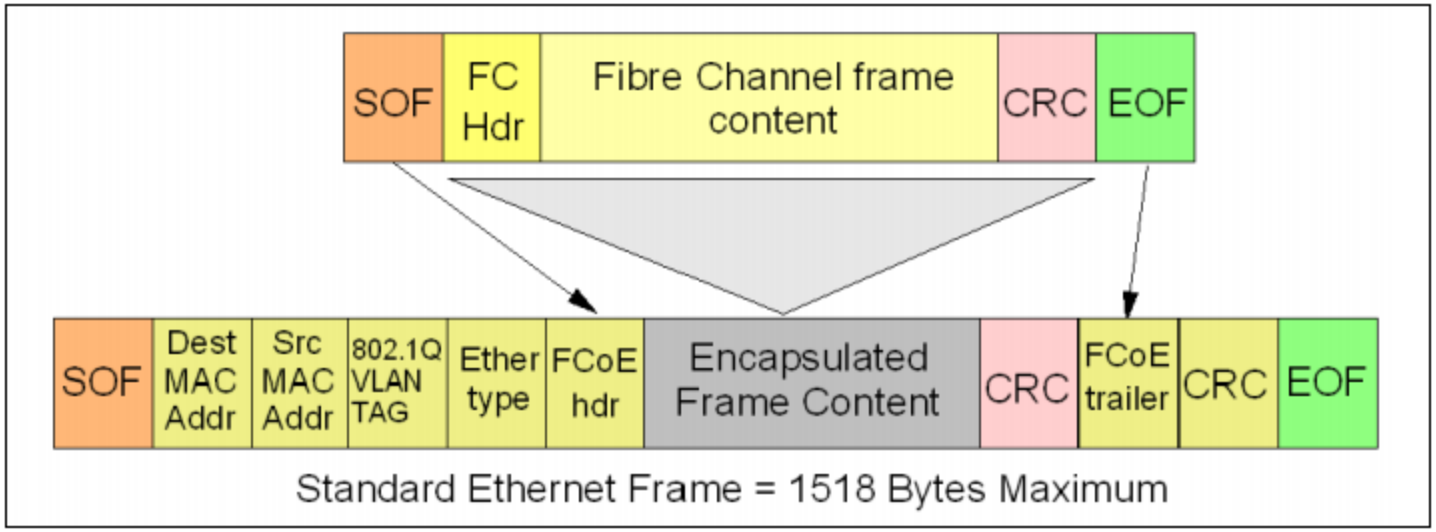
- Traditionally, datacenter required separate storage and Ethernet networks
  - Require higher hardware costs
  - Require two different sets of skills
    - Thus, two sysadms (or a very good single one), and twice personnel expenses
- With appearance of 10 Gbps Ethernet, suddenly converging both networks becomes interesting
  - Convergence requires SCSI payload being encapsulated within network protocols
  - iSCSI is, thus, one form of convergence

FCoE



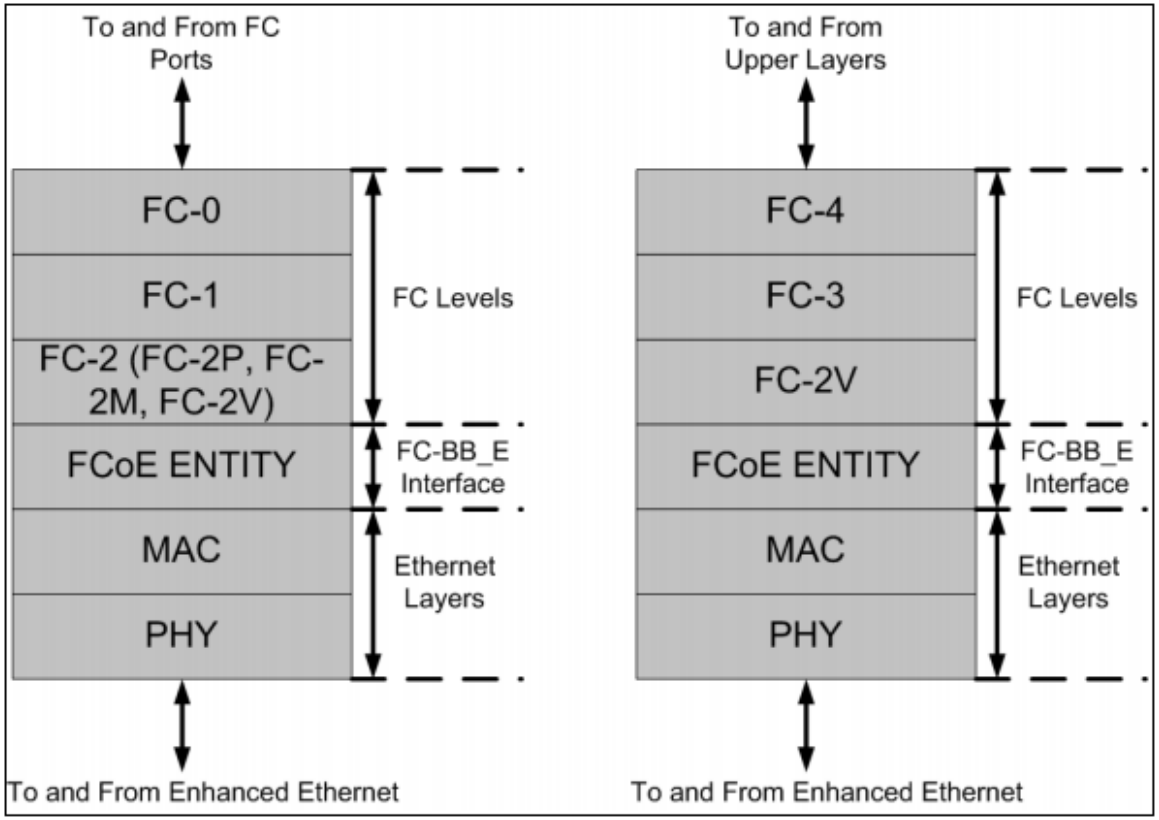
- FCoE = Fibre Channel over Ethernet
- FC with SCSI payload is directly encapsulated over Ethernet

FCoE



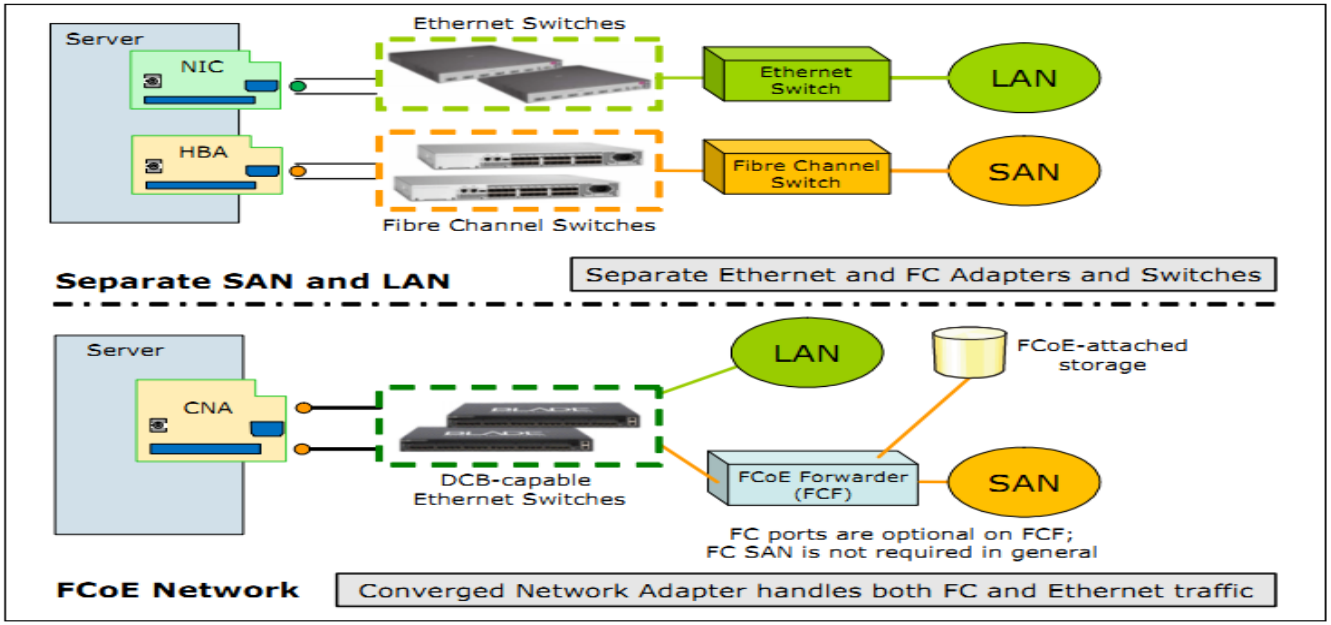
- FCoE frame is encapsulated within payload of Ethernet frame
  - Jumbo frames (payload > 1500 bytes) must be used to avoid fragmentation
  - Note that encapsulated FC frame includes also full FC headers

FCoE



- AS full FC headers are encapsulated, it is straightforward to extract FC frame and forward it to a FC SAN
  - No checksums need to be recalculated, so it is very fast

Shortcomings of FCoE



- Ethernet is connectionless, unconfirmed protocol
  - Congestion-triggered frame loss can create a real mess on FCoE performance
- FCoE assumes lossless Ethernet will be used
  - May require use of DCB (Data Center Bridging) capable switches
    - Improved link-level flow control and management controls to avoid loss of critical frames due to congestion
  - DCB does not come cheap

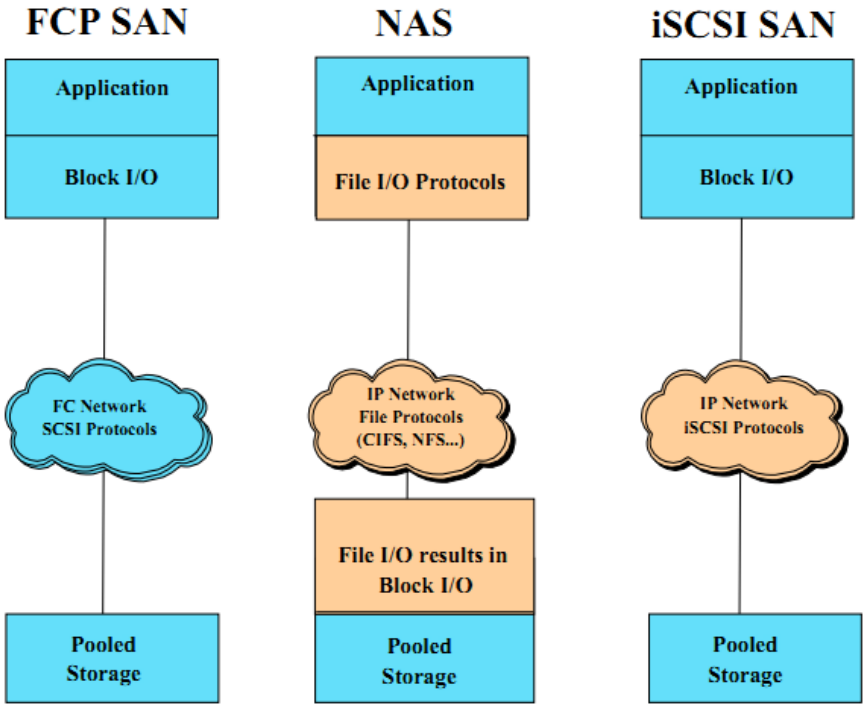
NAS: Network-Attached Storage

Network Attached Storage

- NAS is other way to access storage devices from servers using IP networks
  - NAS = high-performance storage appliance directly attached to IP networks, providing *File Serving* to clients and servers in an heterogeneous environment

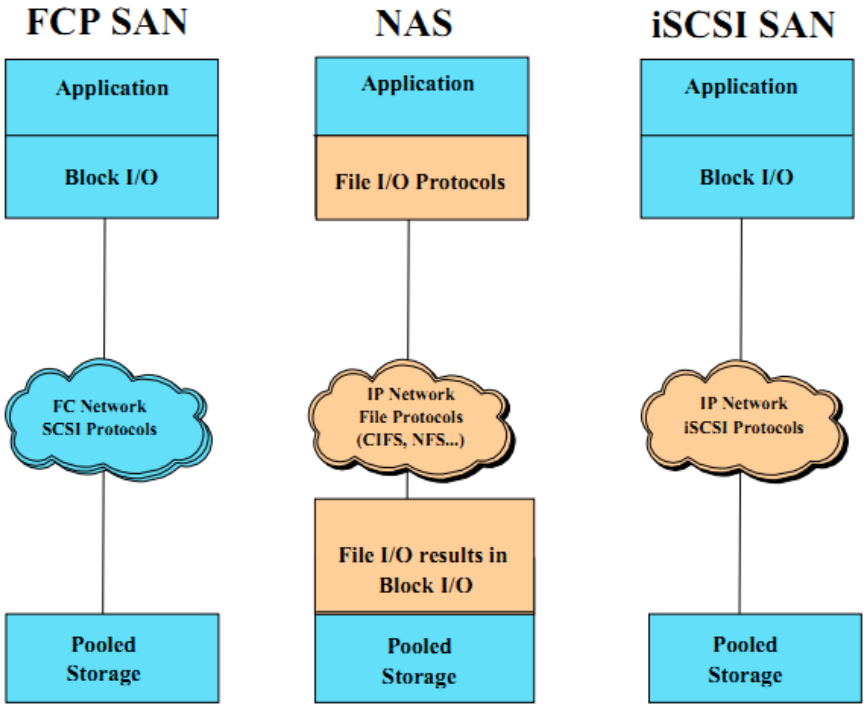
- NAS ≠ SAN
  - Same letters, totally different concepts

Block I/O vs File I/O



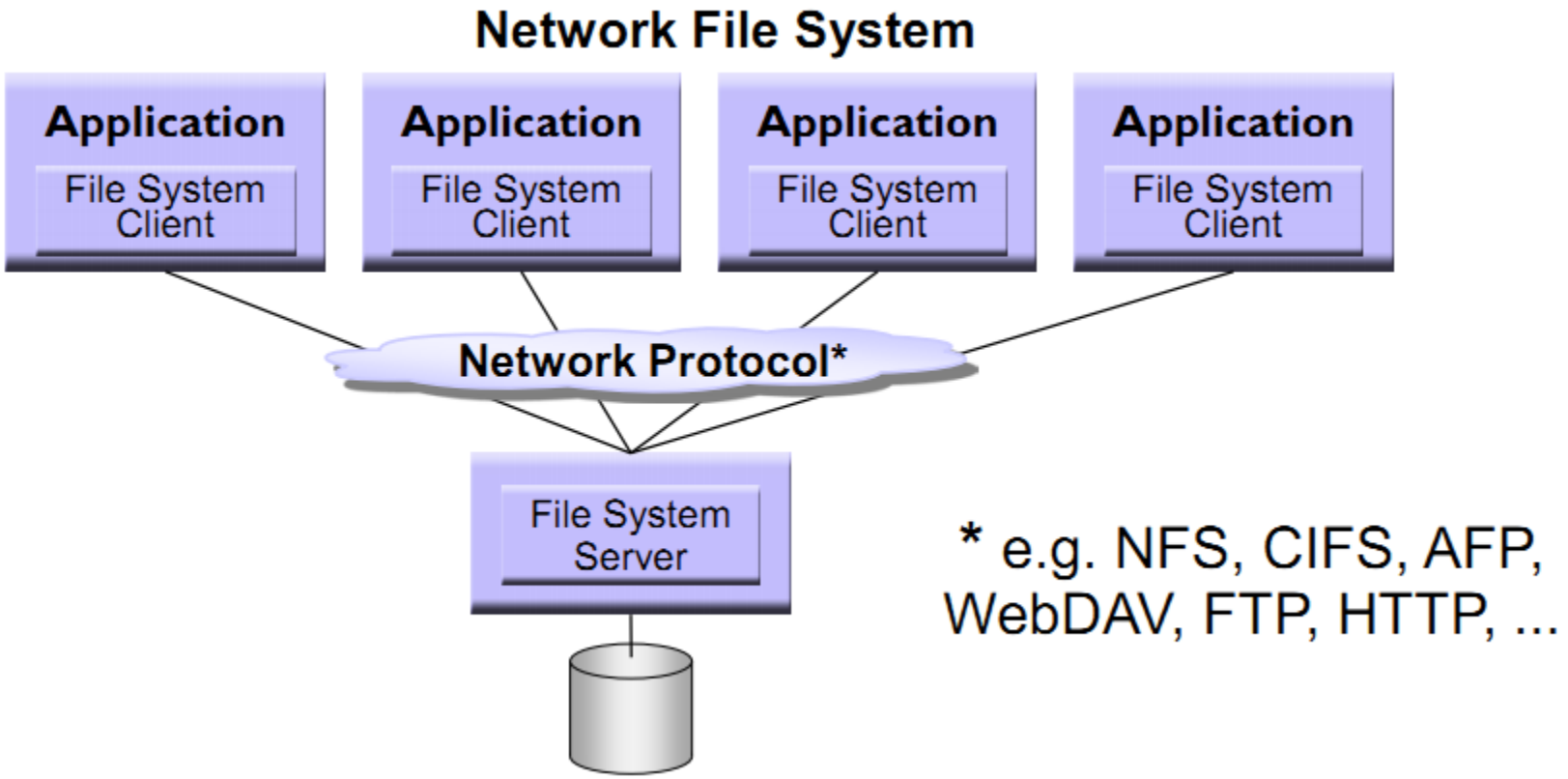
- SAN (FCP), iSCSI, FCoE are all block I/O protocols
  - SCSI payload carried over networked protocol
  - Device addressed as LBAs
    - Data directly written or read on addressed storage block, regardless of file

Block I/O vs File I/O



- NAS appliance requires use of file-sharing protocol
  - Access provided by file-sharing server, and to byte offsets (seeks) within file
  - Block storage of file is not visible to client application

Block I/O vs File I/O



- Client application accesses data through network file system
  - File system redirector transforms file access OS system calls into filesharing protocol system call
  - Redirection is transparent for client application

Block I/O vs File I/O

- NAS server (NAS appliance or NAS Gateway) transforms file I/O to block I/O
  - File sharing protocol call delivers filename and byte offset requested
  - File system in NAS server provides mapping to requested LBAs
  - Block I/O to requested LBAs is then performed on storage

Typical NAS protocols

- NFS:
  - Typical for Linux / Unix environments
  - Transported over UDP or TCP
    - UDP = potentially horrible reliability problems
    - TCP = higher overload
- CIFS (aka SMB)
  - Typical for Windows sharing
    - Can be provided by Linux server using Samba
  - Large protocol stack = large overhead
    - SMB -> NetBIOS -> TCP -> IP -> Ethernet
  - SMB 2.0 slow for bulk transfers
  - SMB 3.0 seems better (?)

NAS & iSAN

- Same gateway can support both iSCSI and NAS
  - Server manages coherent access to data from both file I/O and block I/O
- Co-existence of NAS and iSCSI very valuable for storage virtualization
  - More flexibility to choose how storage will be accessed from virtual machines

What's next?

- We have already seen how the SCSI interface hides the physical implementation of storage, showing to the server just an abstract representation of its contents

- However, the physical implementation details of a storage device are critical to understand the performance (or lack of), timing and latencies of its I/O transactions

- So, we will see now the internals of both magnetic and solid state disk drives