Cloud e Datacenter Networking

Università degli Studi di Napoli Federico II

Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione DIETI

Laurea Magistrale in Ingegneria Informatica

Prof. Roberto Canonico

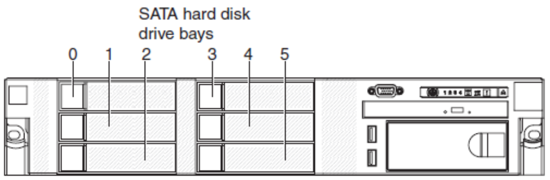
Datacenter: storage systems organization

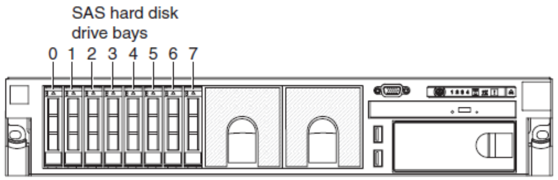
Lesson outline

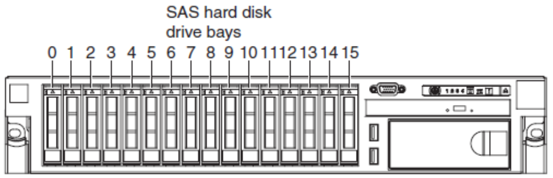
- Storage options for datacenter servers
- Shared storage infrastructures: NAS vs SAN
- Network convergence for storage infrastructures
- iSCSI and FCoE

Storage options for rack servers

- Rack servers may usually be configured with a number of options for internal storage
- Hard disks directly connected to the server's motherboard in the server chassis form the so called *Direct Attached Storage* (DAS)
- Form factors include both 3,5" and 2,5" disks
- Interfaces include:
  - SATA (Serial ATA)
  - SAS (Serial Attached SCSI)
- SAS requires a SCSI controller but supports disks hot-swapping
- More recently, magnetic hard disks are replaced by *Solid State Disks* (SSDs) that guarantee higher throughput and reduced access-time
- SSDs are typically connected by means of an NVMe interface
  - NVMe is an interface specification specifically designed for SSDs

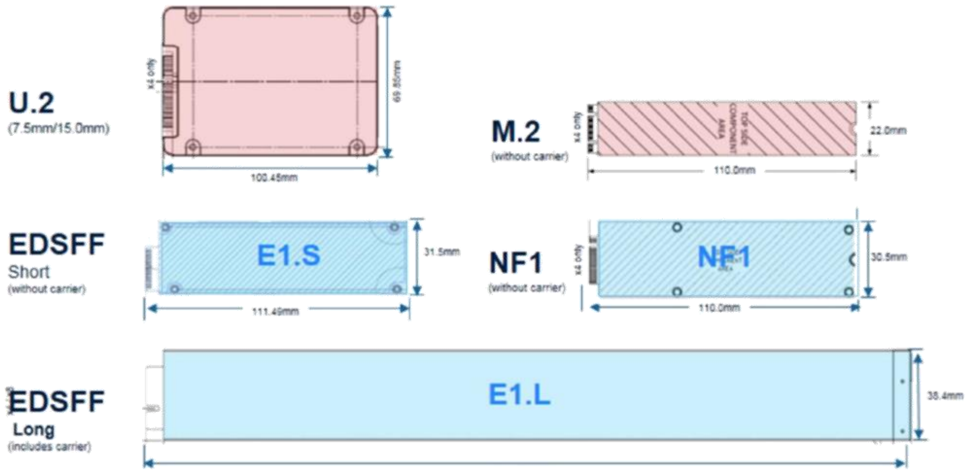






SSD disks for datacenters

- In the last few years, several SSD-based storage devices have been produced for specific use in datacenter infrastructures
- SSDs interface is typically an NVMe (*Non-Volatile Memory Express*) interface
  - a.k.a. *Non-Volatile Memory Host Controller Interface Specification* (NVMHCIS)
- These disks are produced in different form factors:
  - EDSFF (*Enterprise and Data Center SSD Form Factor*) for 1U enclosures
    - E1.L (Long)
    - E1.S (Short)
  - M.2 – a.k.a. *Next Generation Form Factor (NGFF)*
    - M.2 supports PCIe, SATA and USB
  - U.2 – a.k.a. SFF-8639 (2.5-inch)
  - Add In Cards
    - PCIe card form factor



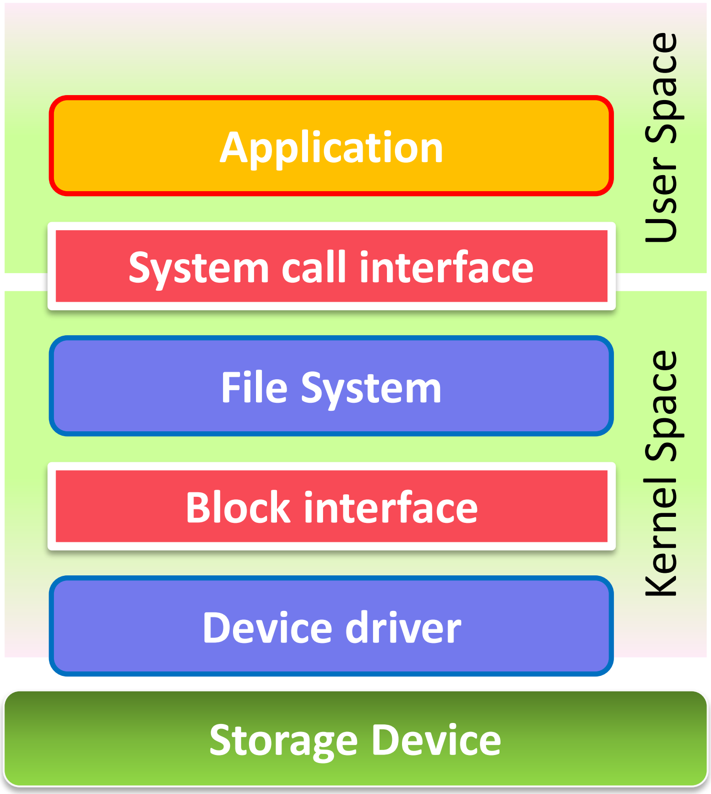
EDSFF Solid State Disks

- E1.L (*Long*)
  - SNIA specification SFF-TA-1007
  - up to 32 SSDs in a single 1U enclosure
  - up to 1PB in a single 1U enclosure
    using 32TB SSDs (e.g. Intel-based DC P4500)

- E1.S (*Short*)
  - SNIA specification SFF-TA-1006
  - up to 32 SSDs in a single 1U enclosure

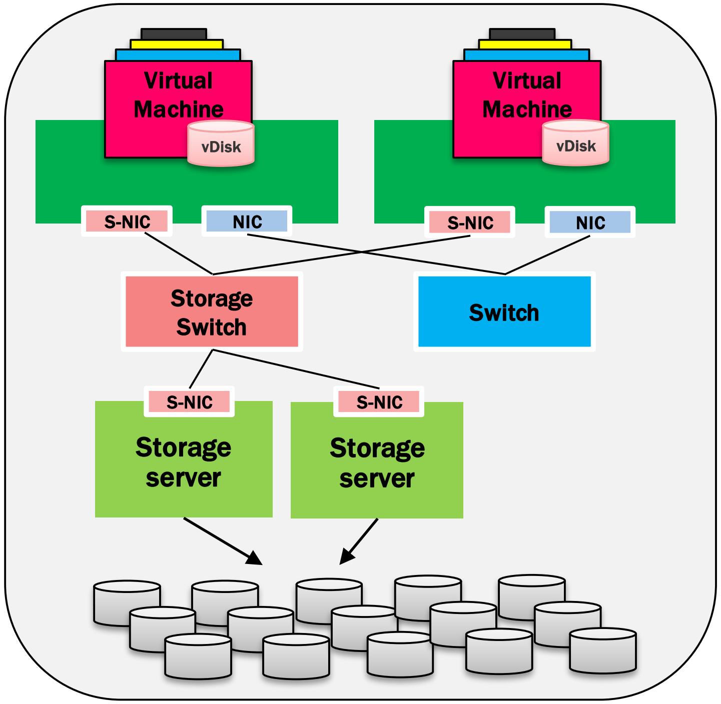
Storage abstractions

- Storage system provide persistent (i.e. non volatile) data storage
- Operating systems provide two kind of storage abstractions
  - **File system**
    - A system call interface to user space applications
  - **Block device**
    - A block device interface to file systems
    - Through interfaces such as ATA, SATA, SCSI, SAS, FC, etc.



Storage systems in a datacenter

- To make more efficient use of storage resources, storage in a datacenter is provided by shared devices connected to servers through a *network*
- Storage is virtualized and resources are shared
- To connect shared storage devices to servers two approaches can be pursued:
  - General purpose (Ethernet)
  - Dedicated technologies (Fibre Channel)
- Typical approach: separate networks for VM-to-VM traffic and VM-to-Storage traffic

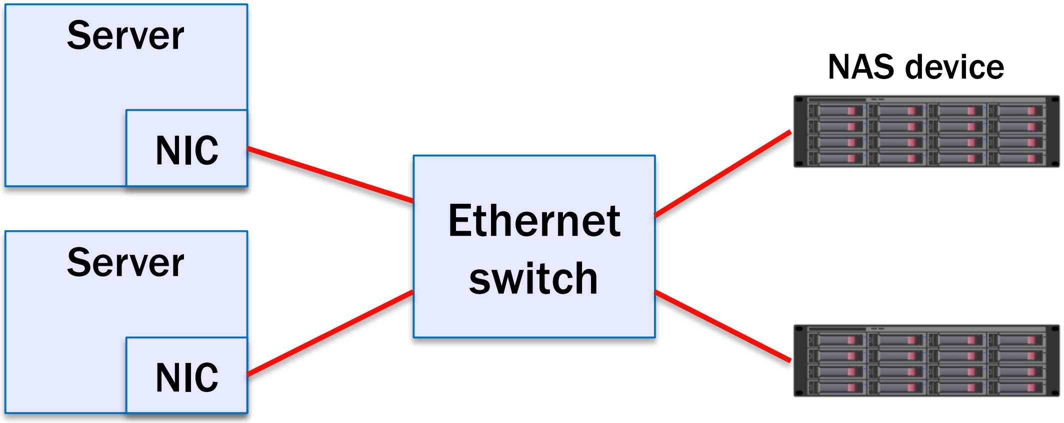


Network Storage types

- File (NAS)
  - Examples: SMB2 (CIFS) (Windows), NFS
  - Typical operations: open, close, read, write, rewind
- Block (SAN)
  - Examples: SCSI over FC/FCoE/iSCSI/SAS/SATA
  - Typical operations: read/write extent of blocks from/to LUN
- Object
  - Examples: T10 OSD, OpenStack, Amazon S3, SNIA CDMI
  - Typical operations: put, get
- Big Data
  - Examples: HDFS
  - Operations: analysis with Map-Reduce

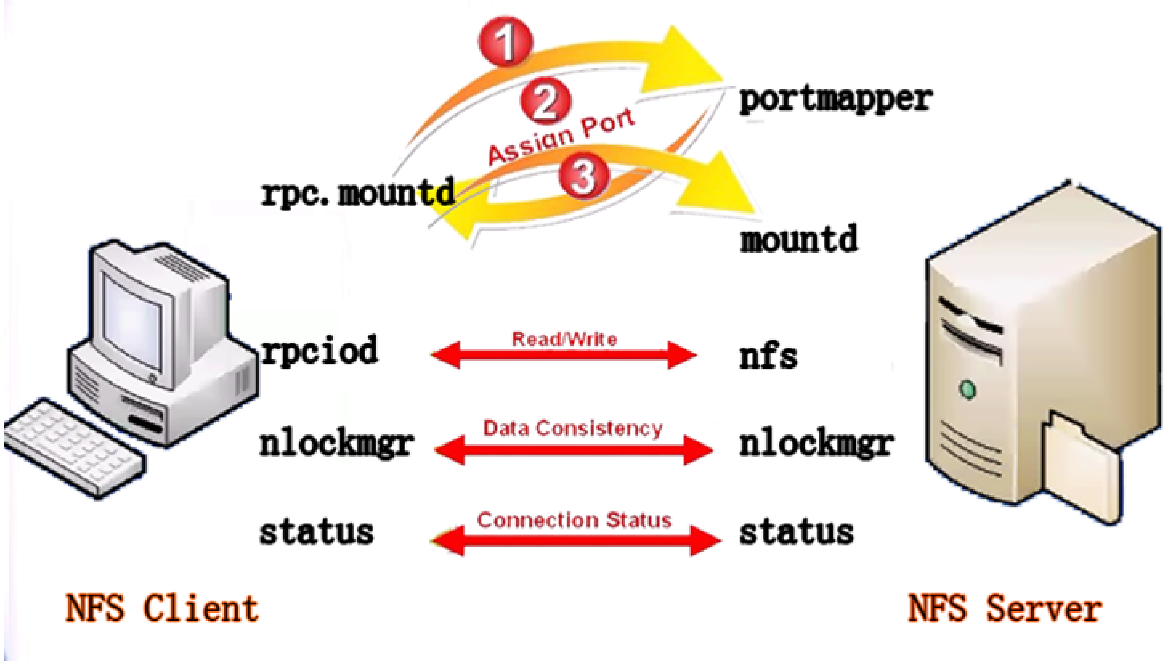
Network Attached Storage: NAS

- A **Network Attached Storage (NAS)** is a storage device that is able to “export” its own filesystem to remote servers through a network file system protocol
- Example of network file system protocols:
  - NFS
  - Server Message Block (SMB or Samba)
- Remote servers access the NAS resources through the fileystem abstraction
- Remote directories need to be “mounted” on the servers’ filesystem
- NAS devices are cheaper than SANs
- Connection between servers and NAS devices is through Ethernet



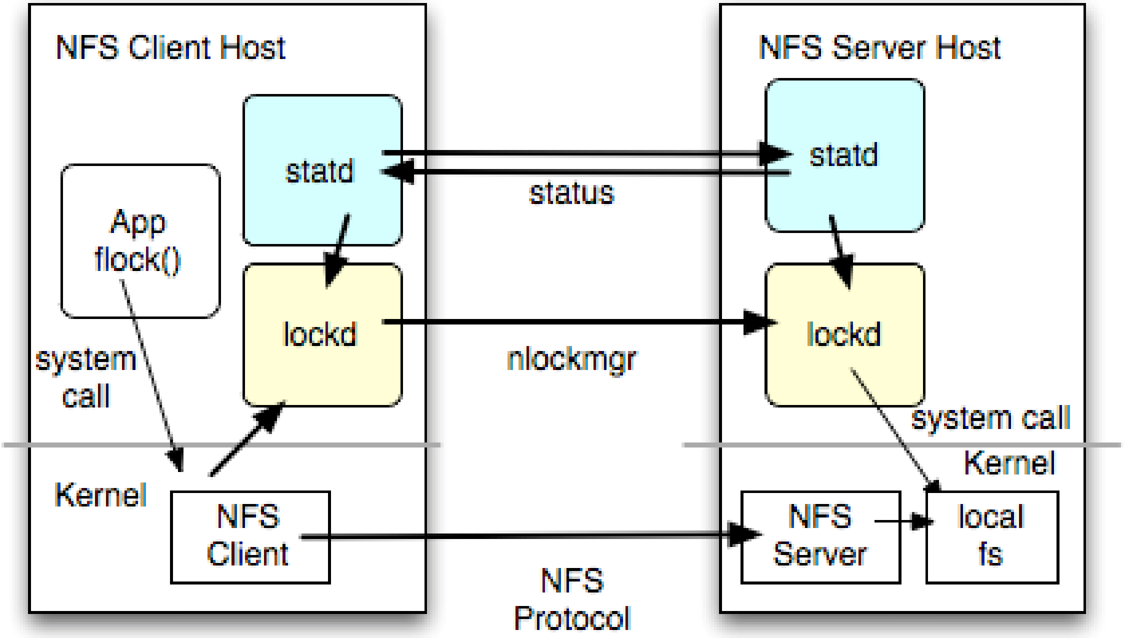
Network File System (NFS)

- NFS is a POSIX-compliant distributed file system defined as an open standard in RFCs
  - Works according to the server-client model
  - NFS builds on the *Remote Procedure Call* (RPC) system
  - In NFSv3, service listens on random TCP port
  - NFS use RPC to get the port of service
- Some features :
  - Shared POSIX file system
  - Implemented in Linux kernel



Consistency and concurrency in NFS

- Lockd offers a write lock to handle concurrent update
- Statd handles the consistency between server and clients



Storage Area Network: SAN

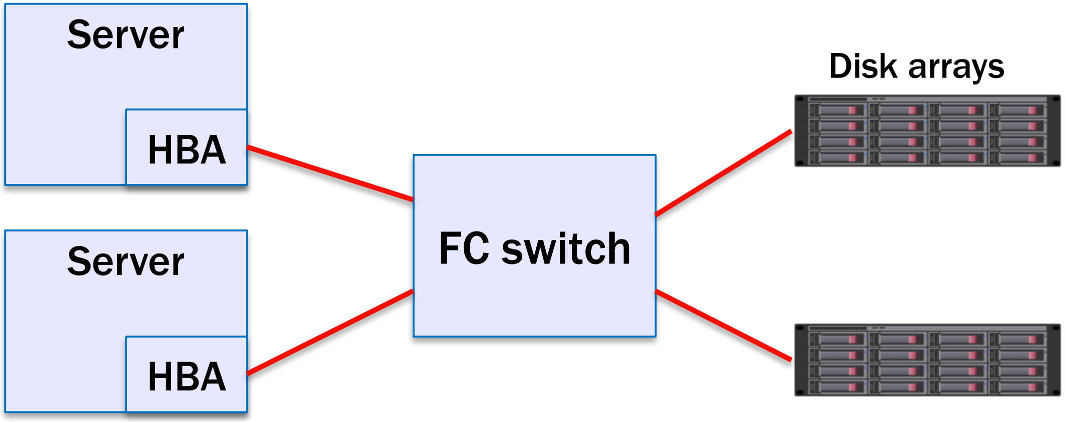
- A **Storage Area Network (SAN)** is a dedicated network that carries data between computer systems and storage devices
- A SAN consists of:

  - *a communication infrastructure*, which provides physical connections, and

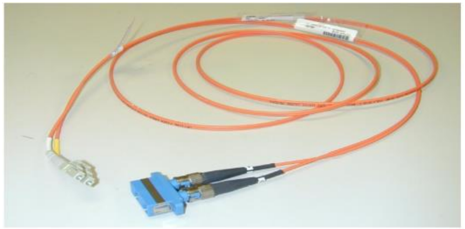
  - *a management layer*, which organizes the connections, storage elements, and computer systems
- Differently from NAS, a SAN provides servers with a block storage abstraction
- A server can attach a remote volume as if it were directly attached
- A SAN supports centralized storage management
  - SANs make it possible to move data between various storage devices, share data between multiple servers, and back up and restore data rapidly and efficiently

Fibre Channel architecture

- Fibre Channel is the reference standard for SANs (*block storage*)
- Operates over copper and fiber optic cables at distances of up to 10 kilometers
- Hosts are equipped with special NICs called *Host Bus Adapters* (HBA)
- Special FC switched are required to interconnect servers with storage devices

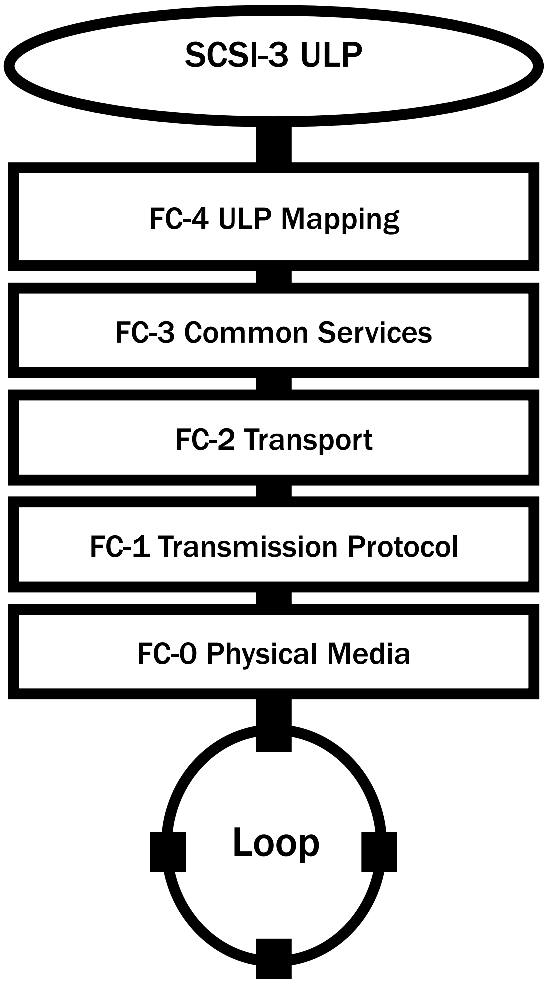


Server
HBA
Server
HBA
FC switch
Disk arrays



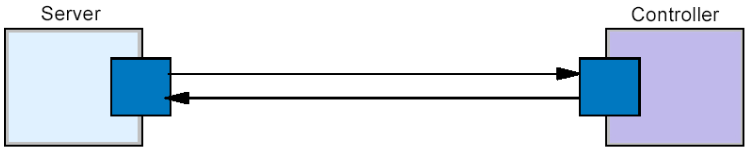
Fibre Channel protocol

- Fibre Channel is a technology based on a complex layered architecture
- Switched network protocol
- 1/2/4/8/16 Gbps + 10 Gbps data rate
- With FC the delivery of data is guaranteed and there’s no loss of data
  - Credit based link level flow control
- FC-4 Protocol Mapping for SCSI:
- defines how to send SCSI information on FC
- defines Data Information Units
  - FCP_CMND (unsolicited command)
  - FCP_XFER_RDY (data descriptor)
  - FCP_DATA (solicited data)
  - FCP_RSP (command status)

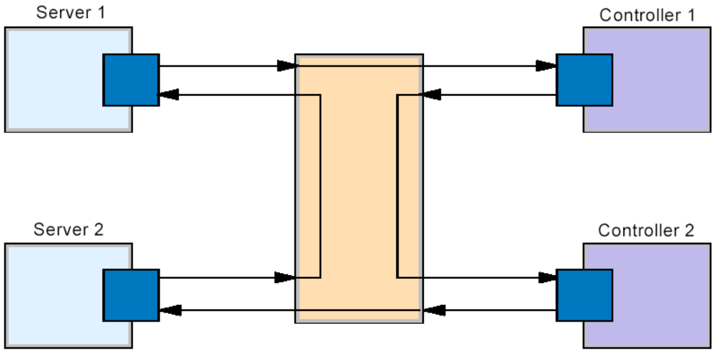


Fibre Channel topologies

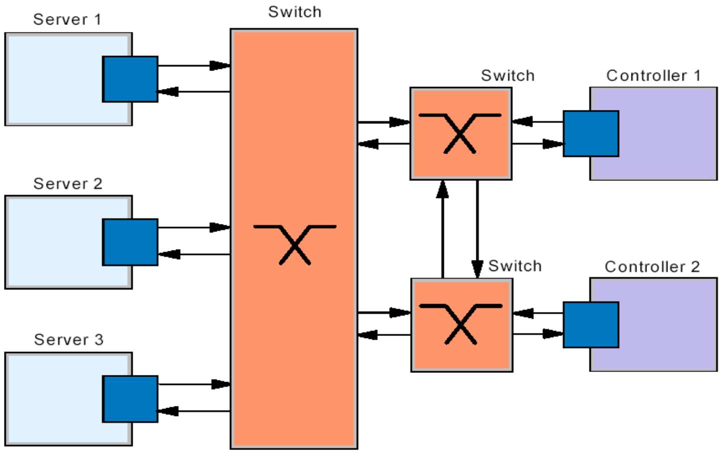
- **Point-to-point**
  - A direct connection between two endpoints
- **Arbitrated loop**
  - This is a ring topology that shares the fiber-channel bandwidth among multiple endpoints
  - The loop is implemented within a hub that interconnects the endpoints
  - An arbitrated scheme is used to determine which endpoint gets control of the loop. The maximum number of ports is 127.
- **Switched fabric**
  - Provides the max flexibility and makes the best use of the aggregated bandwidth by the use of switched connections between endpoints
  - One or more switches are interconnected to create a fabric, to which the endpoints are connected



point-to-point



arbitrated loop



switched fabric

Network convergence for storage protocols

- Fibre Channel requires its own interconnection systems
- To decrease costs (to buy dedicated switch fabrics and to deploy a dedicated cabling system) in modern datacenters are recently applied new technologies that allow to connect SAN systems to servers through the an Ethernet infrastructure
  - This infrastructure may be separated from the Ethernet infrastructure used for server-to-server communication or just be the same
- Communication requirements for a networked storage system:
  - Lossless data transfer
  - Timely delivery

SCSI

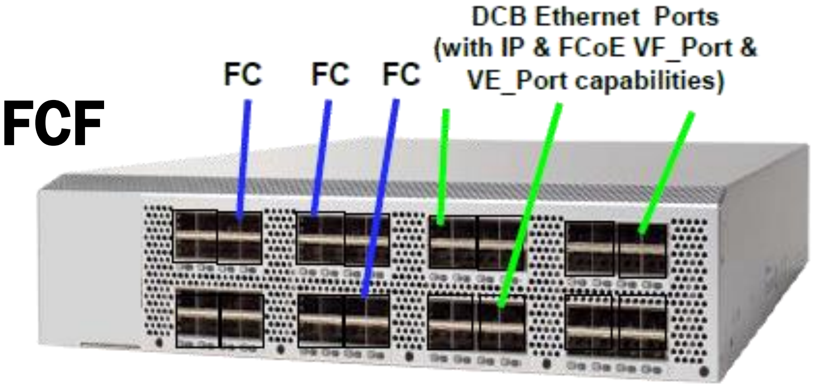
- SCSI is a technology used to connect devices to a host
- The endpoint of most SCSI commands is a "logical unit" (LU)
- Examples of logical units include hard drives, tape drives, CD and DVD drives, printers and processors
- An ***initiator*** creates and sends SCSI commands to the ***target***
- A ***task*** is a linked set of ***SCSI commands***
  - Any SCSI activity is related to a task
- Some LUNs support multiple pending (queued) tasks
  - The target uses a "task tag" to distinguish between tasks
- A SCSI command results in an optional data phase and a response phase
  - In the data phase, information travels either from the initiator to the target, as in a WRITE command, or from target to initiator, as in a READ command
  - In the response phase, the target returns the final status of the operation, including any errors
    - A response terminates a SCSI command

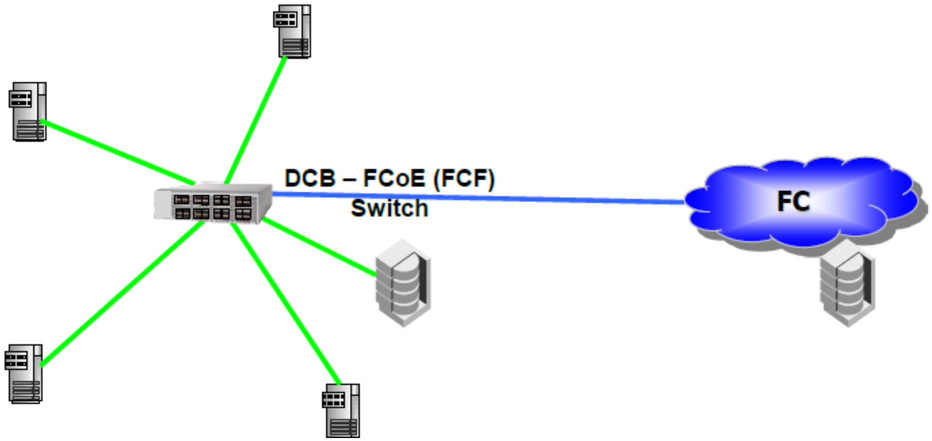
iSCSI

- iSCSI directly implements a SAN across a TCP/IP network
- iSCSI initiator functionality available in most operating systems and hypervisors
- Communication between initiator and target occurs over one or more TCP connections
- The TCP connections are used for sending control messages, SCSI commands, parameters and data within iSCSI Protocol Data Units (iSCSI PDU)
- The group of TCP connections linking an initiator with a target form a *session*
- iSCSI supports ordered command delivery within a session
- All commands (initiator-to-target) and responses (target-to-initiator) numbered
- The targets listen on a well-known TCP port for incoming connections
- The initiator begins the login process by connecting to that well-known TCP port
- As part of the login process, the initiator and target MAY wish to authenticate each other
- Once suitable authentication has occurred, the target MAY authorize the initiator to send SCSI commands

Fibre Channel over Ethernet (FCoE)

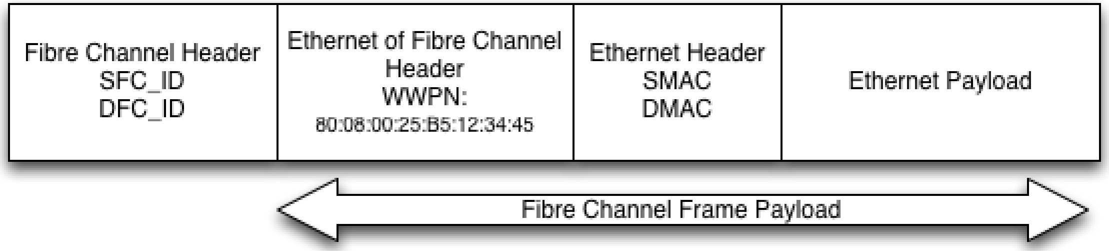
- FCoE is a standard (T11 FC-BB-5) that allows to transmit FC messages as a L3 protocol encapsulated in Ethernet frames (with type=0x8906)
- FC frames usually carry SCSI commands
- FCoE requires specific Ethernet extensions to be implemented
  - Lossless switches and fabrics (e.g. supporting IEEE 802.3 PAUSE)
  - Jumbo frames support strongly recommended
- Traditional FC storage devices can be connected to the Ethernet infrastructure through a switching device called *Fibre Channel Forwarder* (FCF)
- FCFs act as bridges towards traditional FC SAN devices, encapsulating and decapsulating FC frames





Stacking FC and Ethernet the other way round: EoFC

- It is also possible to carry Ethernet frames on a Fiber Channel infrastructure
- *Ethernet over Fiber Channel* (EoFC) provides transmission of Ethernet frames encapsulated in FC PDUs



Fibre Channel Header
SFC_ID
DFC_ID

Ethernet of Fibre Channel Header
WWPN:
80:08:00:25:B5:12:34:45

Ethernet Header
SMAC
DMAC

Ethernet Payload

Fibre Channel Frame Payload

- CNHs (*Converged Network HBAs*) provide Ethernet interfaces to the host
  - The hosts forms Ethernet frames that the CNH encapsulates into FC frames
  - Since standard Ethernet MTU is 1500 bytes, it fits into the maximum 2048 byte Fibre Channel frame; Jumbo Ethernet frames up to 9216 bytes may be transmitted by fragmenting them into multiple 2048-byte FC frames
  - Ethernet MAC addresses are extended with the 80:08 prefix to obtain 64-bits FC WWPN addresses
- The biggest EoFC benefit is the lossless network that Fibre Channel provides