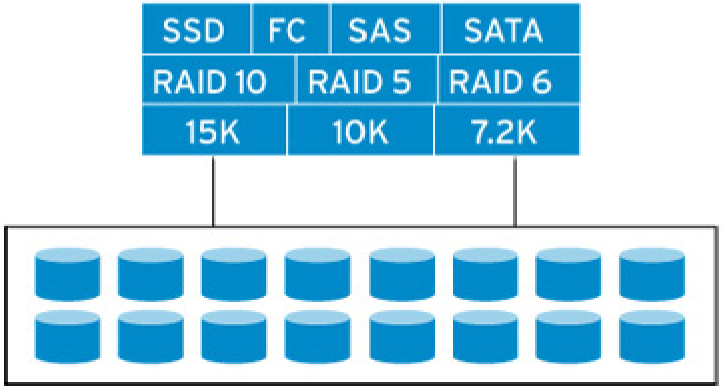
Storage Virtualization



SSD FC SAS SATA

RAID 10 RAID 5 RAID 6

15K 10K 7.2K

Raj Jain
Washington University in Saint Louis
Saint Louis, MO 63130
Jain@cse.wustl.edu

These slides and audio/video recordings of this class lecture are at:

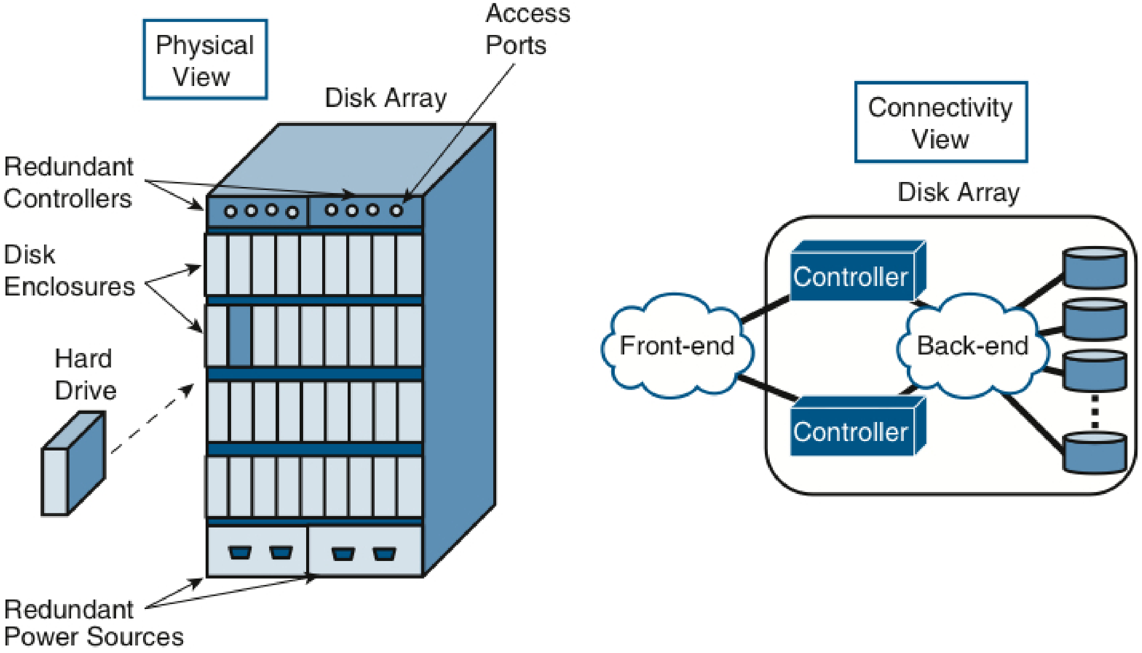
http://www.cse.wustl.edu/~jain/cse570-13/

Overview

1. Storage Interfaces: SCSI and Fibre Channel
2. Storage Area Networks
3. Storage Virtualization
    1. Device Virtualization: RAID
    2. Fabric Virtualization: Storage access over Ethernet or IP
4. SAN vs. NAS

Disk Arrays

- ❑ In data centers, all disks are external to the server ⇒Data accessible by other servers in case of a server failure
- ❑ JBODs (Just a bunch of disks): Difficult to manage
- ❑ Disk Arrays: An easy to manage pool of disks with redundancy



Ref: G. Santana, "Data Center Virtualization Fundamentals," Cisco Press, 2014, ISBN:1587143240

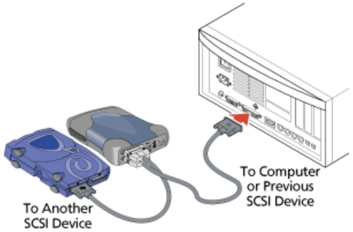
Data Access Methods

Three ways for applications to access data:

- **Block Access**: A fixed number of bytes (block-size), e.g., 1 sector, 4 sectors, 16 sectors
- **File Access**: A set of bytes with name, creation date, and other meta data.
  - May or may not be contiguous.
  - A file system, such as, FAT-32 (File Allocation Table) or NTFS (New Technology File System) defines how the meta-data is stored and files are organized.
  - File systems vary with the operating systems.
- **Record Access**: Used for highly structured data in databases. Each record has a particular format and set of fields. Accessed using Structured Query Language (SQL), Open DataBase Connectivity (ODBC), Java DataBase Connectivity (JDBC)
- Storage systems provide block access. A logical volume manager in the OS provides other "virtual" views, e.g., file or record

SCSI (Small Computer System Interface)

- Used to connect disk drives and tapes to computer
- 8-16 devices on a single bus. Any number of hosts on the bus
  At least one host with host bus adapter (HBA)
- Standard commands, protocols, and optical and electrical interfaces.



- Peer-to-peer: host-to-device, device-to-device, host-to-host
  But most devices implement only targets. Can't be initiators.
- Each device on the SCSI bus has a "ID".
- Each device may consist of multiple logical units (LUNs).
  LUNS are like apartments in a building.
- A direct access (disk) storage is addressed by a Logical Block Address (LBA). Each LB is typically 512 bytes.
- Initially used a parallel interface (Parallel SCSI) ⇒ Skew
  Now Serial Attached SCSI (SAS) for higher speed

Ref: http://en.wikipedia.org/wiki/SCSI

Advanced Technology Attachment (ATA)

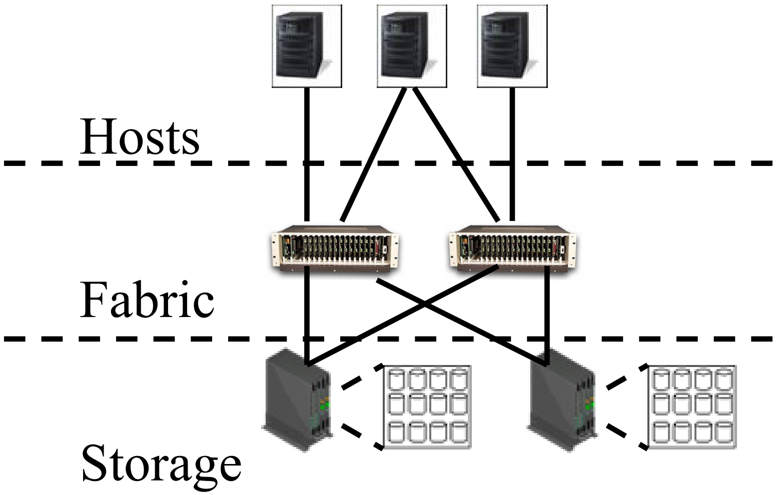
- Parallel Advanced Technology Attachment (PATA):
  - Designed in 1986 for PCs. Controller integrated in the disk ⇒ Integrated Device Electronics (IDE).
  - 133 Mbps using parallel ribbon cables
- ATA Packet Interface (ATAPI): Extended PATA to CD-ROMS, DVD-ROMs, and Tape drives
- Serial Advanced Technology Attachment (SATA):
  - Designed in 2003 for internal hard disks. 6 Gbps.
- PATA Enhancements: ATA-2 (Ultra ATA), ATA-3 (EIDE) SATA Enhancements: external SATA (eSATA), mini SATA (mSATA)

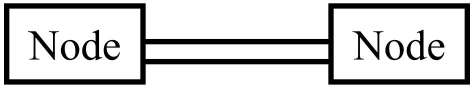
ESCON and FICON

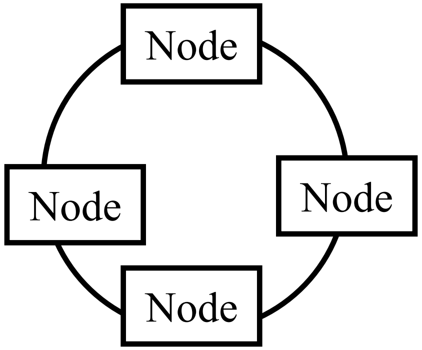
- Enterprise System Connection (ESCON):
  - Designed by IBM for main frames.
  - Includes switches enabling sharing by multiple servers
  - Fibers allowed 17 Mbps over 3-43 KM
  - Half-duplex
- Fiber Connectivity (FICON):
  - Supports point-to-point and cascaded topologies
  - Supports multiple concurrent I/O operations per channel
  - Uses Single Byte Command Code Sets (SBCCS)
  - Uses Fibre Channel as a transport

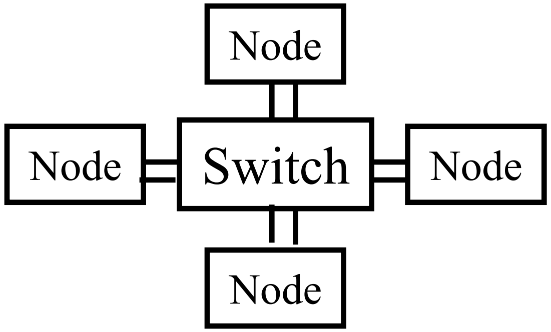
Fibre Chanel

- ANSI T11 standard for high-speed *storage area network* (SAN)
- 2, 4, 8, 16, 32 GBps. Can run on TP or fiber
- Allows point-to-point, arbitrated loop (ring), switched fabric topologies





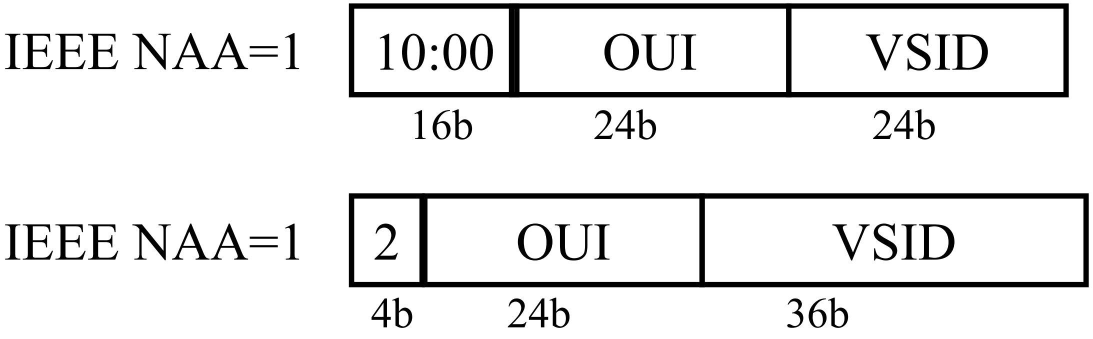




Ref: http://en.wikipedia.org/wiki/Fibre_Channel

Fibre Channel (Cont)

- FC host bus adapters (HBA) have a unique 64-bit World-Wide Name (WWN) similar to 48-bit Ethernet MAC addresses with OUI, and vendor specific identifiers (VSID), e.g., 20000000C8328FE6
- Several different network addressing authorities (NAA)



IEEE NAA=1 | 10:00 (16b) | OUI (24b) | VSID (24b)

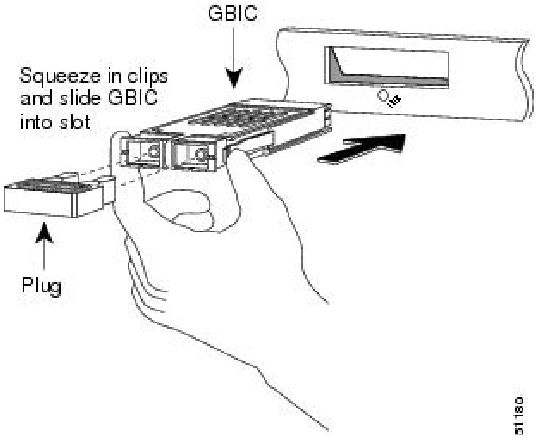
IEEE NAA=1 | 2 (4b) | OUI (24b) | VSID (36b)

Ref: http://en.wikipedia.org/wiki/Fibre_Channel

Fibre Channel Devices

- ❑ Host Bus Adapters (HBA): Network interface card.
- ❑ Gigabit Interface Converter (GBIC):
  Single mode fiber for long-distance.
  Multimode fiber for short distance.
  HBA ports are empty. Plug in GBIC.
- ❑ Hubs: Physical layer Device. Like a active patch panel. Multiple hosts or storage devices. Only one host can talk to one device at a time using an arbitrated loop (FC-AL) protocol.
- ❑ Switches: A link layer device. Forwards FC frames according to destination address.
- ❑ Routers and Gateways:
  Connect FC to other types of storage (SCSI)

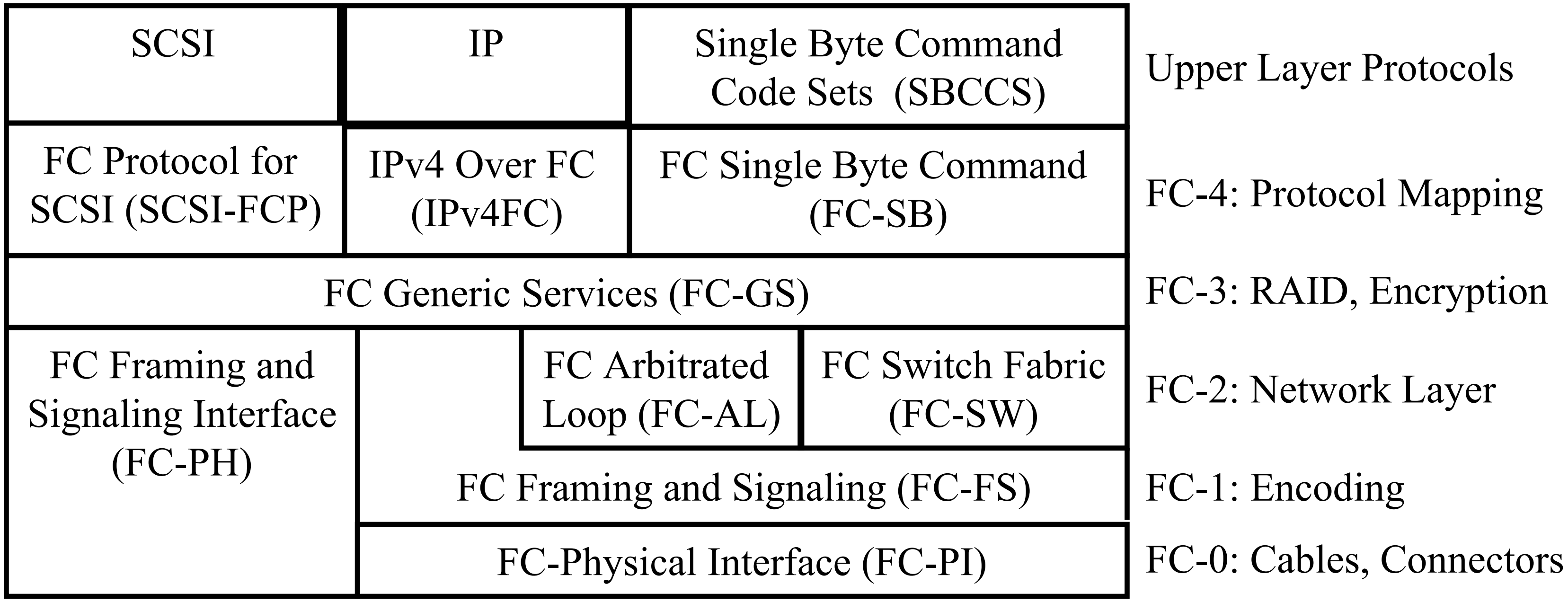
Source: Softel-optic



Source: Cisco

Ref: C. Poelker, A. Nikiti, "Storage Area Networks For Dummies," For Dummies, 2009, ISBN:9780470385135

Fibre Channel Protocol Layers



SCSI | IP | Single Byte Command Code Sets (SBCCS) — Upper Layer Protocols

FC Protocol for SCSI (SCSI-FCP) | IPv4 Over FC (IPv4FC) | FC Single Byte Command (FC-SB) — FC-4: Protocol Mapping

FC Generic Services (FC-GS) — FC-3: RAID, Encryption

FC Framing and Signaling Interface (FC-PH) | FC Arbitrated Loop (FC-AL) | FC Switch Fabric (FC-SW) — FC-2: Network Layer

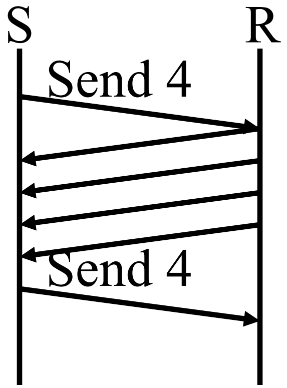
FC Framing and Signaling (FC-FS) — FC-1: Encoding

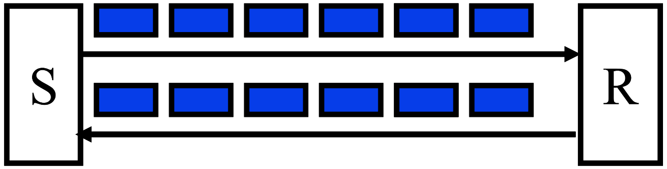
FC-Physical Interface (FC-PI) — FC-0: Cables, Connectors

- ❑ New extensions are named by adding a number, e.g., FC-SW-3 extends FC-SW-2, which extended FC-SW
- ❑ Fibre Channel Shortest Path (FSPF) protocol is used to find routes through the fabric. It is a link-state protocol.
- ❑ Vendor specific equal cost path multiplexing

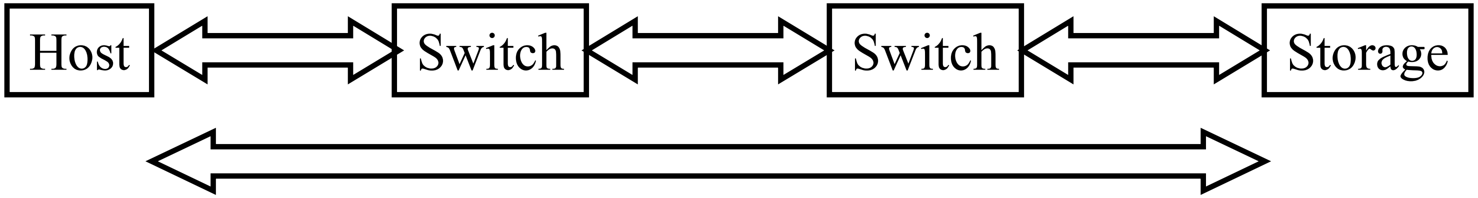
Fibre Channel Flow Control

- Transmitter sends frames only when allowed by the receiver
- Credit-based flow control
- For optimal performance,
  the Credit ≥ Round-trip path delay





- Both Hop-by-Hop and End-to-End



Fibre Channel Classes of Service

- **Class 1**: Connection-oriented dedicated (physical links). Frame order guaranteed. Delivery confirmation. End-to-end flow control.
- **Class 2**: Connectionless. Multiple paths ⇒ order not guaranteed. Hop-by-hop and end-to-end flow control.
- **Class 3**: Datagram service. No delivery confirmation. Only hop-by-hop flow control. Most common.
- **Class 4**: Connection-oriented virtual circuits (fractional links) with delivery confirmation
- **Class 5**: Not yet defined
- **Class 6**: Connection-oriented multicast with delivery confirmation
- **Class F**: Packet-switched delivery with confirmation. For inter-switch communication

Ref: Santana 2014

What is Storage Virtualization?

- Restating Rick F. Van der Lans: *Storage virtualization means that Applications can use storage without any concern for where it resides, what the technical interface is, how it has been implemented, which platform it uses, and how much of it is available*
- Distance: Remote storage devices appear local
- Size: Multiple smaller volume appear as a single large volume
- Spread: Data is spread over multiple physical disks to improve reliability and performance
- File System: Windows, Linux, and UNIX all use the same storage device
- Virtual Interface: A SCSI disk connected to a computer with no SCSI interface
- Advantages: High availability, Disaster recovery, improved performance, sharing (better CapEx)

Benefits of Storage Virtualization

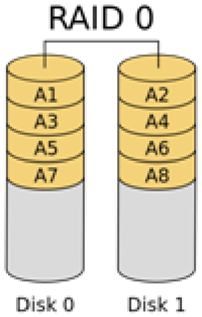
- Much larger distances
- Greater performance
- Increased disk utilization
- Higher availability with multiple access path
- Higher availability due to redundant storage
- Disaster recovery capability
- Continuous on-line back
- Easier testing
- Increased scalability
- Allows thin provisioning (Appears as if there is more disk than physical)

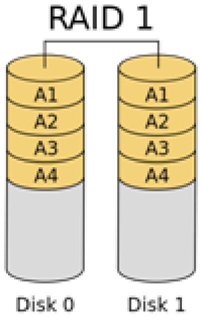
Virtualizing Storage

- Partitions and file systems are "Virtual" views of the storage
- A disk array can be partitioned into virtual (logical) devices with LUNs, File systems assigned to different tenants
- Thin Provisioning: Allocate blocks only when used
  ⇒ Overbooking
- Another way to virtualize storage is use multiple physical disks to look like a single disk using RAID
- RAID (Redundant array of independent disks)
- Originally Redundant array of inexpensive disks
  (as invented by Patterson, Gibson, and Katz)
- Trick: Divide and replicate data among multiple drives
- Provides availability, performance and/or capacity.

RAID Levels

- RAID 0: block-level *striping* without parity. Zero redundancy. Higher performance and capacity.
- RAID 1: *Mirroring* without parity. Higher read performance. Two or more mirrors.
- RAID 2: Bit-level striping with dedicated Hamming code *parity*. Each sequential bit is on a different drive.
  Not used in practice.
- RAID 3: Byte-level striping with dedicated Hamming code parity. Not commonly used.

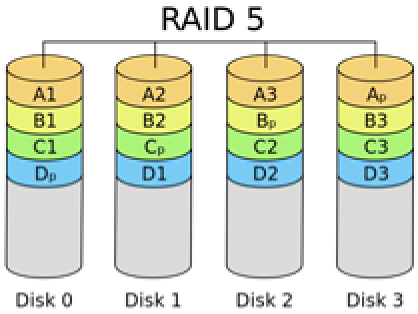




Ref: http://en.wikipedia.org/wiki/RAID

RAID Levels (Cont)

- RAID 4: Block-level striping with dedicated parity. Allows I/O requests to be performed in parallel.
- RAID 5: Block-level striping with *distributed parity*. Masks failure of 1 drive.
- RAID 6: Block-level striping with double distributed parity. Masks up to two failed drives. Better for large drives that take long time to recover.

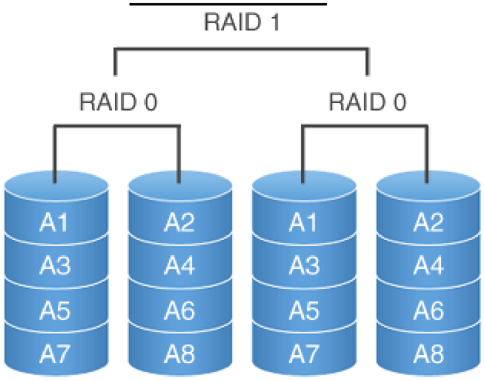


Ref: http://en.wikipedia.org/wiki/RAID

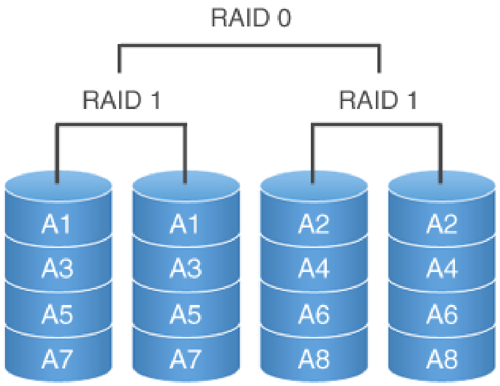
Nested RAIDs

- RAID of RAID drives
- RAID 01: Stripe and then mirror
  = RAID 0+1
  Data is striped across primary disks that are mirrored to secondary disks.
- RAID 10: Mirror then stripe
  = RAID 1+0
- The order of digits is the order in which the set is built.
  RAID 0+1 ⇒ Stripping first and then mirroring
- Mirrored striped set with distributed parity = RAID 5+3 or RAID 53

RAID 01



RAID 10

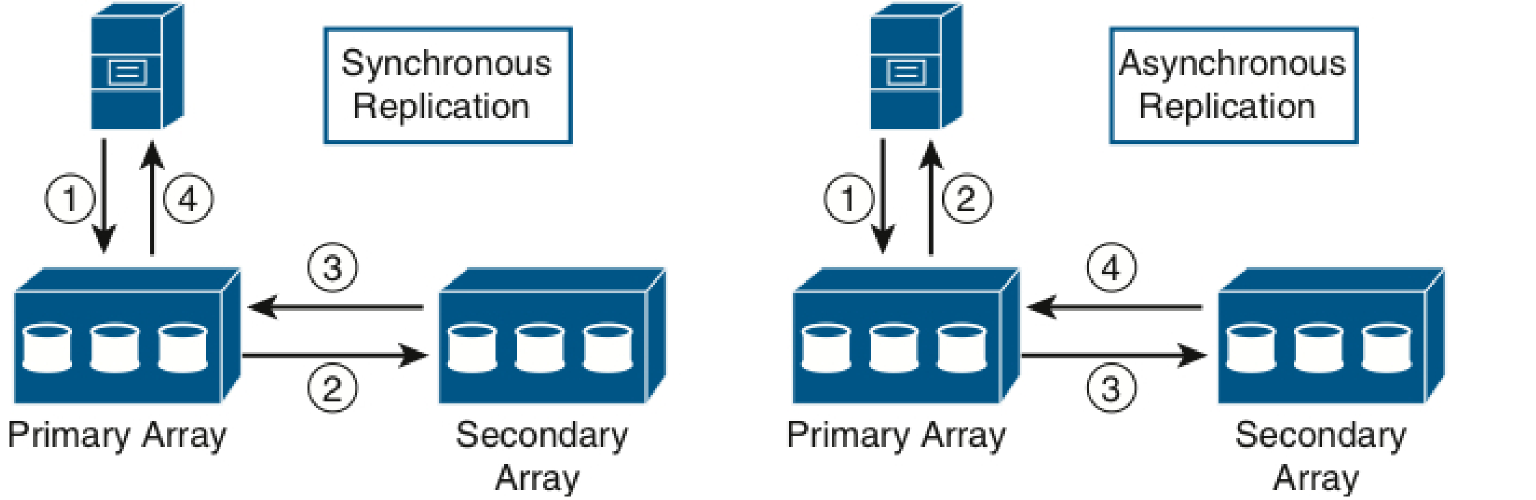


Homework 6

- What is RAID 50?

Synchronous vs. Asynchronous Replication

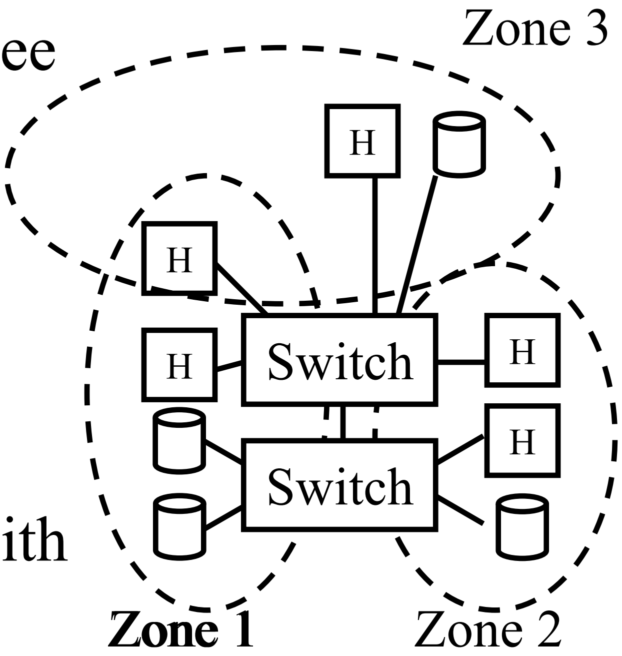
- ❑ Synchronous: Immediate secondary writes.
  Write completes only after finishing on secondary storage
  ⇒ Guaranteed recovery but slow
- ❑ Asynchronous: Delayed secondary writes
  Writes acked to server even before completion on secondary.
  Writes to secondary are queued in the primary



[Source: Santana 2014]

Virtual Storage Area Network (VSAN)

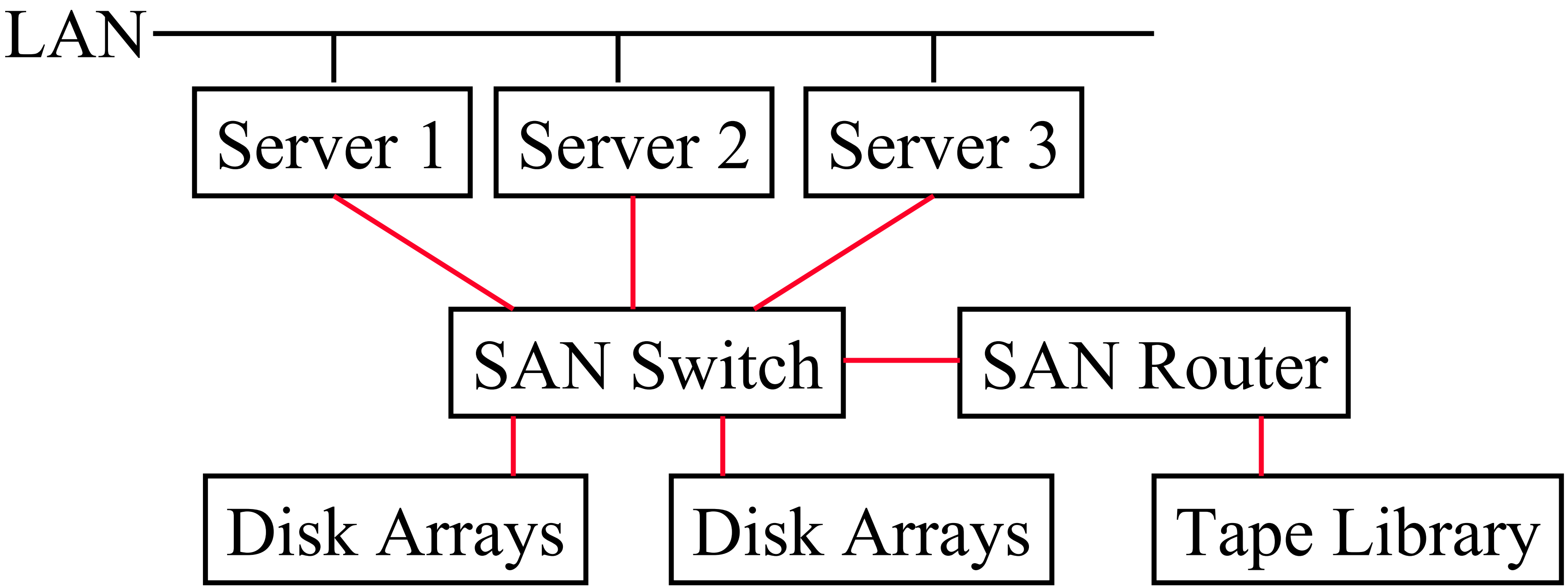
- Zones in a FC SAN provide isolation among tenants. Some switch ports can see only some other switch ports.
- Different zones share a zone server, name server, and login server
  ⇒ Subject to common failures
- Virtual Storage Area Network (VSAN) technology allows different partitions with their own servers. Similar to VLANs.



- Each VSAN provide complete fabric services
  ⇒ Each VSAN can be subdivided in to zones.

Ref: Santana 2014

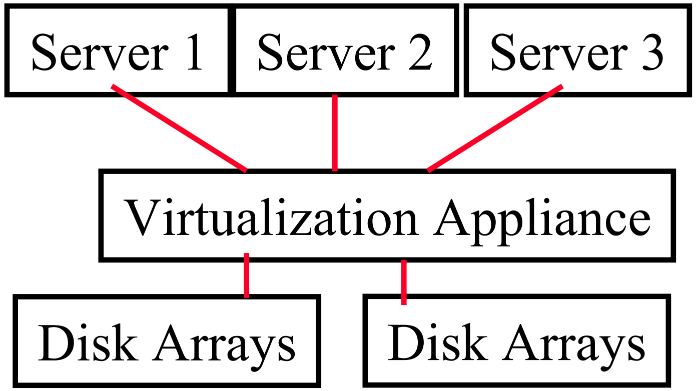
Physical Storage Network



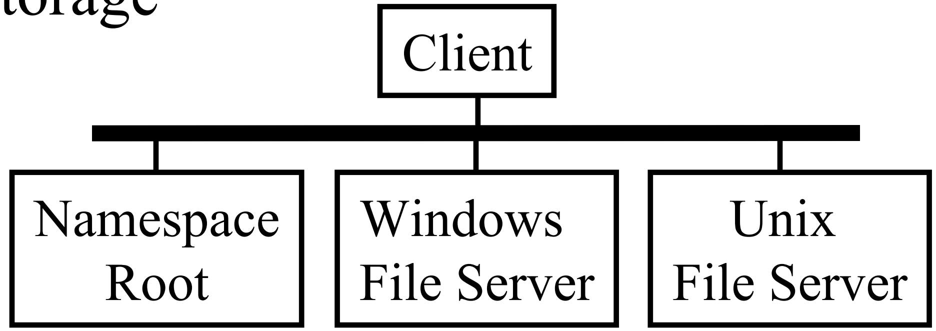
- Each host has a one-to-one relationship with a storage device ⇒ Physical

Virtual Storage Network

- In-Band: All control and data goes to virtualization appliance



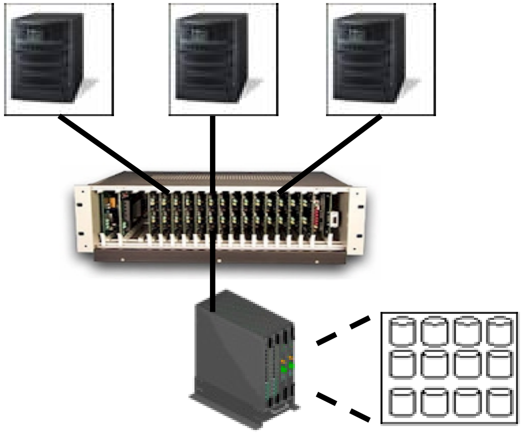
- Out-Band: All control goes to Namespace root. Data goes directly between host and storage



SAN vs. NAS

- Storage Area Network (SAN)
  Network attached storage (NAS)
- SAN: Storage servers connected via special purpose storage network, e.g., Fibre Channel
- NAS: Storage servers accessed over a general purpose network, e.g., Ethernet

iSCSI (Internet Small Computer System Interface)



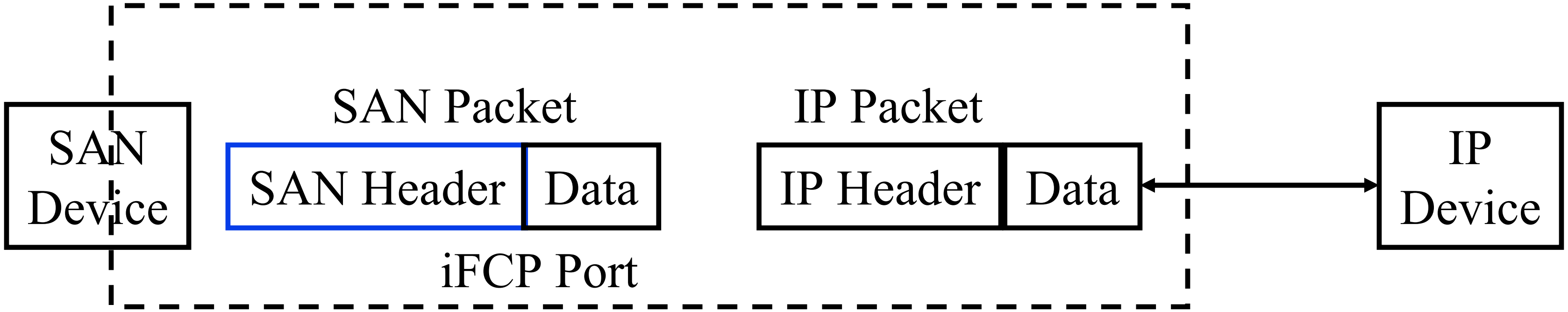
- IETF protocol to carry SCSI commands over traditional TCP/IP/Ethernet
- Requires no dedicated cabling.
- Uses TCP end-to-end congestion control
- Can use the same Ethernet port on the computers to connect to storage devices on different computers
- iSNS (Internet Storage Name Service) can be used to locate storage resources

Ref: http://en.wikipedia.org/wiki/ISCSI

Ref: C. Wolf, E. M. Halter, "Virtualization: From the Desktop to the Enterprise," Apress, 2005, ISBN:1590594959

iFCP (Internet Fiber Channel Protocol)

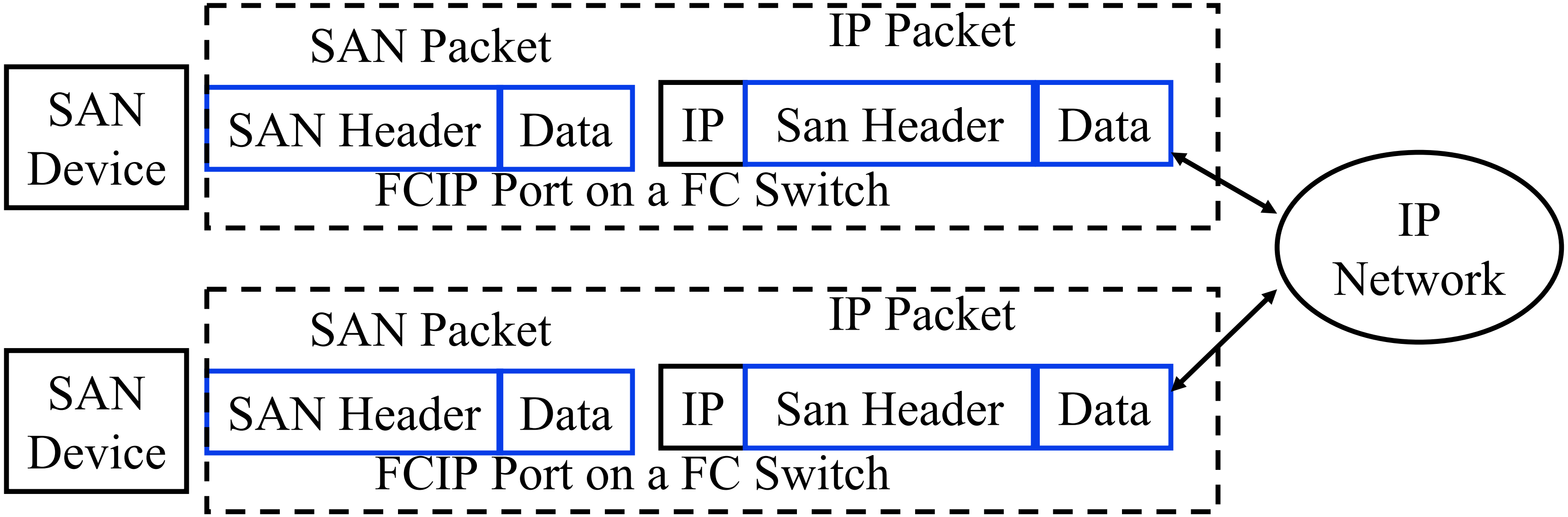
- Interconnect FC devices using TCP/IP
- Can connect native IP based storage and FC devices
- SAN frames are converted to IP packets at the source and sent to the destination
- Uses TCP Congestion Control (end-to-end)
- IP to iFCP



Ref:: RFC 4172

FCIP (Fibre Channel over IP)

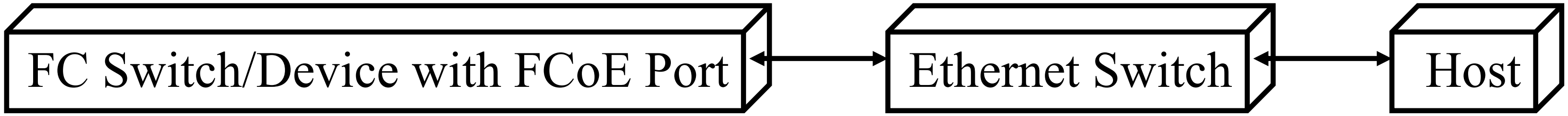
- Tunneling protocol for passing FC frames over TCP/IP.
- SAN packets are encapsulated in IP packets at the source and decapsulated back at the destination
- Doesn't allow to directly interface with a FC device.
- Some FC switches have FCIP ports.



Ref: RFC 3821

FCoE (Fibre Channel over Ethernet)

- Maps FC directly over Ethernet
- Replaces FC0 and FC1 layers with Ethernet
- Allows FC traffic to go over Ethernet without needing FC media.
- FCoE runs directly on Ethernet (unlike iSCSI which runs on TCP) ⇒ Not routable over IP networks ⇒ Extension issues
- Has a dedicated EtherType (0x8906)
- Required extensions to Ethernet to minimize loss during congestion
- Required mapping between FCIDs and Ethernet MAC addresses



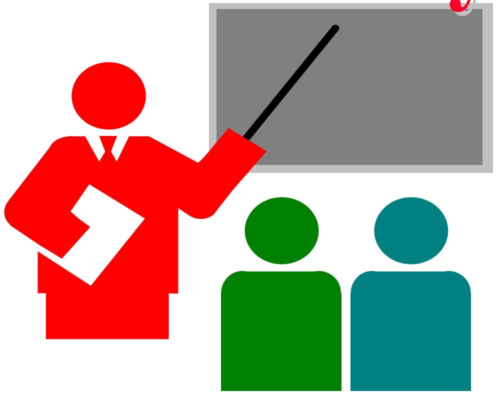
Ref: http://en.wikipedia.org/wiki/FCoE

Virtual File Systems

- Storage access is either block based or file based
- File systems, e.g., NTFS or FAT32 store files on a block based storage.
- Virtual file systems allows files located on multiple network drives to appear as if on a single local drive
  ⇒ Network drives can be replicated, relocated, reconstructed
- Windows DFS
- Linux DFS: Open source implementation of Windows DFS on Linux
- AFS: Andrew File System from CMU (Andrew Carnegie)
- Parallel Virtual File System (PVFS) distributes data across multiple servers to provide concurrent access for parallel task

Ref: Wolf2005

Summary



1. SCSI is a common interface used on storage devices
2. Fibre channel is a storage area network
3. RAID allows data to be partitioned over multiple drives for performance and fault tolerance
4. iSCSI, iFCP, FCIP, FCoE are protocols for interconnecting storage over Ethernet/IP.
5. SAN is FC based. NAS is Ethernet based.
6. Virtual file systems allow files to be accessed in multiple views from the same storage system.

Acronyms

AFS	Andrew File System
ATA	Advanced Technology Attachment
ATAPI	Advanced Technology Attachment Programming Interface
ANSI	American National Standards Institute
CapEx	Capital Expenditure
CMU	Carnegie Mellon University
DFS	Distributed File System
EIDE	Enhanced Integrated Device Electronics
eSATA	External Serial Advanced Technology Interface
FAT	File Allocation Table
FC	Fibre Channel
FC-AL	Fibre Channel Arbitrated Loop
FC-FS	Fibre Chanel Framing and Signaling
FC-GS	Fibre Chanel generic services
FC-PH	Fibre Chanel Framing and signaling interface
FC-PI	Fibre Chanel physical Interface

Acronyms (Cont)

❑ FC-SB	Fibre Chanel Single Byte Command
❑ FC-SW	Fibre Chanel Switch Fabric
❑ FC-PI	Fibre Chanel physical
❑ FCID	Fibre Channel Identifier
❑ FCIP	Fibre Channel over IP
❑ FCoE	Fibre Channel over Ethernet
❑ FSPF	Fibre Channel Shortest Path
❑ GBIC	Gigabit Interface Converter
❑ HBA	Host Bus Adapters
❑ IDE	Integrated Device Electronics
❑ IETF	Internet Engineering Task Force
❑ iFCP	Internet Fibre Channel Protocol
❑ IPv4FC	IPv4 over Fibre Channel
❑ iSCSI	Internet Small Computer System Interface
❑ iSNS	Internet Storage Name Service
❑ JDBC	Java DataBase Connectivity

Acronyms (Cont)

LB	Logical Block
LBA	Logical Block Address
LUN	Logical Unit Number
MAC	Media Access Control
mSATA	Mini Serial Advanced Technology Interface
NAS	Network attached storage
NTFS	New Technology File System
ODBC	Open DataBase Connectivity
OS	Operating System
OUI	Organizationally Unique Identifier
PATA	Parallel Advanced Technology Attachment
PHY	Physical Layer
RAID	Redundant Array of Independent Disks
SAN	Storage Area Network
SATA	Serial Advanced Technology Interface
SBCCS	Single Byte Command Code Sets

Acronyms (Cont)

SCSI	Small Computer System Interface
SCSI-FCP	SCSI over Fibre Channel Protocol
SQL	Structured Query Language
TP	Twisted Pair
VLANs	Virtual Local Area Network
VSAN	Virtual Storage Area Network
WWN	World-Wide Name

Reading List

- G. Santana, “Data Center Virtualization Fundamentals,” Cisco Press, 2014, ISBN:1587143240 (Chapter 9 and 10) (Safari Book)
- C. Poelker, A. Nikiti, "Storage Area Networks For Dummies," For Dummies, 2009, ISBN:9780470385135 (Safari Book)
- C. Wolf, E. M. Halter, "Virtualization: From the Desktop to the Enterprise," Apress, 2005, ISBN:1590594959 (Not available on Safari⇒Optional)

Wikipedia Links

- http://en.wikipedia.org/wiki/Arbitrated_loop
- http://en.wikipedia.org/wiki/Block_(data_storage)
- http://en.wikipedia.org/wiki/Direct-attached_storage
- http://en.wikipedia.org/wiki/Fibre_Channel_electrical_interface
- http://en.wikipedia.org/wiki/Fibre_Channel_network_protocols
- http://en.wikipedia.org/wiki/Fibre_Channel_over_Ethernet
- http://en.wikipedia.org/wiki/Fibre_Channel_switch
- http://en.wikipedia.org/wiki/Fibre_Channel_zoning
- http://en.wikipedia.org/wiki/Hierarchical_storage_management
- http://en.wikipedia.org/wiki/Internet_Fibre_Channel_Protocol
- http://en.wikipedia.org/wiki/Internet_Storage_Name_Service
- http://en.wikipedia.org/wiki/ISCSI
- http://en.wikipedia.org/wiki/Logical_unit_number
- http://en.wikipedia.org/wiki/Nested_RAID_levels
- http://en.wikipedia.org/wiki/Network-attached_storage

Wikipedia Links (Cont)

- http://en.wikipedia.org/wiki/Non-RAID_drive_architectures
- http://en.wikipedia.org/wiki/Non-standard_RAID_levels
- http://en.wikipedia.org/wiki/Parallel_Virtual_File_System
- http://en.wikipedia.org/wiki/SCSI
- http://en.wikipedia.org/wiki/Standard_RAID_levels
- http://en.wikipedia.org/wiki/Storage_area_network
- http://en.wikipedia.org/wiki/Storage_hypervisor
- http://en.wikipedia.org/wiki/Storage_virtualization
- http://en.wikipedia.org/wiki/Switched_fabric
- http://en.wikipedia.org/wiki/Thin_provisioning
- http://en.wikipedia.org/wiki/Virtual_file_system