SIC

Sistemas e Infraestruturas de Comunicação

Storage Solutions &Technologies

Images: Old backup tapes from energy utility

Back to basics: how to protect your data?

Images Source: pctechnotes.com; www.notcot.com

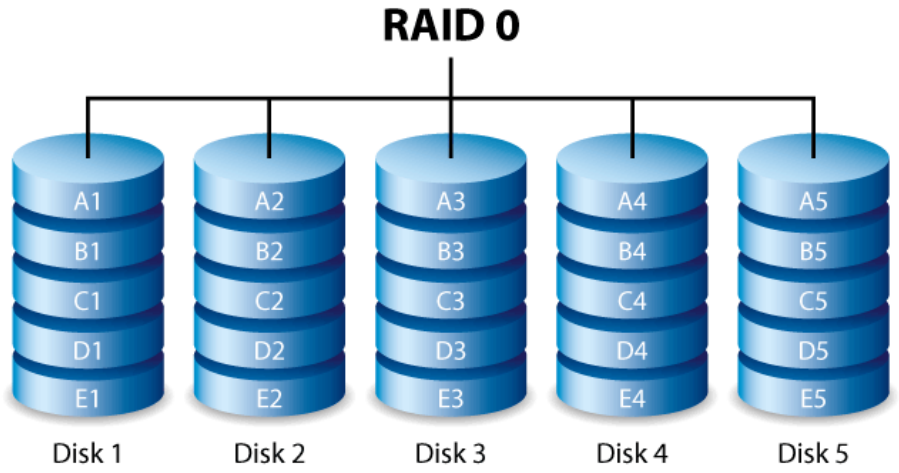
RAID (Redundant Array of Independent Disks)

- *RAID 0; RAID 1; Nested RAID; RAID 3; RAID 4, RAID 5, RAID 6…*

RAID 0

- ***Stripping***: data is striped across the HDDs in a RAID set
- The stripe size is specified at a host level for software RAID and is vendor-specific for hardware RAID
- When the number of drives in the array increases, performance improves because more data can be read or written simultaneously
- Used in applications that need high I/O throughput
- No data protection and availability if drives fail

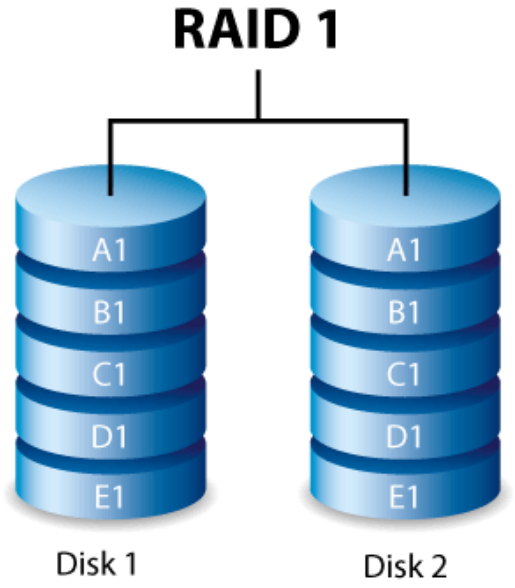
Images Source: seagate.com



RAID 1

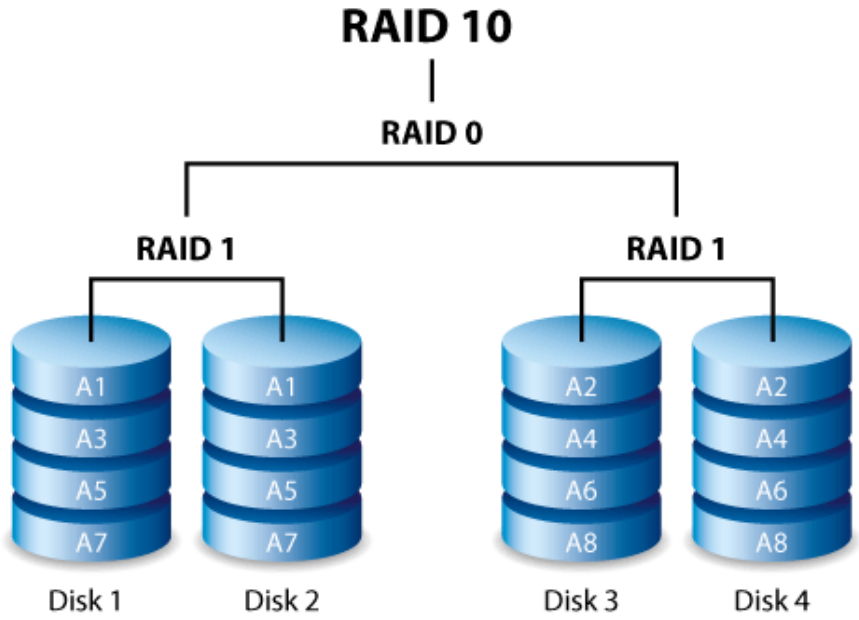
- ***Mirroring*** is applied to stored data in different HDDs, yielding multiple copies of data (usually two copies).
- Complete data redundancy & faster recovery from disk failures.
- Duplication of data
  - Needed storage capacity is 2x the amount of stored data.
- Used in mission-critical applications that cannot afford data loss.

Images Source: seagate.com



Nested RAID (RAID 10)

- RAID 1+0 (also called ***striped mirror***)
- Data is first mirrored and then both copies of data are striped across multiple HDDs in a RAID set
- Typical applications:
  - High-load Online Transaction Processing (OLTP),
  - Databases requiring high I/O, random access, high availability



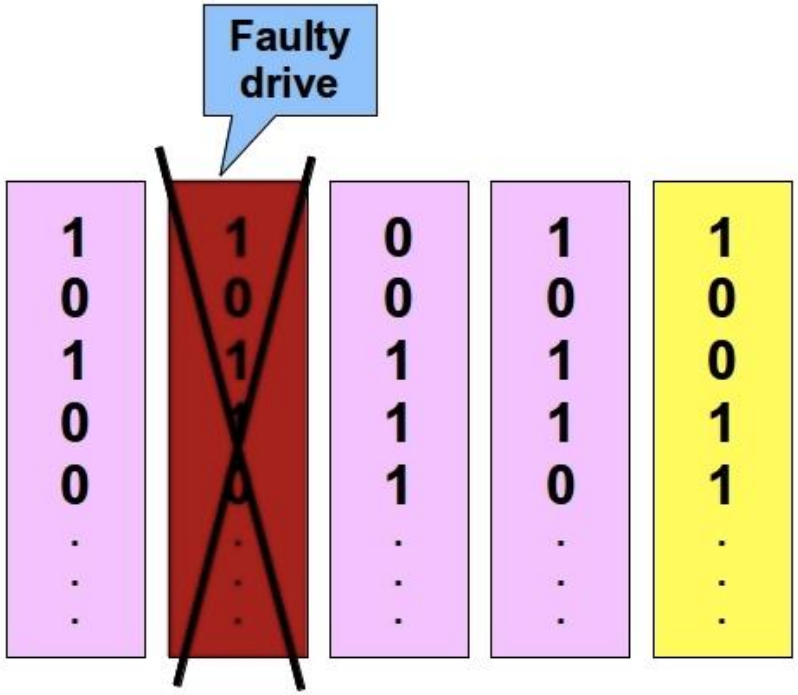
Images Source: seagate.com

RAID redundancy: parity

- For instance: 4+1 drives
- Drive 2 fails:

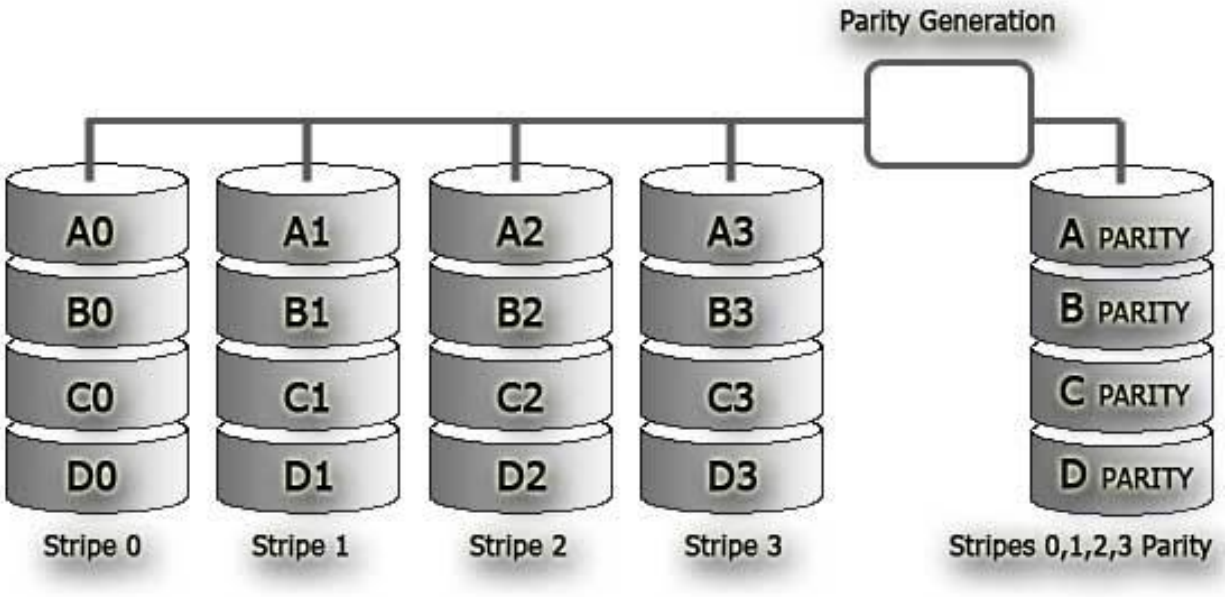
  10100 *XOR* ? *XOR* 00111 *XOR* 10110 = 10011

  ? = 10110



RAID 3

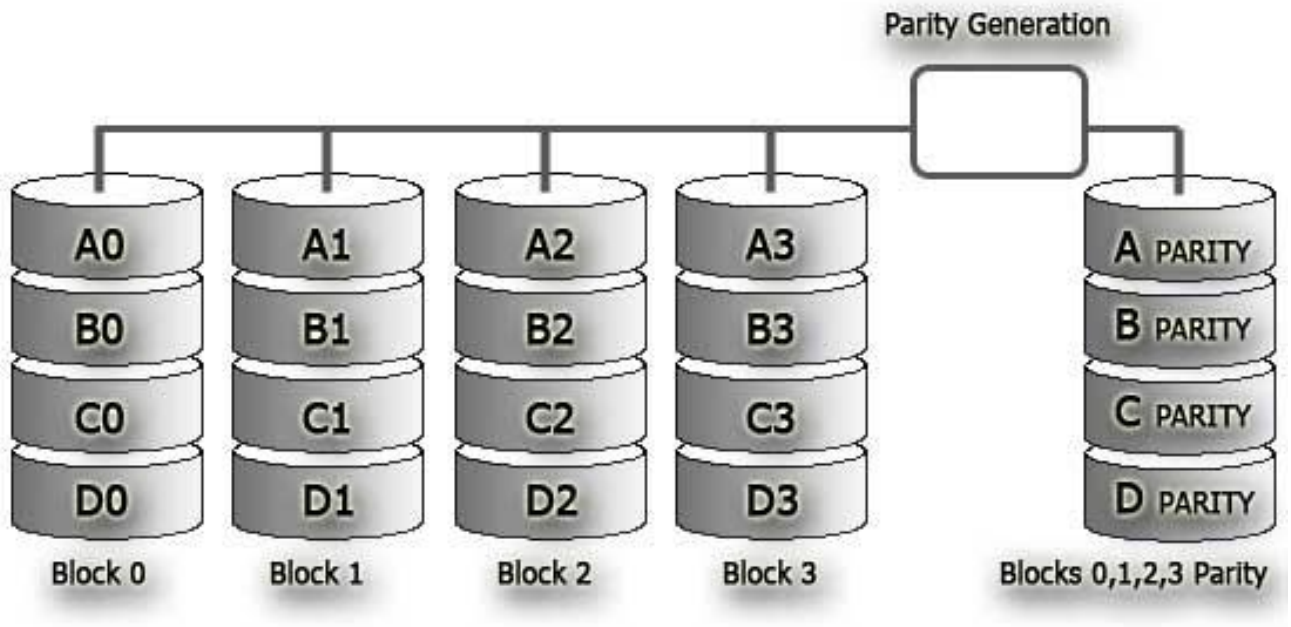
- Striping for performance; parity for improved fault tolerance
- Striping is performed at a byte-level
- Parity information is stored on a dedicated drive
  - Data can be reconstructed if a drive fails
- Used in applications involving large sequential data access, such as video streaming
- Requires synchronized disks for enhanced performance



Images Source: broadberry.co.uk

RAID 4

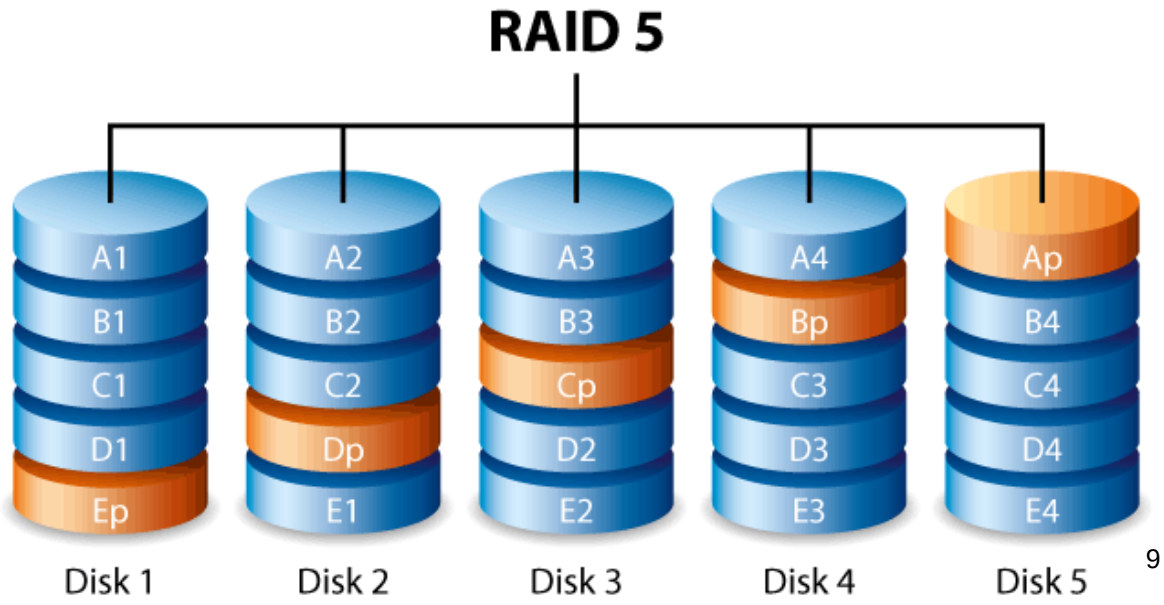
- Stripes data across all disks except the parity disk
- Striping is performed at block-level
- Parity information is stored on a dedicated disk
- Unlike RAID 3, data disks can be accessed independently, so that specific data elements can be read or written on a single disk without read or write of an entire stripe



Images Source: broadberry.co.uk

RAID 5

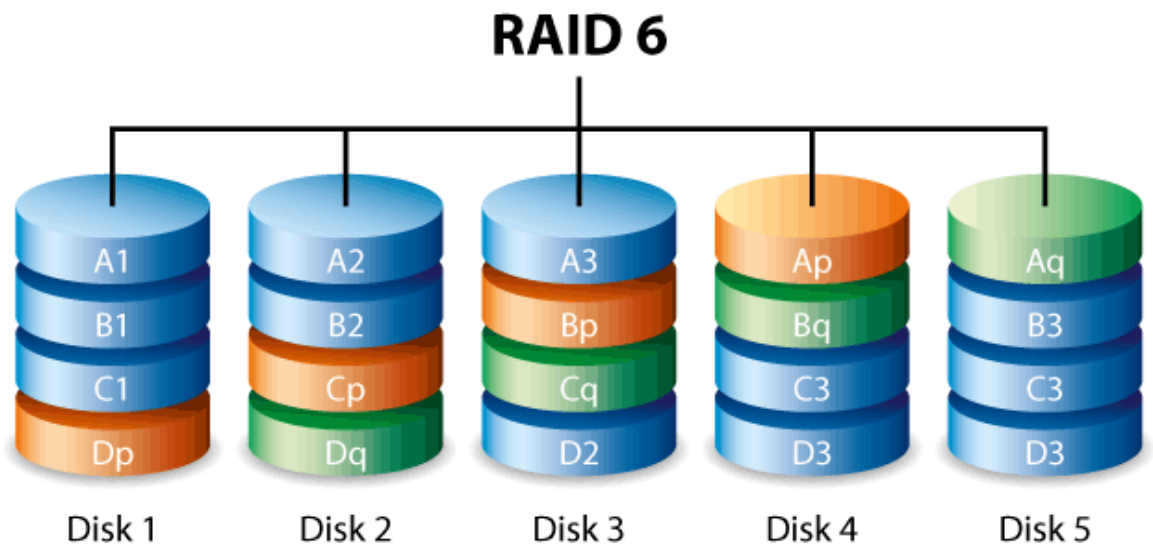
- The difference between RAID 4 and RAID 5 is the parity location
- While RAID 4 uses a dedicated drive for parity, RAID 5 distributes parity across all disks, overcoming write bottlenecks
- Typical applications: messaging, medium-performance media serving, RDBMS implementations with optimized data access



Images Source: seagate.com

RAID 6

- RAID 5 + additional parity element (two parities)
- Survives the failure of 2 disks in the RAID group

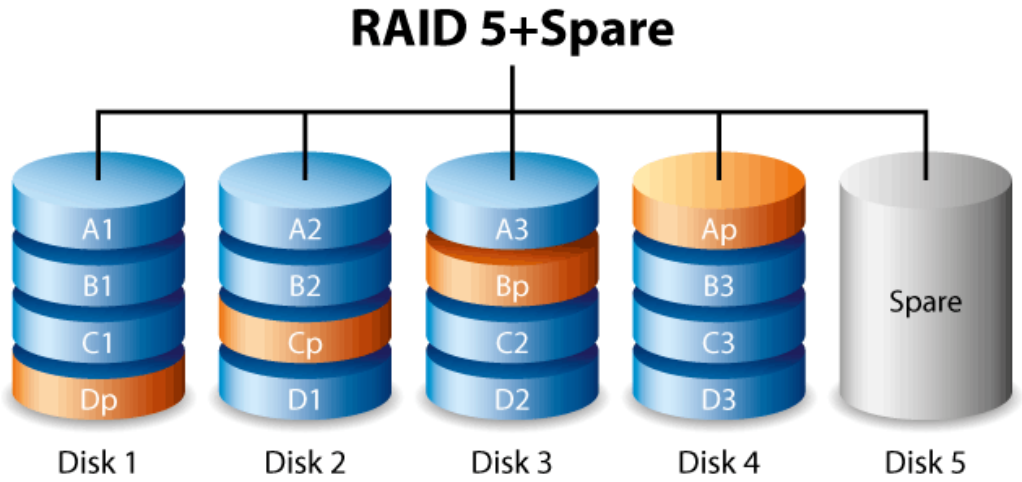


Images Source: seagate.com

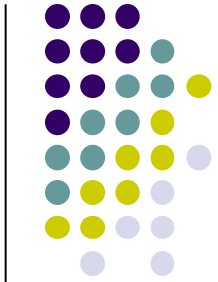
Hot Spare

- Spare HDD in a RAID array that temporarily replaces a failed HDD of a RAID set.
- When the failed HDD is replaced with a new HDD:
  - The hot spare replaces the new HDD permanently, and a new hot spare must be configured on the array.
  - Or data from the hot spare is copied to it, and the hot spare returns to an idle state, ready to replace the next failed drive.
- Hot spares should be large enough to host data from the failed drive.
- Some systems use multiple hot spares to improve data availability.
- Hot spares can be automatic or user-initiated.

Images Source: seagate.com



Intelligent Storage Systems



Intelligent Storage Systems are RAID arrays that are:

- Highly optimized for I/O processing
- Have large amounts of cache for improving I/O performance
- Have operating environments that provide:
  - Intelligent cache management
  - Array resource allocation
  - Connectivity for heterogeneous hosts
  - Advanced array-based local and remote replication options

Key Elements:

- Front End + Cache + Backend + Physical Disks

Intelligent Storage Systems

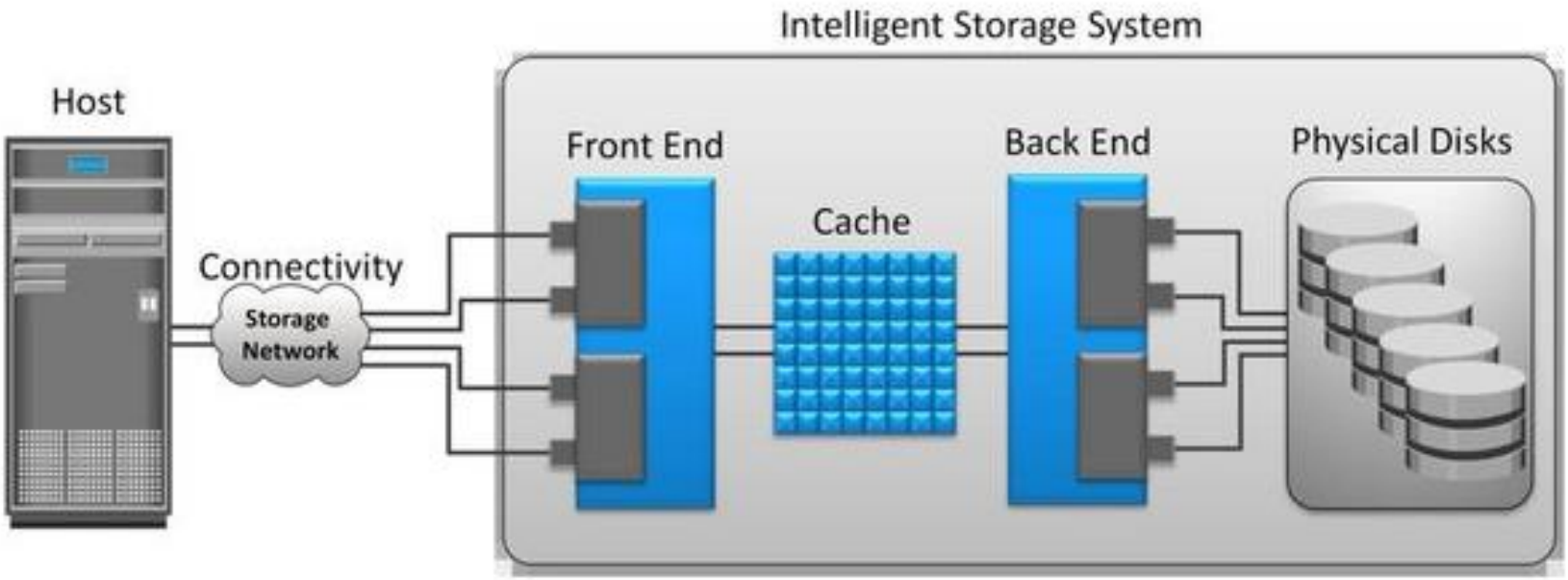


Image Source: emc.com

Interesting video materials:

- Trainsignal, brief: https://www.youtube.com/watch?v=6rK8Nh_4y-U
- EMC, long but good: https://www.youtube.com/watch?v=xOk9-ZwW6-o

Intelligent Storage Systems
Front-end

- Front-end ports + front-end controllers
- Interface between the storage system and the host
- Front-end **ports** enable hosts to connect to the system and have processing logic that executes the appropriate transport protocol (e.g., SCSI, Fibre Channel, iSCSI) for storage connections
- Front-end controllers route data to and from the cache via the internal data bus. When the cache receives write data, controller acknowledges

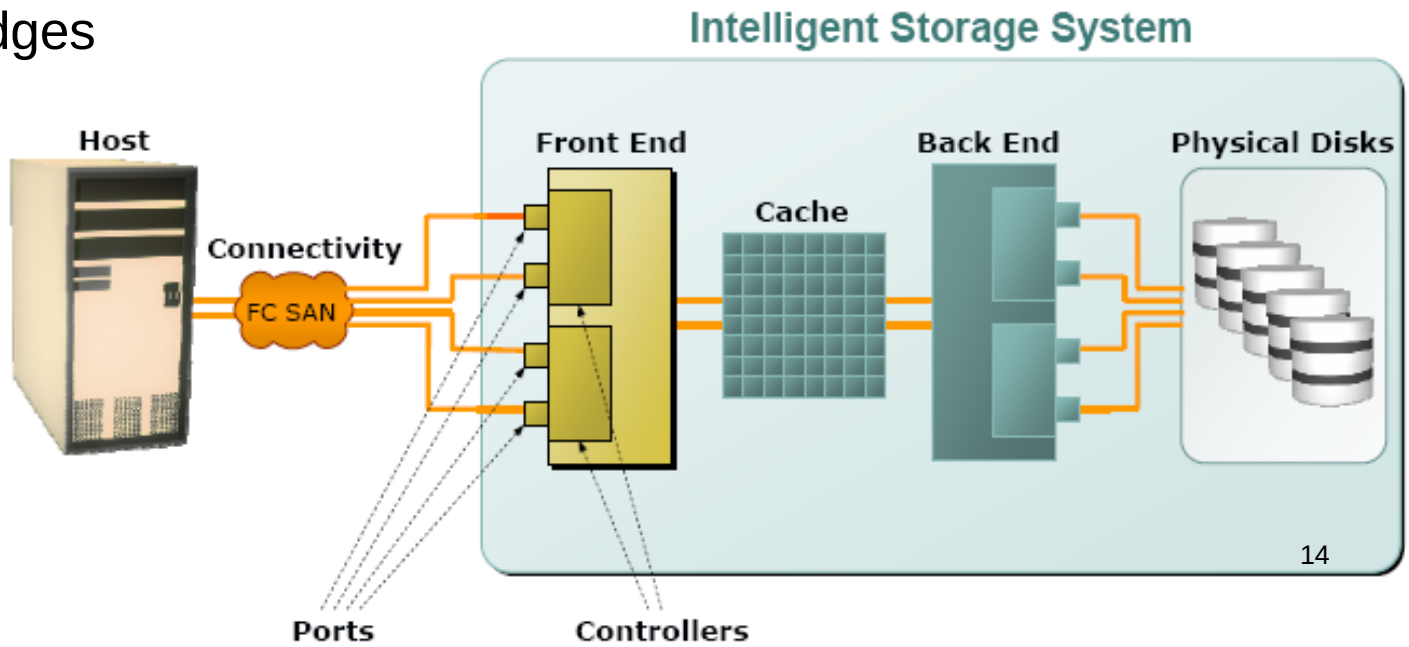
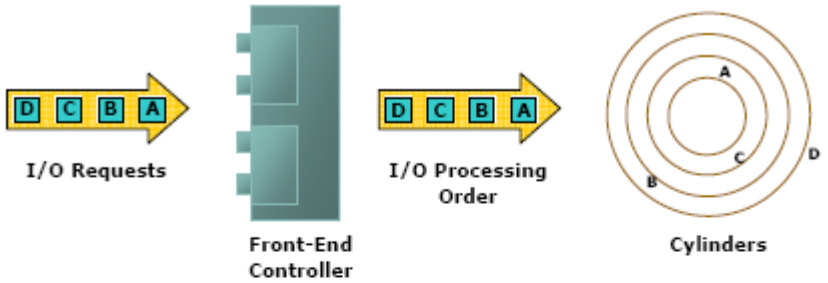


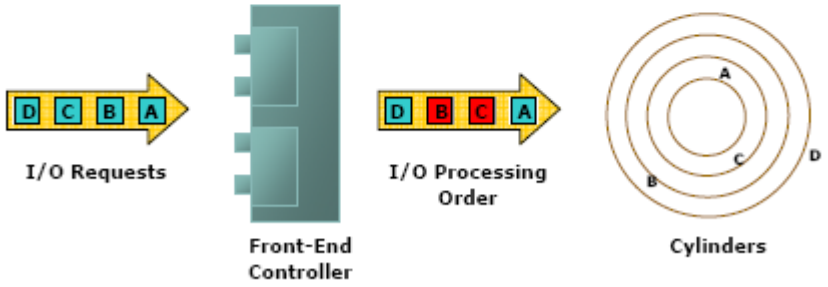
Image Source: emc.com

Intelligent Storage Systems
Front-end

- **Controllers** optimize I/O processing by using ***command queuing*** algorithms, a technique implemented on front-end controllers
- ***Command queuing*** determines the order of execution of received commands to reduce drive head movements and improve disk performance



Without Optimization (FIFO)



With command queuing

Image Source: emc.com

Intelligent Storage Systems
Cache

- Enhances the I/O performance by isolating hosts from the mechanical delays associated with physical disks (slowest components of an intelligent storage system).
- Accessing data from a physical disk usually takes a few milliseconds, Accessing data from cache takes less than a millisecond.
- Write data is placed in cache and then written to disk.

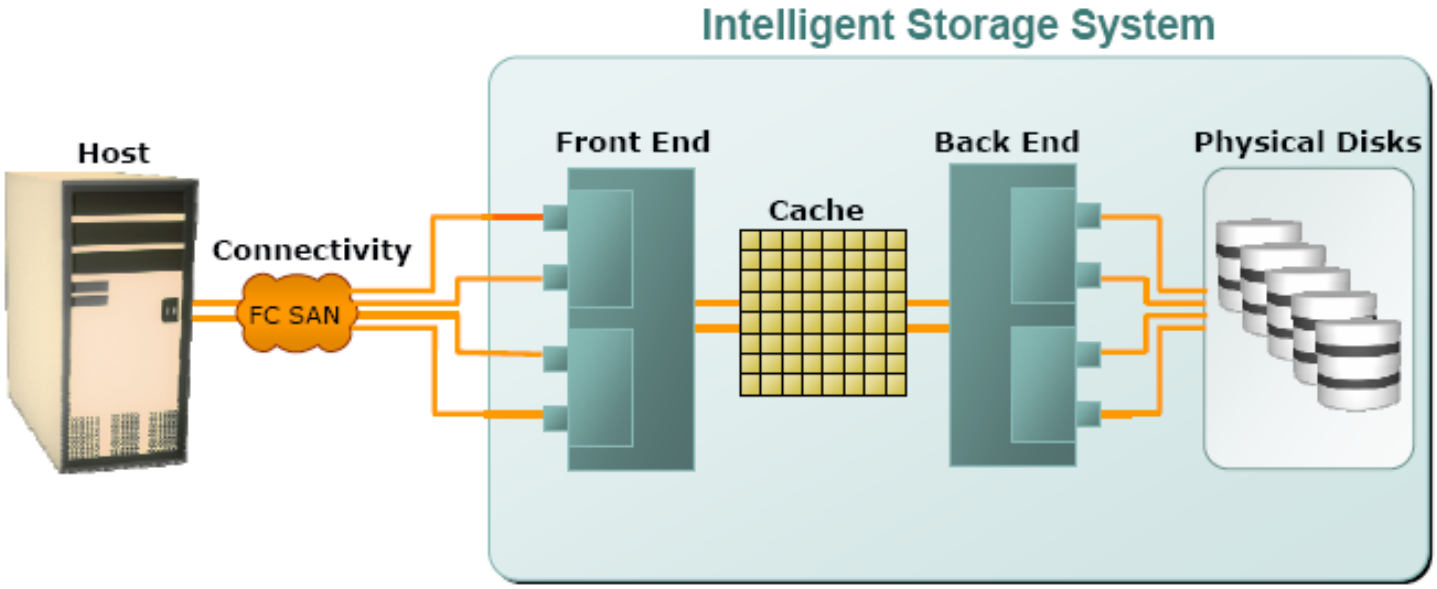
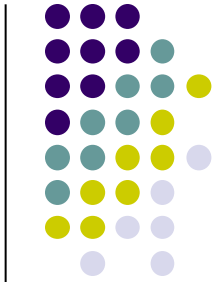


Image Source: emc.com

Intelligent Storage Systems
Protecting cache data



Cache failure on READ: no problem (replica on disk)

Cache failure on WRITE: risk of data loss!!!

Cache mirroring:

- Each write to cache is held in 2 different memory locations on two independent memory cards (introduces coherency problems, though)

Cache vaulting:

- Cache is exposed to the risk of uncommitted data loss due to power failure

Use battery power to write the cache content to the disk (use a set of physical disks to dump the contents of cache on power failure)

Intelligent Storage Systems
Backend

- Back-end ports + back-end controllers
- Physical disks connect to ports on the back end
- The back-end controller communicates with the disks for reads and writes. It also provides additional (limited) temporary data storage.
- Algorithms implemented on back-end controllers provide error detection and correction, along with RAID functionality.
- Multiple controllers and dual-ported disks may be used for resilience and load balancing

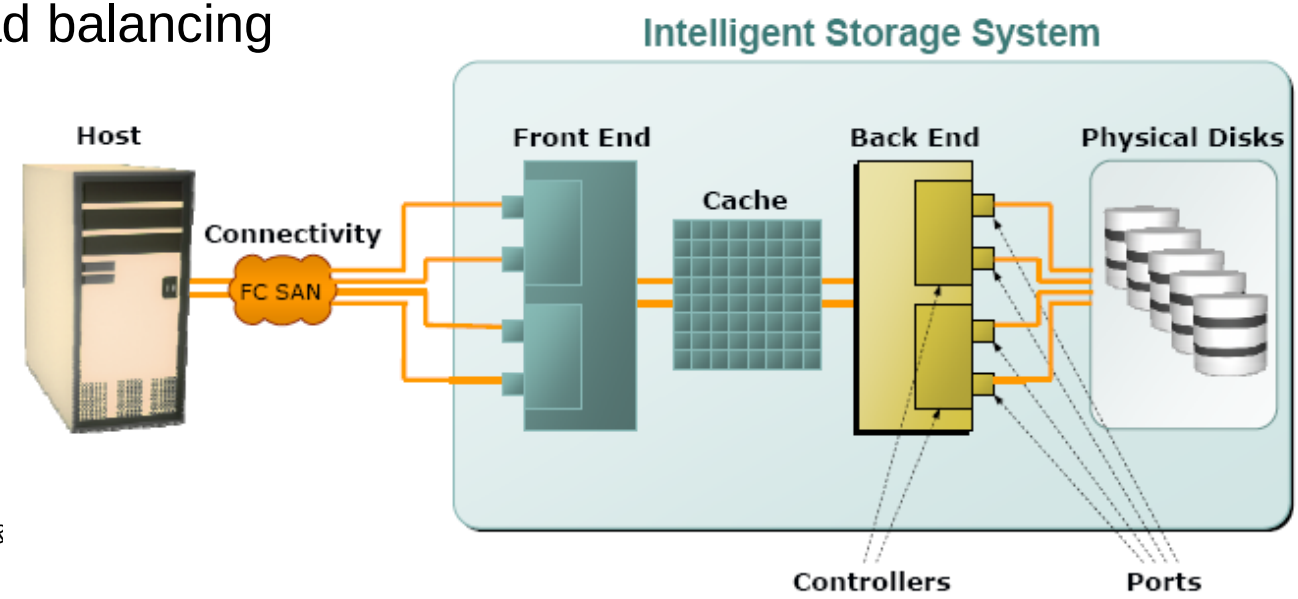


Image Source: emc.com

Intelligent Storage Systems
Physical Disks

- Disks connected to the back-end
- Typical interfaces: SCSI, Fibre Channel, SATA

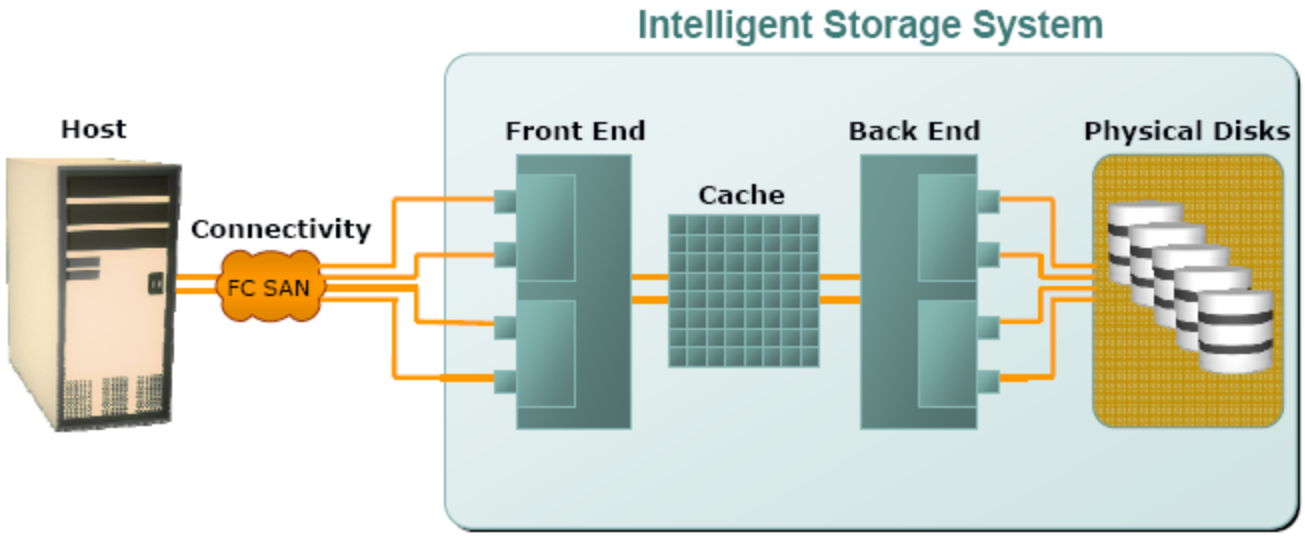


Image Source:
emc.com

LUNs – Logical Unit Numbers

- Physical drives or groups of RAID-protected drives can be logically split into volumes known as logical volumes.
- A LUN identifies a *logical unit* or *logical volume*

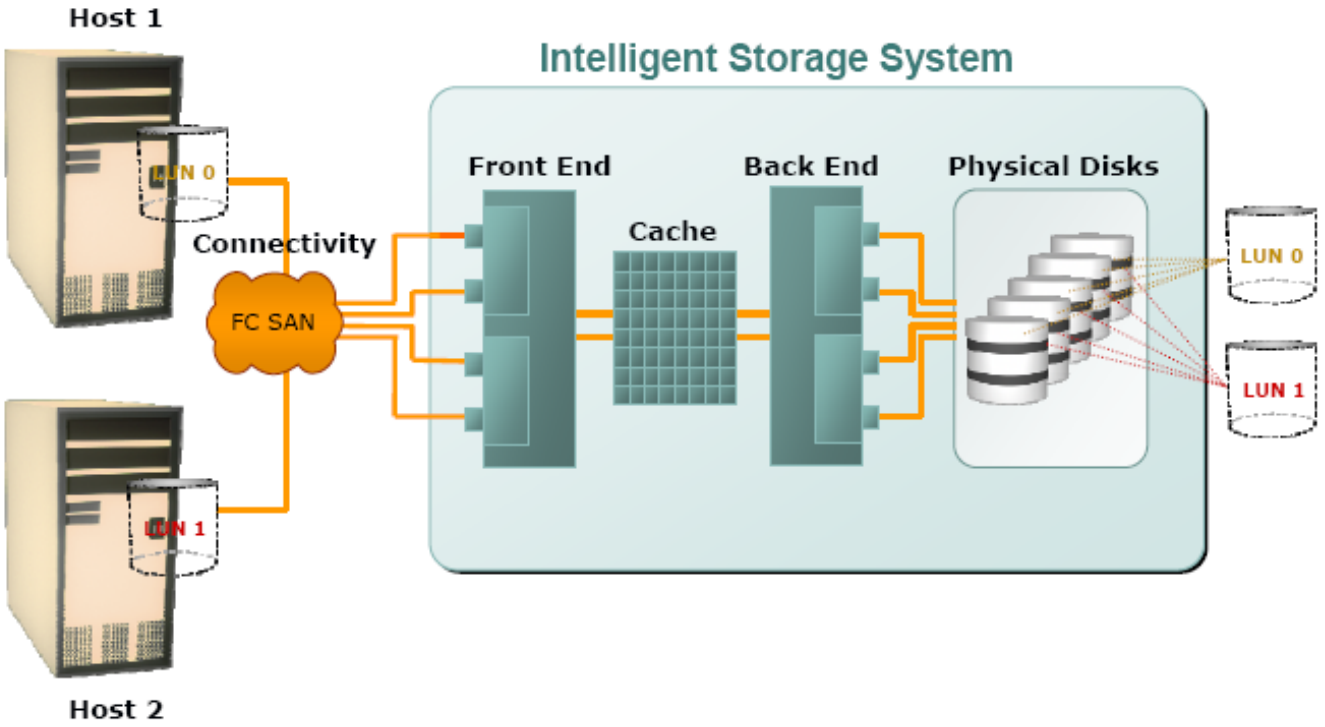


Image Source: emc.com

DAS – Directly Attached Storage

"Digital storage system directly attached to a server or workstation, without a storage network in between" [Wikipedia]

Applications access directly, using **block-level access protocols**

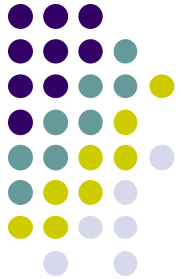
May be internal:

- Laptop, desktop, or server-internal disk
  (limited capacity, limited space available)

or external DAS (but still directly attached):

- Storage unit (possibly) shared by servers on the same rack
  - http://www.dell.com/us/business/p/direct-attached-storage?~ck=bt
  - Typical communications protocols: SCSI, Fibre Channel (FC)
- Less distance and device count limitations
- Centralized management of storage devices

DAS – connectivity & management



Internal DAS:

- ATA/IDE, SATA, parallel SCSI

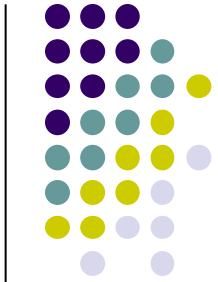
External DAS:

- Serial SCSI (parallel SCSI possible for short distances)
- Fibre Channel (FC)

Management (LUN creation, file system layout, data addressing…)

- Internal:
  - Disk partitioning (Volume management), file system layout
- External:
  - Disk-array based management

DAS Benefits & Limitations



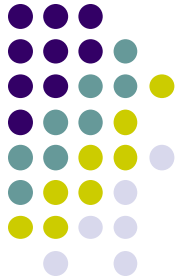
Benefits:

- OK for local data provisioning
- Quick and simple deployment and management (on small environments)
- Relatively low CAPEX

Limitations:

- Limited scalability
  - Number of connectivity ports to hosts,
  - Number of addressable disks,
  - Distance limits
- Maintenance requires downtime (for internal DAS)
- Limited ability to share resources:
  - Array front-end port, storage space
  - Islands of over and under-utilized storage pools
  - Each DAS is shared across a limited number of servers, at most

How to overcome DAS limitations?



Need to overcome limited scalability

- # of connectivity ports to hosts, # of addressable disks, distance limits
- Share storage devices across local rack servers
- Improve scale and efficiency of management operations

Networked Storage

- Storage Area Networks (SAN)
  - Fibre-Channel (FC) SAN
  - IP-based SAN (e.g., iSCSI)
- Network Attached Storage (NAS)

SAN (Storage Area Network)

- Dedicated high-speed network of servers and shared storage devices
- Provides block-level data access (as in DAS)
- Consolidate storage resources (less distance restrictions, more devices)
- Higher scalability (hundreds or thousands of devices)
- Popular SAN technologies: fiber channel, iSCSI

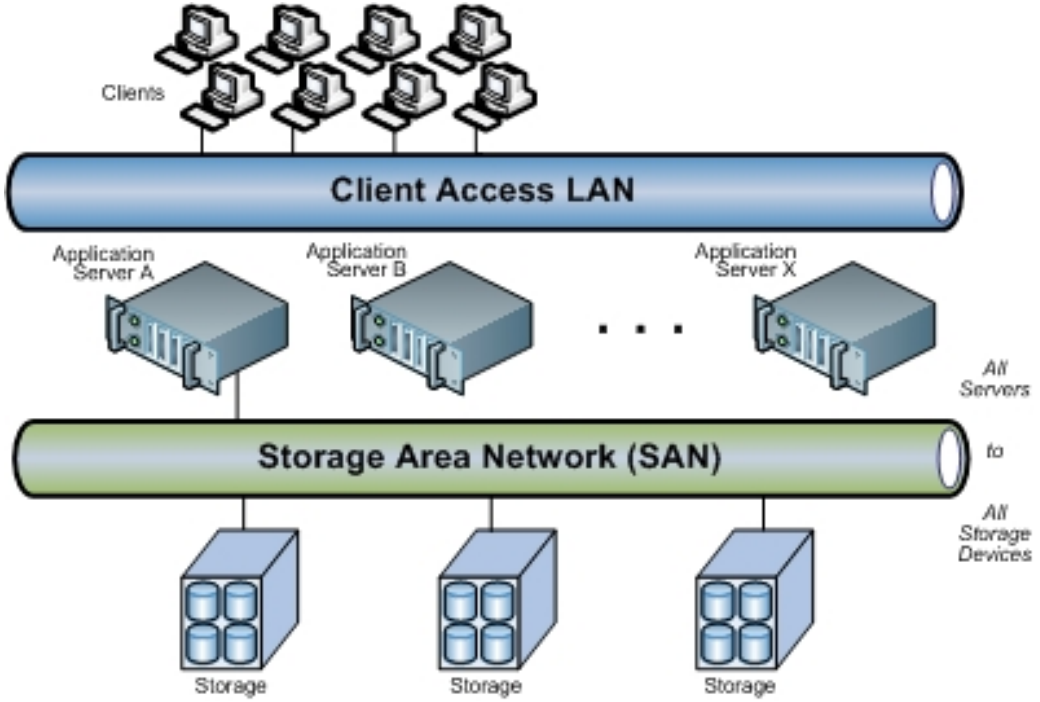


Image source: http://www.snia.org/education/storage_networking_primer/san/what_san

[fibre channel]

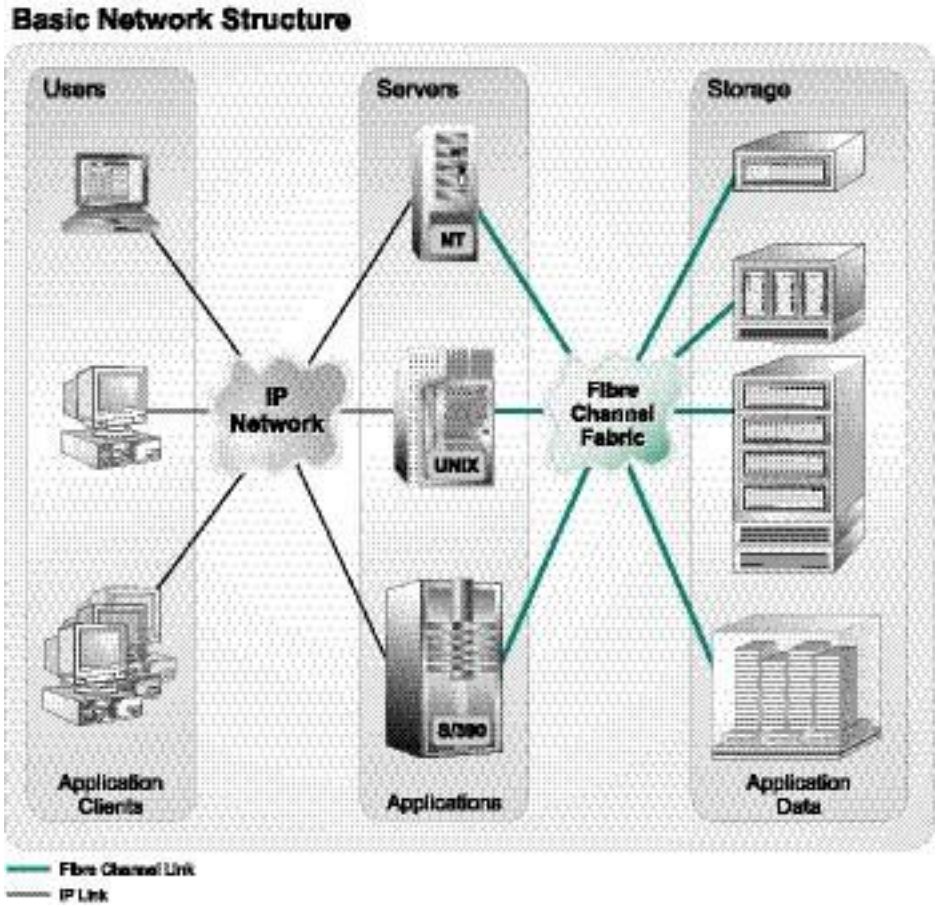
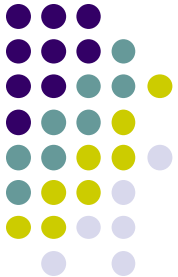


Image source & a lot of further reading materials:
Fibre Channel Industry Association *http://fibrechannel.org/*
Prefer video sources? *https://www.youtube.com/watch?v=mHLkD5yTujo*

[fibre channel "native" variants]

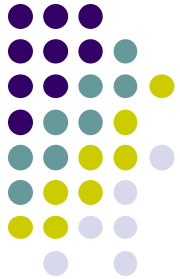


NAME	Throughput (full duplex; MB/s)*	Line Rate (Gbaud)	Market Availability
1GFC	200	1.0625	1997
2GFC	400	2.125	2001
4GFC	800	4.25	2004
8GFC	1,600	8.5	2005
16GFC	3,200	14.025	2011
32GFC	6,400	28.05	2016
128GFC	25,600	4x28.05	2016
256GFC	51,200	4x57.8	2019
…			

Careful when comparing with Ethernet: full duplex; MB/s instead of Mbps)
http://fibrechannel.org/roadmap.html

Only discussed at high level

How About NAS?



Your old-style LAN file-sharing taken to the next level

- ~~Ethernet or~~ IP-based sharing of storage resources
- **File-level data access** and sharing (the key distinction from SAN)
  - Protocols such as NFS or CIFS/SMB
- Storage servers (low-cost, custom-built) or specialized NAS devices

- Typically lower costs, when compared to SAN
  - Scale economies, usage of lower-cost PC and Ethernet components
- Scalable and highly available (when using clustering), up to SAN levels
- May integrate higher-level security (user-based AAA) – **is this an advantage?**
- Centralized storage and management (up to SAN levels)
- (in theory) Higher overheads than SAN
  - Ethernet and IP features not necessary for storage applications
  - File-level access instead of block-level access

NAS key components

File sharing protocols + IP network + NAS devices

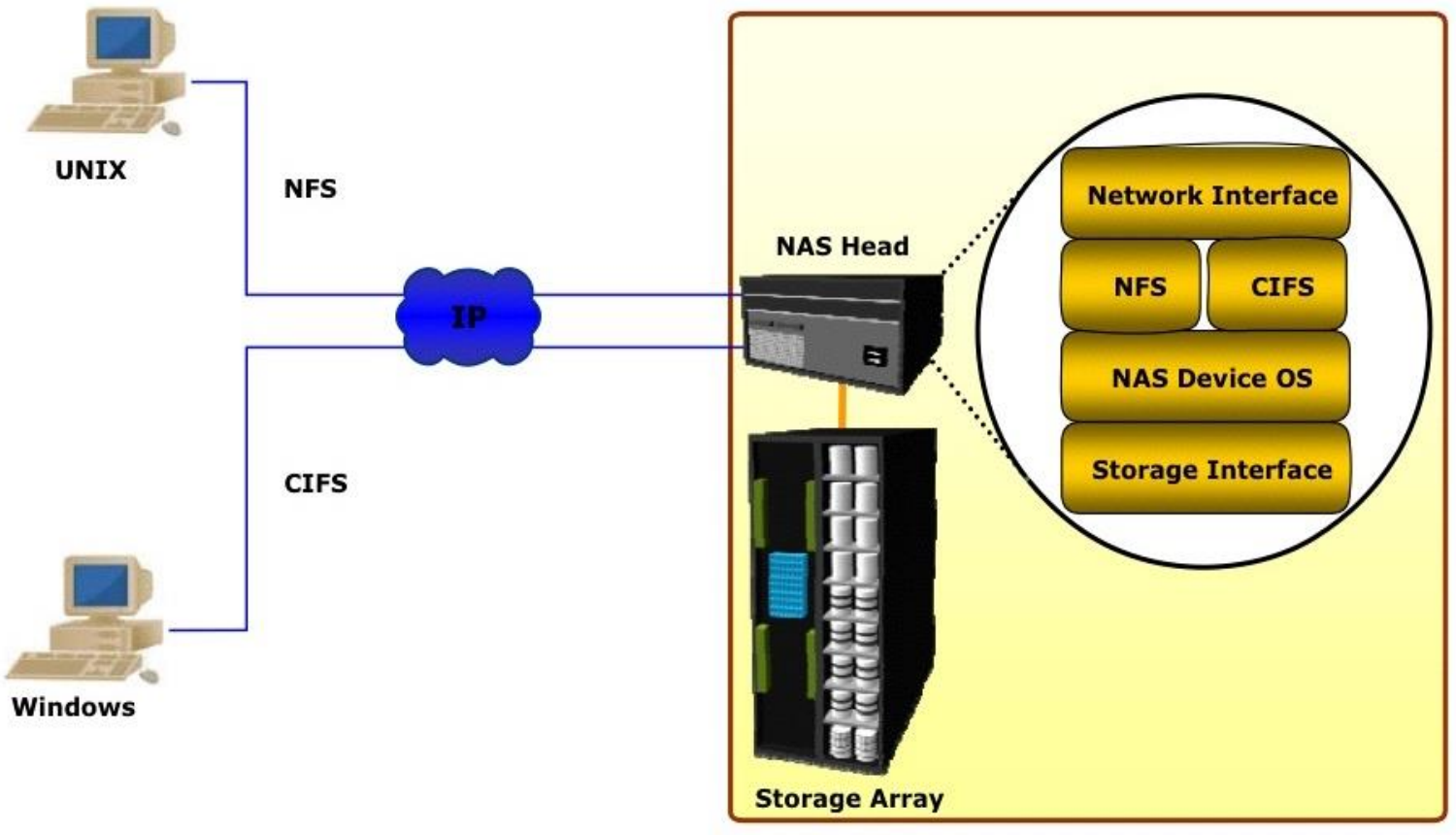


Image Source: emc.com

IP-based SAN

Mix advantages from SAN and IP-based NAS

- Implement SAN infrastructures on top of lower-cost Ethernet/IP networks
  - iSCSI (Internet Small Computer System Interface)
  - FCoE (Fibre Channel over Ethernet)
  - FCIP (Fibre Channel over IP)
- Replace file-level data access with block-level data access (like FC)

- Possible overhead penalties, compared with FC, can be potentially offset by lower cost, higher speed general-purpose networking equipment.

iSCSI – SCSI over IP

- IP encapsulation at host/server:
  - Traditional Ethernet NIC + software
  - Or… specialized Ethernet NIC with iSCSI initiator
- The same applies on the storage side, though professional storage array systems typically use specialized hardware
- Carries block-level messages over IP (TCP/IP packets)
- Gateways to FC are possible

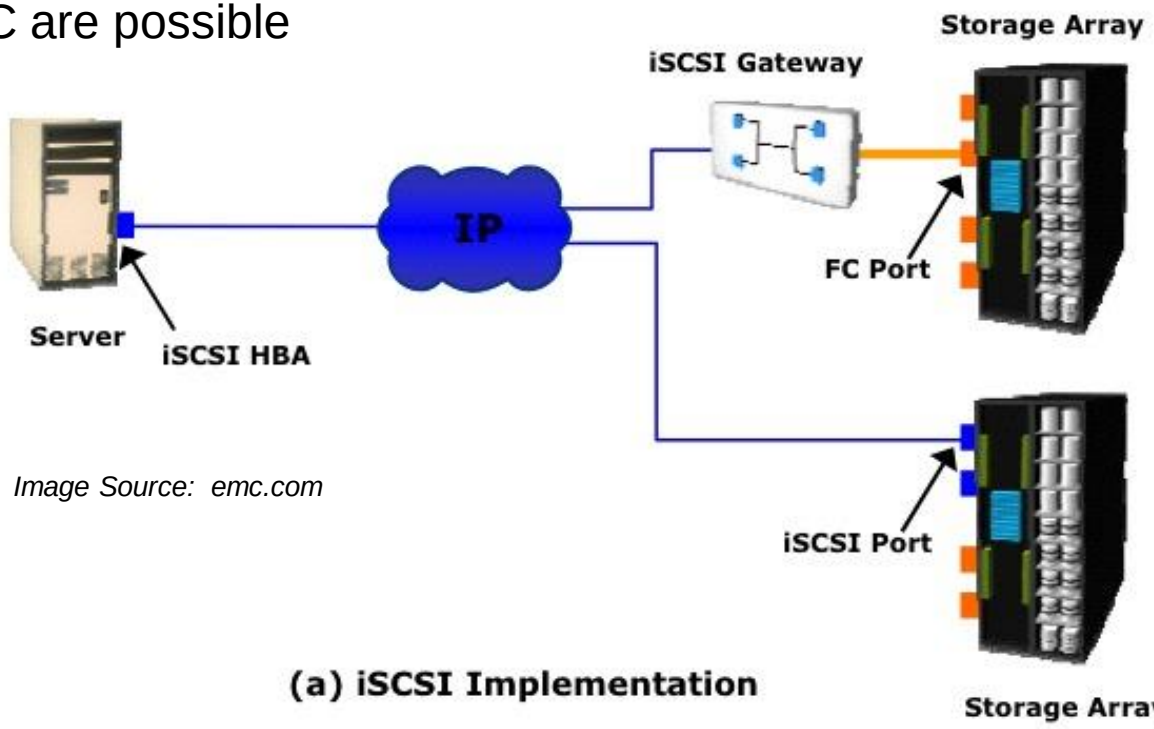
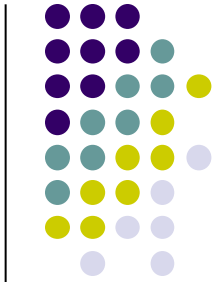


Image Source: emc.com

(a) iSCSI Implementation

iSCSI – key components



iSCSI host initiators

- Host computer using a NIC or iSCSI HBA to connect to storage
- iSCSI initiator software may be required

iSCSI targets

- Storage array with embedded iSCSI capable network port
- Or… FC-iSCSI bridge

LAN for IP storage network

- Interconnected Ethernet switches and/or routers
- General-purpose or iSCSI-optimized equipment
  - http://www.dell.com/us/business/p/powerconnect-5424/pd
  - http://www.computerweekly.com/feature/iSCSI-switch-selection-and-configuration

Credits & Further Reading

Several slides were inspired by/include content from:

- The various links provided in the slides

Information Storage and Management, 2nd Edition, EMC Education Services

- http://books.google.pt/books?id=PU7gkW9ArxIC&printsec=frontcover#v=onepage&q&f=false