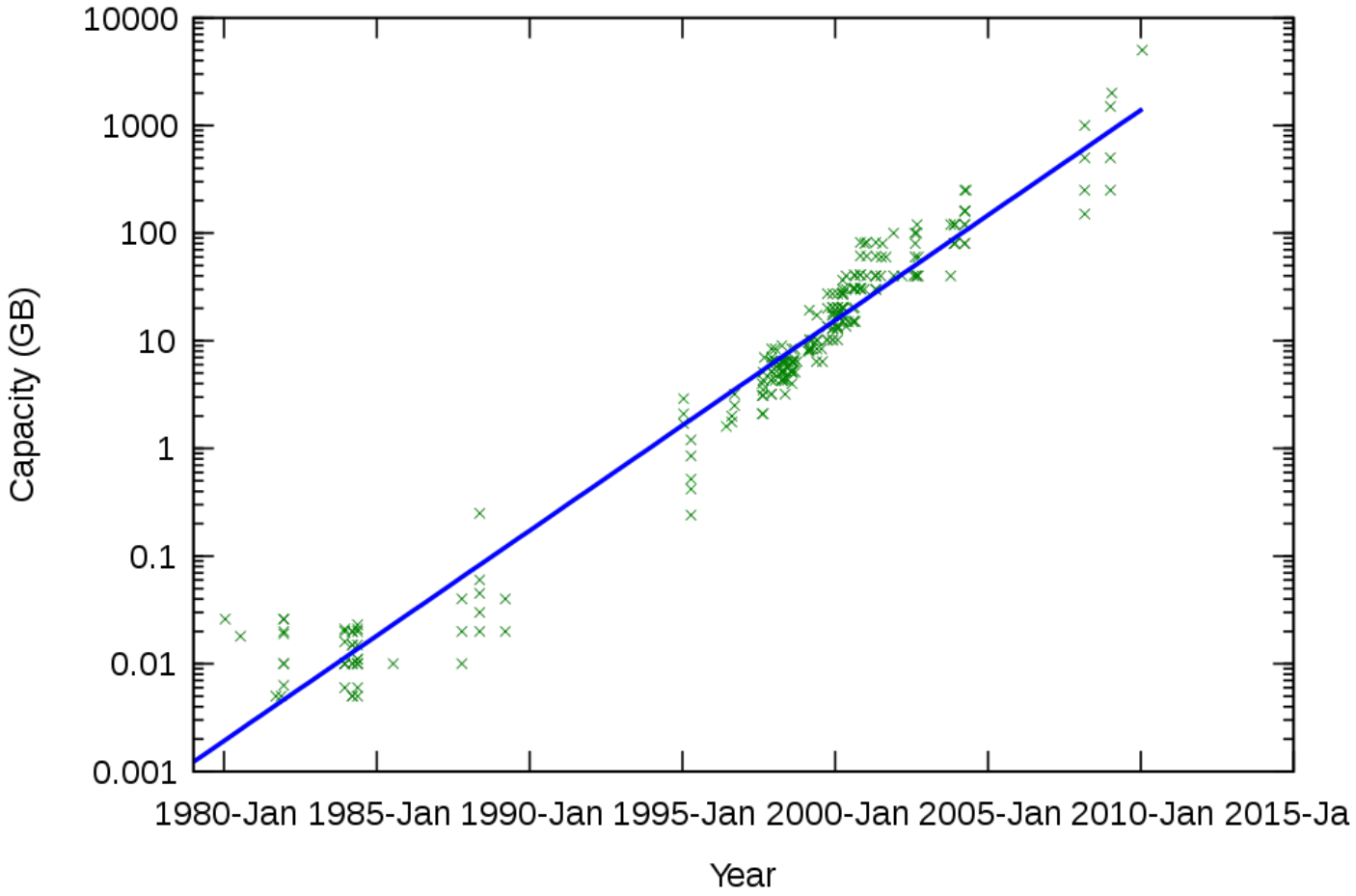
Storage

GIRS 2011

Introduction

- Modern information systems store vast amounts of information:
  - Youtube: 20h of video per minute (2011)
  - Flickr: 38400 photos per hour (2011)
  - Facebook: 1.3 million photos per hour (2009)
- HD capacity increased
  - First 80Gb hard disk (2000)
  - First 3Tb hard disk (2010)
- Prices have dropped
  - $19.70/Gb in 2001, $0.06/Gb in 2010

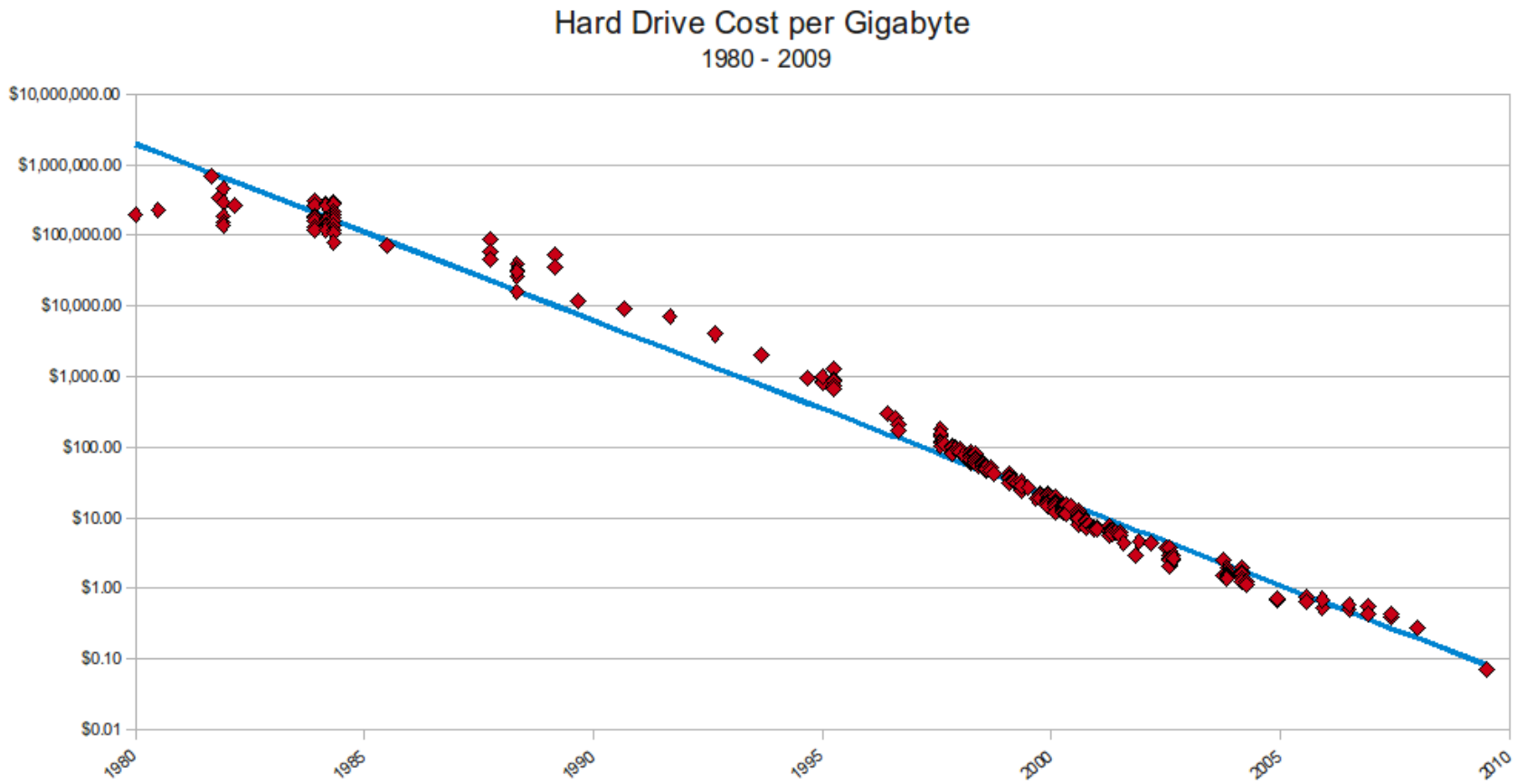
Hard Disk Capacity



http://en.wikipedia.org/wiki/File:Hard_drive_capacity_over_time.svg

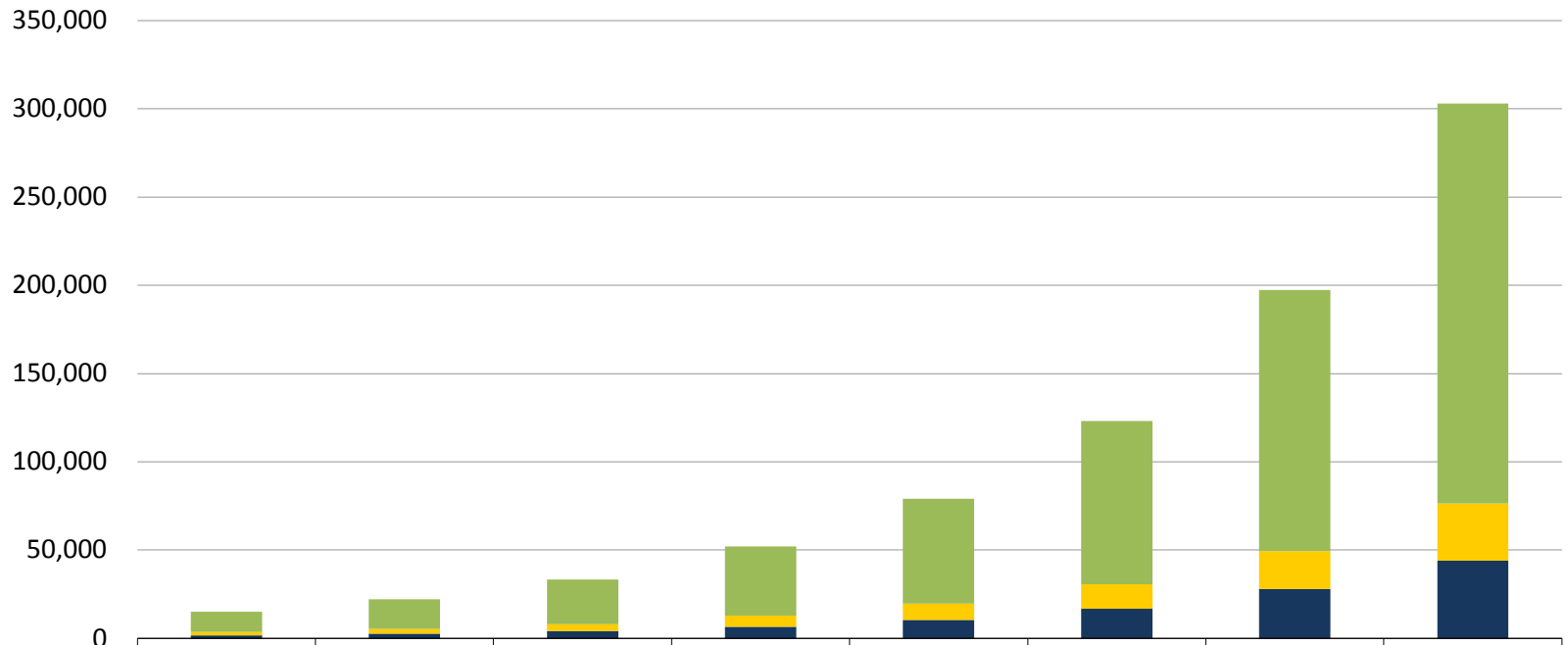
Ryders Law: capacity doubles every 12 months

Hard Disk Price per MB



http://www.mkomo.com/cost-per-gigabyte

Total Archived Capacity, by Content Type, Worldwide, 2008-2015 (Petabytes)



	2008	2009	2010	2011	2012	2013	2014	2015
File	11,430	16,737	25,127	39,237	59,600	92,536	147,885	226,716
Database	1,952	2,782	4,065	6,179	9,140	13,824	21,532	32,188
E-mail	1,652	2,552	4,025	6,575	10,411	16,796	27,817	44,091

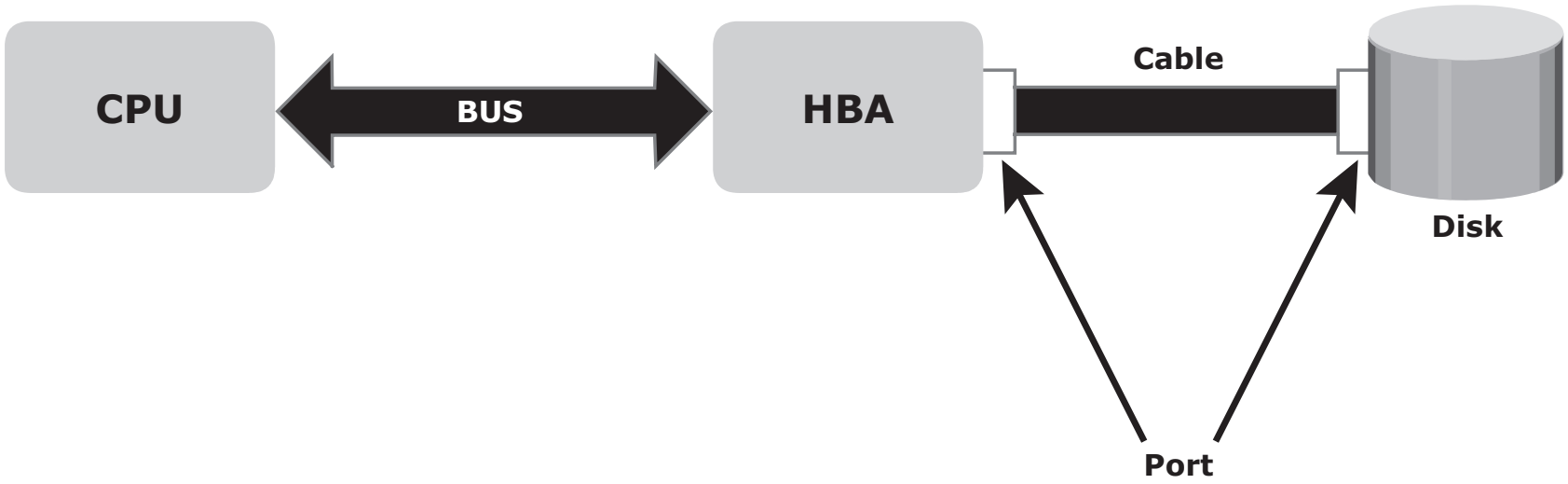
Source: Enterprise Strategy Group, 2010.

Local Storage

- Directly attached to computation resources

- Optical Disks (CD, DVD, Blue Ray)
- Magnetic Mediums (Tape, Hard disk)
- Flash Disks (SSD, Flash Pens)

Local Storage



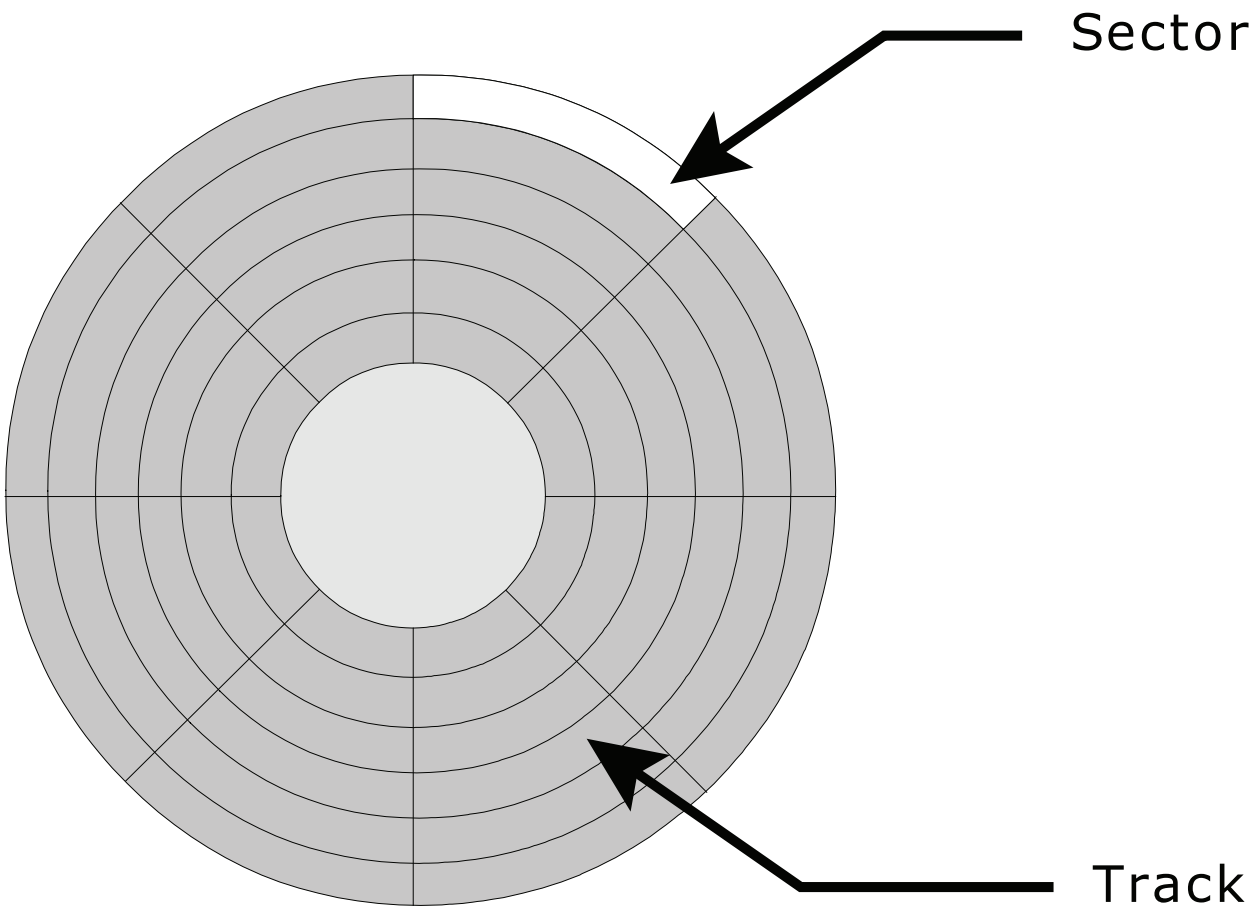
Optical Disks

- Flavours: CD, DVD, UMD, UDO, BluRay,
- Advantages
  - Low cost per GB (~0.05€)
  - Portable
  - Long Life Time (up to 50y)
  - EMI resistant
  - Easily available
- Disadvantages
  - Limited Capacity per Disk
  - Low Access Speed
- Usage
  - Long Term Archive
  - Distribution, Migration

Magnetic Mediums: Hard disks

- Advantages
  - Low cost per GB (0.018€)
  - High access speed
  - Convenient
  - Omnipresent
- Disadvantages
  - Limited capacity per device
  - Not that portable
  - Moving parts
  - Sensible to temperature and shock
- Usage
  - Common storage

Magnetic Mediums: Hard disks



Magnetic Mediums: Hard disks

- Rotational medias imply variable performance
  - Limit Bandwidth and IOPS
- Seek time: Time to move head across different tracks
  - Full Seek, Average Seek, Track to Track
  - Average between 3 and 15ms
- Rational Latency: Head must wait for cylinder to rotate until sector is available
  - 5.4K,7.2K,10K and 10K RPM
  - Between 5.5 and 2ms
- Bandwidth: Transfer speed after sector is located
  - Limited by disk hardware, transport bus, CPU, RAM, driver, etc..

- Rs = Seek + Rotational + Transfer
- Disk suitability to application depends on BW and Latency
  - Single user environment favors bandwidth
    - Copy single file, read single file
  - Multi user environment favors IOPs
    - Same as web servers and databases

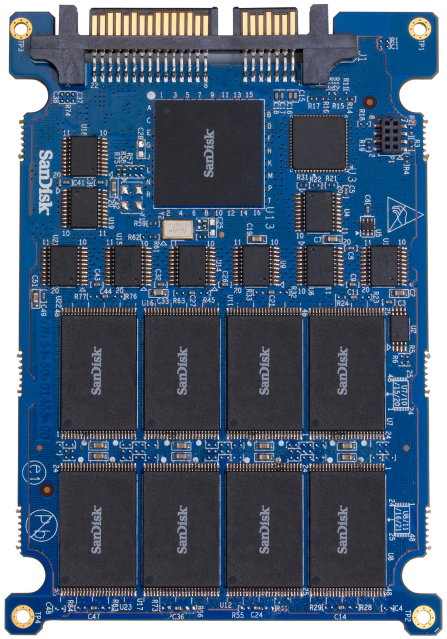
BLOCK SIZE	RS = E+L+X	IOPS = 1/RS
4 KB	5 ms + (0.5 / 15,000 rpm) + 4K / 40MB = 5 + 2 + 0.1 = 7.1	140
8 KB	5 ms + (0.5 / 15,000 rpm) + 8K / 40MB = 5 + 2 + 0.2 = 7.2	139
16 KB	5 ms + (0.5 / 15,000 rpm) + 16K / 40MB = 5 + 2 + 0.4 = 7.4	135
32 KB	5 ms + (0.5 / 15,000 rpm) + 32K / 40MB = 5 + 2 + 0.8 = 7.8	128
64 KB	5 ms + (0.5 / 15,000 rpm) + 64K / 40MB = 5 + 2+ 1.6 = 8.6	116

Magnetic Mediums: Tapes

- Advantages
  - Excellent storage capabilities
  - High Capacity (>3TB)
  - Low cost per GB (~0.018€)
- Disadvantages
  - Very high seek time
  - Unpractical to handle
- Usage
  - Short/Medium/Long Term Archive
  - Distribution, Migration

Flash Disks

- Advantages
  - No moving parts
  - Low seek time
  - High Performance
- Disadvantages
  - Limited storage capacity
  - High cost per GB (7.5€)
  - Limited number of writes.
    - Require Wear-Leveling
- Usage
  - Data Bases
  - Video Processing
  - Data acquisition (Real time/High Speed)

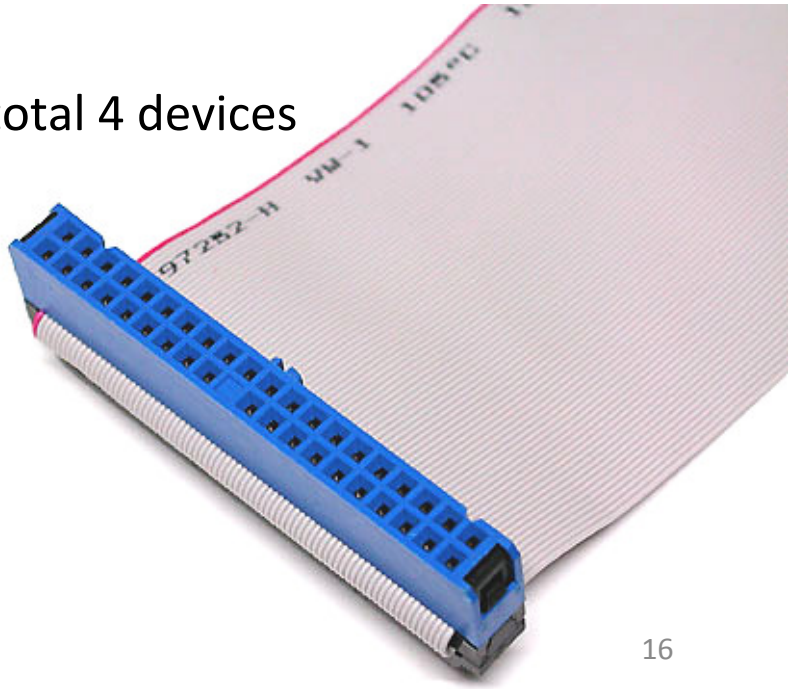


Storage Access Protocols

- Local
  - PATA
  - SCSI
  - SATA
  - SAS
- Network
  - FC
  - iSCSI
  - AoE
  - CIFS
  - NFS

Parallel AT Attachment (PATA)

- Developed by Western Digital in 1986.
- Transfer speed: 16MByte/s – 133MByte/s
- Allows direct block addressing (512bytes)
- Maximum device size: 137GB, 144PB (using ATA-6)
- Two devices per channel
  - Computers usually have two channels: total 4 devices
- Almost obsolete



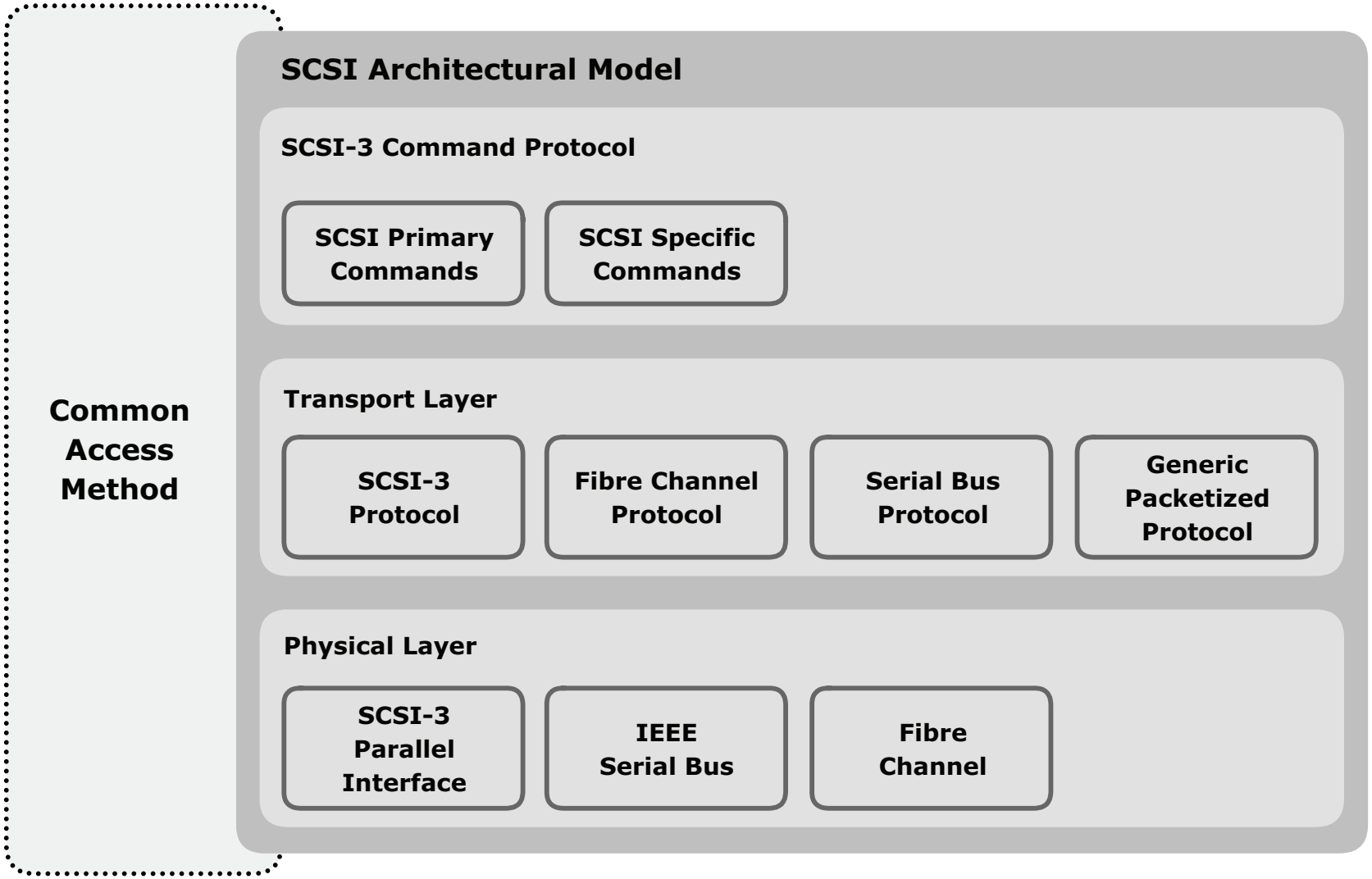
Small Computer System Interface (SCSI)

- Developed by Shugart Associates em 1978
- Transfer speed: 5MB/s to 640MB/s
- Layered protocol: Transport, Link, Phy
- Bus with up to 8 devices
  - Each device has an ID plus a LUN (Logical Unit Number)
  - Two roles: Initiator and Target
    - Most devices act as targets
- SCSI Commands
  - Initiator sends commands to targets
  - Around 60 commands

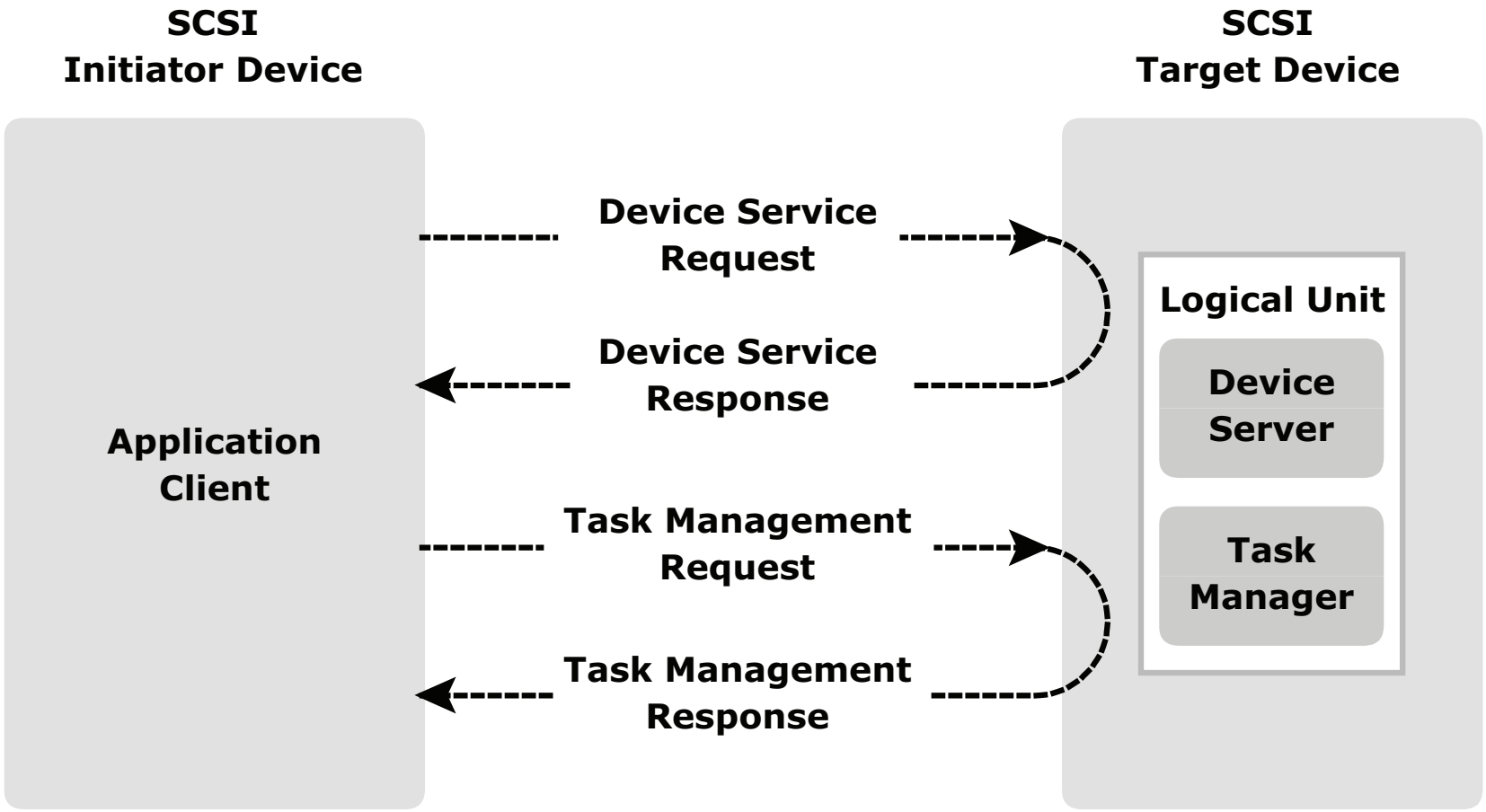
Small Computer System Interface (SCSI)

- Advantages
  - Suports many types of devices: Printers, Scanners, Flash Devices
  - 8 devices per channel
  - Robustness (rich set of commands)
  - Low CPU Usage
- Disadvantages
  - Expensive devices
  - More complex than PATA
  - Somewhat in decline
- Aplicações
  - High performance servers

Small Computer System Interface (SCSI)



Small Computer System Interface (SCSI)



Serial ATA (SATA)

- Developed to replace PATA
  - Cheaper cables, hot-swapping, Higher bandwidth, higher effiency, richer protocol
- Transfer speed: 150MB/s a 600MB/s
- Layered protocol: Transport, Link, Phy
  - Transport: responsible for creation of SCSI frames(control + dados)
  - Link: Packet coding
  - Phy: Connection and maintenance of electrical levels.
- NCQ (Native Command Queueing)
  - Allows to reorder requests to optimize access
    - Reduce Hard Disk head movement
- Power management functions

Serial ATA (SATA)

- Advantages
  - Higher transfer speed
  - Low cost
- Disadvantages
  - Sensible to transmission errors
  - 1 device per channel
  - Serial protocol limits continuous transfer rate
- Usage
  - Desktop's, Laptops
  - Entry level servers and NAS

Serial Attached SCSI (SAS)

- SCSI evolution
  - Better performance, simpler interface
  - Migration from parallel interfaces to serial interfaces
- New PHY
  - No need for resistive terminators
- Unique Universal Addressing (SAS Addresses)
- Connectors are compatible with SATA
- Single controller supports up to 16,384 devices
  - Total bandwidth of 6Gbit/s (750MiB/s)

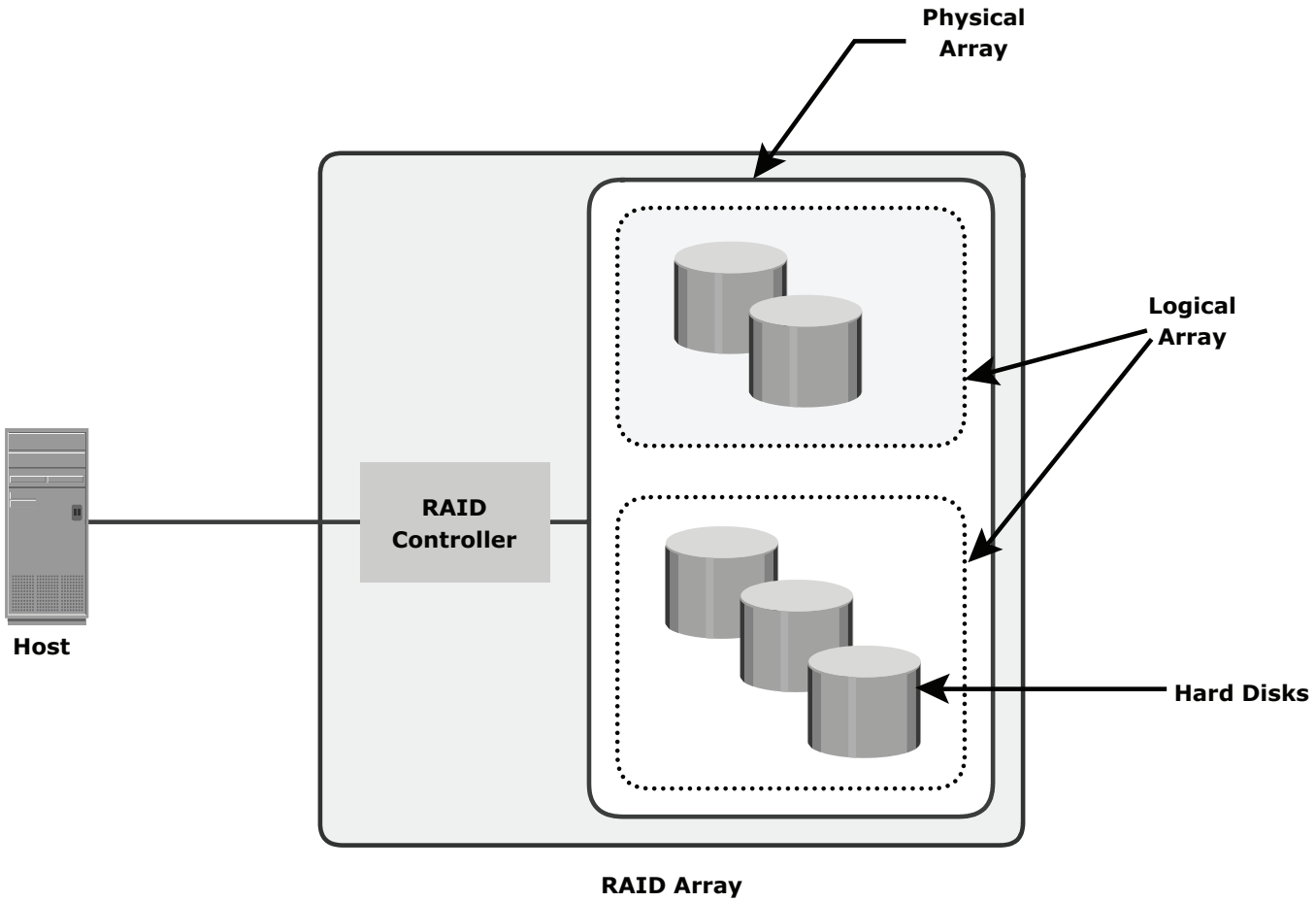
Serial Attached SCSI (SAS)

- Advantages
  - Lower power consumption
  - Higher performance
    - 10K and 15K RPM hard disks
  - Simpler than SCSI
- Disadvantages
  - Higher cost (than SATA)
- Usage
  - Medium to High level servers
  - Medium performance storage environments

Redundant Array of Inexpensive (or Independent) Disks (RAID)

- Array is composed by several Member Disks
  - Only one device is available to the operating system
- Arrays can be tailored towards:
  - Increased performance
  - Increased redundancy
  - Increased capacity
  - Increased error recovery performance

RAID


Physical Array
Logical Array
RAID Controller
Host
Hard Disks
RAID Array

RAID

- Software RAID
  - Use software to group several devices
  - May have a reasonable impact in CPU
  - Requires explicit support by the OS
- Hardware RAID
  - Uses dedicated hardware
  - Can potentially have better performance
    - Without using CPU time
  - Indepdendent of the operating system
    - Except for RAID card drivers

RAID

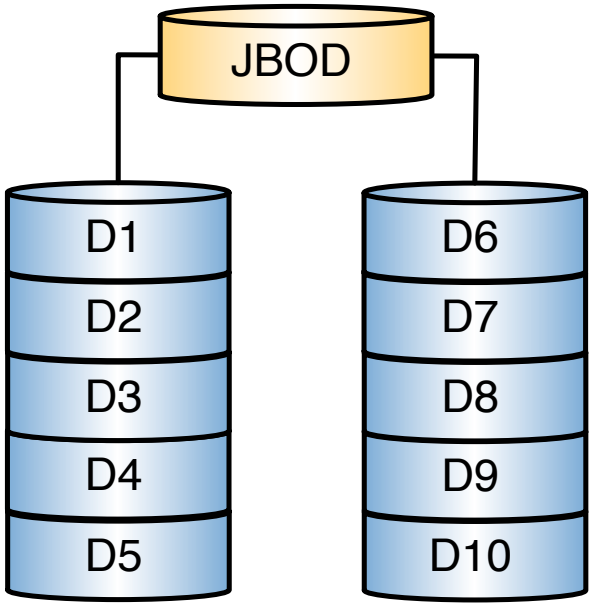
- Advantages
  - Can be very cheap
  - Can be optimized towards several metrics
- Disadvantages
  - “False sense of safety”
  - Duration of fail recovery process
  - Can become very expensive
- Usage
  - Agregation of HD and SSD
  - High performance video storage
  - High capacity, redundant storage

RAID

- RAID can provide redundancy but is not a backup solution!
  - Only one version of data is kept
  - Devices can fail (hard disks)
  - RAID controllers can fail
  - Array logical structure can become corrupt
  - Humans can corrupt data
  - Etc...

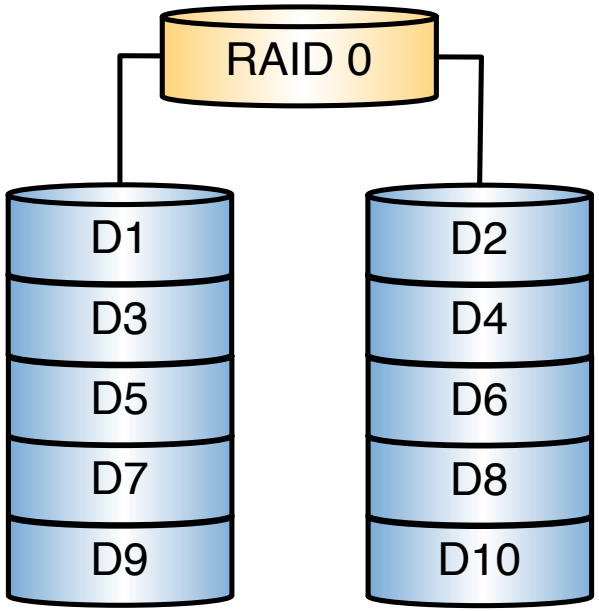
Just a Bunch Of Disks (JBOD)

- Provides an array concatenating all devices
  - Devices may be heterogeneous
- Performance: Pt=Pd x N
  - When accessing sectors of all devices (not sequential access!)
- Failure: none
- Capacity: sum of all devices



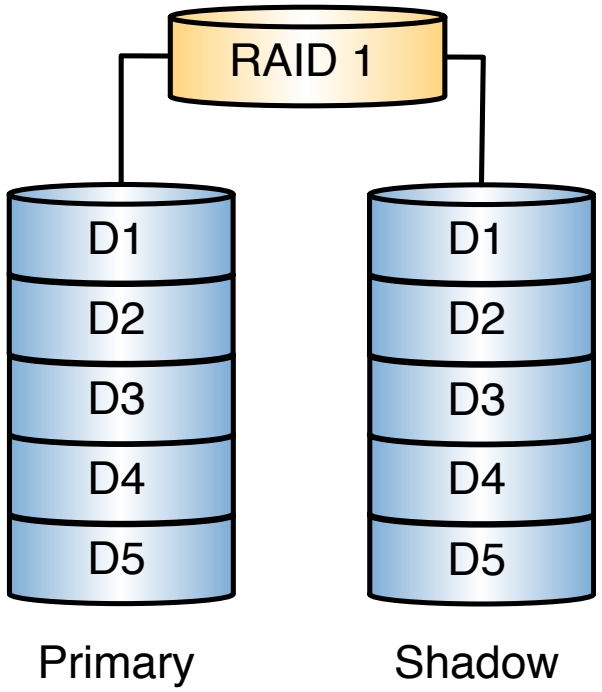
RAID 0 (Stripe Set)

- Provides an array concatenating all devices
  - Devices should have similar capacity
- Performance: N (read and write)
- Failure: none
- Capacity:
  - Homogeneous: N * C
  - Heterogeneous: N * Csmallest



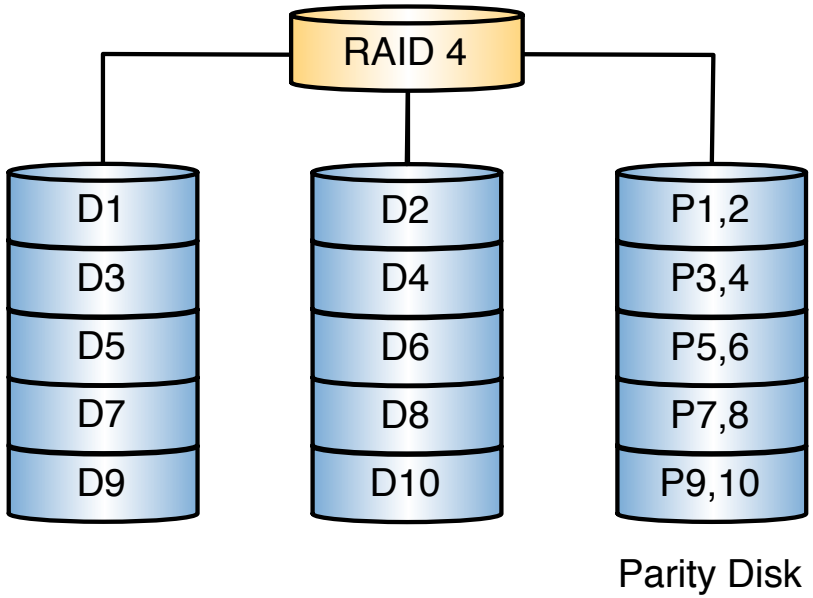
RAID 1 (mirror)

- Considers 1 primary and one or more shadow devices
  - Devices are exact copies
  - Requires min of 2 devices
    - No upper limit
- Performance: <1x write, Nx reads
  - Write implies updating all devices
  - Read can be provided by any device
- Failure: N-1 failures supported
- Capacity: Csmallest



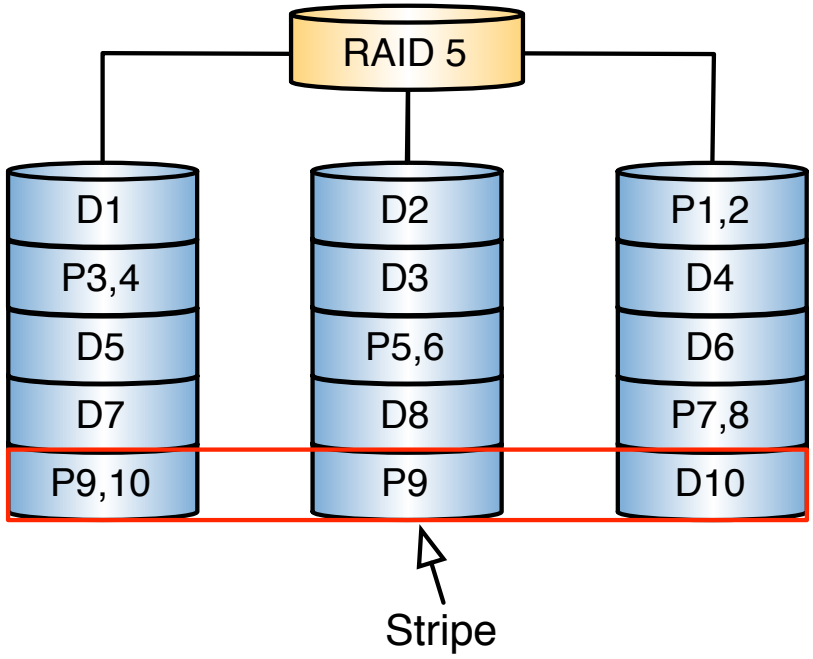
RAID 4 (Independent Data Disks with Shared Parity Disk)

- 3 or more disks
  - N-1 data devices, 1 parity device
  - Parity is Block based
    - RAID 2 is bit based, RAID 3 is byte based
  - High CPU usage
- Performance: <<1x write, Nx read
  - Parity disk is a bottleneck!
  - Especially bad for random writes
- Failure: 1 device
  - Parity can be used to provide missing data
- Capacity: (N-1) x Csmallest



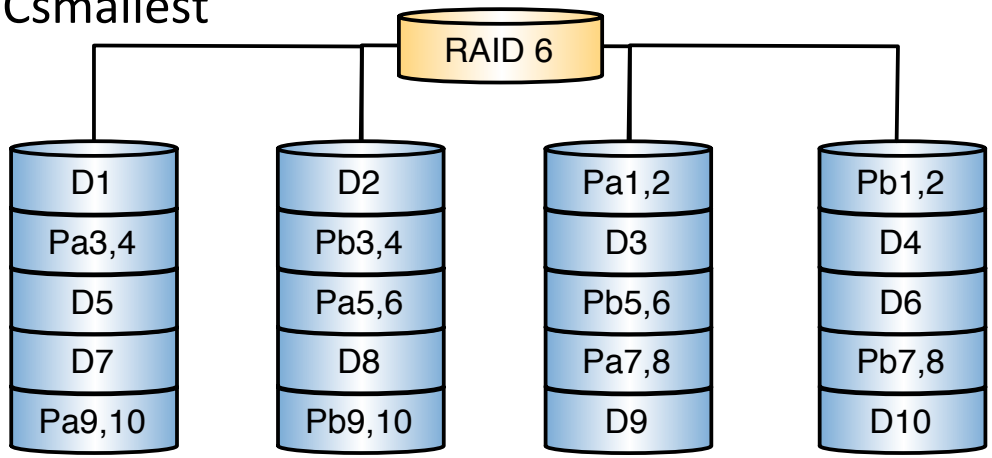
RAID 5 (Stripe Set with Parity)

- Similar to RAID 4 but without dedicated parity disk
  - Parity is split among devices
  - Higher CPU usage (uses XOR)
  - Min 3 disks
- Performance: <<1x write, Nx read
  - WB cache can improve write performance
- Failure: 1 disk
  - Ex: D1 XOR P1,2 recovers D2
- Capacity: (N-1) x Csmallest



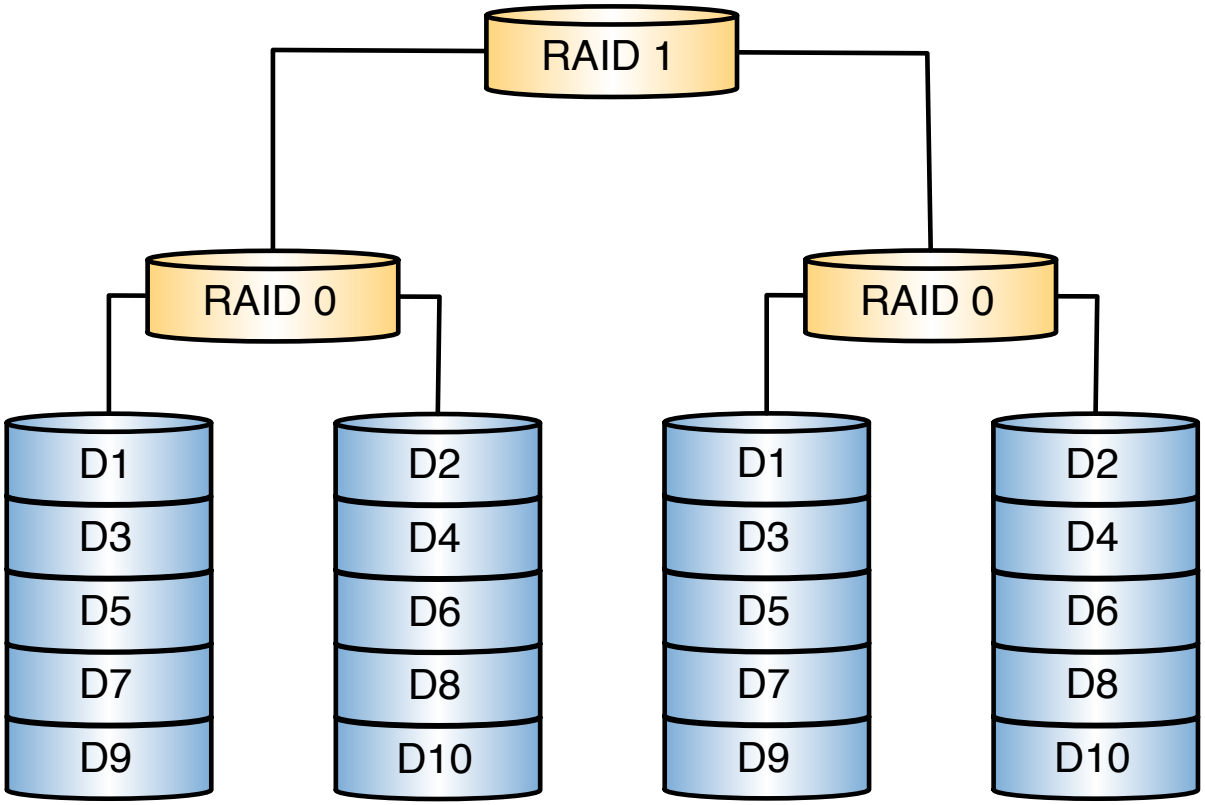
RAID 6 (Stripe Set with 2 Parity Blocks)

- Similar to RAID 5
  - Two parity blocks
  - Min 3 disks
  - Faster to rebuild
  - Even Higher CPU usage (uses 2 parity functions)
- Performance: <<<1x write, Nx read
  - WB cache can improve write performance
- Failure: 1 disk
  - Ex: D1 XOR P1,2 recovers D2
- Capacity: (N-2) x Csmallest



Nested RAID

- Combine several RAID levels
  - Performance, Capacity and Failure support vary with levels



Remote Storage

- Block based
  - Individual blocks made available remotely
  - Allows mapping local device in external hosts
- File based
  - Individual files made available
  - Allows access to files and directories of local filesystem
- Object based
  - No file or block semantics
  - Allows putting getting objects in remote system

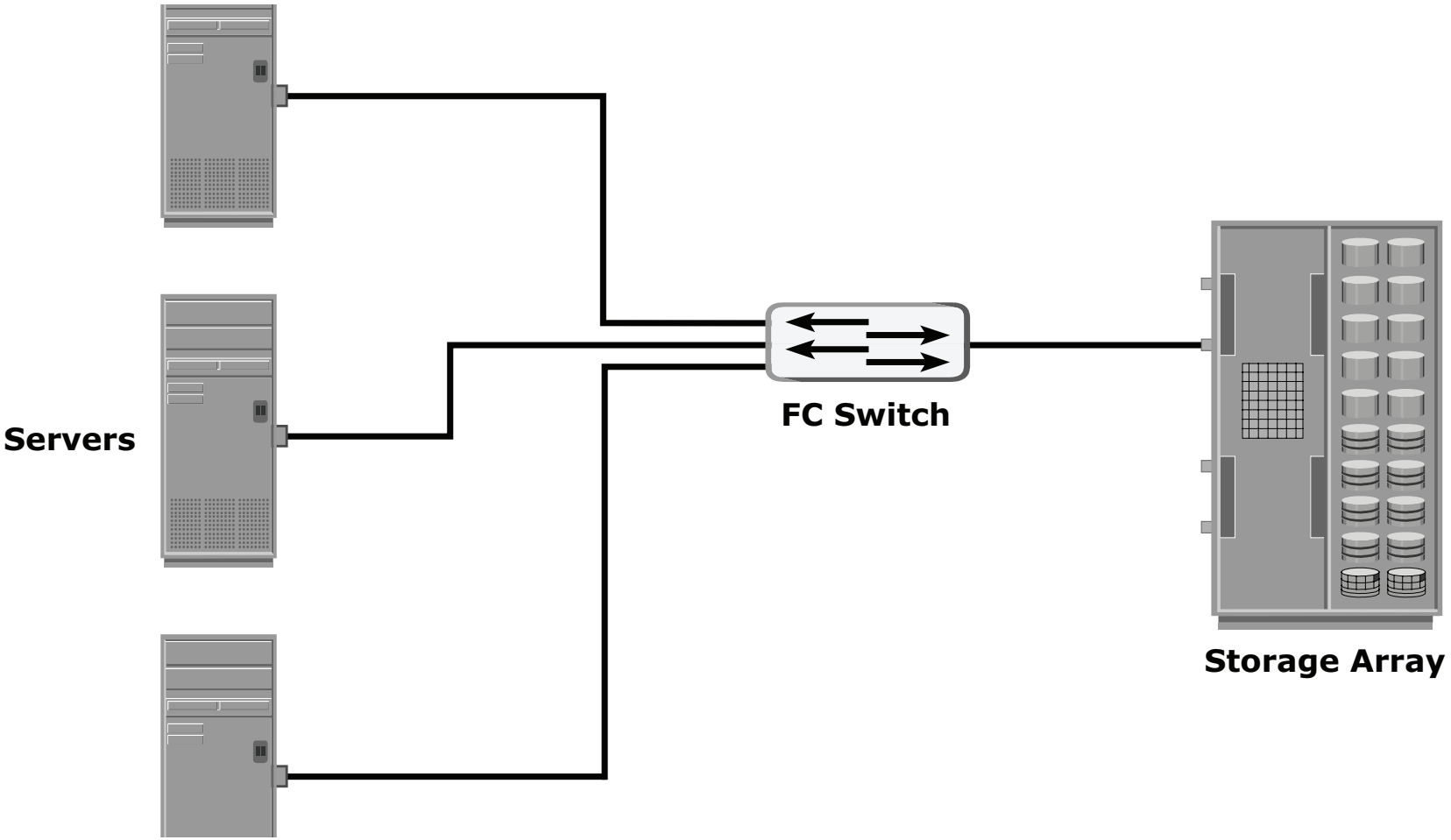
Fiber Channel (FC)

- Developed in 1988
- Can use both Fiber and Copper cables
  - Can be sent over Ethernet (FCoE)
- Uses SCSI transport protocol
- Layered protocol (5 layers)
  - PHY, Data Link, Network Layer, Common Services e Protocol Mapping.
- Unique, 56bytes identifier
- Connection types:
  - Point-to-Point
  - Arbitrated Loop
  - Switched Fabric

Fiber Channel (FC)

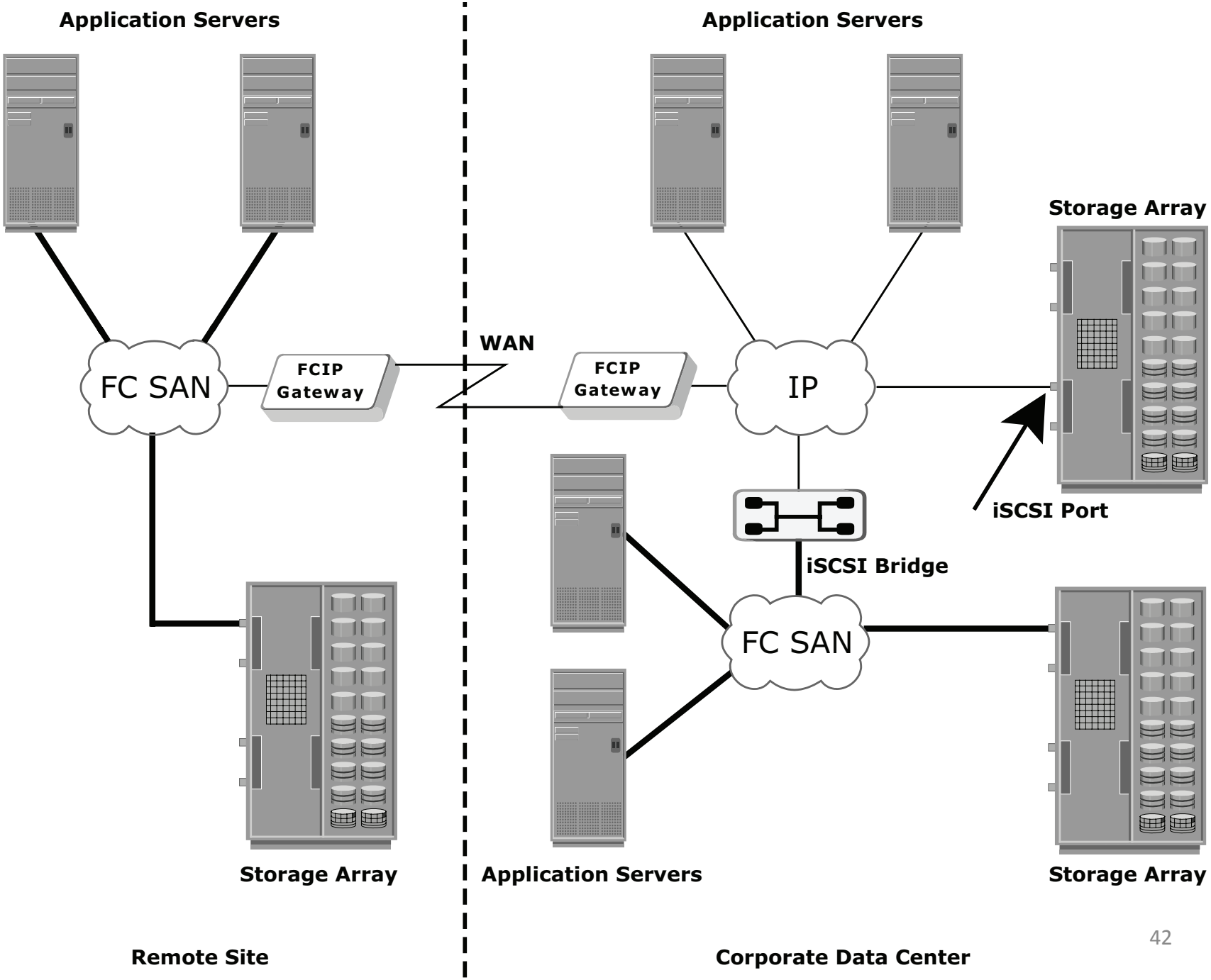
- Advantages:
  - High transfer performance (1.6TB/s)
  - High range (up to 10km)
- Disadvantages
  - High cost
  - Complex management
- Usage
  - Used in higher performance SAN
    - Standard de facto
  - Large DB
  - OffSite storage

Fiber Channel (FC)



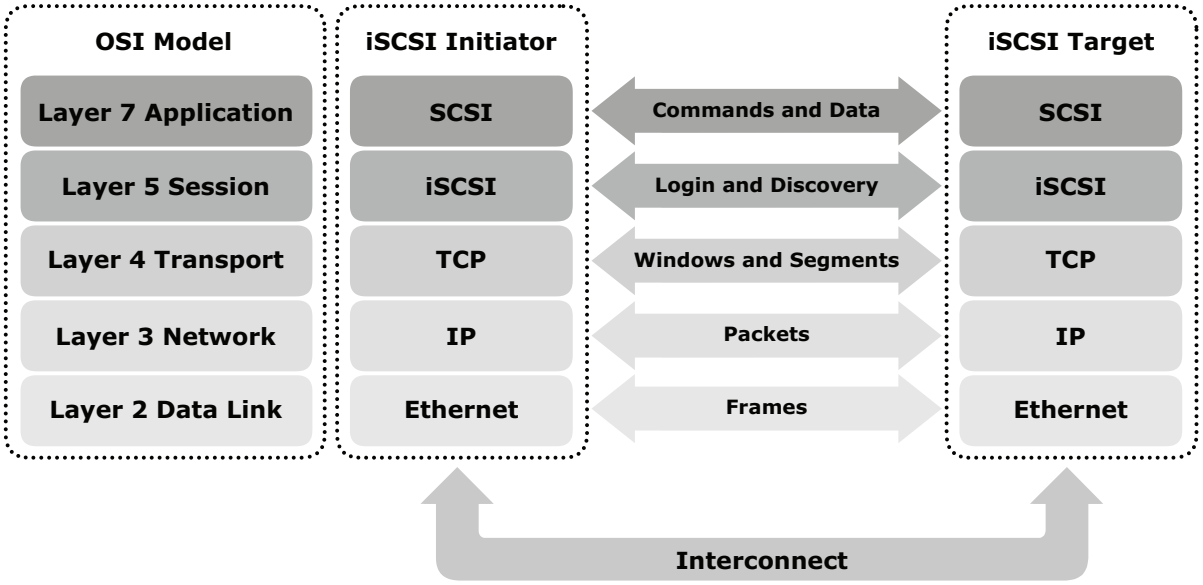
Internet Small Computer System Interface (iSCSI)

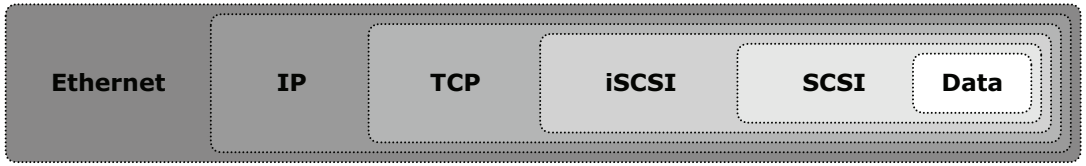
- Transports SCSI protocol over TCP/IP networks
  - Remote devices appear as local SCSI devices
  - Block based
- Initiator: uses a remote device
  - May be sent explicitely to the OS
  - May be masked by hypervisor to guests
- Target: provides devices
  - May be an actual storage or a gateway


Application Servers
Application Servers
Storage Array
FC SAN
FCIP Gateway
WAN
FCIP Gateway
IP
iSCSI Port
iSCSI Bridge
FC SAN
Storage Array
Application Servers
Storage Array
Remote Site
Corporate Data Center

Internet Small Computer System Interface (iSCSI)

- Requires the establishment of iSCSI sessions
  - Can be authenticated with PAP/CHAP
    - Ex: against LDAP directory or internal database





Internet Small Computer System Interface (iSCSI)

- Advantages
  - Uses IP (WAN)
  - Flexible in terms of distance (>100Km)
  - May take advantage of IP multipath
- Disadvantages
  - Lower performance than FC
    - Limited by Etherent/IP/TCP stack
  - Complex management
- Usage
  - Distributed BD
  - OffSite Backups
  - Headless Desktops
  - Virtualized environments

ATA over Ethernet (AoE)

- Similar to iSCSI, but transports ATA over Ethernet
  - Much simpler design.... as ATA
- Advantages
  - Low cost
  - Low overhead (without TCP and IP)
  - Simple to implement
- Disadvantages
  - Limited to ATA devices
  - Limited to the same ethernet segment
  - Lack of error recovery (stateless)
  - Inferior authorization control
- Usage
  - Low end servers

CIFS

- Created for Windows systems
  - Variation from SMB
- Operates over TCP and is stateful
  - Allows session restore (reopen files)
  - If server is stopped, all clients are disconnected
  - Supports file locking
- Provides:
  - File search, open, read, write, close
  - Modify attributes

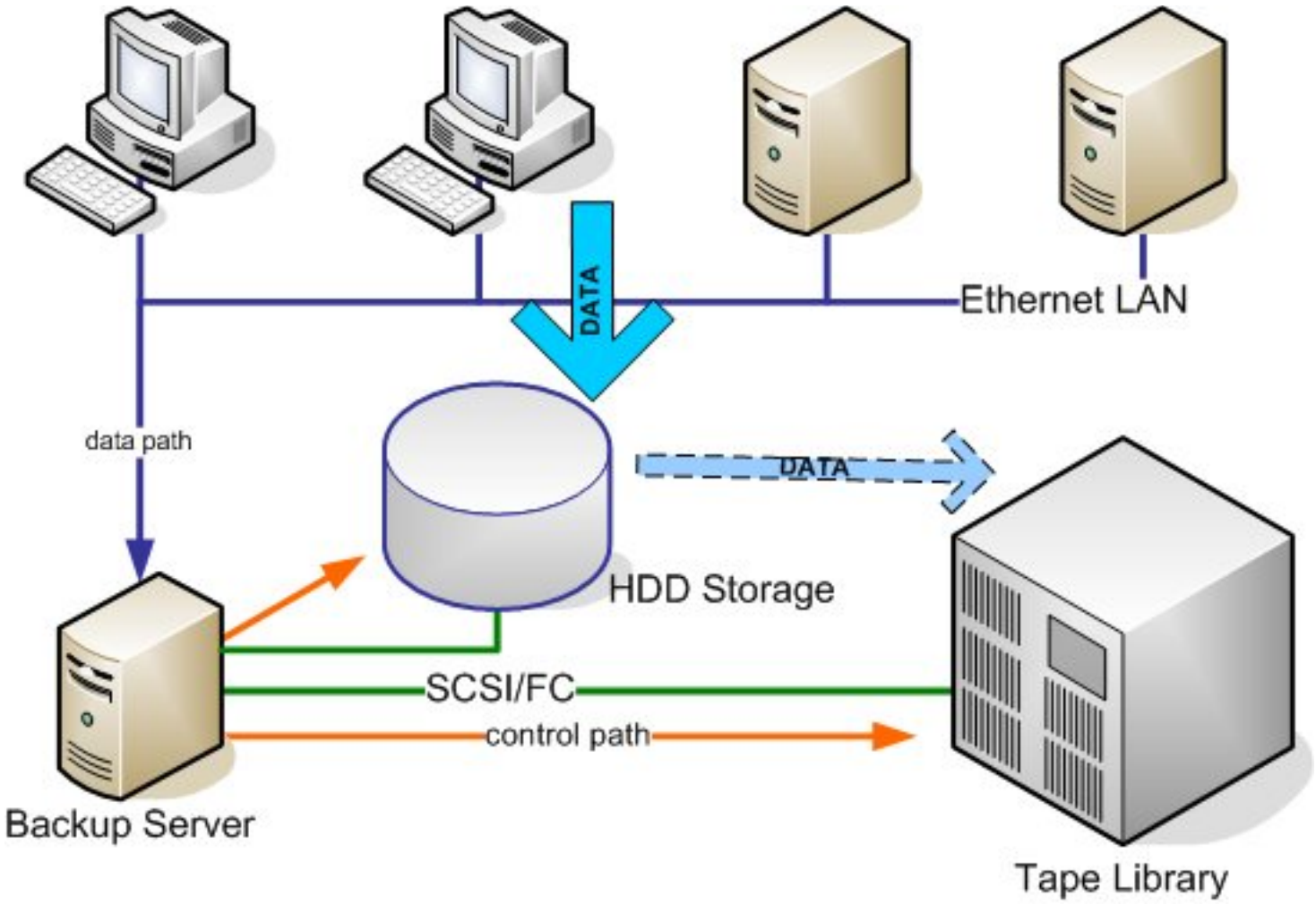
NFS

- Created for Unix systems
  - Based on RPCs
  - v2 uses UDP and is stateless, v3 uses UDP and TCP, and is stateless
  - NFSv4 uses TCP and is stateful
  - Stateless operation is very resilient
    - RPC invocation has all information required
    - No state information
- Provides:
  - File search, open, read, write, close
  - Modify attributes
  - Modify links and Directories
- No locking provided!

Directly Attached Storage (DAS)

- Storage device connected to host through Point to Point interconnect (FC or SCSI)
  - USB and Firewire in SOHO environments
- Advantages
  - Simplicity of use
  - Storage may be used by applications
  - Portability
- Disadvantages
  - More management points
  - No resource sharing between systems
  - Difficult to manage
- Usage
  - Server backup
  - Information storage
  - Personal computers

Data-to-Disk-to-Tape (D2D2T)



Data-to-Disk-to-Tape (D2D2T)

- Advantages
  - Redundancy
  - High performance for recent data
  - High capacity (multiple tapes)
- Disadvantages
  - Complexity
  - Access to old data (due to low tape seek time)
- Usage
  - Incremental Backup's
  - Offsite storage
  - Archival

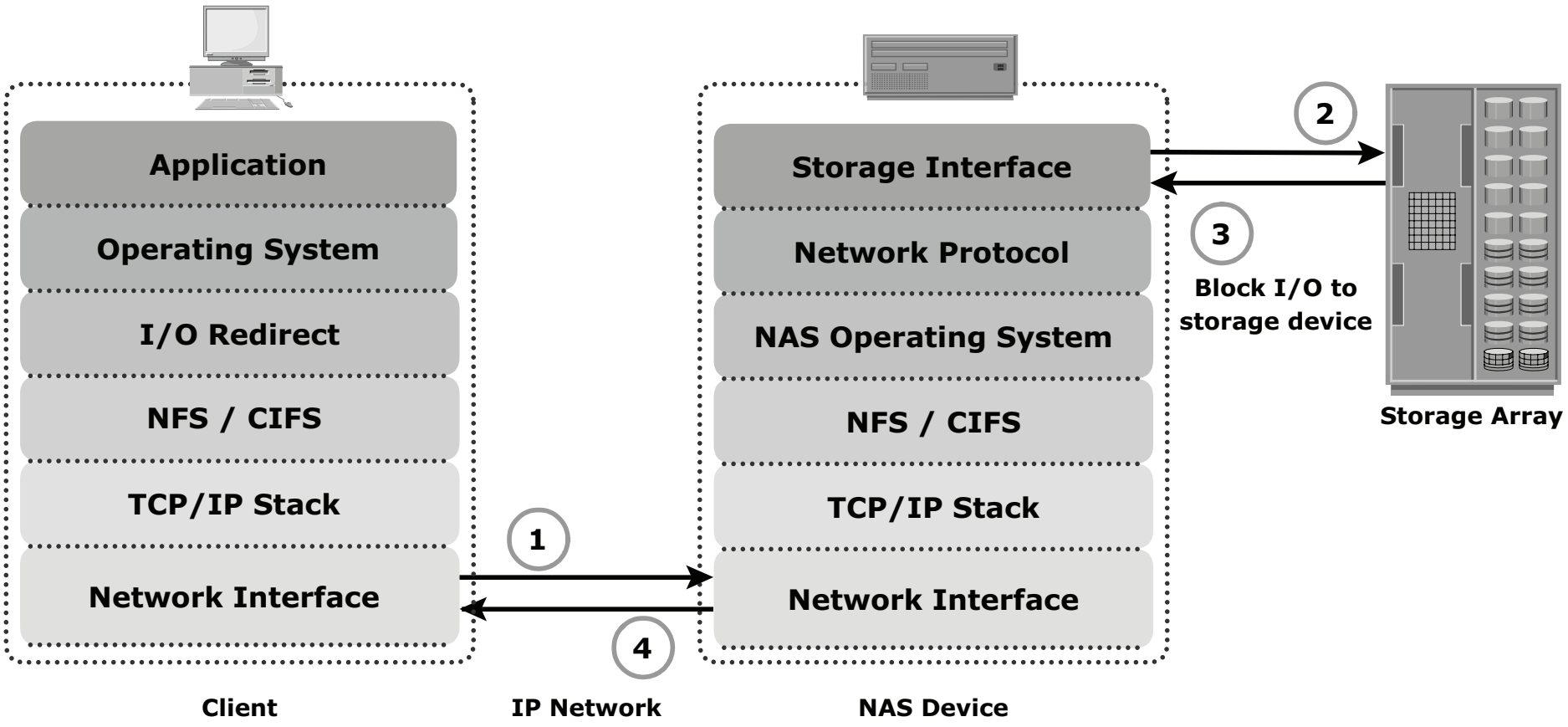
Network Attached Storage (NAS)

- Storage device exports volumes to hosts
  - Uses high level protocols: CIFS, FTP, WebDav, NFS
  - Based in TCP/IP
  - File based access
- Advantages
  - Fast access to shared volumes
  - Possible to create shared volumes
  - NAS may be redundant (into SAN)
  - Single point of management
  - Higher consolidation

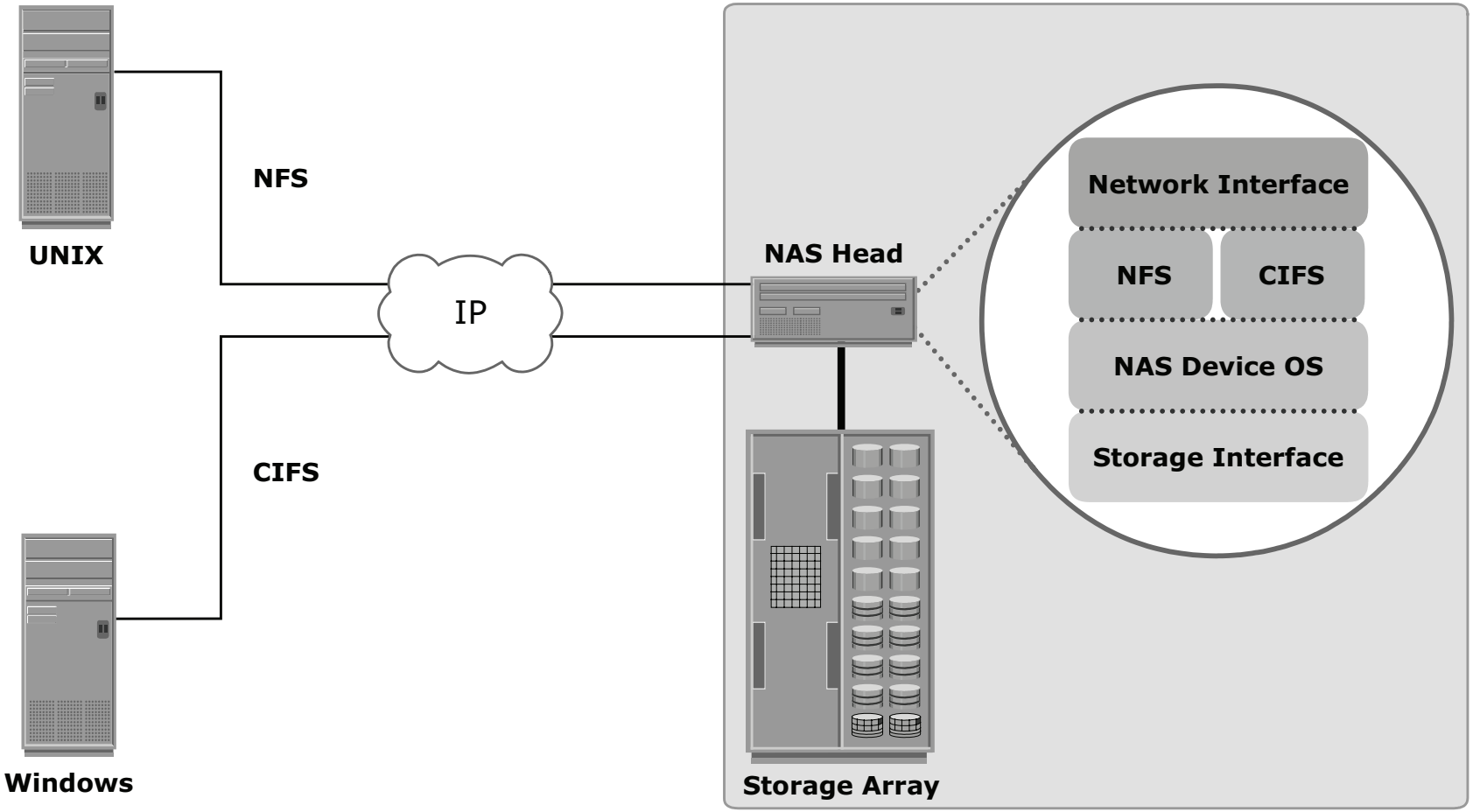
Network Attached Storage (NAS)

- Disadvantages
  - File based access may be too slow
  - File based access may be inapropriate
- Applications
  - Shared storage
    - Ex: ARCA.UA.PT
  - Backup area
  - Data storage to production servers

Network Attached Storage (NAS)

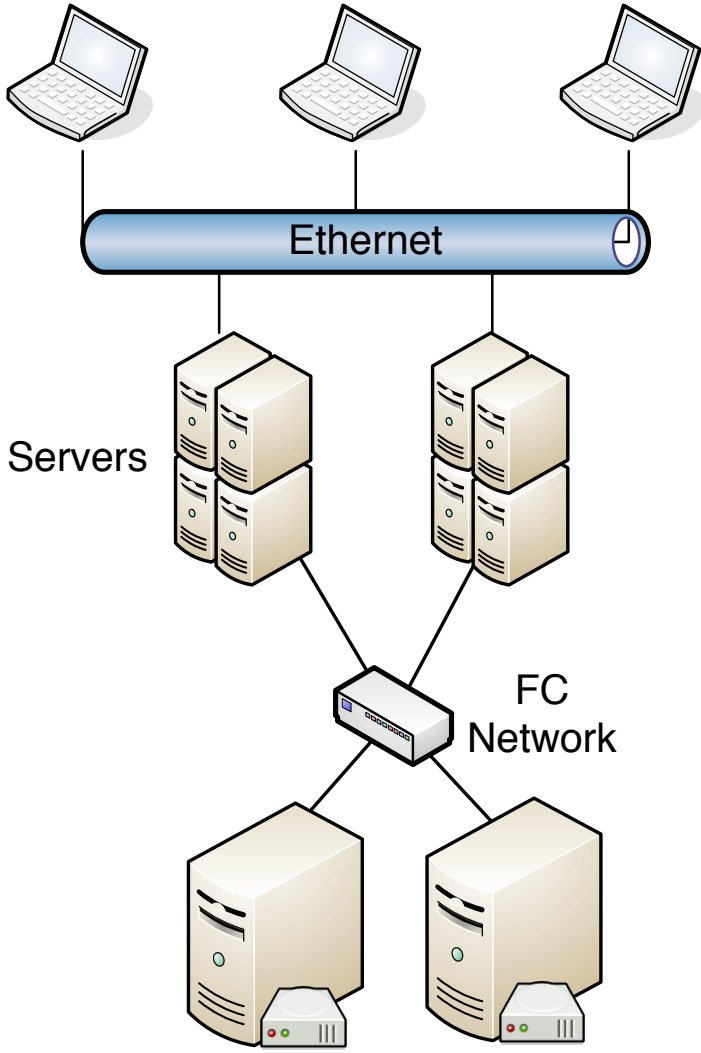


Network Attached Storage (NAS)



Storage Area Network (SAN)

- Several systems providing block based access to volumes
  - Uses **FC (or FCIP)**
  - Clients may not access storage directly
- Advantages
  - Able to store huge amounts of information
  - High levels of availability
  - Fault tolerant and Scalable
  - Virtualized infrastructure
    - Actual infrastructure is different from perceived
  - Block oriented
    - Good for SQL, Mail, etc...



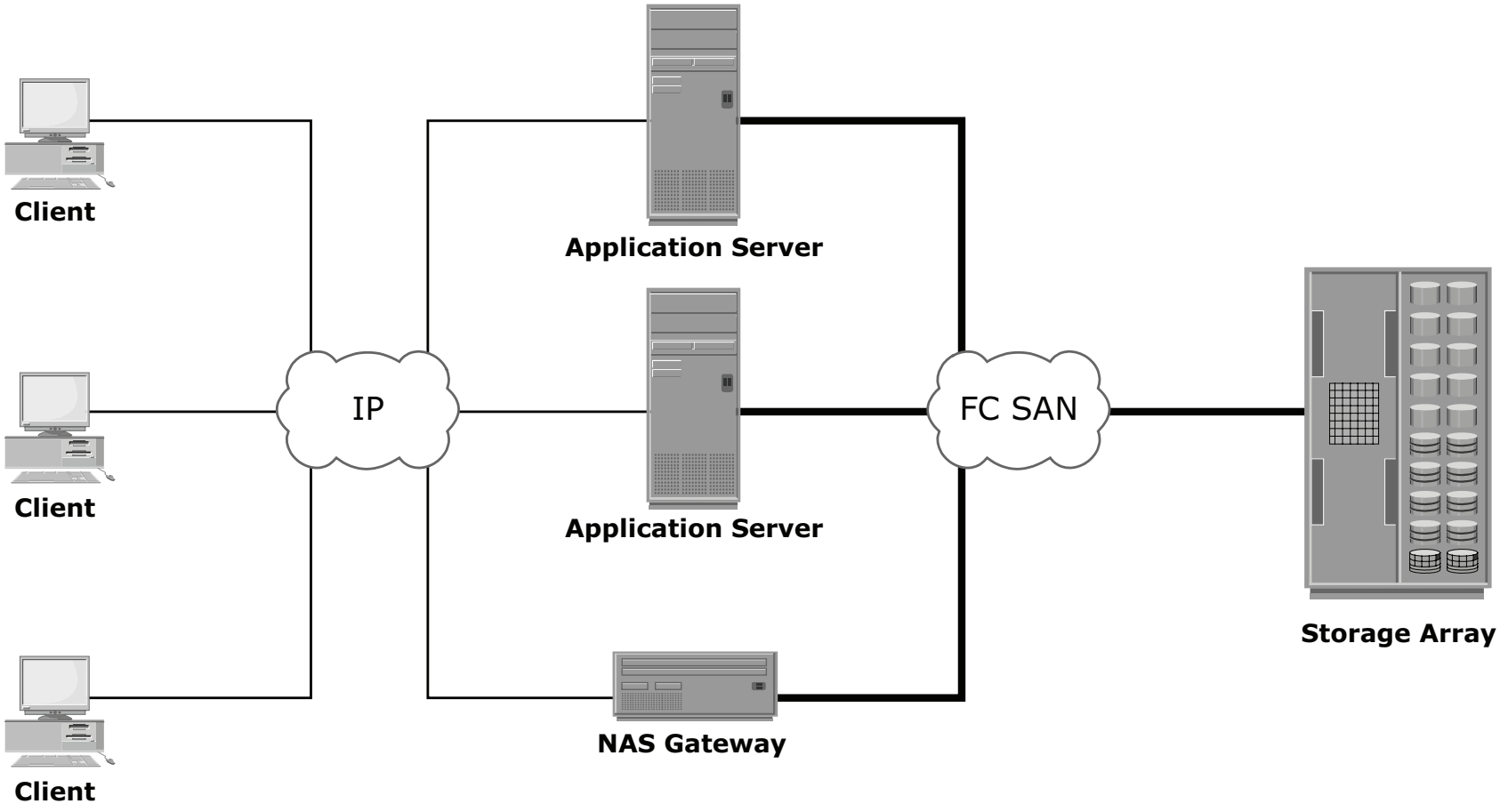
Storage Area Network (SAN)

- Disadvantages
  - High cost due to expensive hardware
    - Using FC
  - More difficult to manage
    - Decouple from real and virtual infrastructure
  - Head nodes may become bottlenecks
- Usage
  - Big databases
  - Storage for workstations and servers
  - Storage for virtualized environments
    - Not so popular in clouds...

IP SAN

- SAN over IP instead of FC
  - Block based
- iSCSI is the standard the facto
  - Makes devices available to servers (VMs)
  - FCoIP also greatly used
- Improved performance due to IP
  - Multipath
- May use broadcast/multicast to keep replicas

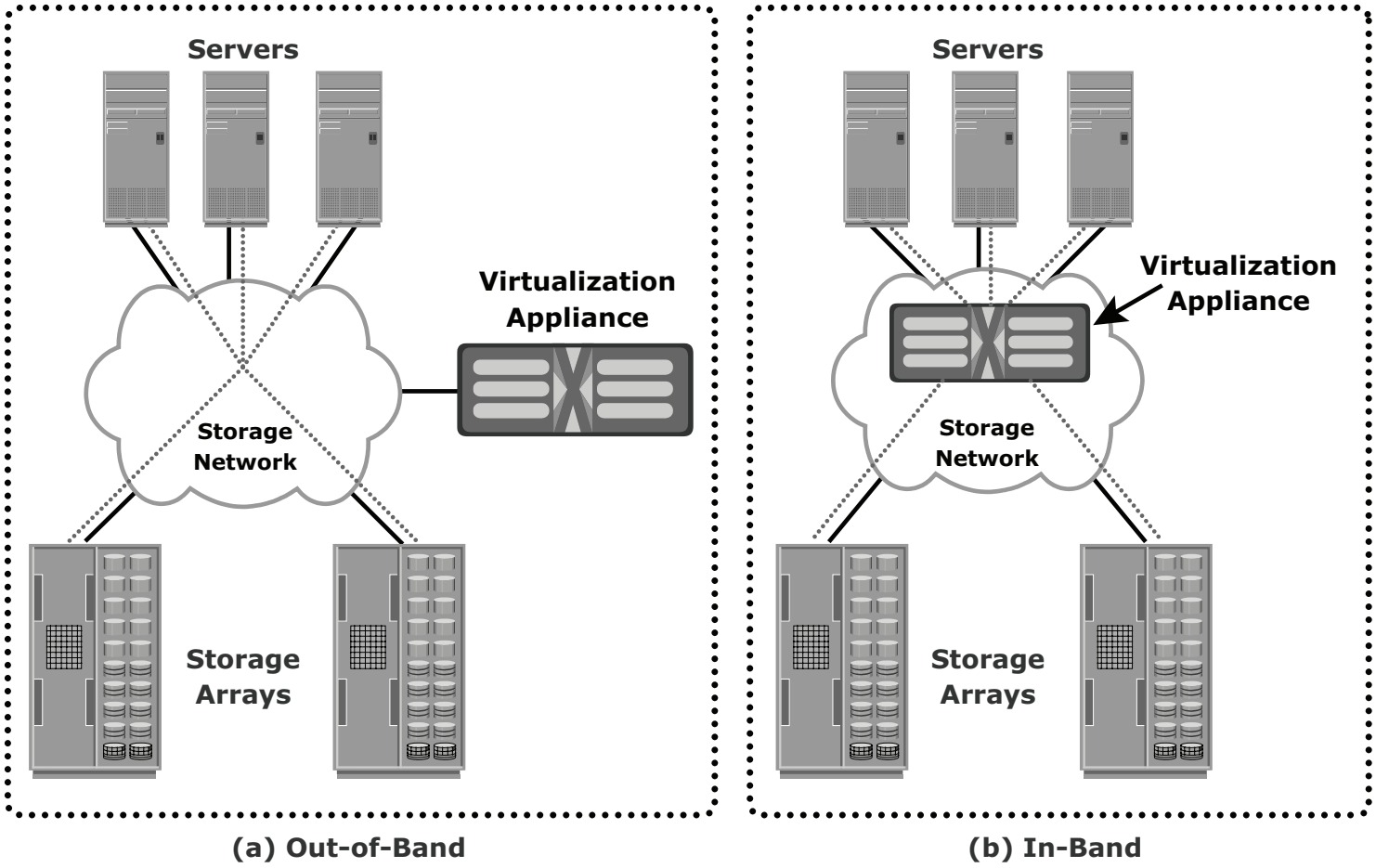
Network Attached Storage (NAS)

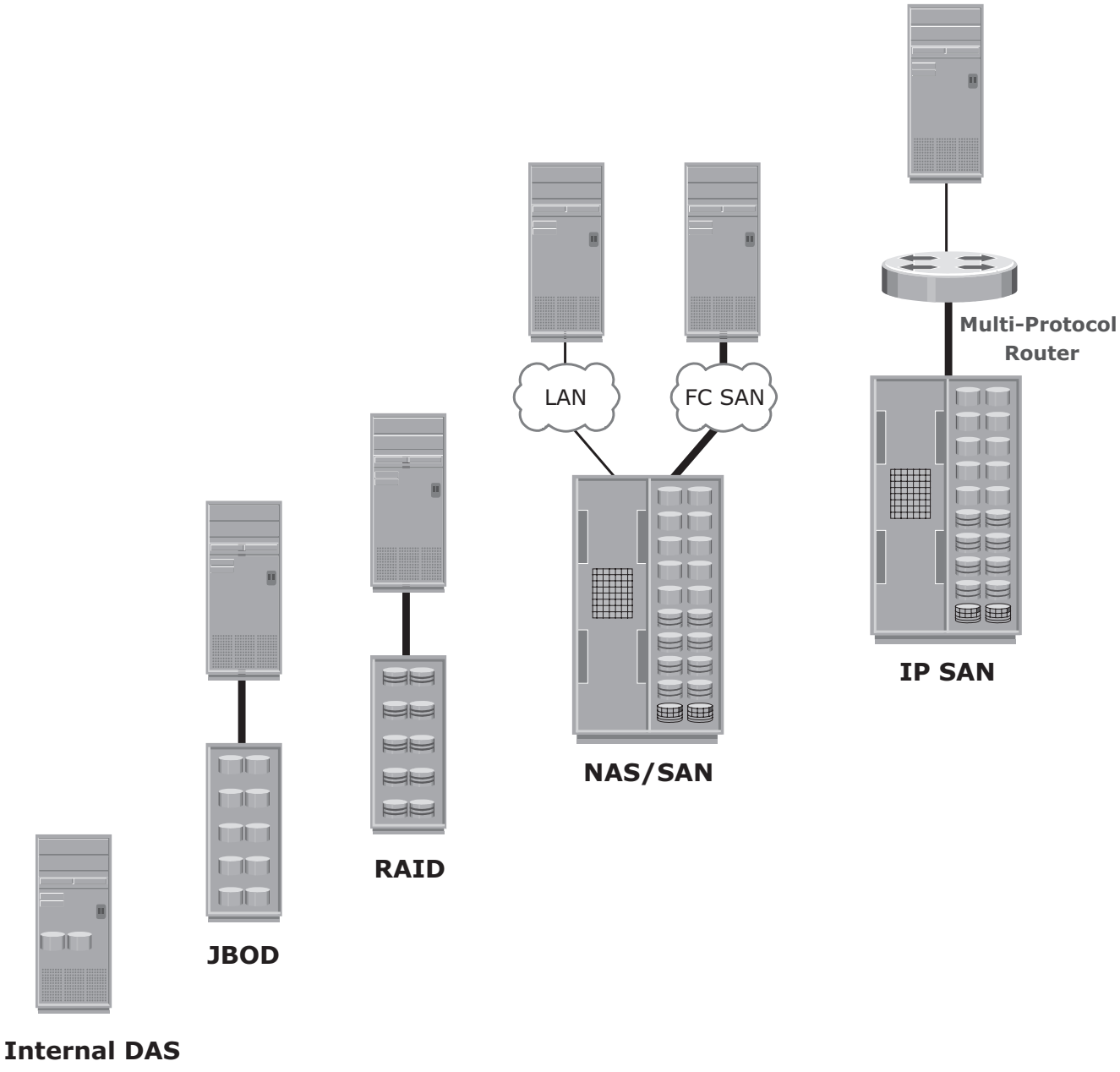


Virtual SAN

- Entire SAN may not be visible to all clients
  - Virtualization layer exposes virtual SAN
  - May be block or file based
- Two Methods:
  - In Band: Appliance provides virtual iSCSI targets mapped to real resources
    - Resources may be files, LV, real disks or other targets
    - Increased flexibility but also latency
    - May adapt between technologies
  - Out of Band: Appliance translates targets
    - No latency added
    - Increased scalability

Virtual SAN

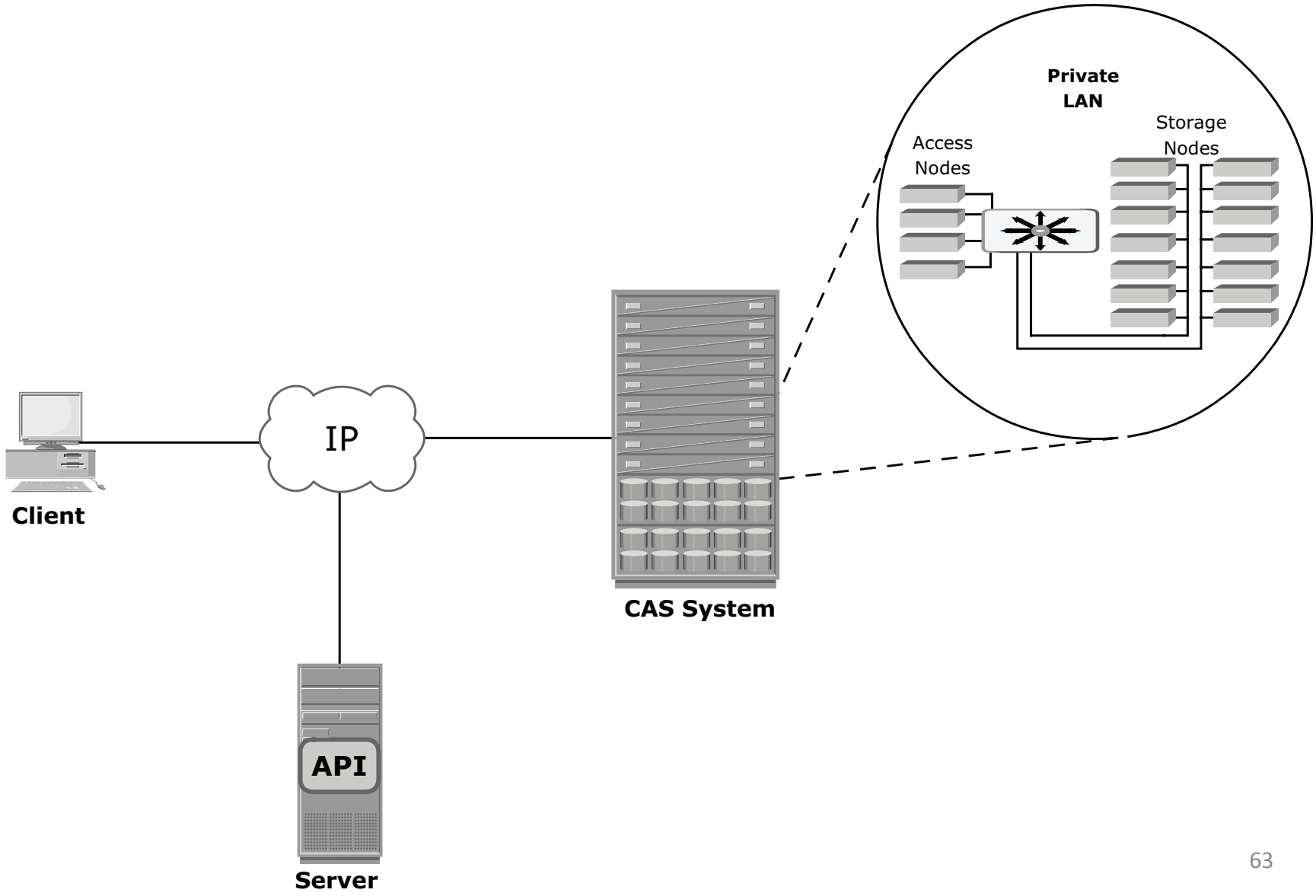

Servers
Virtualization Appliance
Storage Network
Storage Arrays
(a) Out-of-Band
Servers
Virtualization Appliance
Storage Network
Storage Arrays
(b) In-Band


Internal DAS
JBOD
RAID
LAN
FC SAN
NAS/SAN
Multi-Protocol
Router
IP SAN
Time

Content Addressable Storage (CAS)

- Information is accessed as Objects
  - By ID based on Content (e.x. SHA-1)
  - Integrity guaranteed by constant ID
  - Flat address space
- Can rely on low power, cheap, SATA based storage nodes
  - Redundancy provided by simple duplication
  - Simple load balancing
  - Highly Scalable (Ex. Using DHT)
- Can use different interfaces
  - With different primitives (ex. Read, Search, Write)
  - Different technologies (ex. SOAP, REST)
- Usages:
  - Health care: store patient records and images
  - Financial Institution: store records, check images

Content Addressable Storage (CAS)



Further reading

- EMC Educational Services, “Information Storage and Management”, Wiley India.