Management of Information Systems

BME VIK TMIT

Software Engineering, BSc

Management of Information Systems

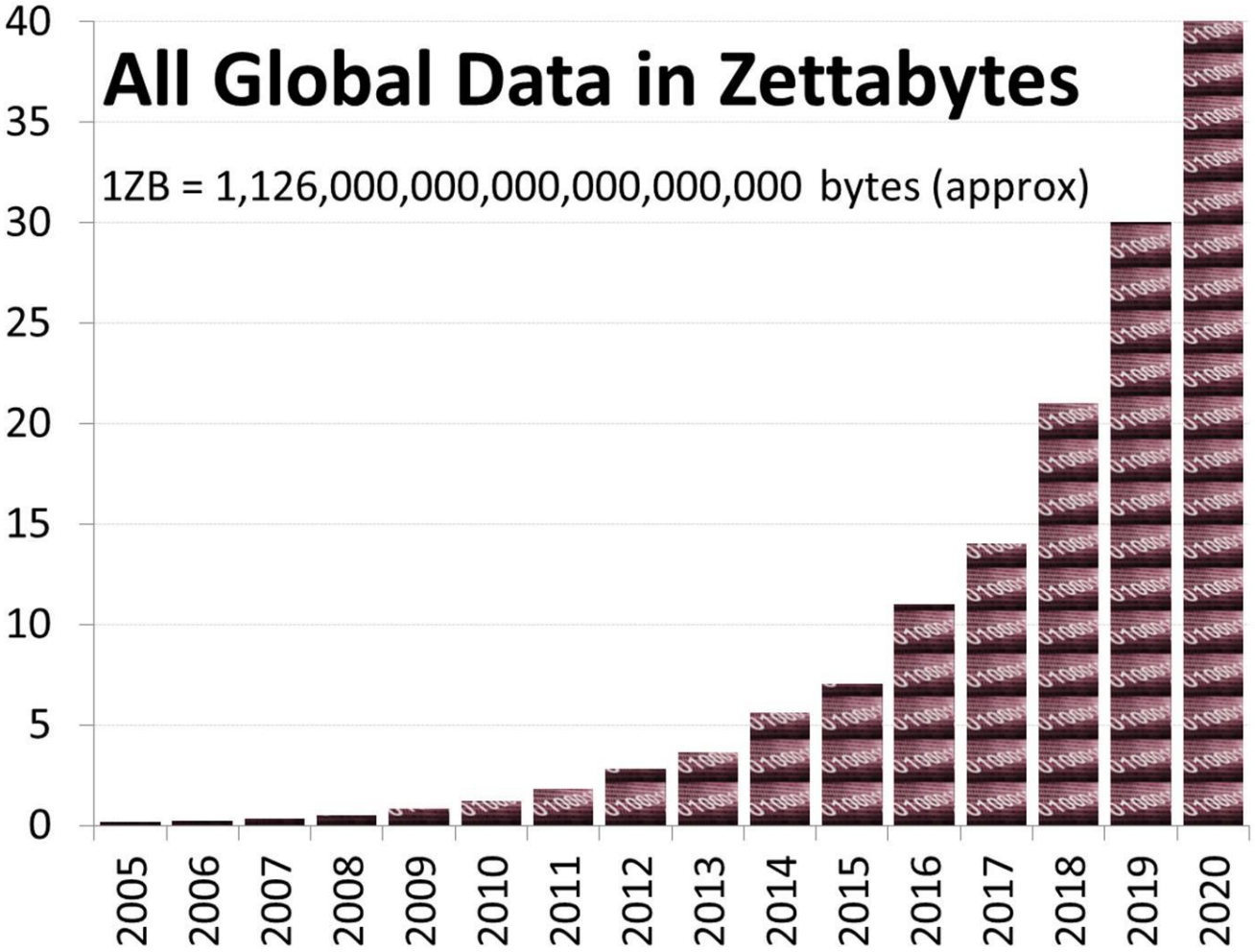
5. Data Management in Industrial Environments

Data Management

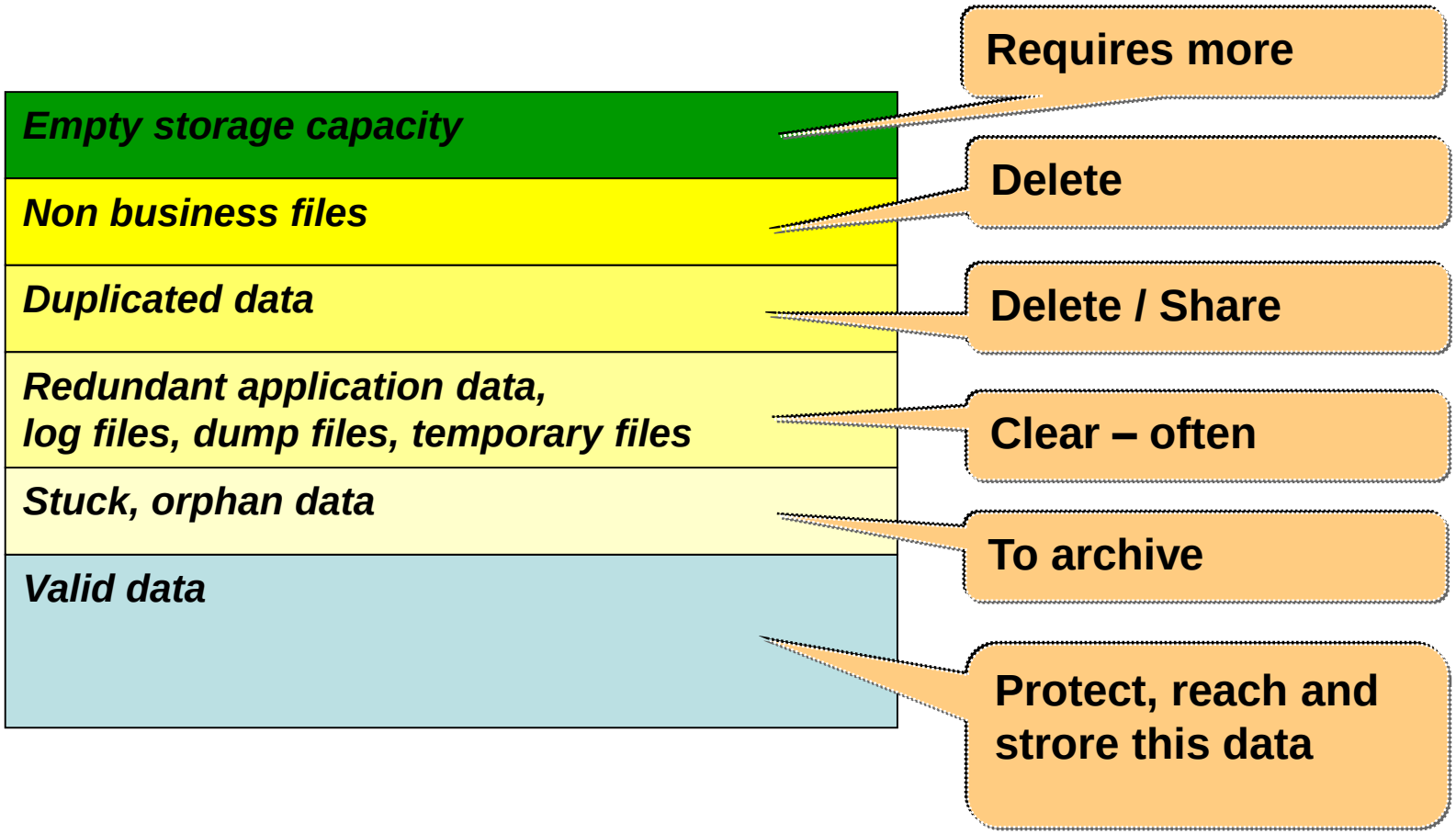
- Types of data
- Data storage types
- Reliability of disks (RAID)
- Data storage systems (DAS, SAN, NAS, IP SAN) and transmission technologies
- Virtualisation

DATA GENERATED

- Zetta = 10^{21}
- In the past two years as much data genereated as much till then



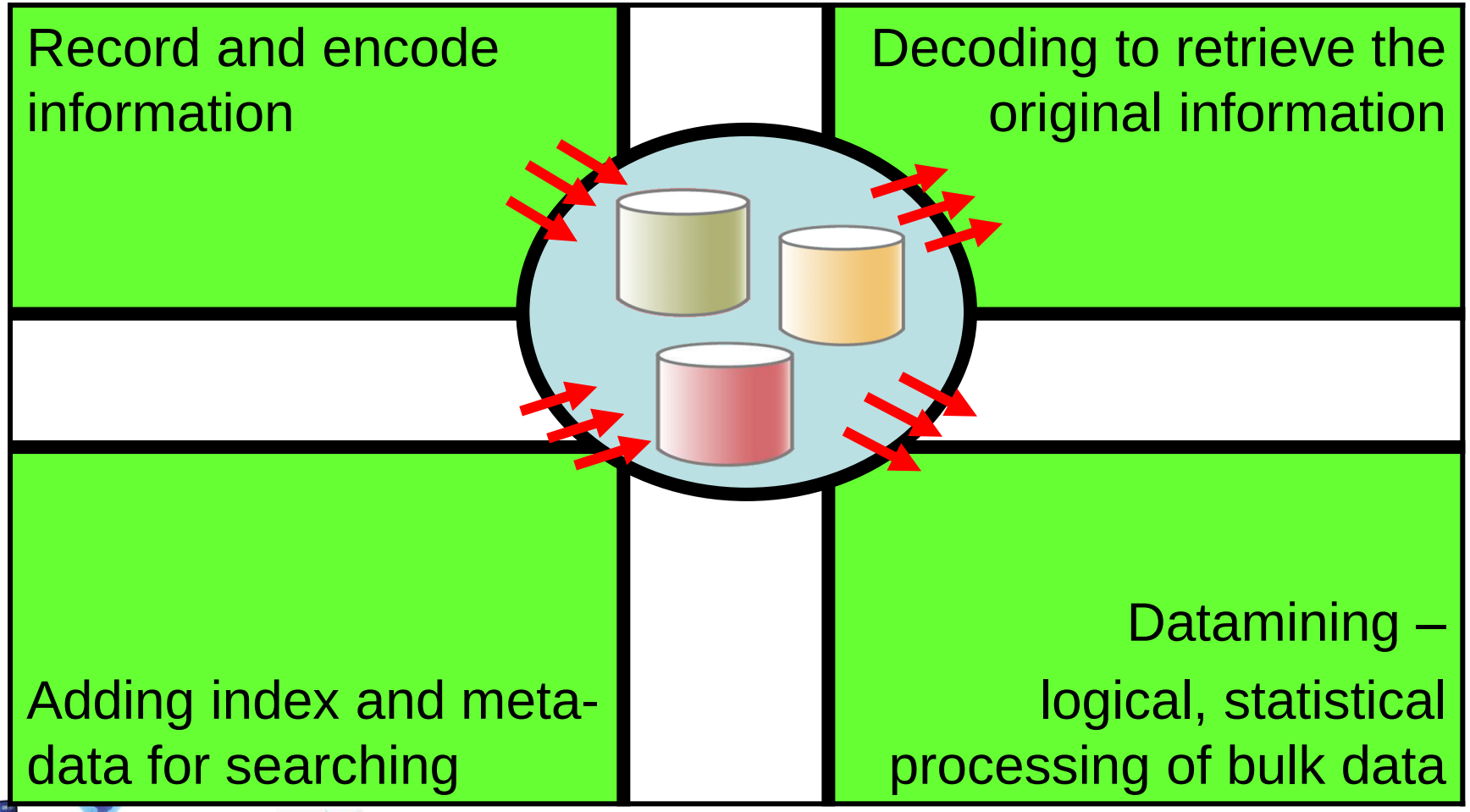
Data Types


Empty storage capacity
Non business files
Duplicated data
Redundant application data,
log files, dump files, temporary files
Stuck, orphan data
Valid data
Requires more
Delete
Delete / Share
Clear – often
To archive
Protect, reach and
strore this data

Data Stream Types

Structured	**Unstructured**
Known source and goal	Unknown
Server to Server	Client to client, Client to Server
Business devices	Personal devices
Special applications	E-mail, Web
Private networks	Public networks
Can be estimated	Hard to estimate, statistic
Database	File servers
It can be estimated, and planned	**It can only be regulated by policies**

Information and Data Life Cycle, Data Types

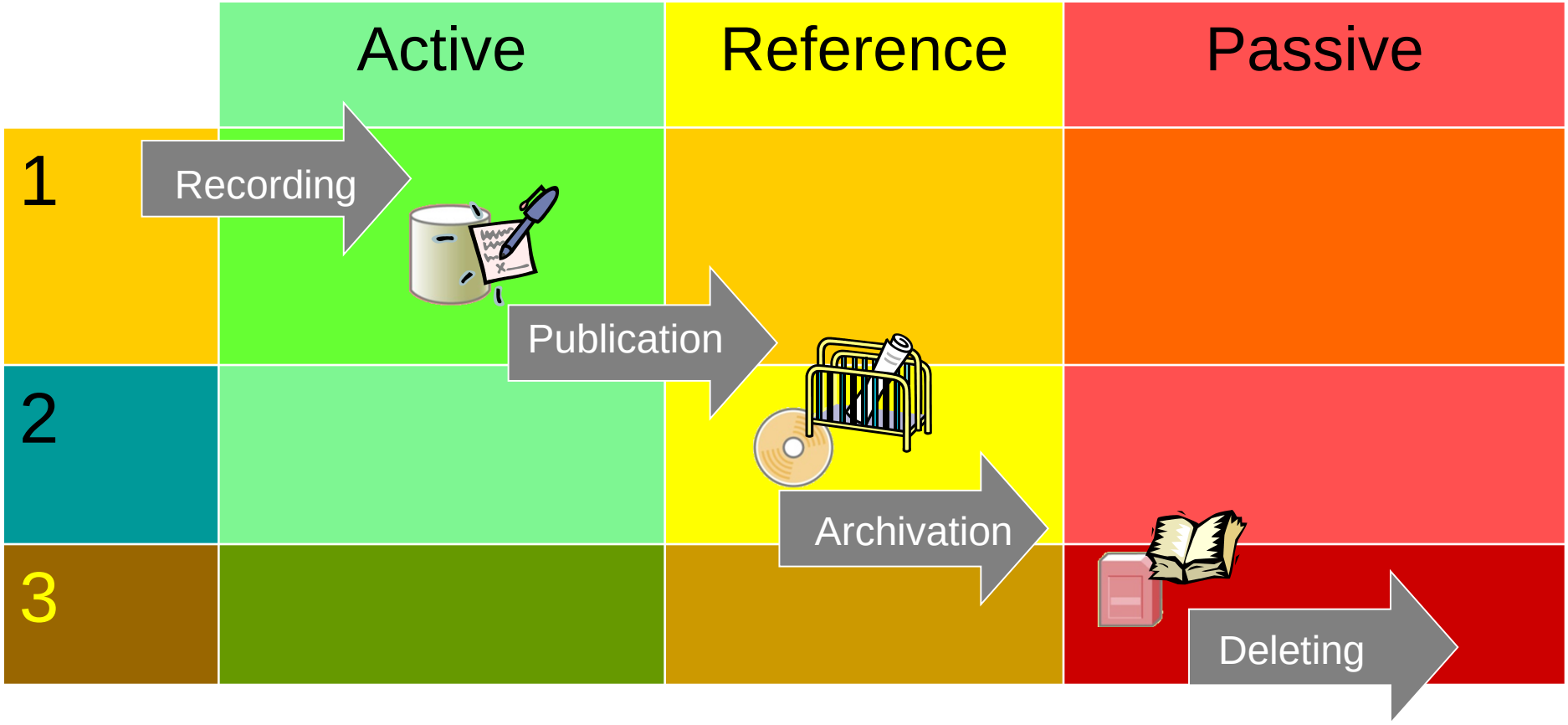


IMPORTANCE OF DATA - CHANGING

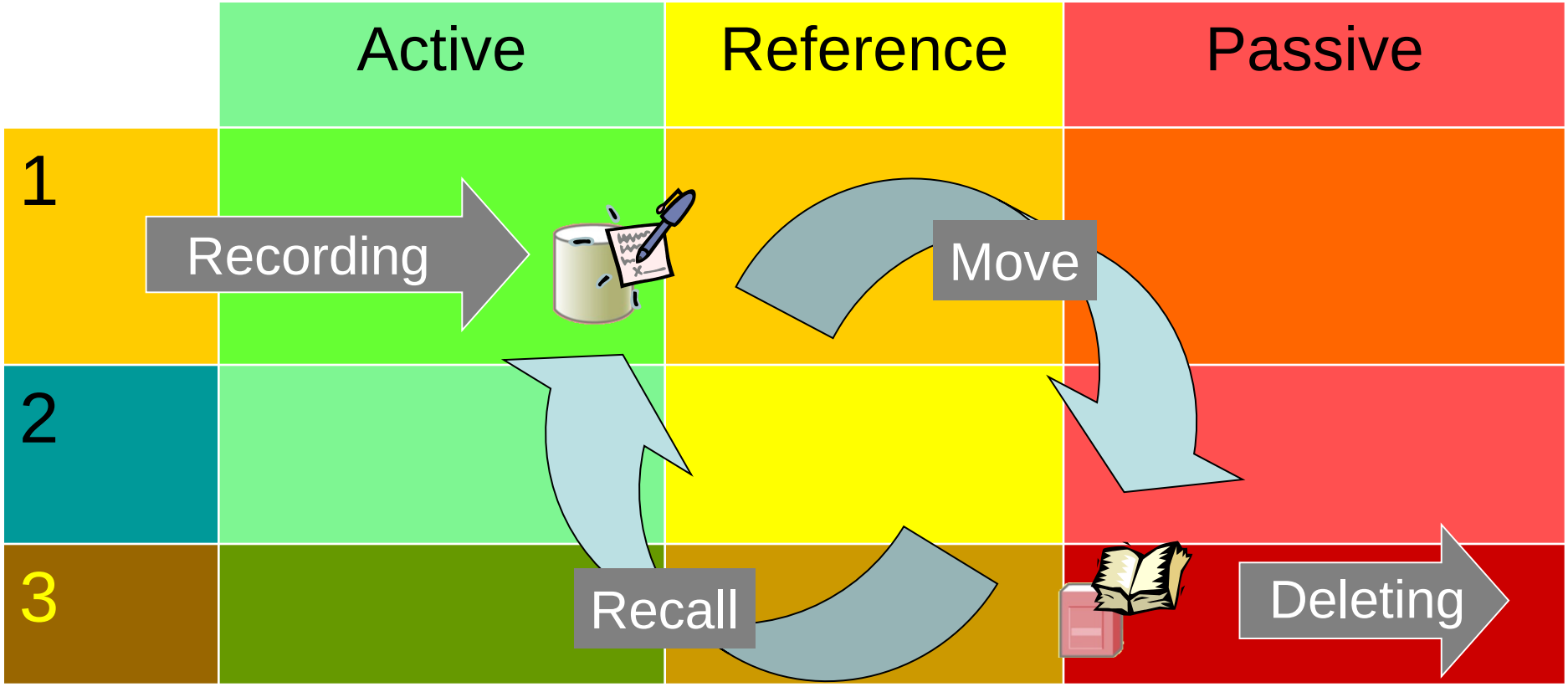
Active	Reference	Passive
Application DB Online Transaction Processing (OLTP) Document handling Software development	Reports Transaction reports Public documents, Web sites, e-mail messages	Archives Back-ups Security video records

Recording → Publication → Archivation → Deleting

Optimal Storage Technology: Select Accoring to Data Value

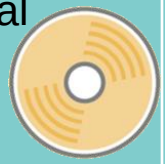


Hierarchycal Storage Management (HSM)

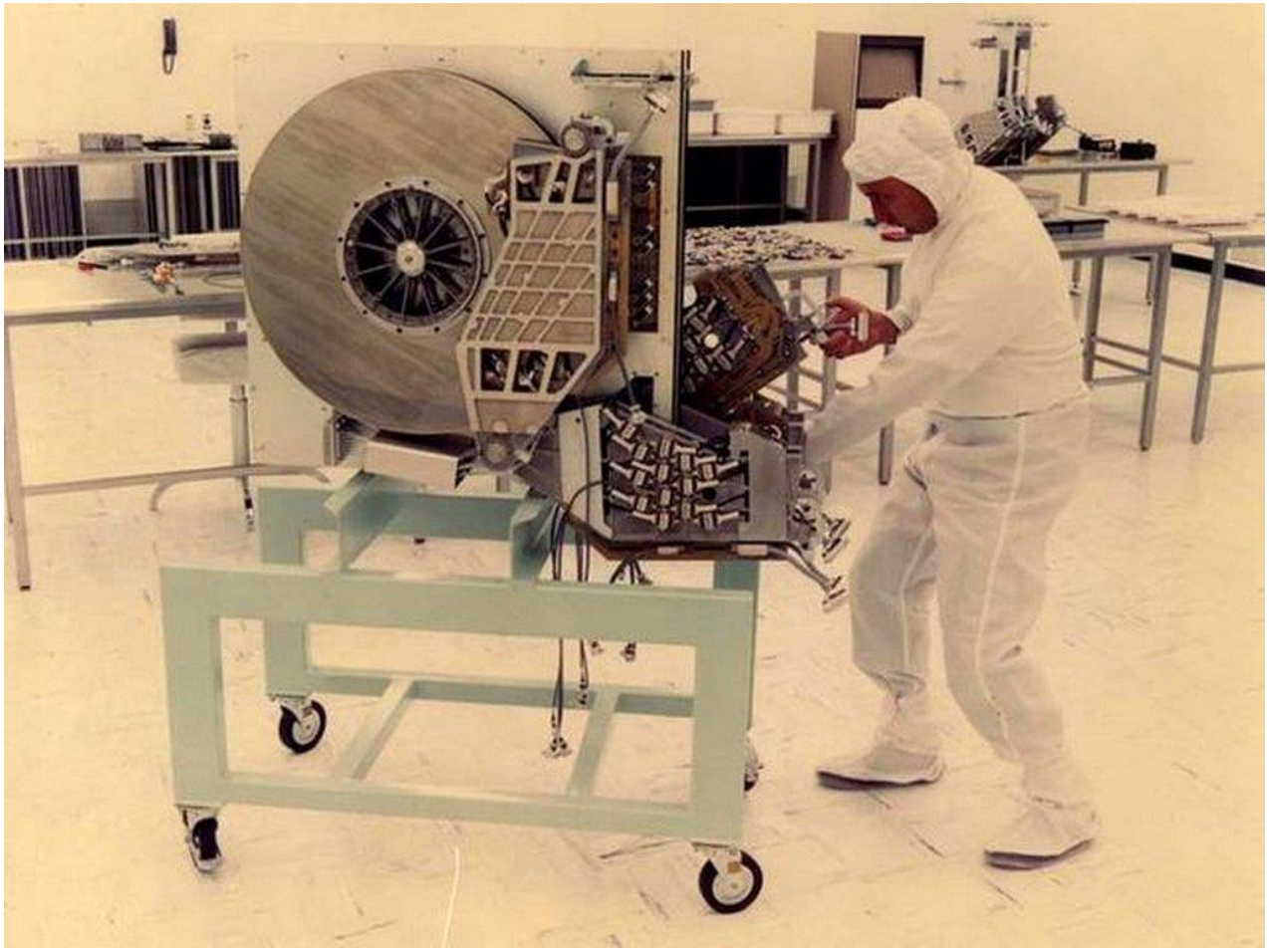


HSM parameters: size, type of data and the time of the last access

Storage Devices

	Advantages	Problems
Disk 	• „Immediate” data access • „Random” read/write (R/W)	• Power supply, cooling • Lifetime only 3-4 years! • Moving parts • Disk replacement
Flash memo-ry 	• EEPROM, pen drive, SSD • No moving parts • Fast	• (Yet) expensive • Small(er)
Optical 	• Secondary storage –WORM (*Write Once Read Many*)	• Cannot keep up with the development of disk and tape • SOHO device (*Small Office Home Office*)
Tape 	• 10-20x cheaper than the disks • Lifetime: 30+ years	• Non-immediate data access • Serial read/write (R/W)

???



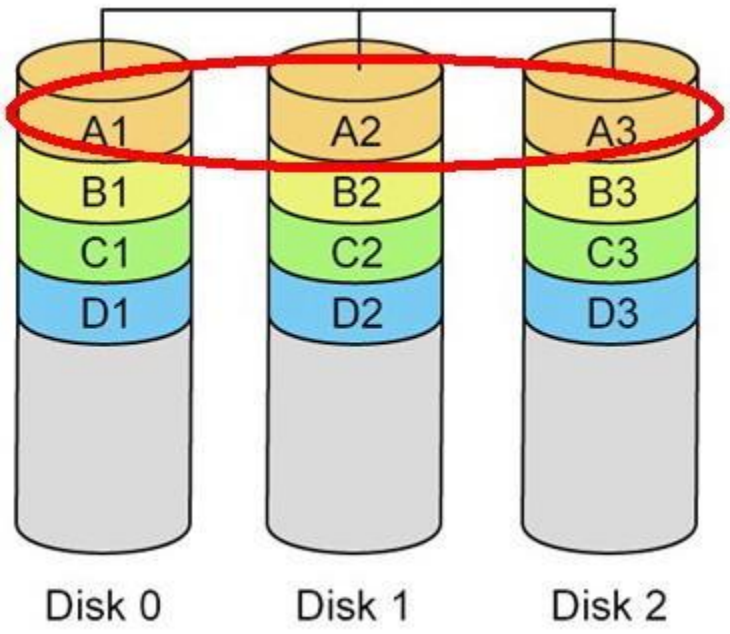
250 MB HDD -1979

RELIABILITY OF DISKS

- Data stored on disks – protect from demage
- Originally one highly reliable disk
  - SLED: Single Large Expensive Drive
    - Expensive…
  - MTBF (Mean Time Between Failure)
    - Appr. 750 000 hours (appr. 85 years)
    - But a large(??) disk array with 1000 disk
    - MTBF: 750 000 hours / 1000 = appr. 1 month

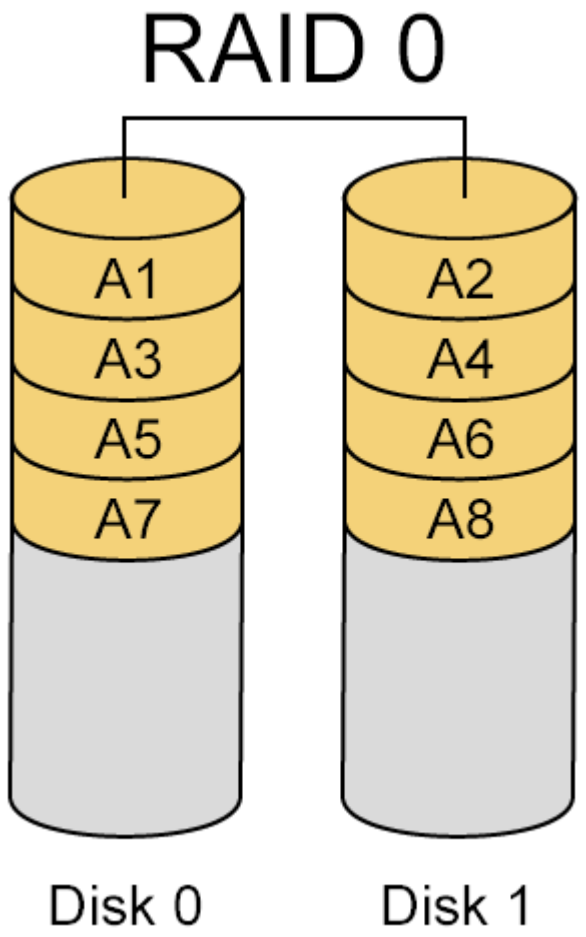
RAID – Redundant Array of Independent (Inexpensive) Disks

- To protect disks against physical failures
  - 1987 Berkeley University (California)
- RAID 0, 1-5, 6
- Disks are divided into stripes
  - Same blocks on all the disks



RAID 0 - STRIPING

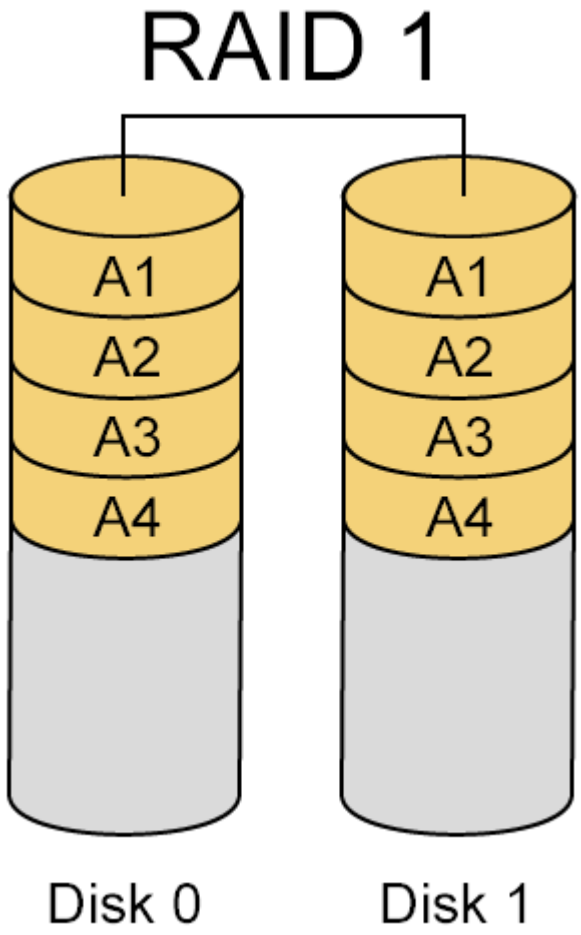
- Goal:
  - Increase the capacity
  - Increase speed
    - (parallel read/write operations)
  - NOT to increase the reliability



RAID 1 – Disk Mirroring

- Parallel operations
- High reliability
- Large overhead in size (2x!)

- Can we have approximately the same reliability with smaller overhead?

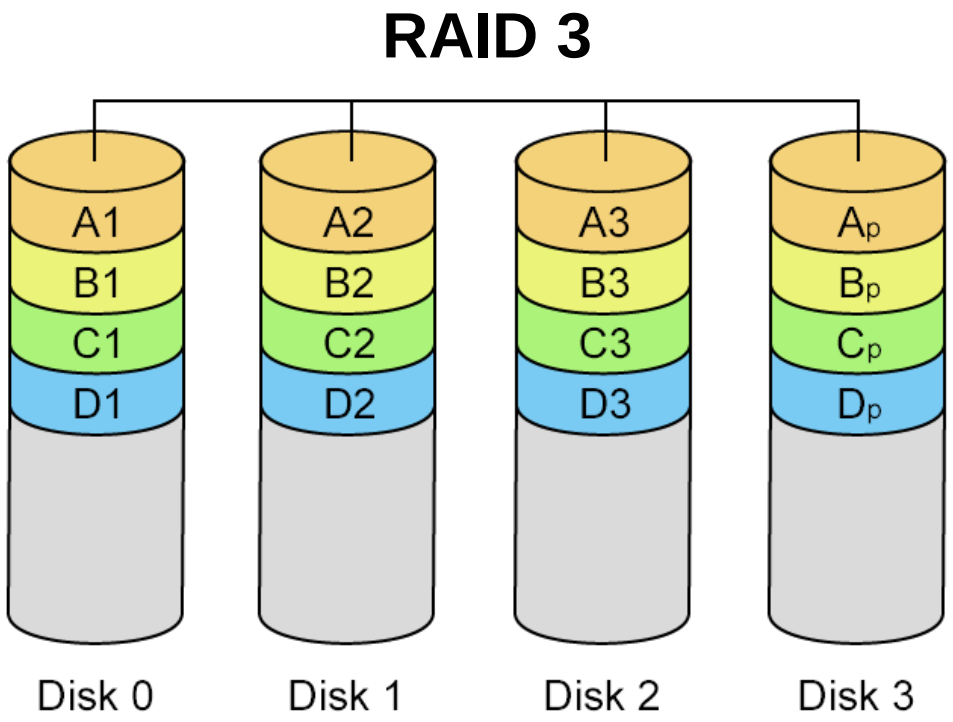


RAID 2 – Error Correcting Code

- ECC – (Error Correcting Code) are stored on certain disks
  - Suitable for detecting and correcting errors
- Not used nowadays, because ECCs are used *inside* the disks

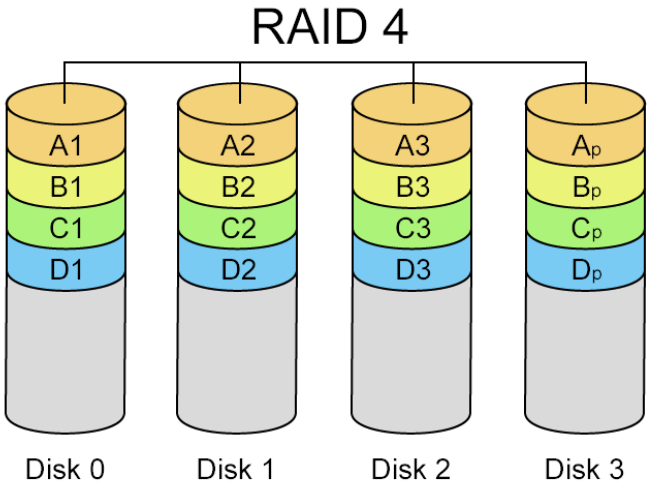
RAID 3 – Parity Disk

- +1 disk: parity (XOR)
- One disk fails: content can be reproduced from the others by XOR – time(!), slow
- Cannot detect disk errors
- Operations on whole stripes
  - Small stripes
  - Single user
  - For large files (video)
- Typical: 2+1, 5+1, 8+1, 14+1
- Parity disk constrains



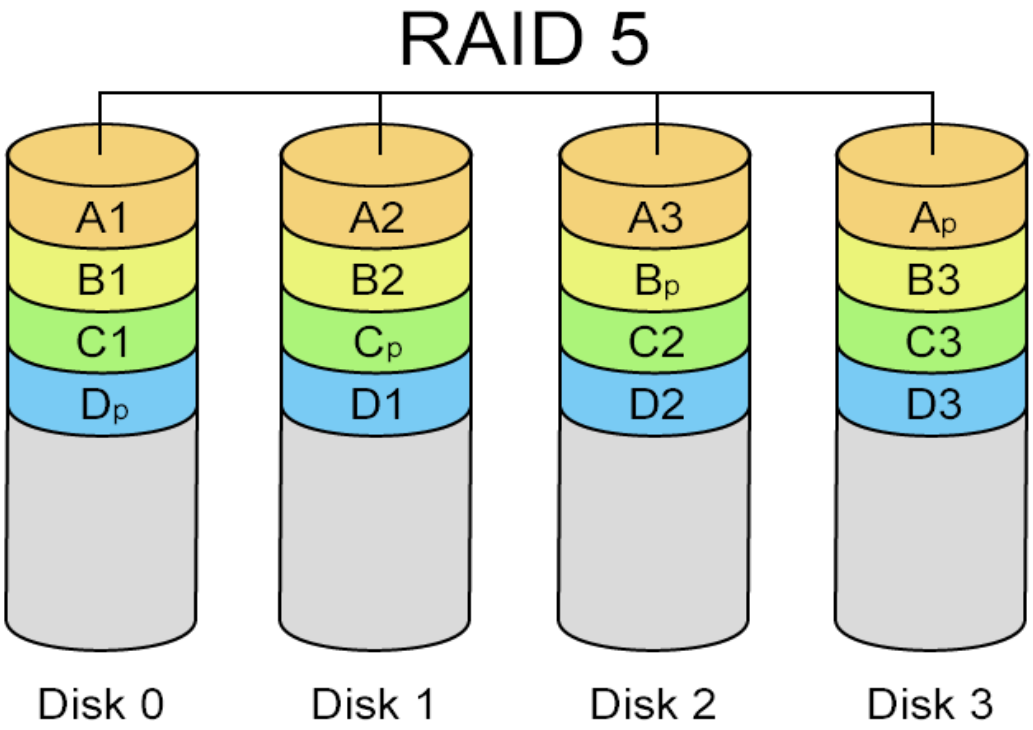
RAID 4

- Disks can be accessed directly
  - Allows parallel service
  - Large stripes
  - Parity disk constrains very much
- Not used in practice



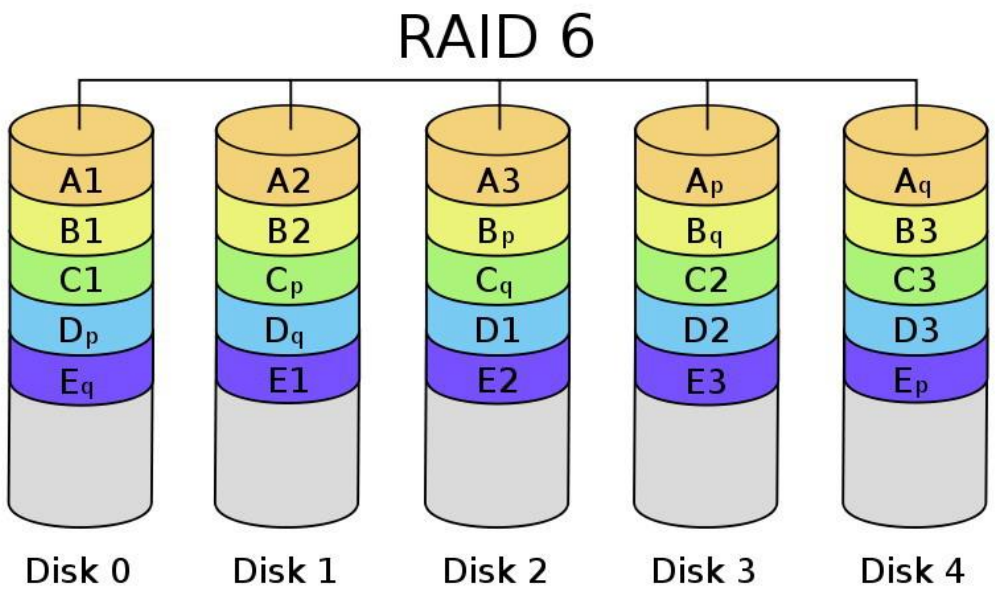
RAID 5 – Distributed Parity

- Parity distributed amongst disks
- Eliminates the bottleneck of parity disk
- Each disk can be accessed directly
- Configureable stripe size

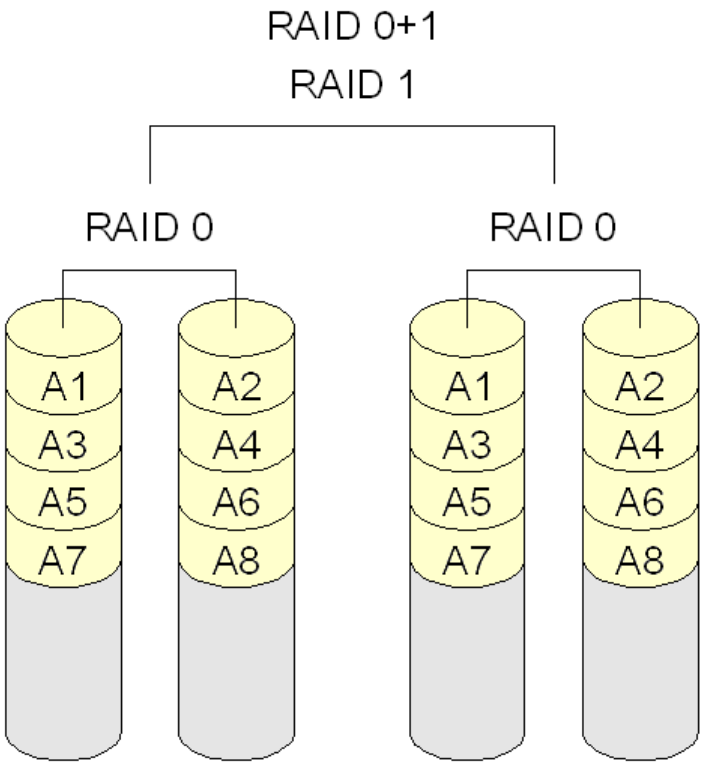


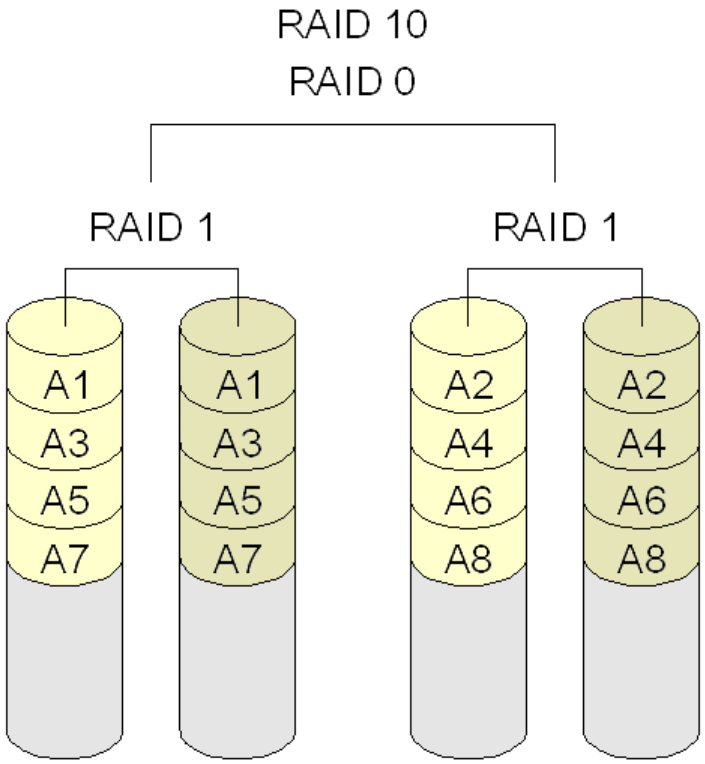
RAID 6 – Double Parity

- Row (XOR – P) and Coloumn (Reed-Solomon Code – Q) parity
  - Distributed among disks
  - Protects against double failures – but very slow



RAID 01, RAID 10





Same?

Error: No mirror on whole stripe (!!)

Restore: Whole stripe (2 disks !!)

Speed: High (HW striping)

No mirror only on half

Only the wrong disk

Slower (SW striping)

RAID

- In practice: 0, 1, 5 are used frequently
- RAID can only protect against PHYSICAL errors!!!!
  - Against logical errors: back-up
    - Next time

FLASH MEMORY

- Non-volatile memory (EEPROM)
  - Erasable by a special signal (‘flash’)
  - Pen-drives
  - SSD
    - Solid State Drive
  - AFA
    - All-Flash Array
  - Hybrid arrays
    - SSD + Disks

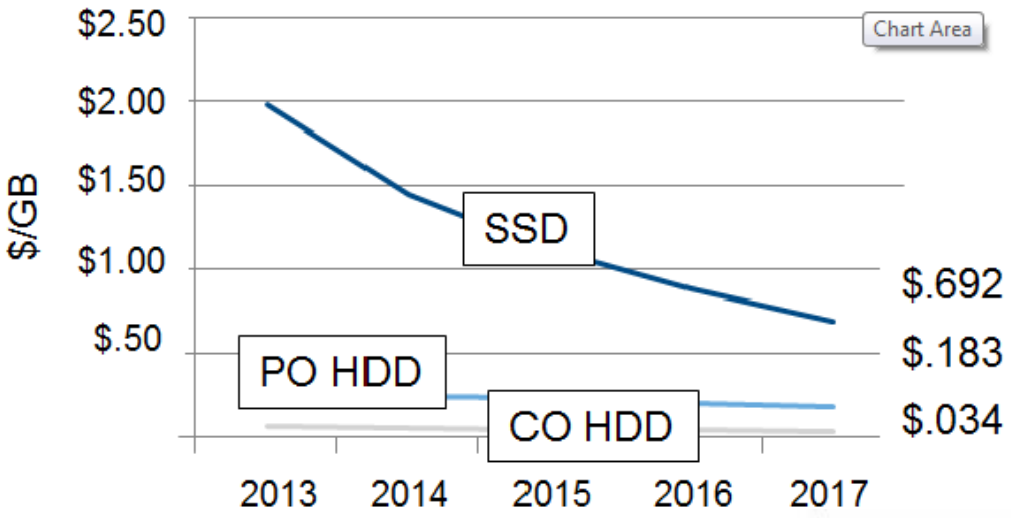
ALL-FLASH ARRAY

FlashStack

ORACLE SAP Microsoft SQL Server

25 TB 250 TB 500 TB 1.5 PB

PRICE

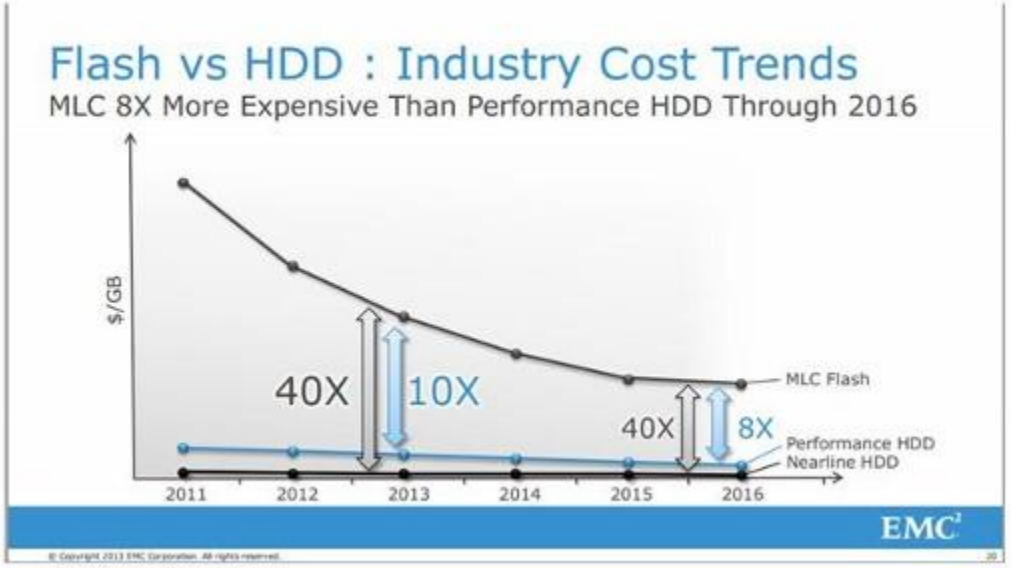


PO: Power optimised

CO: Cost optimised

TCO: Total Cost of Ownership

- ~20-25% of the TCO (HDD)
- ~80% of the TCO (SSD)



LATENCY

- Latency 1000:1 (HDD:SSD)
  - Speed of the connections (SATA, SCSI) is the bottleneck
  - Faster boot time
  - Data reduction technologies can be more effective
    - 6:1 (SSD) – 2:1 (HDD)
      - If it is not done by a higher layer – like in VM
    - Effective usable capacity
    - In SSD typically done by controller while in HDD extra SW
  - Faster – more user requests can be served

Power, Cooling, Size

- Power, cooling
  - No moving part
  - Less power consumption (1:10)
  - Less cooling requirement
    - 50-90% less than of HDD
- Smaller
  - Especially when data reduction is considered
  - 20-40* data density than of HDD
  - Up to 90% higher rack utilisation

All-Flash Array

RELIABILITY

- Smaller maintenance cost
- Reliability
  - 100:1 (unrecoverable bit error rate)
    - 10^{-18} : 10^{-16}
    - 10^{-18} means 1 error in several hundred years
    - In early EEPROMs number of writing cycles were limited (~1000)
    - Today's SSDs – for more than 200 years
  - RAID (1 in older, 5 in newer versions)

HYBRID ARRAYS

- HDDs better in price/GB
- SSDs better in price/IOPS
  - Input/output operations per second
- Add a thin slice of flash to an HDD array
  - SSD: 2% - 5% of total capacity
  - available IOPS may double
  - reading latency from 10+ ms to 3-5 ms
    - Not constant – may cause problems in some applications
  - Only by 10% - 20% more expensive
  - But appr. 2x faster
  - RAID 1

COMPARISION

	Avg upfront per GB HW Cost	1yr avg per GB Power & Cooling	Avg upfront per usable GB Costs*	~ 3 yr per usable GB TCO
100% HDD Storage System	$0.50	$0.50	$0.63	$2.13
Hybrid Storage System	$1.60	$0.38	$0.59	$1.72
100% Flash SSD Storage Systems	$10.00	$0.17	$3.64	$4.13
100% Flash SSD Storage Systems	$5.00	$0.17	$1.82	$2.31
100% Flash SSD Storage Systems	$4.00	$0.17	$1.45	$1.95
100% Flash SSD Storage Systems	$3.00	$0.17	$1.09	$1.59

** HDD cost/usable GB goes up because of formating and RAID overhead.*

SSD based dedupe/compression increases usable GB ~ 2.75 x.

Hybrid systems a little less to account for the HDD overhead.

SSDs are typically overprovisioned from 20 to 50% to account for load balancing & garbage collection

Storage Networks

- DAS – Direct-Attached Storage
- SAN – Storage Area Network
- NAS – Network-Attached Storage
- IP SAN (iSCSI)

DAS – Direct-Attached Storage

- Storage is connected directly to server
  - Block level access
  - Mainly in small(er) systems
- Two subtypes
  - Internal DAS
  - External DAS

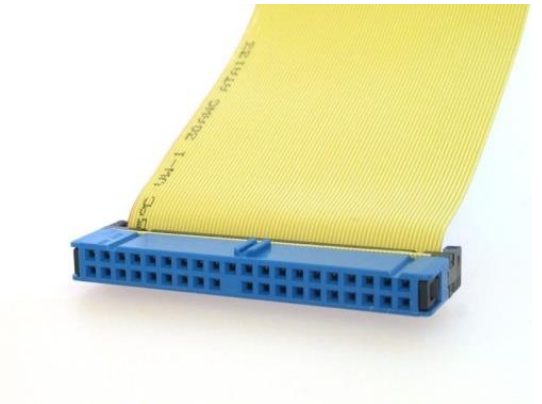
INTERNAL DAS

- Storage is connected directly to server by an internal parallel or serial bus
  - Limited distance
    - < 1m
  - Limited number of devices can be connected
    - (P)ATA or SATA connectors
  - Cables require large space inside the server
    - Complicated maintenance

(P)ATA

- (P)ATA – Parallel Advanced Technology Attachment
  - Half duplex

- 40 & 80 wire/cable
  - 40 wire limited to UDMA 33 MB/s and below
  - 80 wire allowed for UDMA 66, 100, 133 MB/s
  - Development stopped in 2004 (because of the space requirement of the cable)

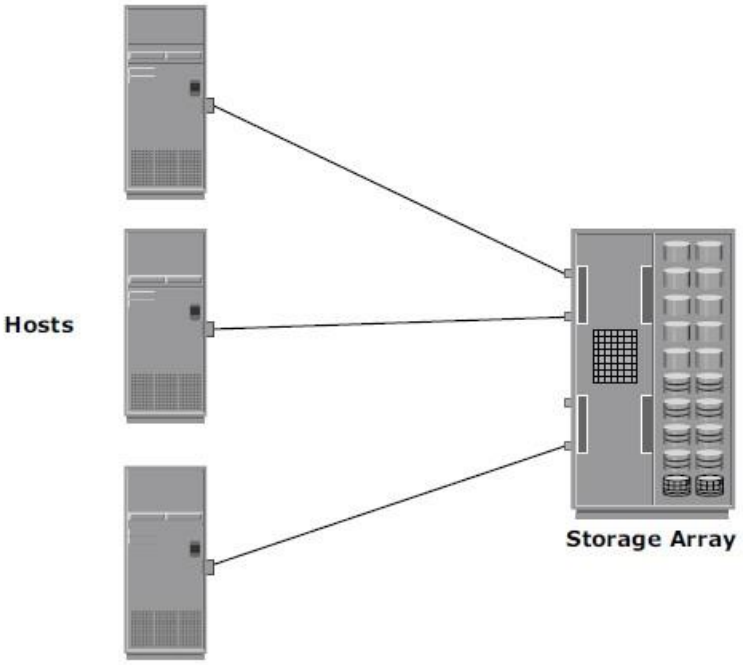


SATA

- Serial Advanced Technology Attachment (SATA)
- Point-to-point connection between the SATA host adapter and the SATA device
  - Half duplex
- New connecting interface
- Higher transmission speed
    - (P)ATA 66/100/133 MB/s
    - SATA 150/300/600 MB/s
- The connecting cable has 4 wires, max. length 1 m

External DAS

- Server is connected directly to an external storage
  - Larger distance
  - Number of devices - typically not (as much) limited
  - SCSI (or FC) connection



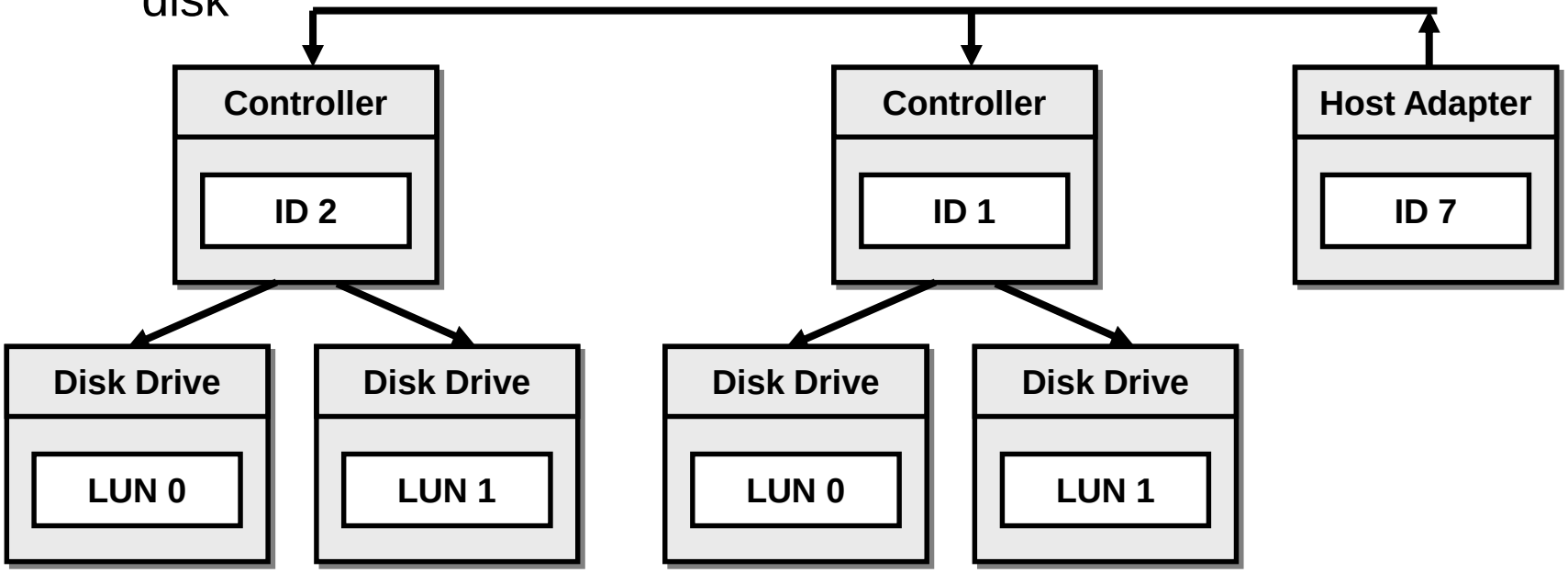
SCSI INTERFACE

- SCSI
  - Small Computer System Interface (SCSI)
  - Standardised I/O bus

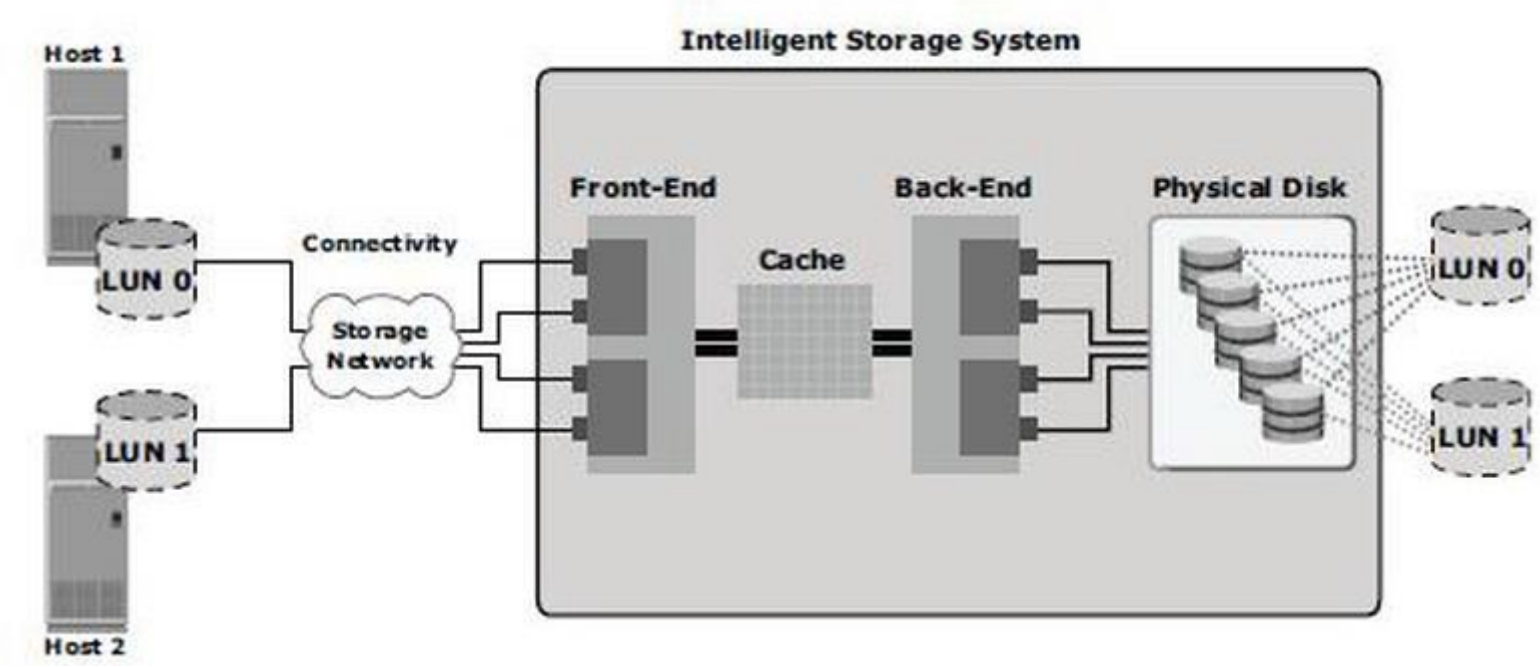
- Devices
  - Disk Drives
  - Tape Drives
  - Removable Media Drives
  - CD/DVD Drives
  - Printers

LOGICAL UNIT NUMBER (LUN)

- NOT(!!!) a number – Group of storages
- LUN is an SCSI group address method
  - Areas on different physical disks can be handled as one virtual disk



LUN


Intelligent Storage System
Host 1
LUN 0
LUN 1
Host 2
Connectivity
Storage Network
Front-End
Cache
Back-End
Physical Disk
LUN 0
LUN 1

SAS - SERIAL ATTACHED SCSI

- Serial Attached SCSI (SAS)
  - Improvement of the parallel SCSI adapter
  - Transmission speed 3, 6 or 12 Gbit/s
  - Full duplex, dual port drives, higher reliability
  - More drives can be addressed from one controller port

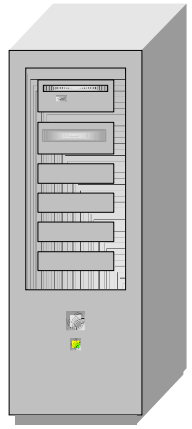
DIRECT-ATTACHED STORAGE (DAS)

Advantages

- Better than to store data on the client
- Limited redundancy
- Low cost, simple

Disadvantages

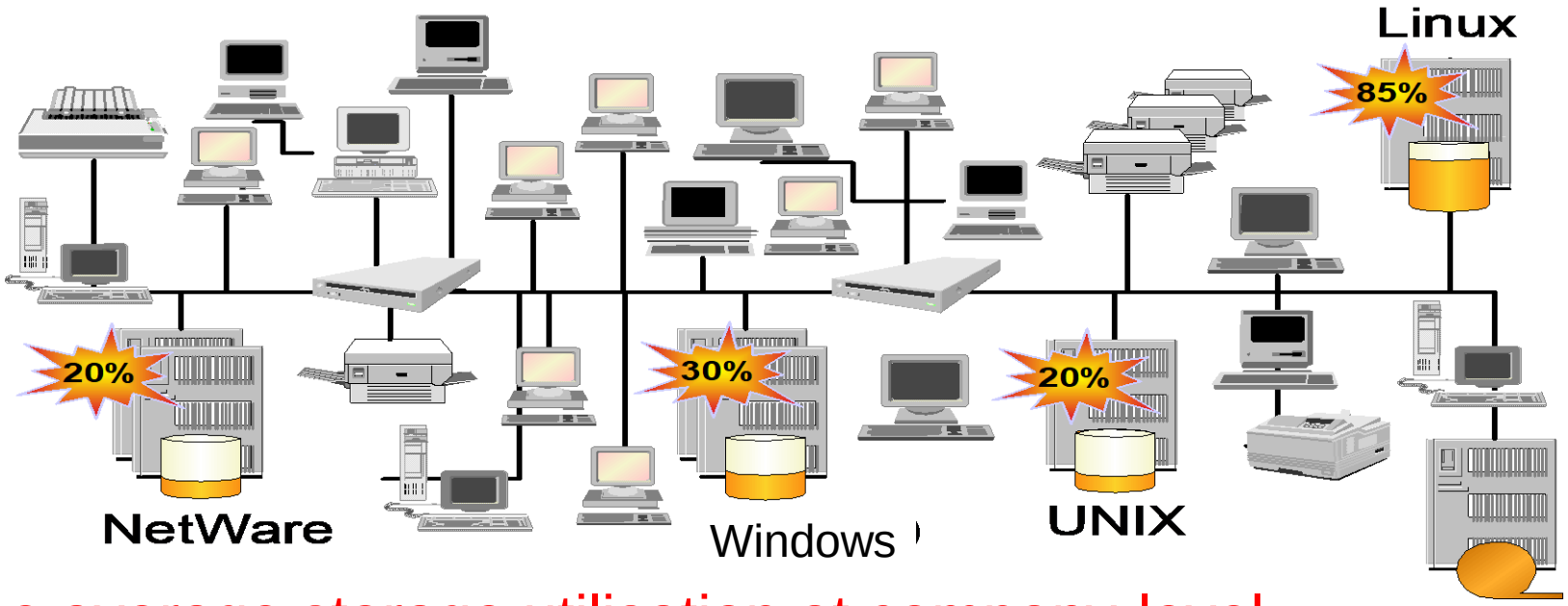
- Difficult management
- Low utilisation
- High cost of back-up
- Difficult data sharing
- Non well scalable
- Limited number of devices



DAS Device

DEDICATED STORAGE DEVICES

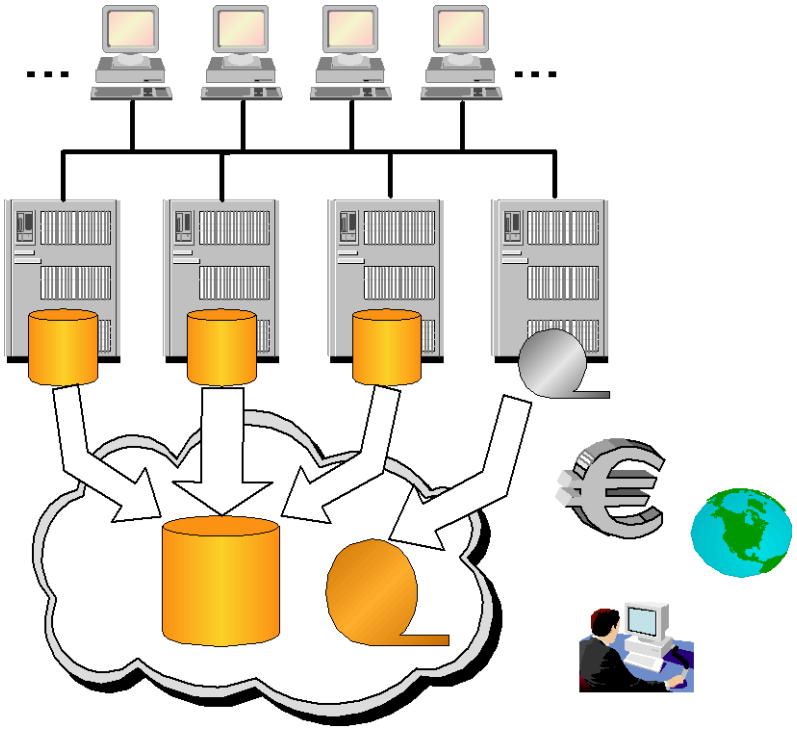
- Separated servers and storages – information islands, under separated management
- Unefficient source utilisation, high costs



The average storage utilisation at company-level is typically only 40-60%

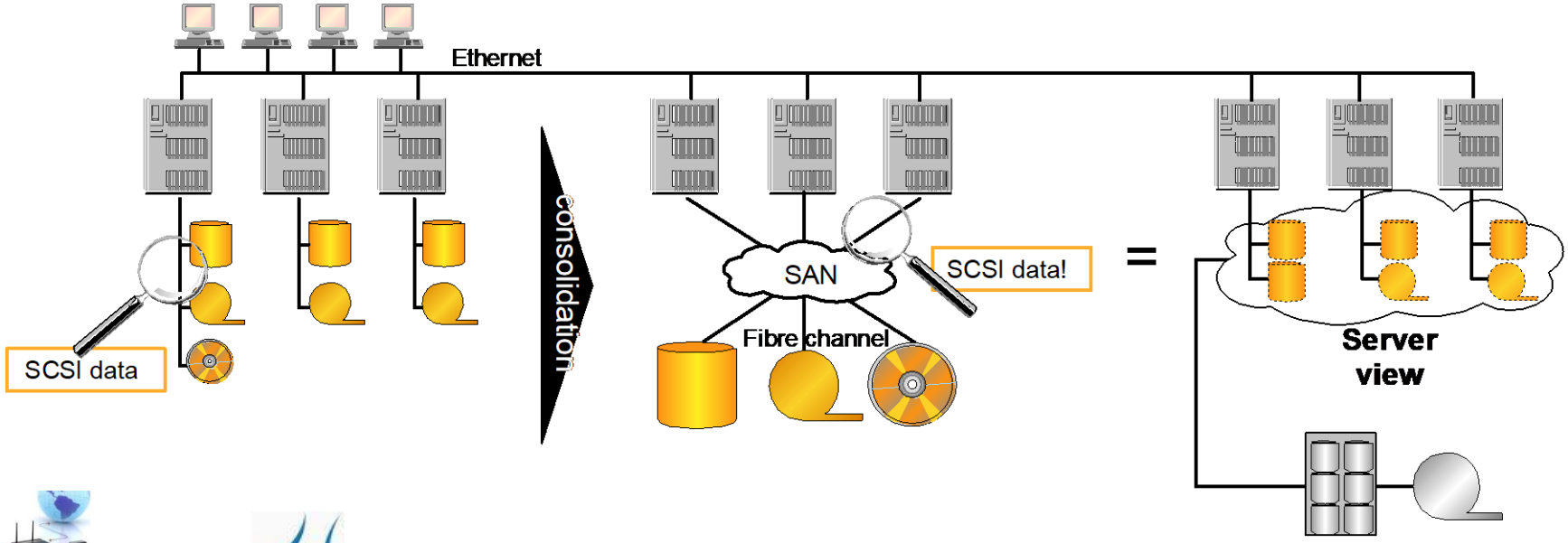
Consolidated Storage Devices

- Consolidated devices, management, data
- Lower complexity, cheaper
- High availability, scalable, disaster tolerate systems can be built



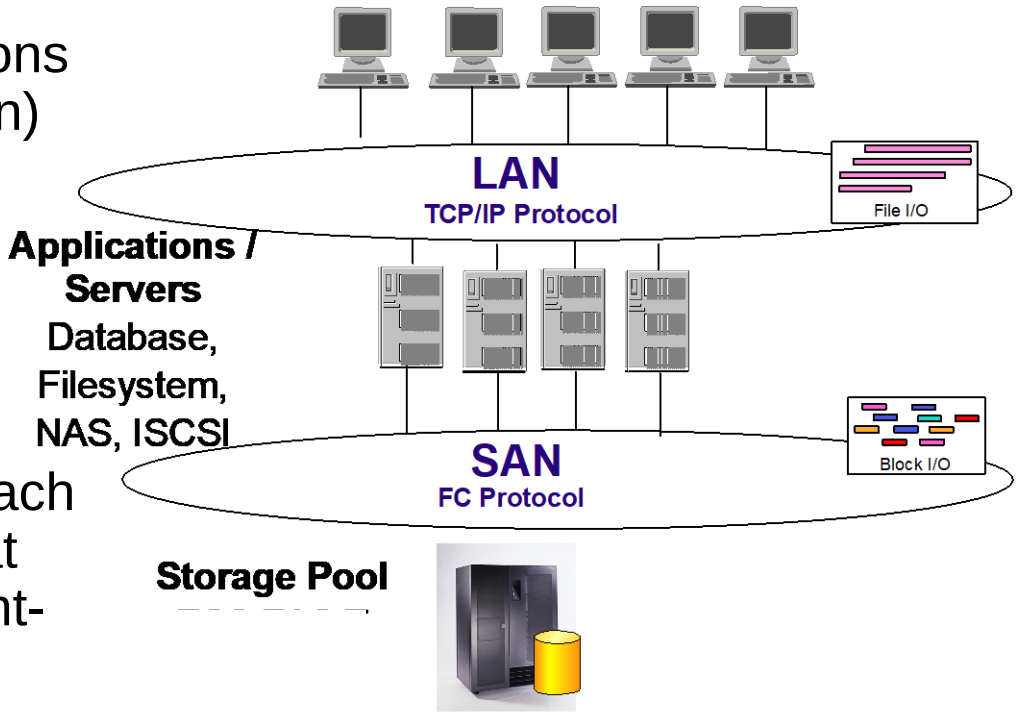
SAN - Storage Area Network

- Network dedicated to data transmission
- The storage devices are physically independent from the servers, more servers can reach the same device
- Data handling protocol not changed, servers can treat them as their own dedicated storage
  - Block access



ADVANTAGES OF SAN NETWORKS

- **Resources**
  - Increase resource efficiency,
  - Scalability
  - Higher level system functions can be installed (replication)
- **Management**
  - higher efficiency
  - higher service levels
- **Information access**
  - Application servers can reach any data on the network, at any time – typically in point-to-point connection

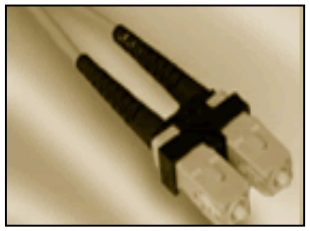


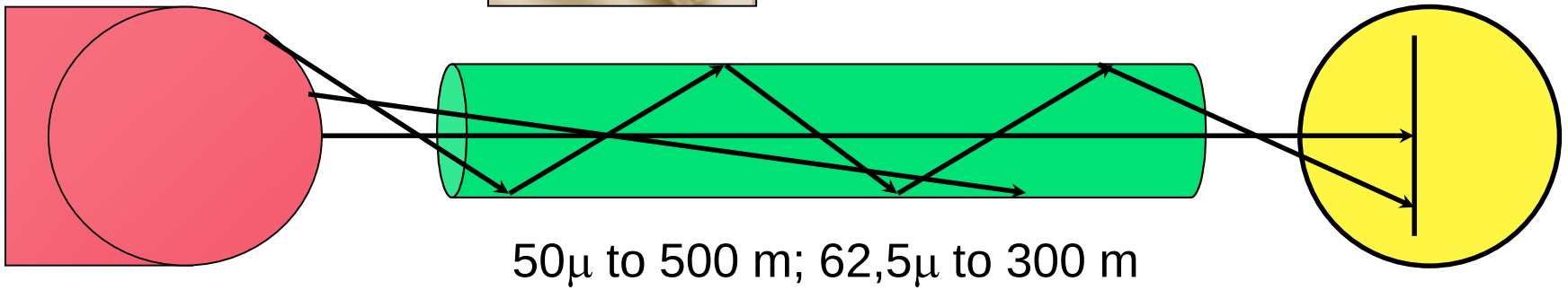
Transmission Technology: Fibre Channel Protocol

- Scalable
  - Large number of devices
  - Long distance
  - Transmission solution for numerous protocols
    - SCSI-3 (remote disks can be accessed in the same way as locals)
    - IP, ATM, …
- Point-to-point or Switched network topology
- Different devices, speeds
  - Different types of copper wire, fiber optic
  - Max. speed: 128 Gb/s

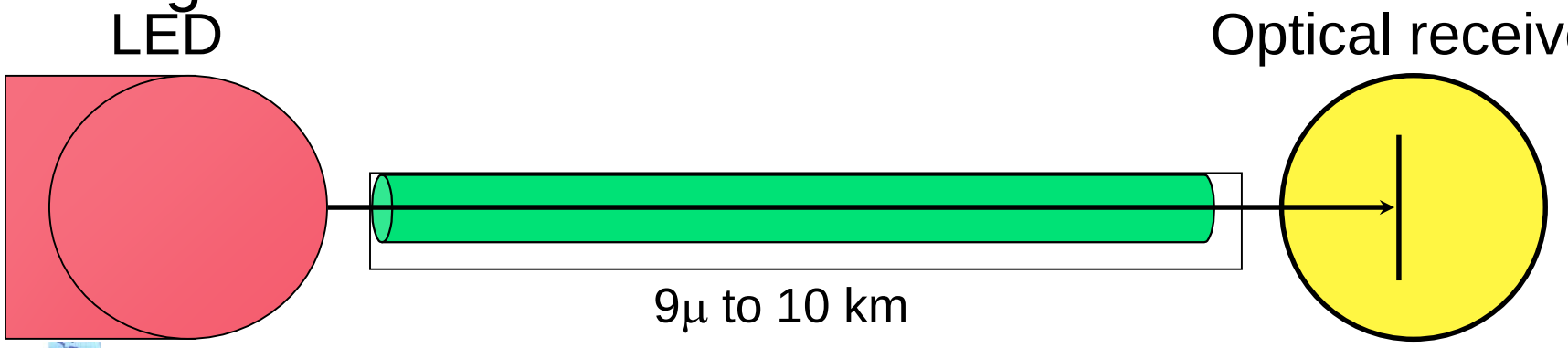
FIBER OPTIC

- Multimode



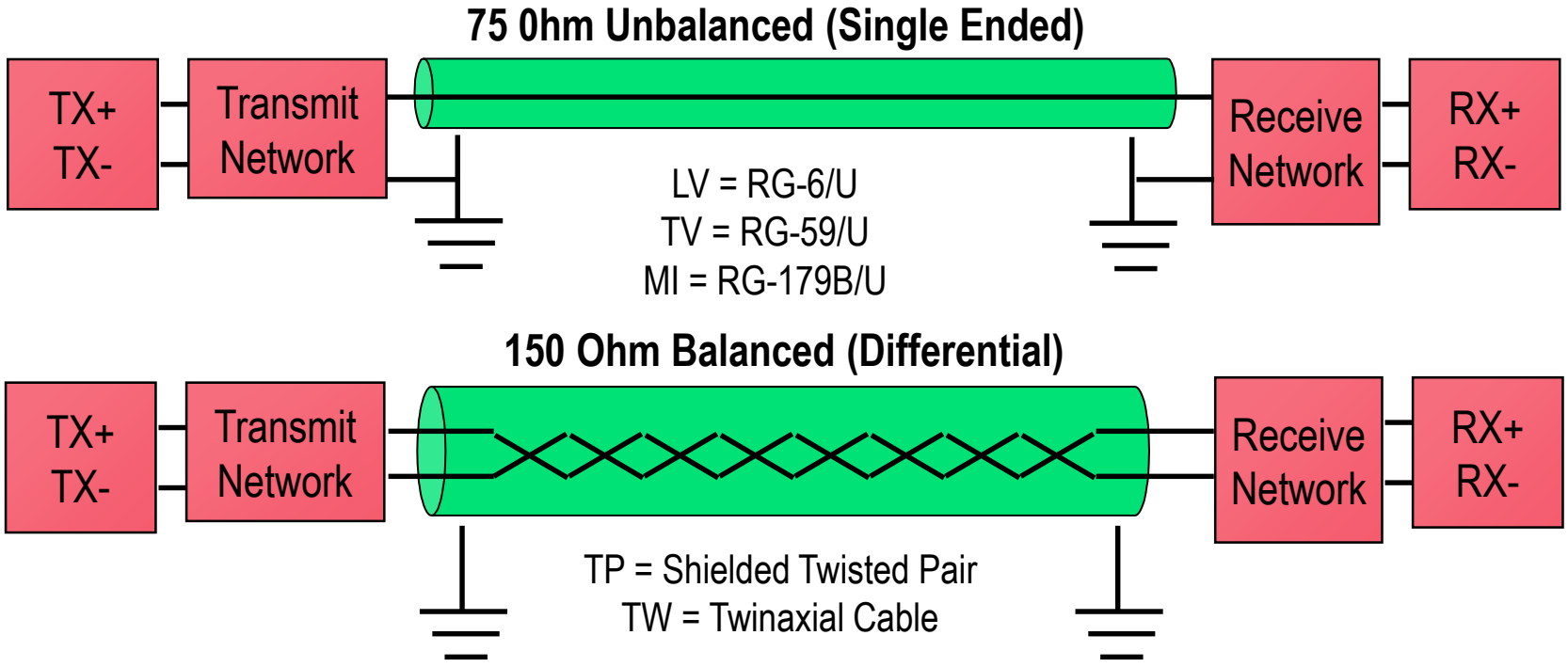


- Single mode



COPPER WIRE

- Mainly at Back-ends
- Max. 15 m
  - Better signal-to-noise ratio than that of the fiber

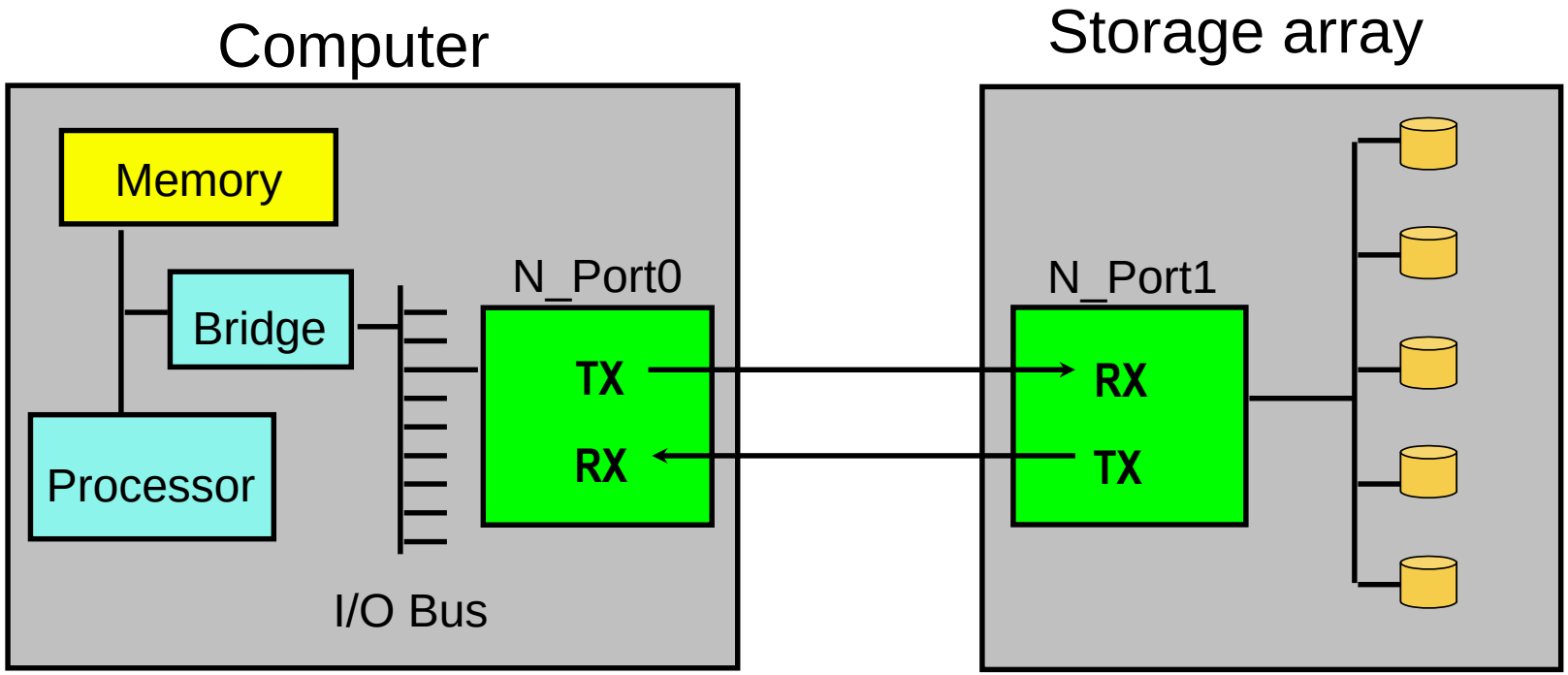


Fibre Channel Protocol

- 3 different topologies
  - Point-to-point
  - Arbitrated loop (not used nowadays)
  - Switched Fabric
- FC interconnects ports
  - Port: any entity communicating over FC, not a physical port
    - N ports (Node)
      - Disk
      - HBA (Host Bus Adapter) in computers
    - F ports
      - FC Switch

POINT-TO-POINT

- DAS
  - SCSI: max. 1,5 GB/s (12 Gb/s)
  - FC: max. 16 GB/s (128 Gb/s)



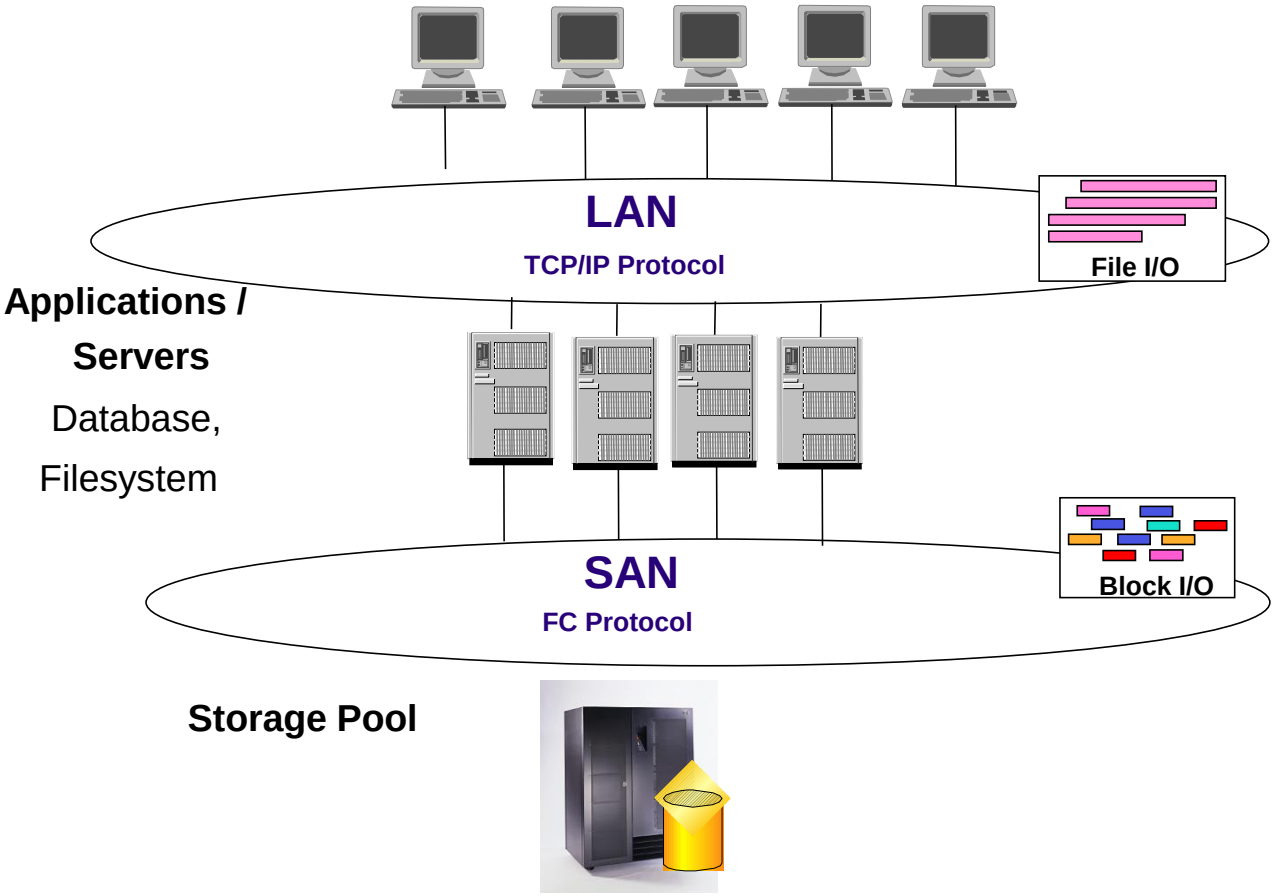
SUMMARY OF HDD TECHNOLOGIES

	SATA III		SAS-3		Fibre Channel
Teljesítmény	Half-duplex	≠	Full-duplex with Link Aggregation	=	Full Duplex
	6 Gb/s		12 Gb/s	≠	128 Gb/s
Interfész	1 m internal cable	≠	10 m internal and external cables		15 m copper cable 500m/10 km optic
	Multipliers 15 HDD max	≠	Expanders >128 devices	=	16 Million (Fabric)
Kialakítás	Single-port HDDs	≠	Dual-port HDDs	=	Dual-port HDDs
	Single-host	≠	Multi-initiator	=	Multi-initiator
Driver sw.	Software transparent with Parallel ATA	≠	Software transparent with Parallel SCSI	=	Software transparent with Parallel SCSI

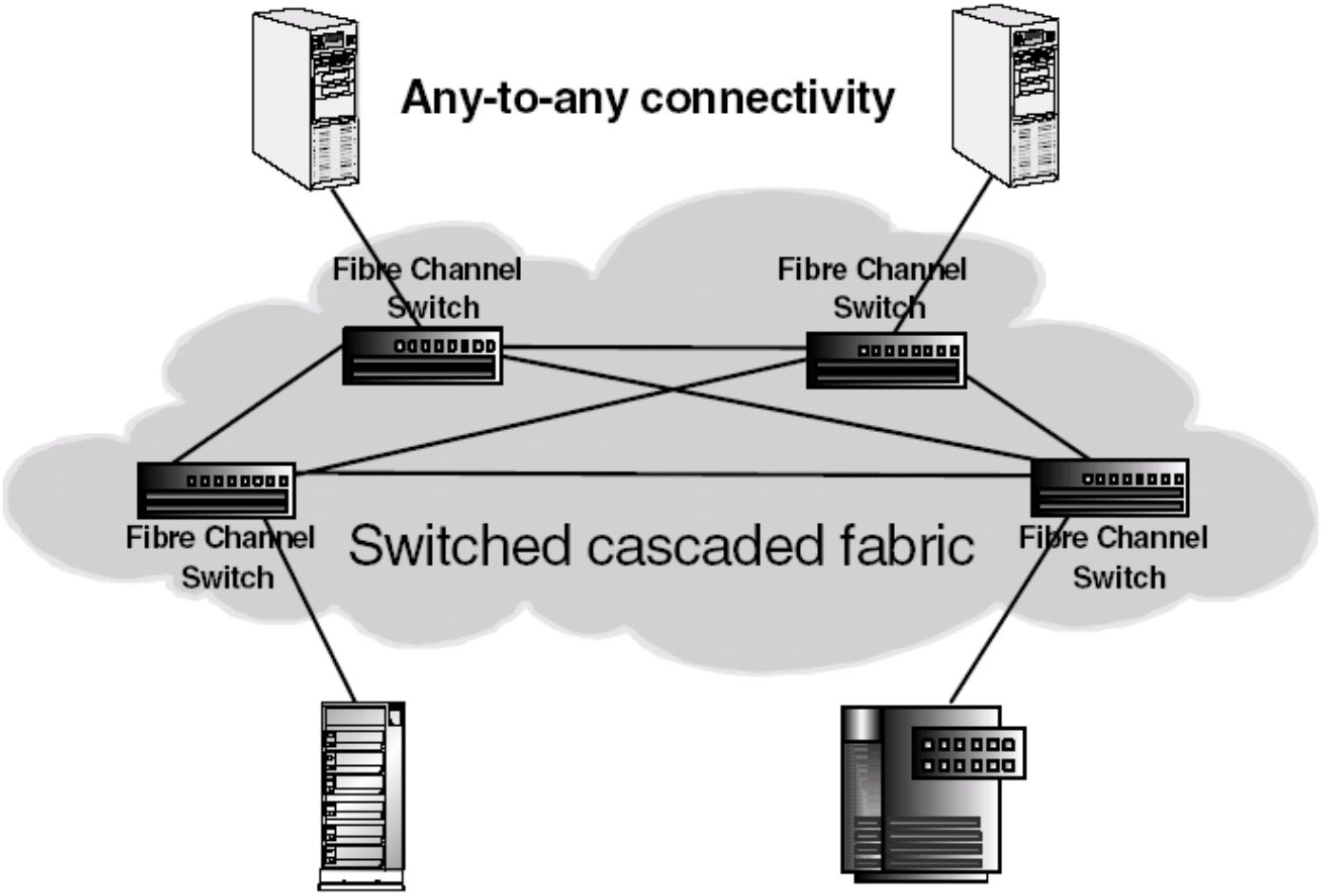
Data Transmission Speed

SCSI 1	**12.0 Mbit/s**	**1.5 MB/s**
Fast SCSI 2	**80 Mbit/s**	**10 MB/s**
Fast Wide SCSI 2	**160 Mbit/s**	**20 MB/s**
Ultra DMA ATA 33	**264 Mbit/s**	**33 MB/s**
Ultra Wide SCSI 40	**320 Mbit/s**	**40 MB/s**
Ultra DMA ATA 66	**528 Mbit/s**	**66 MB/s**
Ultra-2 SCSI 80	**640 Mbit/s**	**80 MB/s**
Ultra DMA ATA 100	**800 Mbit/s**	**100 MB/s**
Ultra DMA ATA 133	**1064 Mbit/s**	**133 MB/s**
Serial ATA I (SATA-150)	**1200 Mbit/s**	**150 MB/s**
Ultra-3 SCSI 160	**1280 Mbit/s**	**160 MB/s**
Fibre Channel	**800 ,1600, 3200 Mbit/s**	**100 ,200, 400 MB/s**
Serial ATA II (SATA-300)	**2400 Mbit/s**	**300 MB/s**
Ultra-320 SCSI	**2560 Mbit/s**	**320 MB/s**
Ultra-640 SCSI	**5120 Mbit/s**	**640 MB/s**
SAS	**3000 -6000 Mbit/s**	**375- 750 MB/s**

SAN (RECALL)

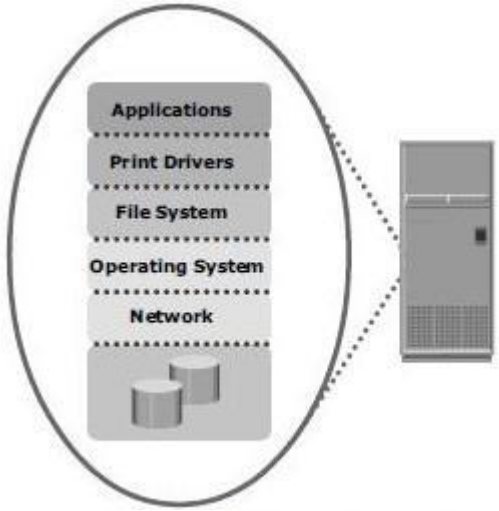

LAN
TCP/IP Protocol
File I/O
Applications / Servers
Database,
Filesystem
SAN
FC Protocol
Block I/O
Storage Pool

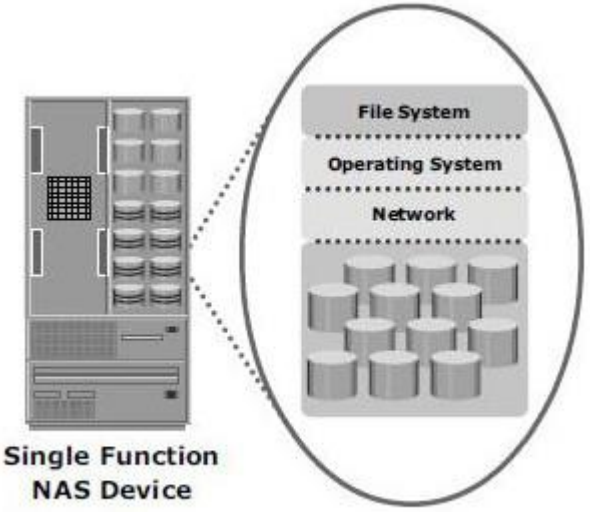
SWITCHED SAN NETWORK



NAS – Network-Attached Storage

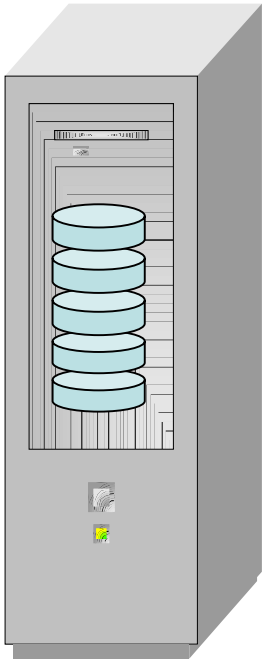
- Disk, which is connected to an IP network
  - Dedicated file server
    - with op. sys. optimised for I/O operations





NAS – Network Attached Storage

- IP (LAN ,WAN) connection, internal SCSI structure
- Internal RAID - error tolerant
- File level access (Block access not supported!!)
- Easy installation
- Scalability inside the device
- Performance limitations (LAN bandwith, protocol overhead)



NAS Protocols

- NFS (Network File System) – UNIX
  - A protocol above UDP, specialised for file operations
- CIFS (Common Internet File System)
  - Op. sys. independent
    - Server
    - Client
  - Above TCP/IP
- FTP (File Transfer Protocol)

IP SAN (iFCP, iSCSI)

- SAN: block level access, FC network
- NAS: file level access, IP network
- IP SAN: block level access, IP network
  - iFCP (Internet Fibre Channel Protocol)
    - Congestion ctrl, error detection&recovery by TCP
  - iSCSI (Internet Small Computer System Interface)
    - SCSI commands over existing LAN/VLAN/WAN
    - Emulates the SCSI bus over IP networks
    - Though any SCSI device can be connected, typically used for server <-> data storage

IP SAN

- For storage consolidation without the need of a dedicated network
  - Low-cost equivalent of Fibre Channel, but performance highly depends on ‘other’ traffic
- For disaster recovery
  - To mirror storages between (remote) data centers
    - As ‘hot stand-by’
    - Over WAN

SAN OR NAS SUMMARY 1.

SAN (Storage Area Network)

Centralized, high performance network, dedicated exclusively for data storage

- Interconnects servers and data storage devices,
- Contains network and switching elementes and supporting software solutions

Advantages:

- Scalable, extendable
- High data transmission speed (4 -10 Gb/s)
- Maintenance: can be centralized; supports hierarchial storage – BUT in case of a complex, heterogenous network: complicated management

SAN OR NAS SUMMARY 2.

NAS (Network Attached Storage)

Data storage device connected to the network; supports data sharing among servers and clients

Advantages:

- Scalable, extendable, but limited bandwidth (LAN)
- Easy to install and maintain
- Typically in small/medium environments

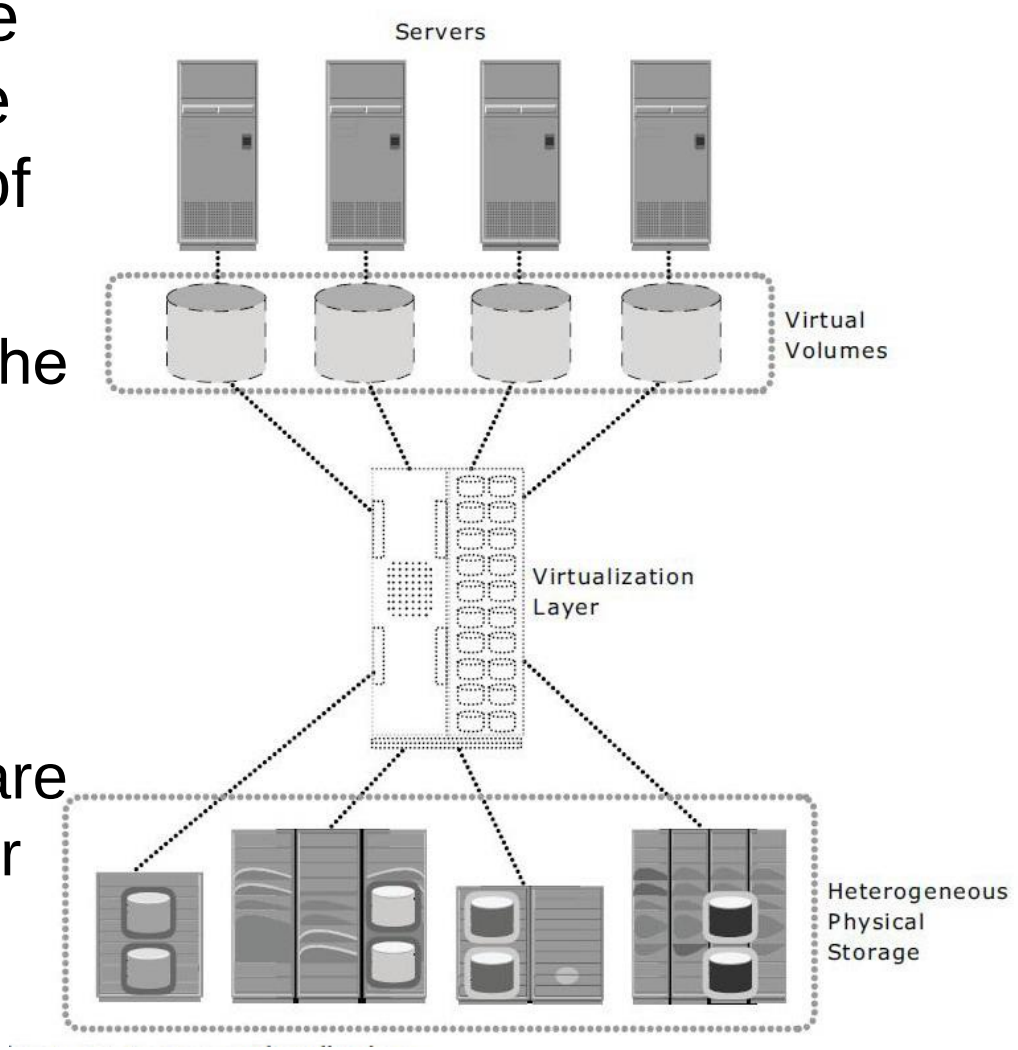
Disadvantage: speed, shared with other traffic

VIRTUALISATION

- **Storage virtualisation** refers to the process of abstracting logical storage from physical storage. The term is today used to describe this abstraction at any layer in the storage software and hardware stack
- The virtualization can be realized in a different system levels

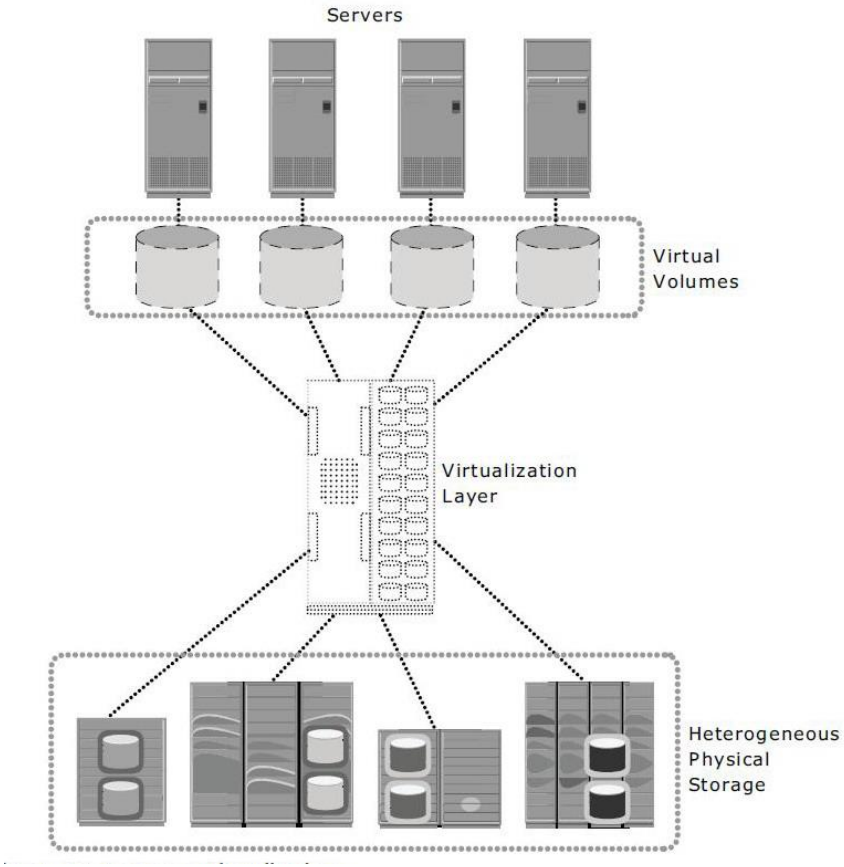
STORAGE VIRTUALISATION

- Virtualisation Appliance or Virtualisation Engine hides the differencies of the disks
  - „translates” between the two formats
  - disks can be shared
    - better utilisation
  - replacement or modification of disks are not seen for the server



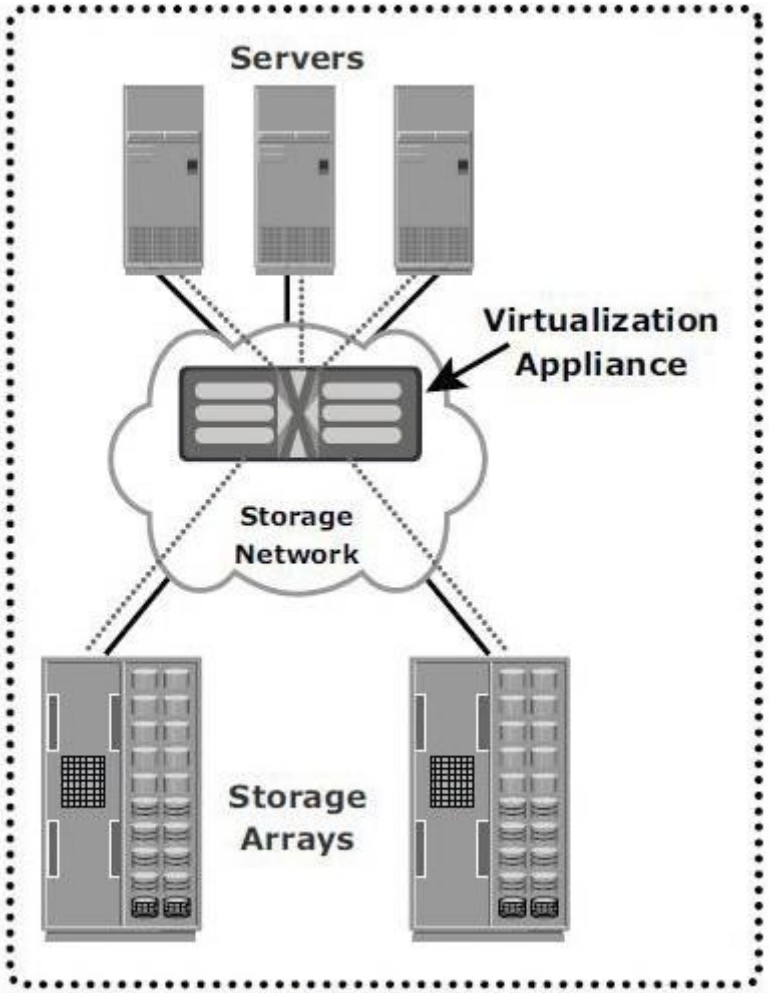
VIRTUALISATION APPLIANCE

- Meta-data
  - mapping table between physical and logical addresses
- Server: LUN=1, LBA=32
  - LBA Logical Block Address
- VM: from table, this corresponds to the physical LUN=4, LBA=0
- Requests the data from physical disk
- Data is transmitted to server as if it came from LUN=1, LBA=32



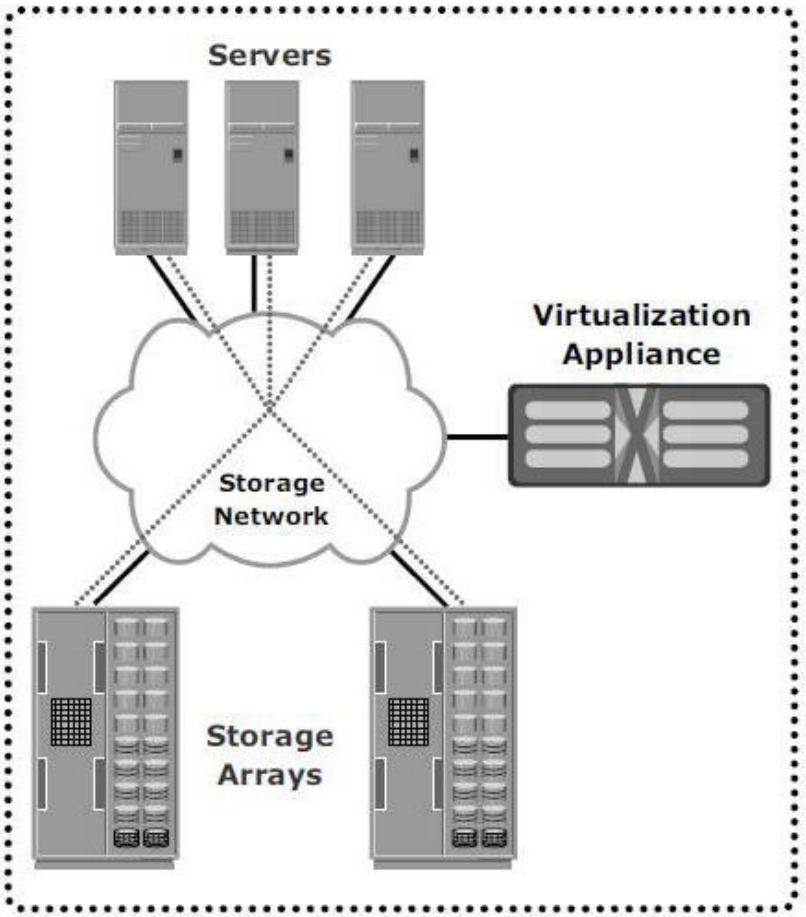
VIRTUALISATION CONFIGURATION – IN-BAND

- Virtualisation Appliance in data path
  - no need for a special software on server
  - but slower since data goes through the Virtualisation Appliance



VIRTUALISATION CONFIGURATION – OUT-OF-BAND

- Separate control and data paths
  - Special software on server:
    - first asks the physical location of the data from Virtualisation Appliance
    - then reaches the data directly
  - Faster data transfer since no additional layer in data path



VIRTUALISATION

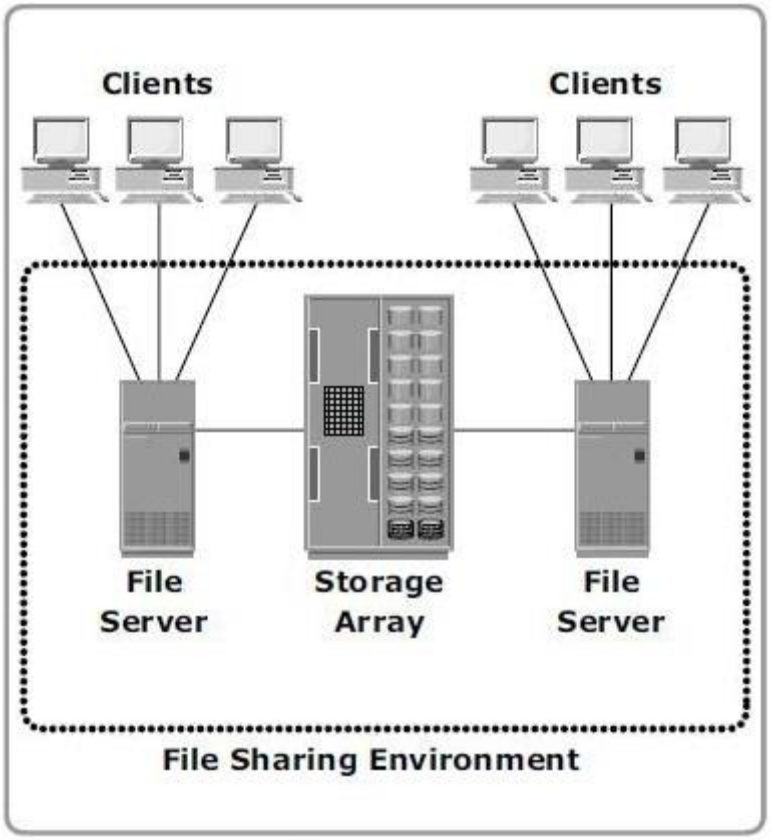
Real sources …	Virtual sources…
• Disks (different types, different vendors)	• Virtual LUNs, they seem as same type of same vendor
• Tipically fix sized	• Size can be modified dinamically
• Different vendor configurations & back-up services	• Centralised management and services
• Migration to new technologies is difficult	• Migration without disturbing the applications

Levels of Virtualisation

- Block level virtualisation
  - discussed till this point
  - *server* wants to have an access to a *data block*
  - knows its (logical) address
- File level virtualisation
  - a *client* wants to have an access to a *file* on a file server
  - must know on which

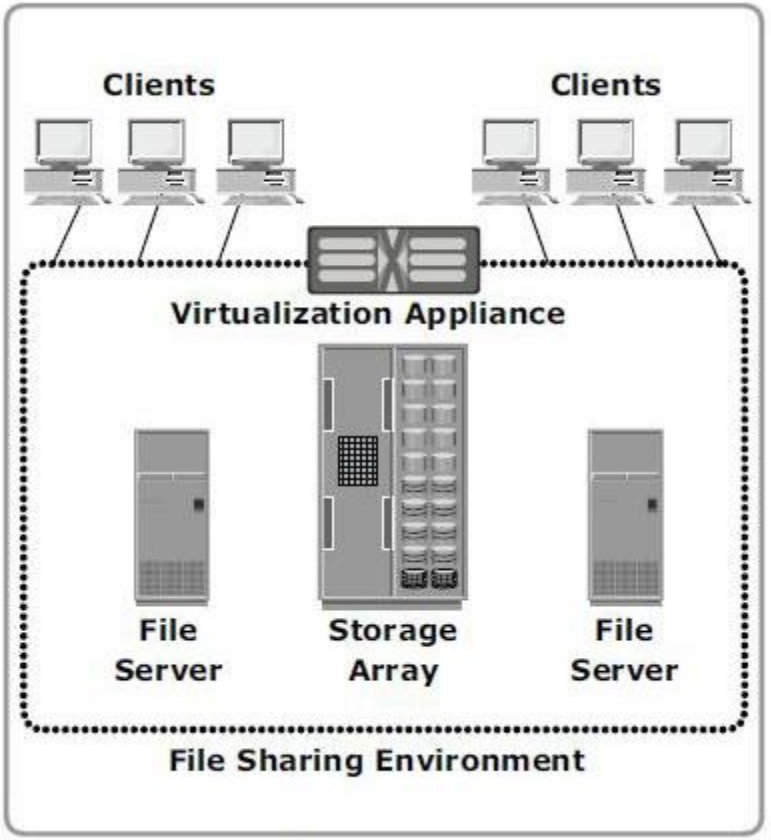
File Level Virtualisation

- Without File Level Virtualisation
  - if a client/host wants to have an access to a file on a file server
  - must know on which
  - one server may be empty while other full
  - file movement affects the client, too



File Level Virtualisation

- Virtualised file server
  - client does not have to know on which server the file is
  - simpler
    - load sharing
    - file movement
    - extension
- Cloud computing



Uploading Data to Cloud

- Time…
  - 1 PB on 1 MB/s line
    - ~ 32 years
- On pen-drive/HDD + transport
- 100 petabyte
  - Film archive of a film studio
  - NASA satellite pictures
  - 2000 years of mp3
  - 200x the genom of all humans