

MANAGEMENT OF INFORMATION SYSTEMS

BME VIK TMIT
SOFTWARE ENGINEERING, BSc



BME VIK TMIT

MANAGEMENT OF INFORMATION SYSTEMS

5. DATA MANAGEMENT IN INDUSTRIAL ENVIRONMENTS



BME VIK TMIT

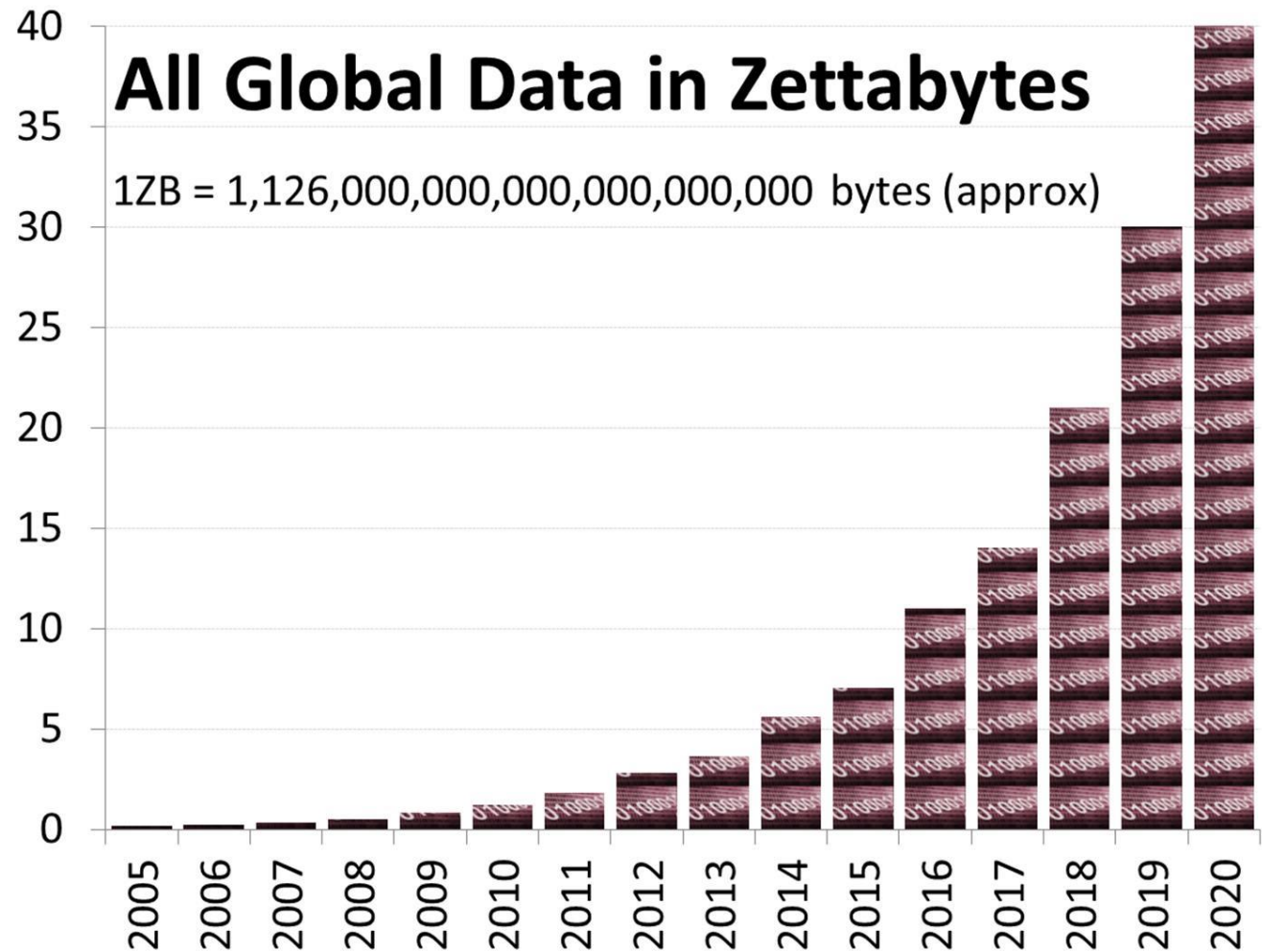
DATA MANAGEMENT

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

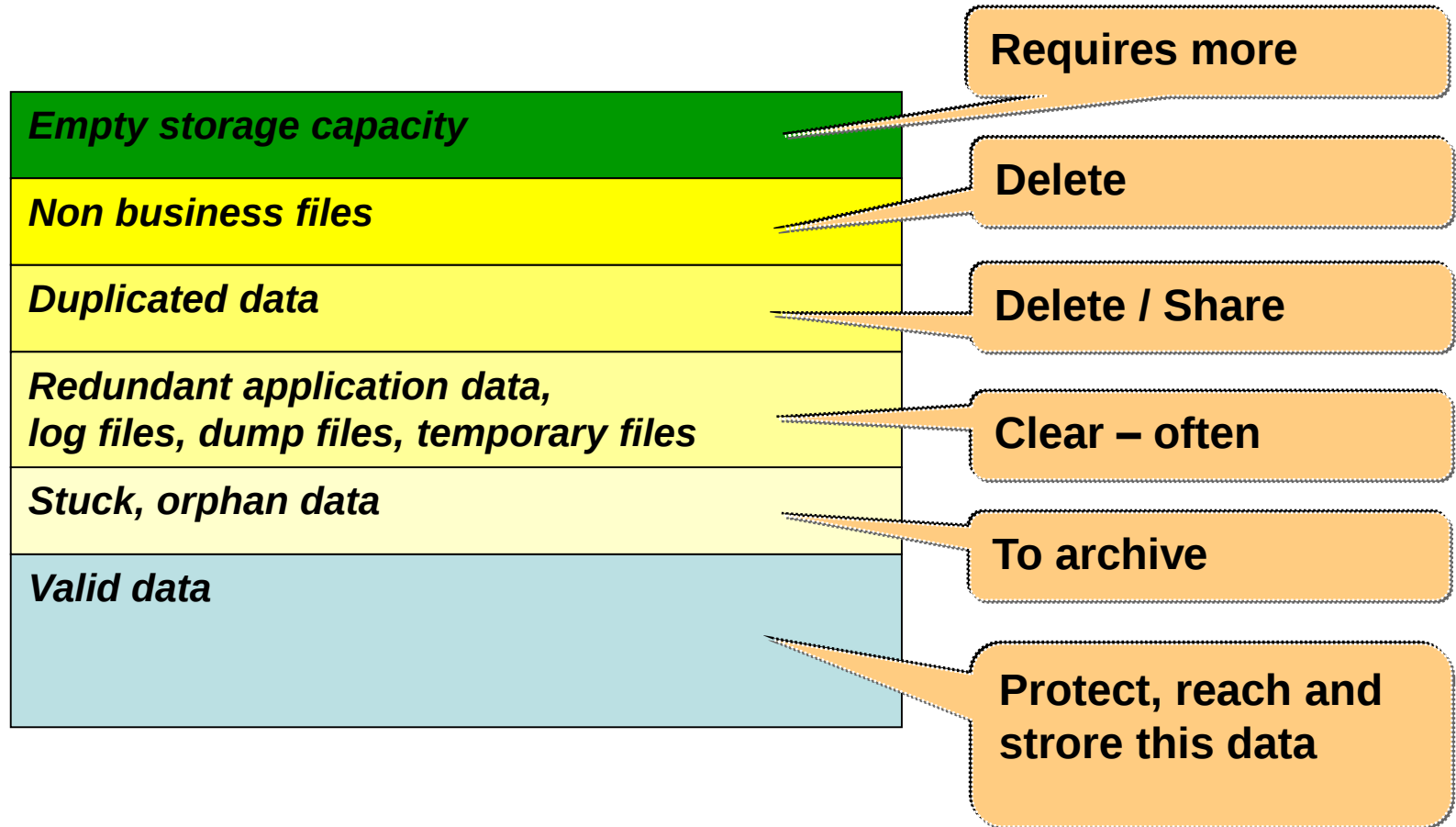


DATA GENERATED

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



DATA TYPES



DATA STREAM TYPES

Structured

Known source and goal

Server to Server

Business devices

Special applications

Strong security

Private networks

Can be estimated

Database

**It can be estimated, and
planned**

Unstructured

Unknown

Client to client, Client to Server

Personal devices

E-mail, Web

Weaker security

Public networks

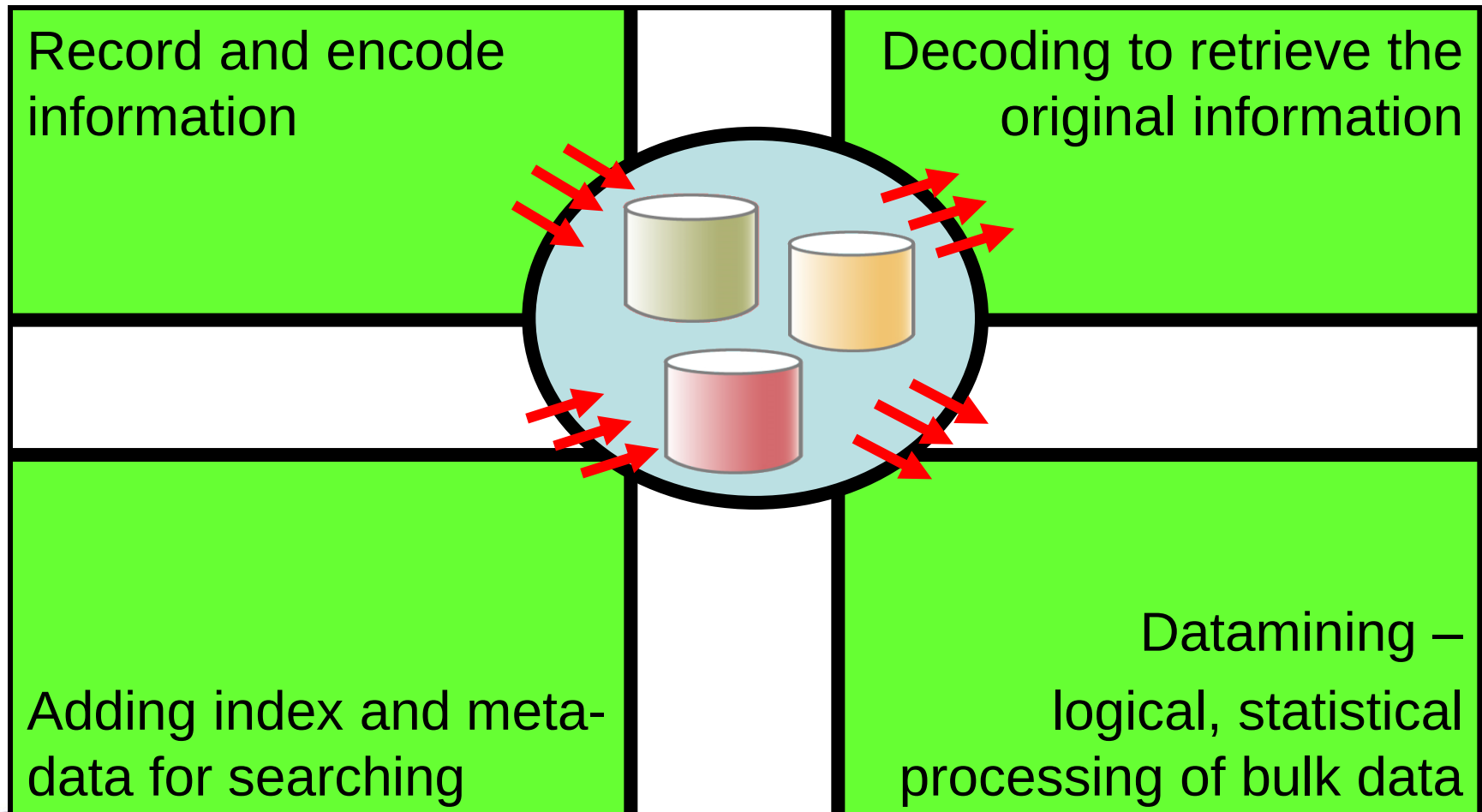
Hard to estimate, statistic

File servers

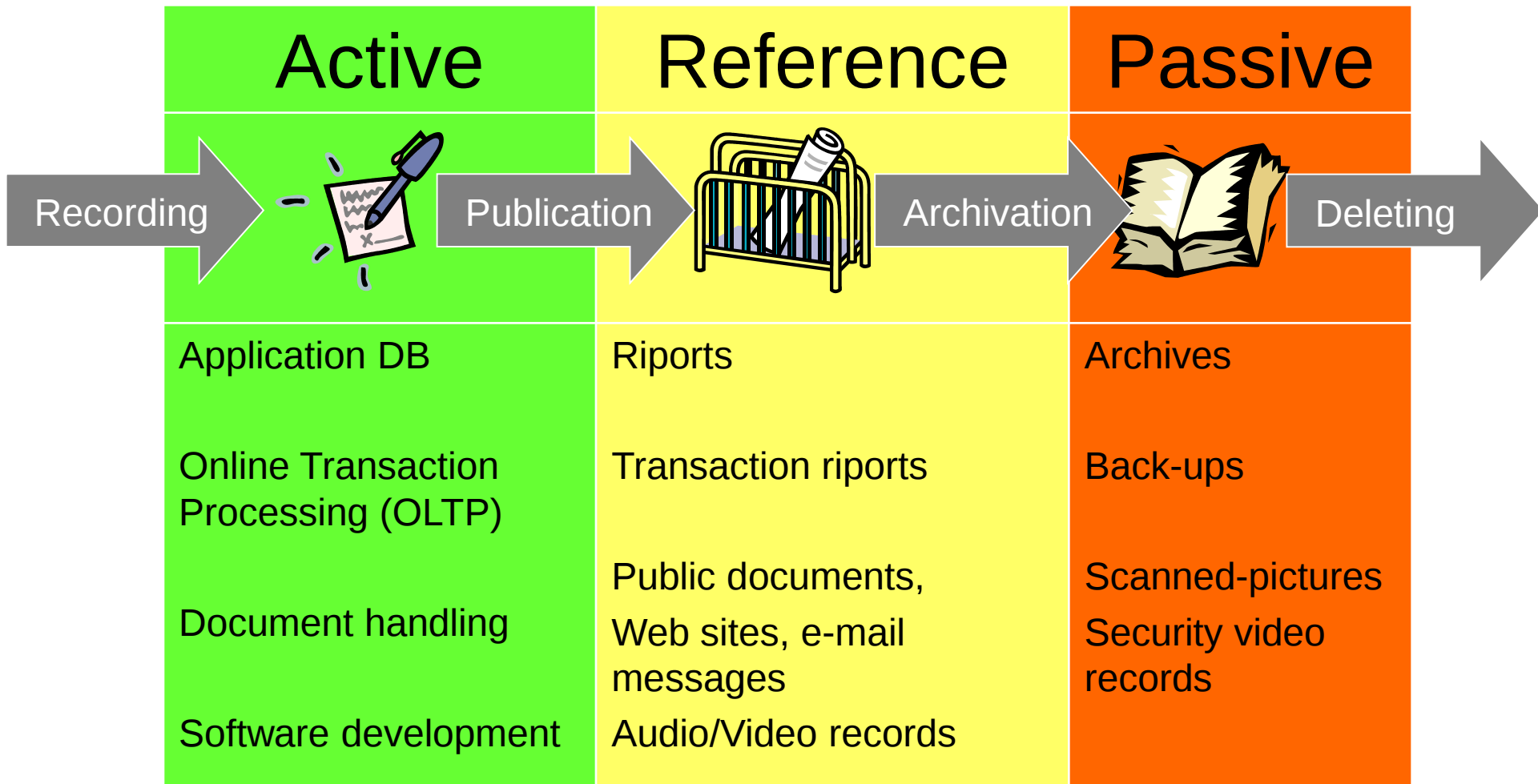
It can only be regulated by policies



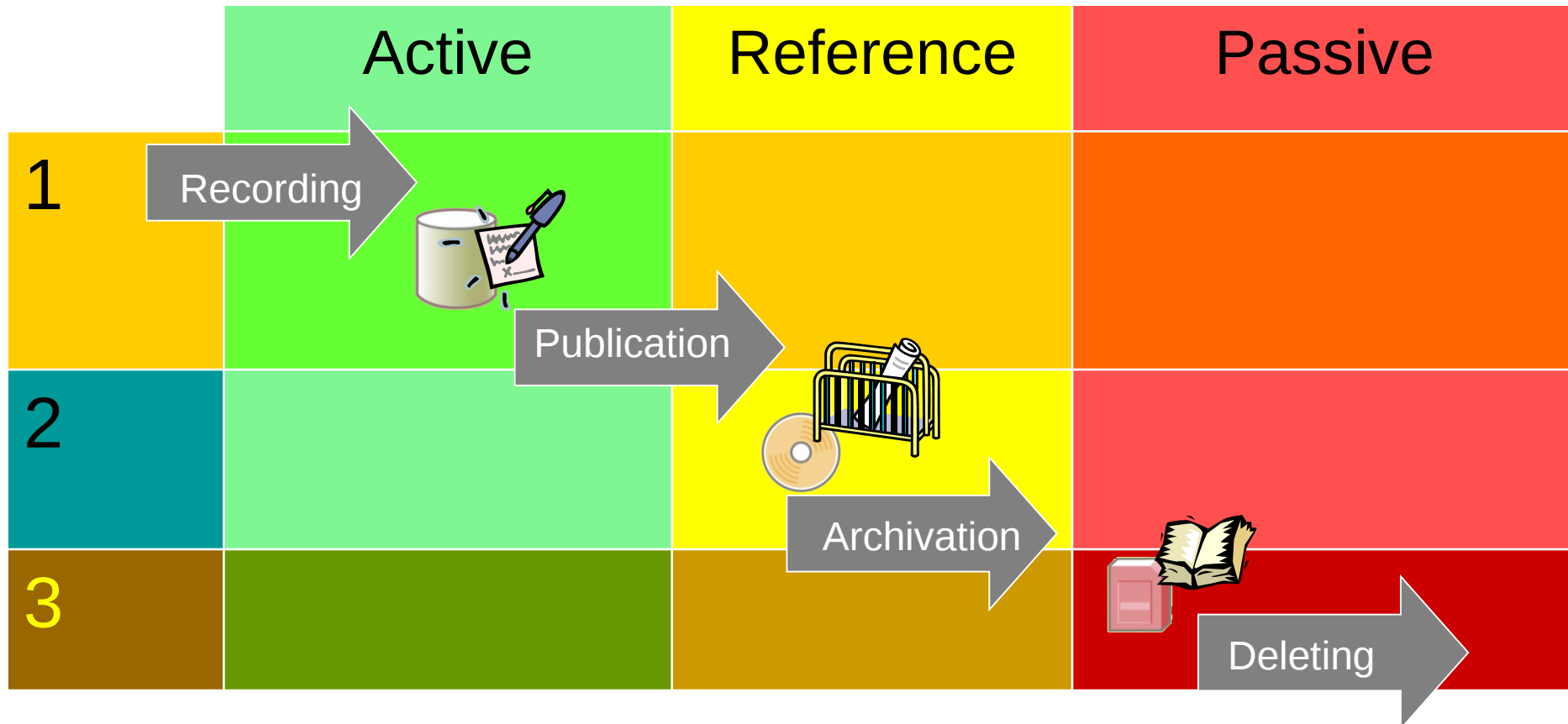
INFORMATION AND DATA LIFE CYCLE, DATA TYPES



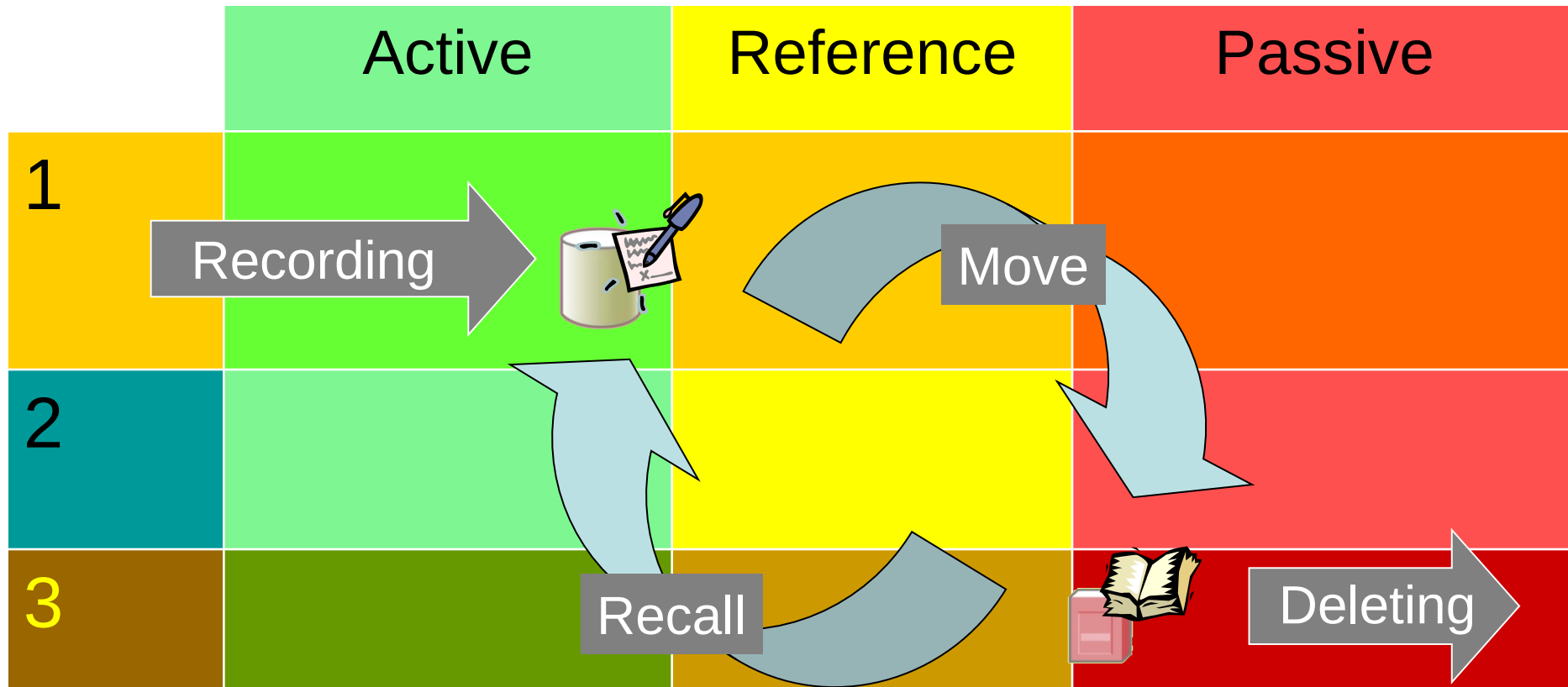
IMPORTANCY OF DATA - CHANGING



OPTIMAL STORAGE TECHNOLOGY: SELECT ACCORING TO DATA VALUE



HIERARCHICAL 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 3-4 years! • Moving parts • Disk replacement
Flash memory 	• EEPROM, pen drive, SSD • No moving parts • Fast	• (Yet) expensive • Small(er)
Optical 	• Secondary storage – WORM (<i>Write Once Read Many</i>)	• Cannot keep up with the development of disk and tape • SOHO device (<i>Small Office Home Office</i>)
Tape 	• 10-20x cheaper than the disks • Lifetime: 30+ years	• Non-immediate data access • Serial read/write (R/W)



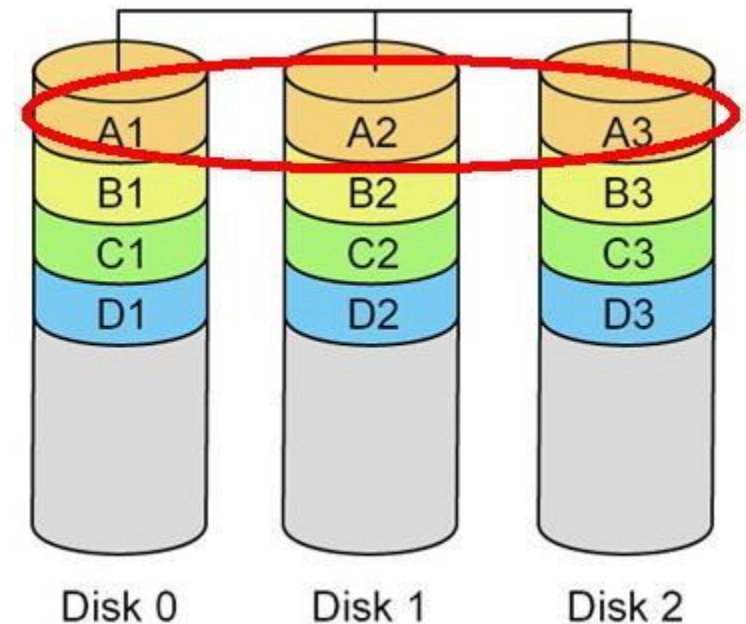
RELIABILITY OF DISKS

- Data stored on disks – protect from damage
- 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: $750\,000 \text{ hours} / 1000 = \text{appr. 1 month}$



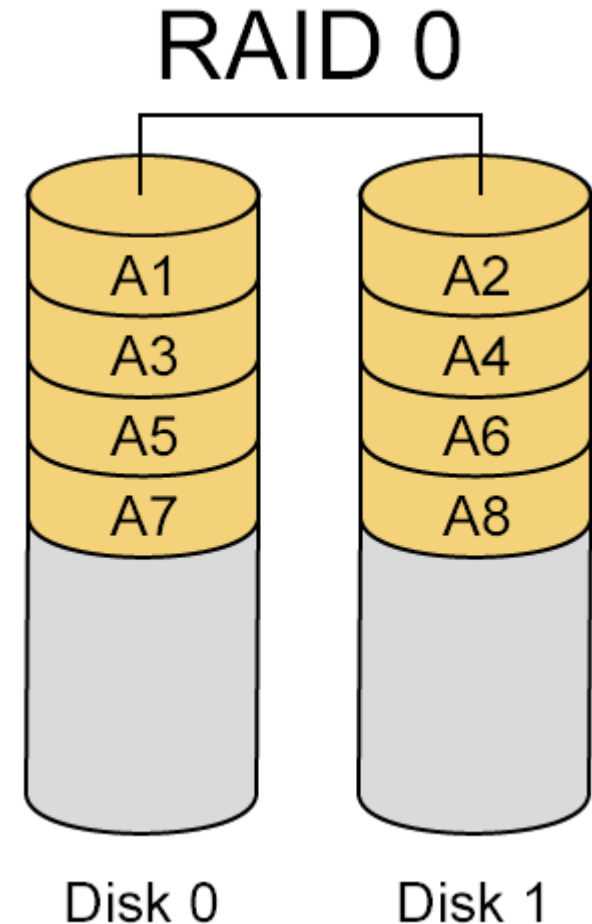
RAID – REDUNDANT ARRAY OF INDEPENDENT (INEXPENSIVE) DISKS

- 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



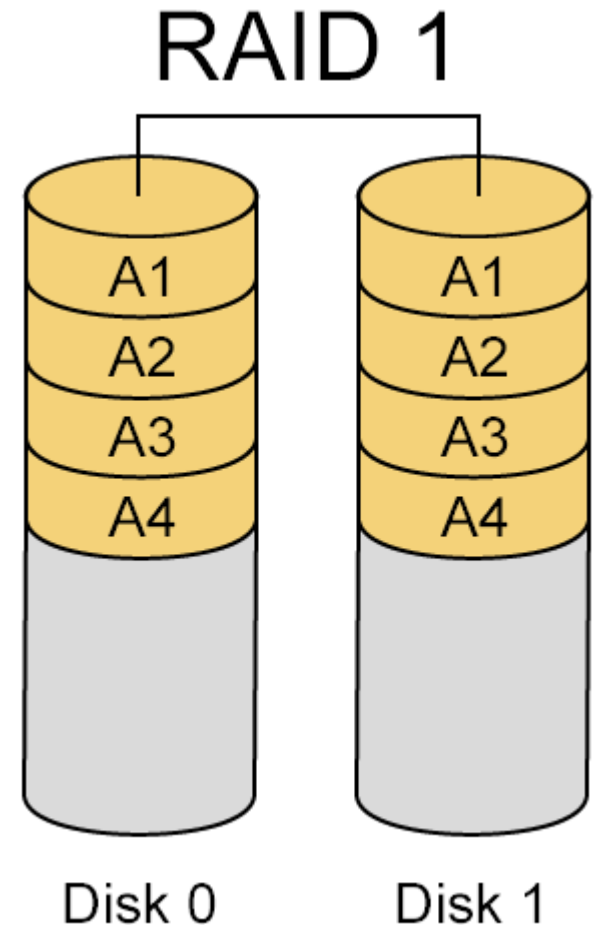
RAID 0 - STRIPING

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



RAID 1 – DISK MIRRORING

- Parallel operations
- High reliability
- Large overhead in size (2x!)
- Can we have approximately the same reliability with smaller overhead?



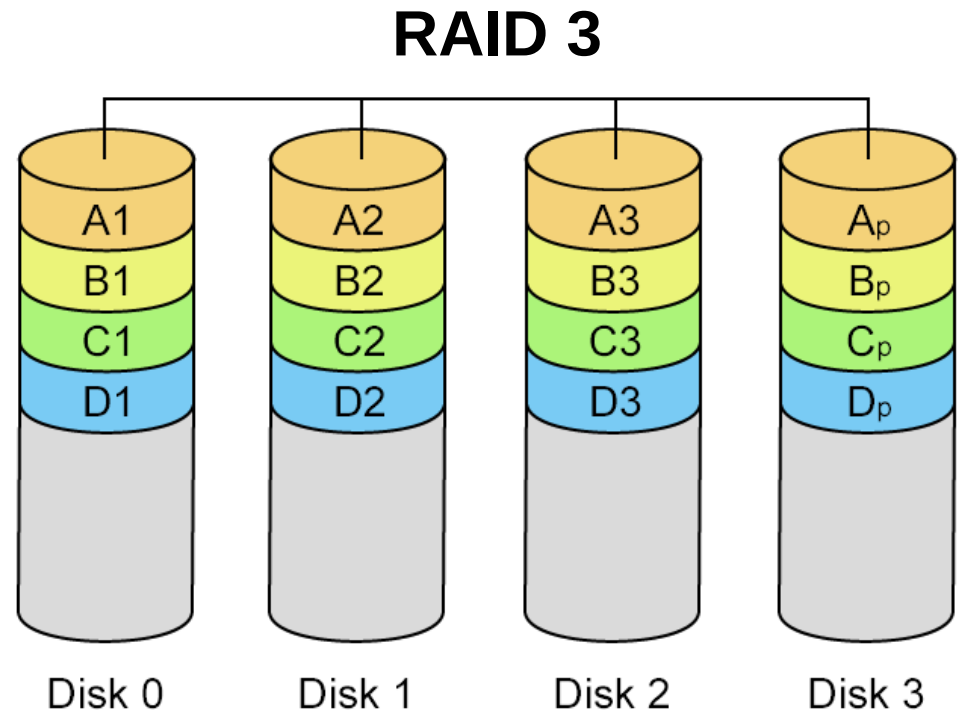
RAID 2 – ERROR CORRECTING CODE

- 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



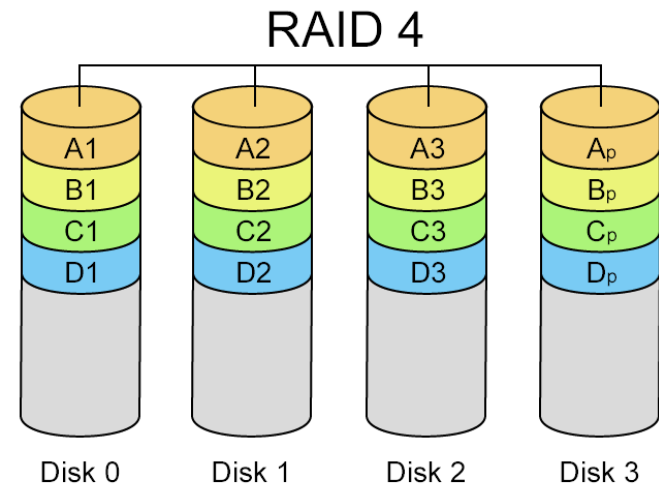
RAID 3 – PARITY DISK

- +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



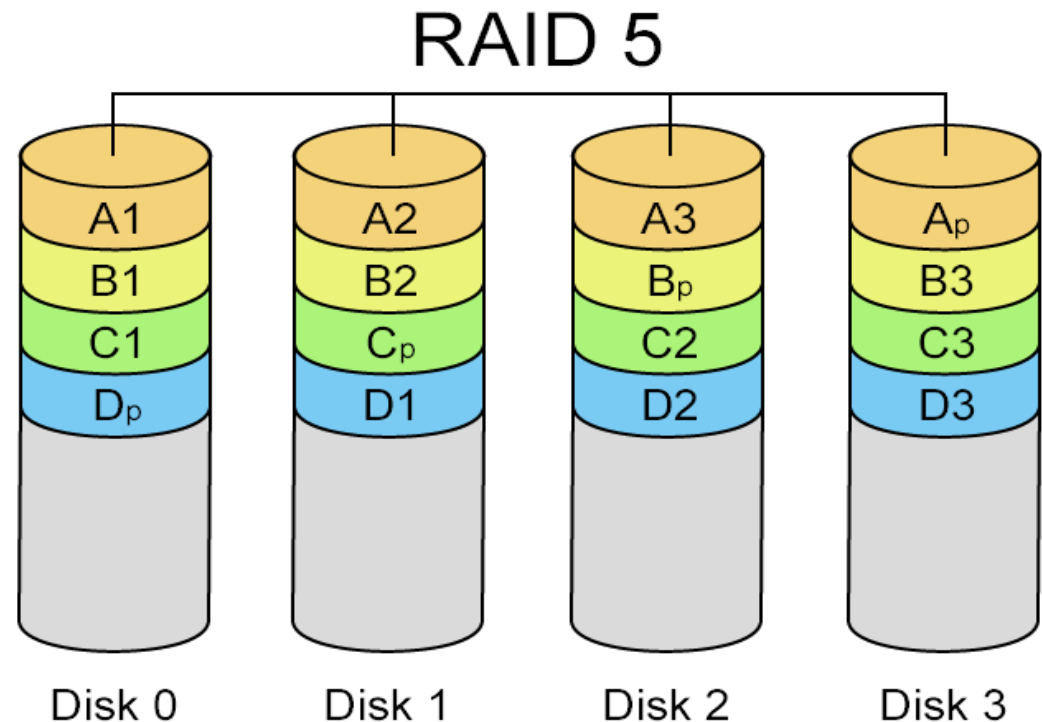
RAID 4

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



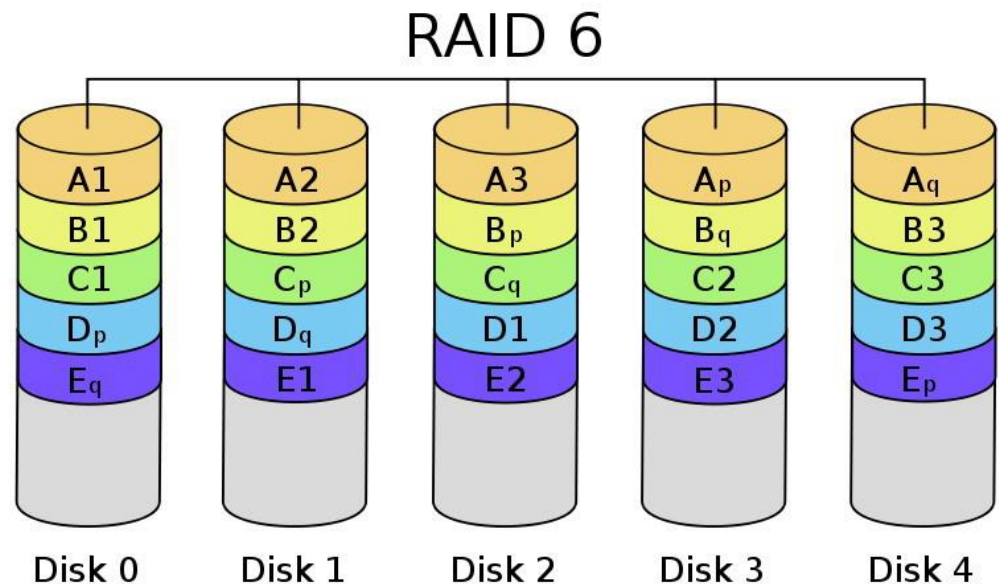
RAID 5 – DISTRIBUTED PARITY

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

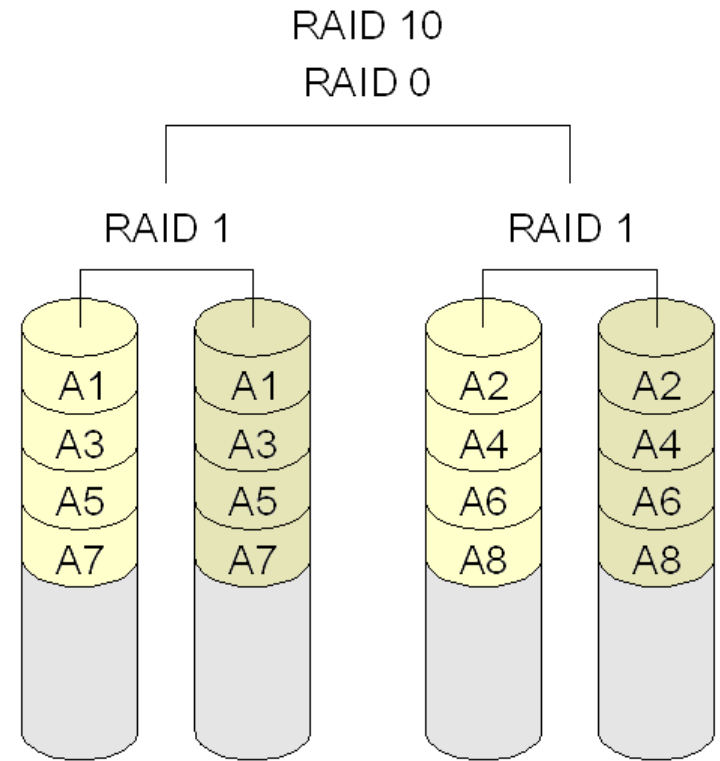
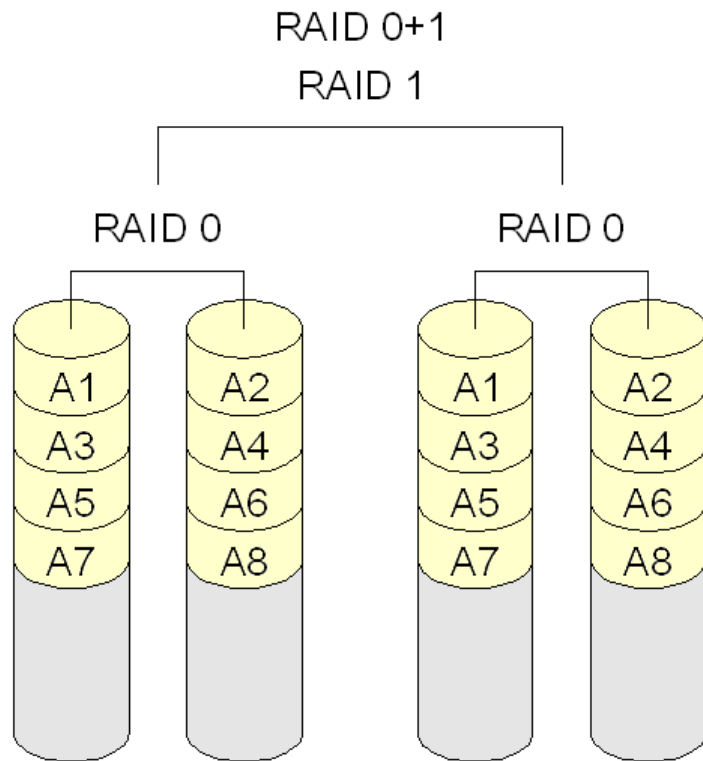


RAID 6 – DOUBLE PARITY

- Row (XOR – P) and Column (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



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

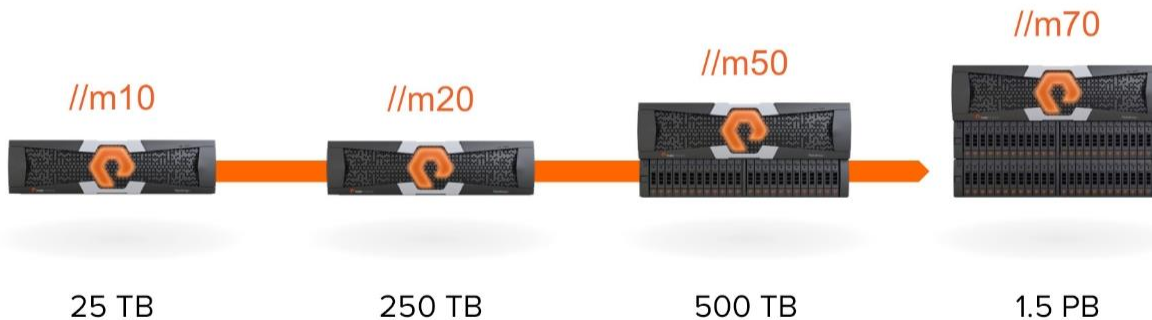




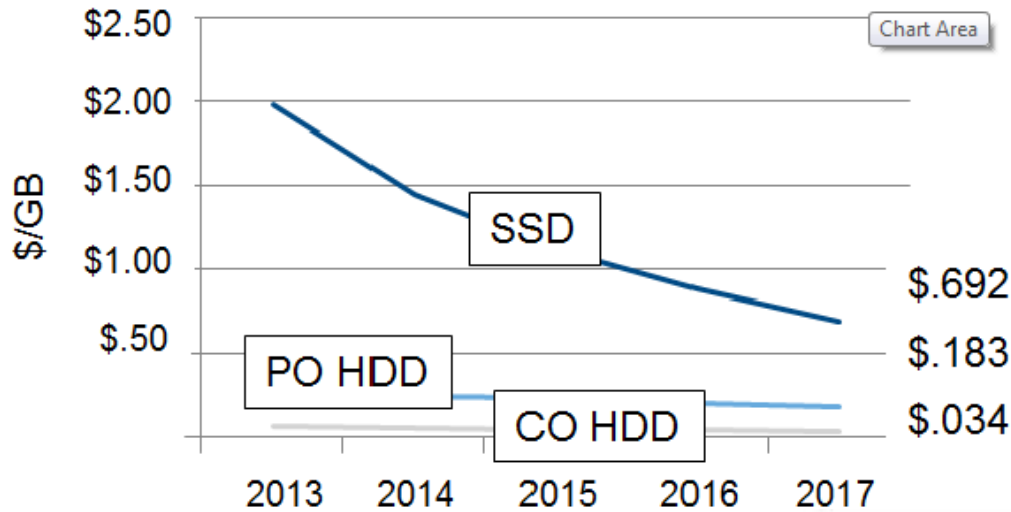
ALL-FLASH ARRAY



ORACLE[®]
SAP[®] Microsoft SQL Server



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
 - More satisfied users

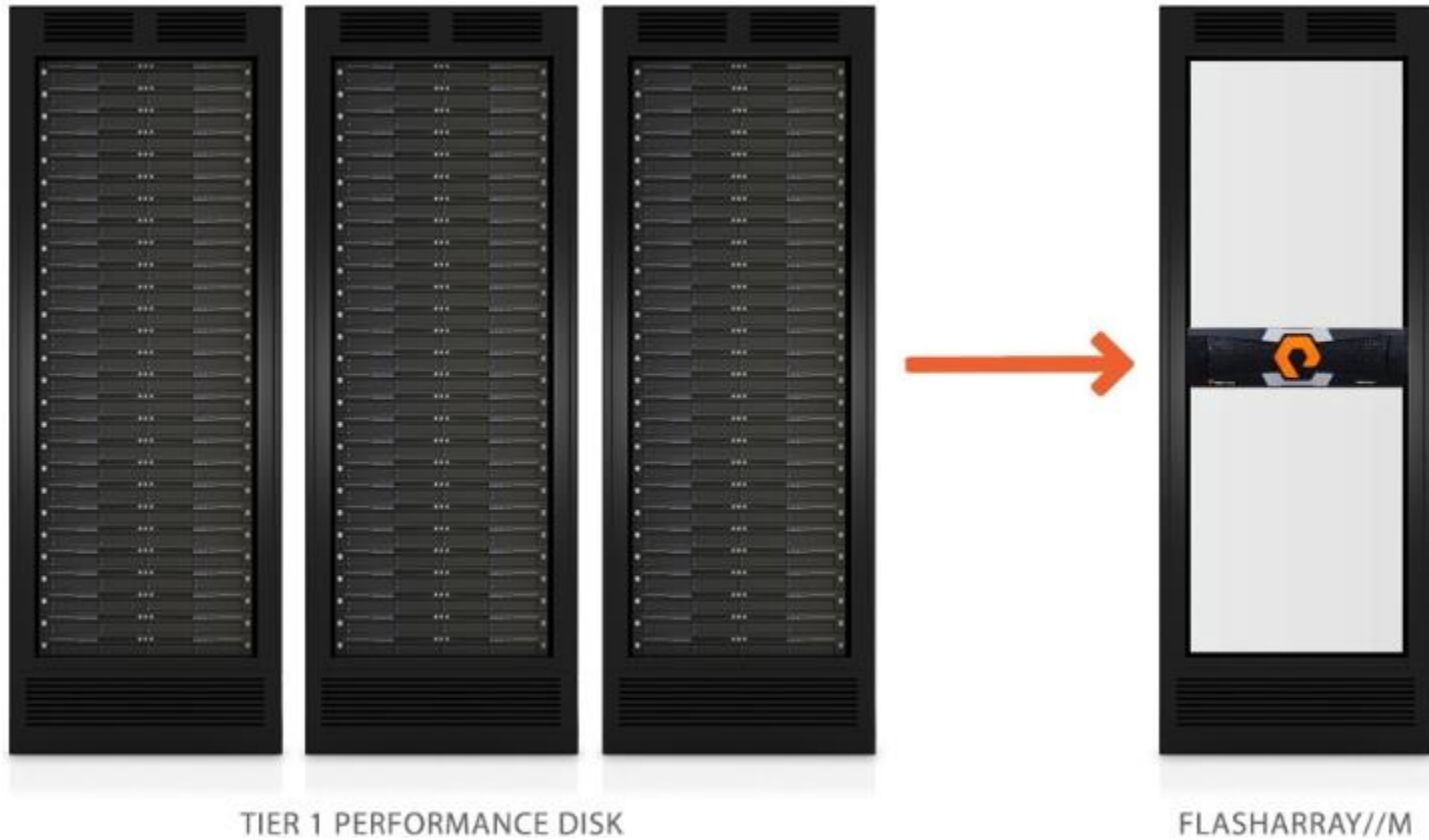


POWER, COOLING, SIZE

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



ALL-FLASH ARRAY



RELIABILITY

- Less 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 writing cycles were quite limited (~ 1000)
 - Nowadays 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 apprx. 2x faster
 - RAID 1



COMPARISON

	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 formatting 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
 - 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

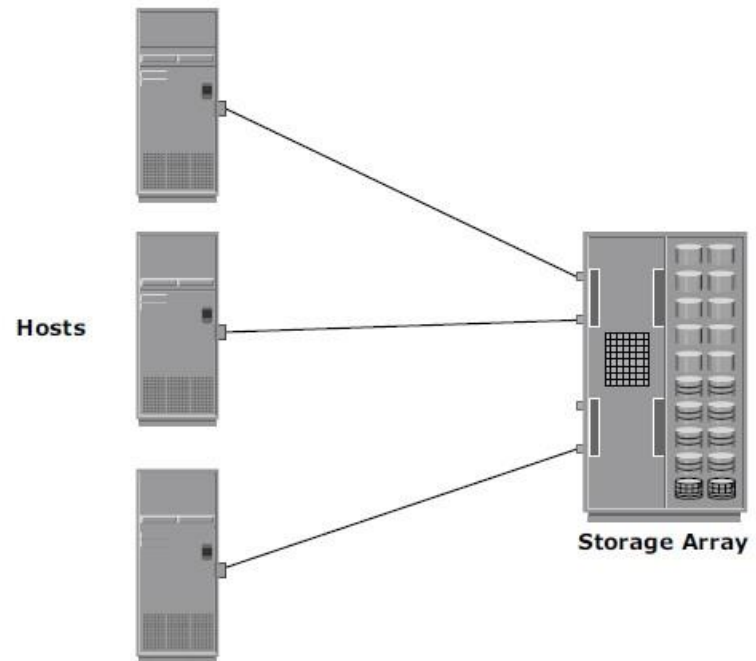


- Serial Advanced Technology Attachment (SATA)
- Point-to-point connection between the SATA host adapter and the SATA device
 - Half duplex
- New connecting interface
- Higher transmission speed
 - (P)ATA 66/100/133 MB/s
 - SATA 150/300/600 MB/s
- The connecting cable has 4 wires, max. length 1 m



EXTERNAL DAS

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



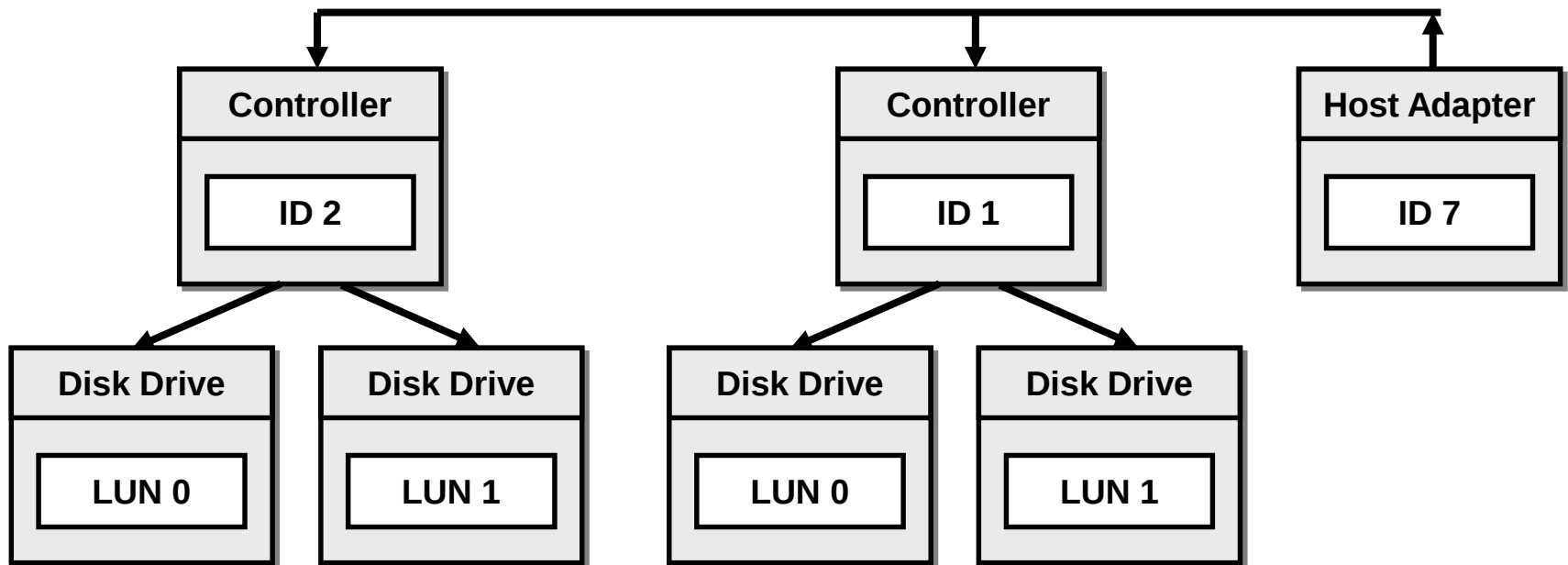
SCSI INTERFACE

- SCSI
 - Small Computer System Interface (SCSI)
 - Standardised I/O bus
- Devices
 - Disk Drives
 - Tape Drives
 - Removable Media Drives
 - CD/DVD Drives
 - Printers

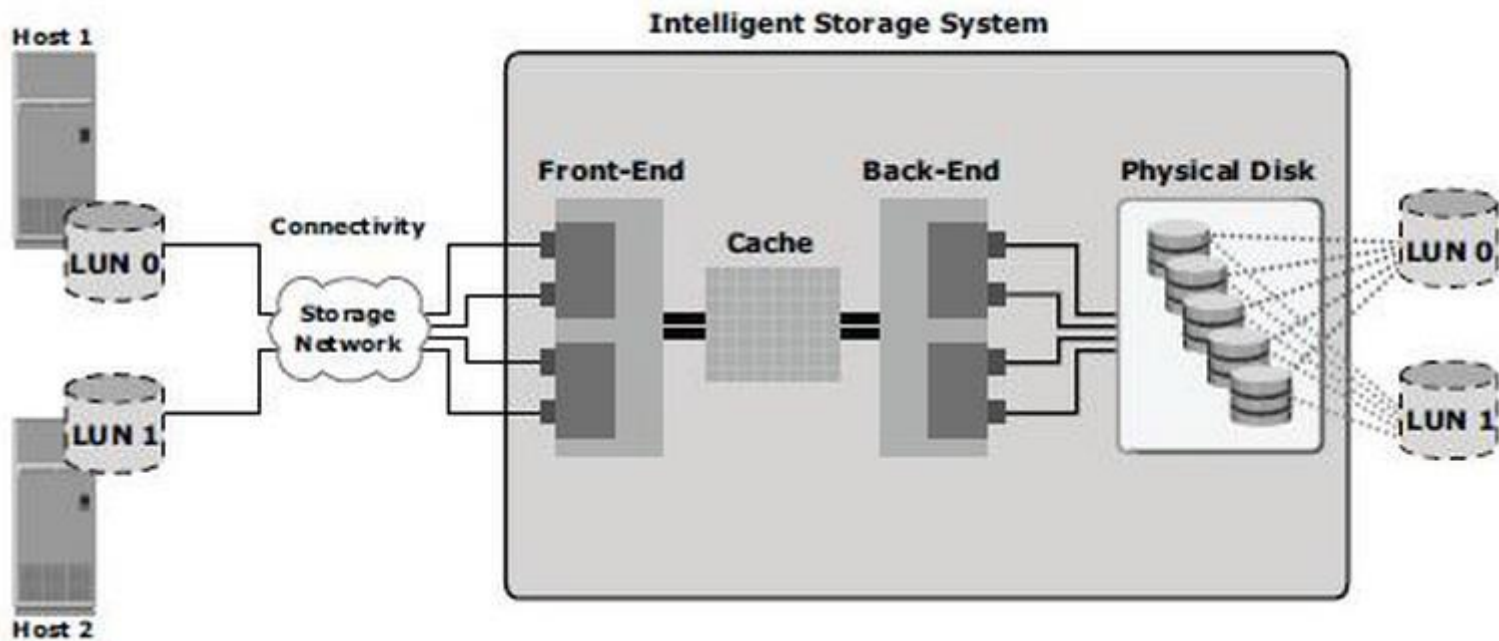


LOGICAL UNIT NUMBER (LUN)

- NOT(!!!) a number – Group of storages
- LUN is an SCSI group address method



LUN



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



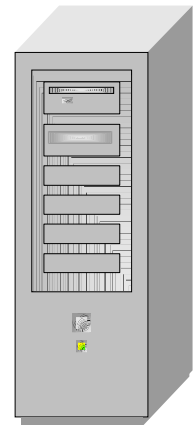
DIRECT-ATTACHED STORAGE (DAS)

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
- Low throughput

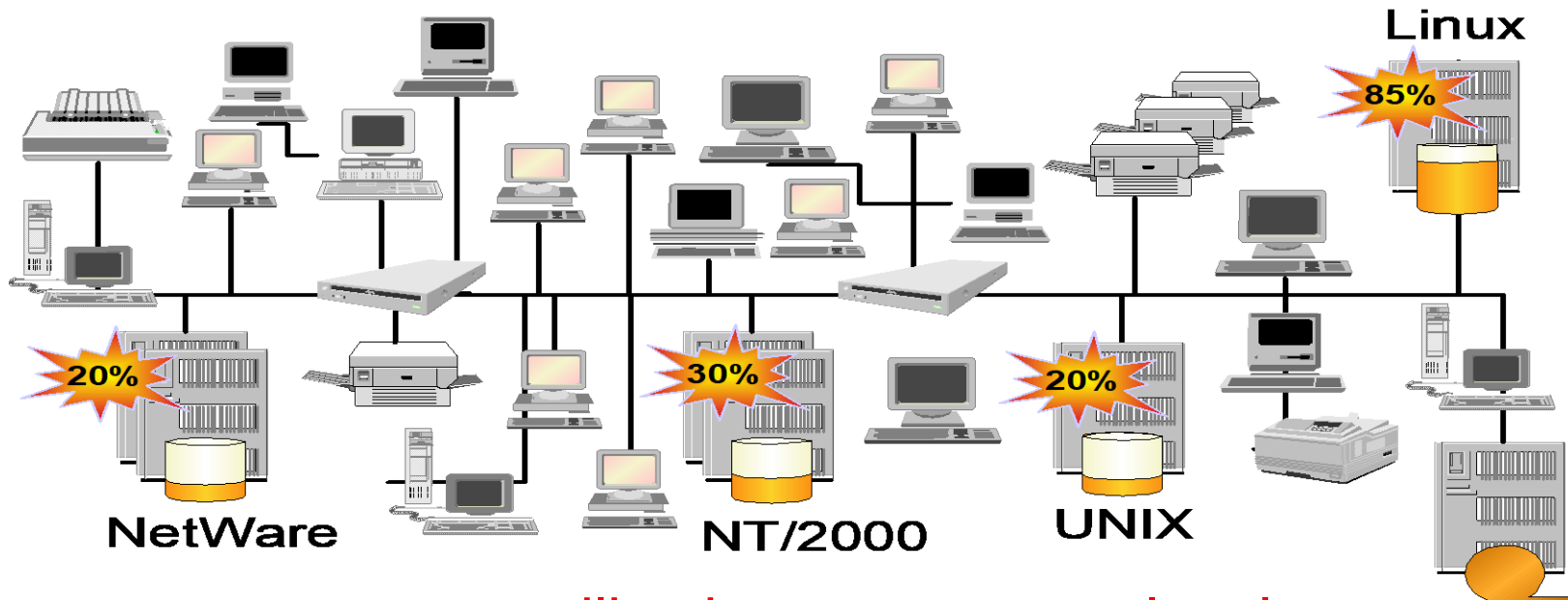


DAS Device



DEDICATED STORAGE DEVICES

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

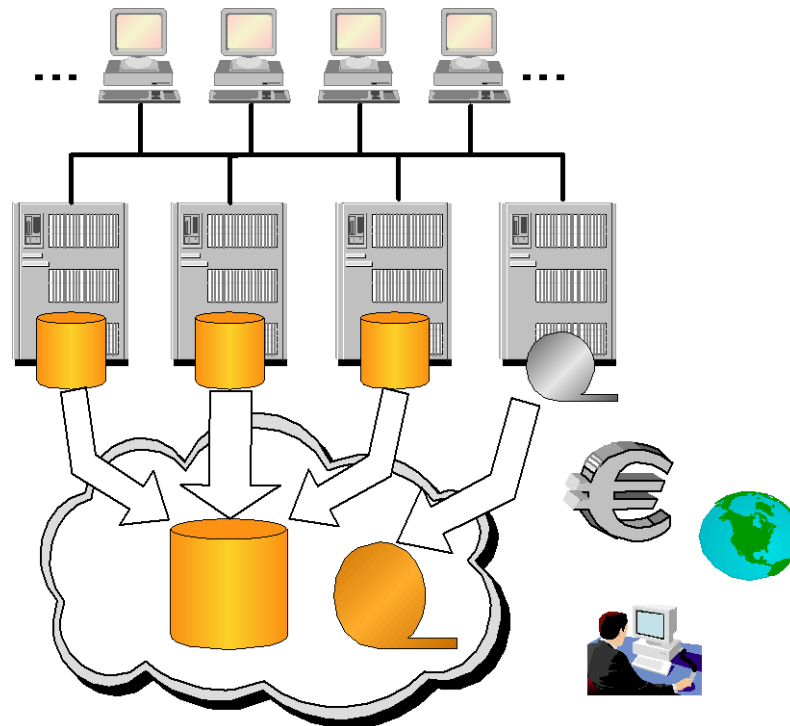


The average storage utilisation at company-level is typically only 40-60%



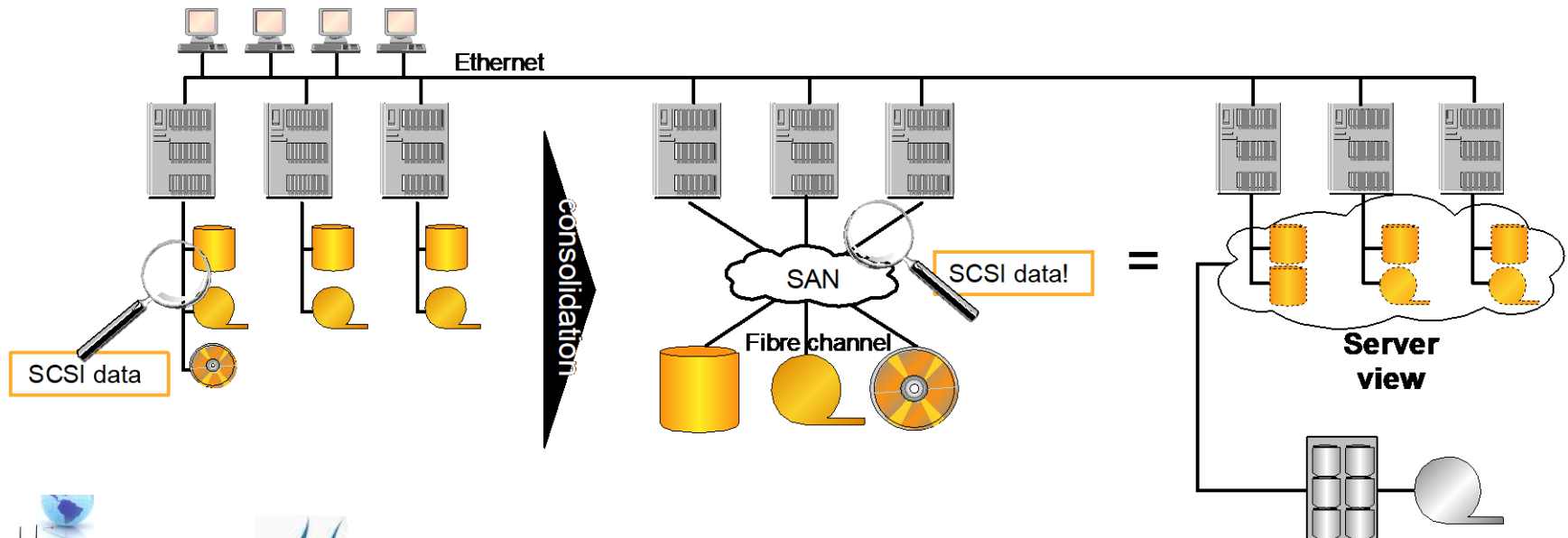
CONSOLIDATED STORAGE DEVICES

- 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**

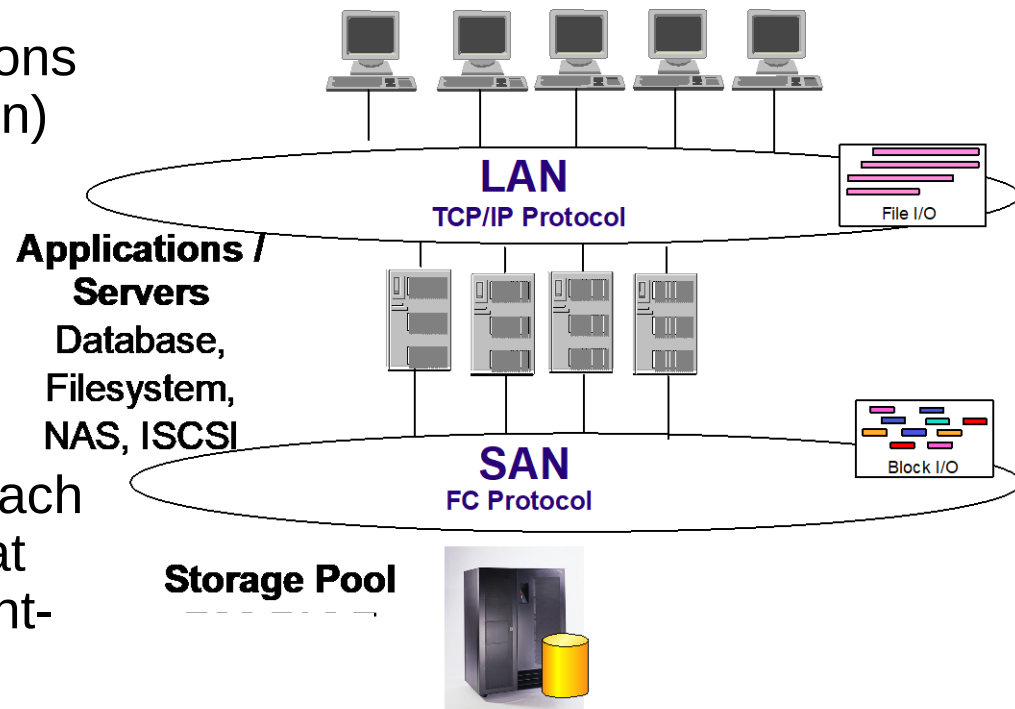
- Increase resource efficiency,
- Scalability
- Higher level system functions can be installed (replication)

- **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



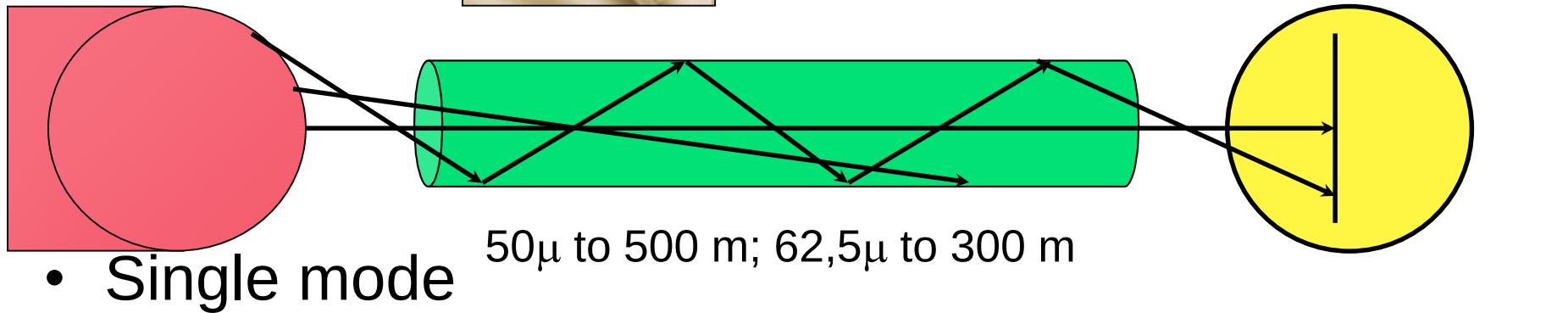
TRANSMISSION TECHNOLOGY: FIBRE CHANNEL PROTOCOL

- 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, fibre optic
 - Max. speed: 128 Gb/s

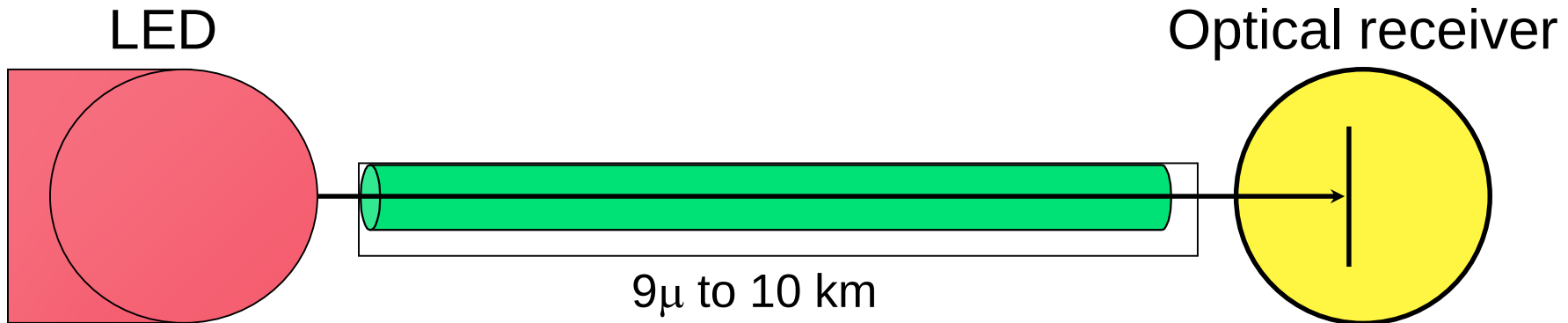


FIBER OPTIC

- Multimode LED

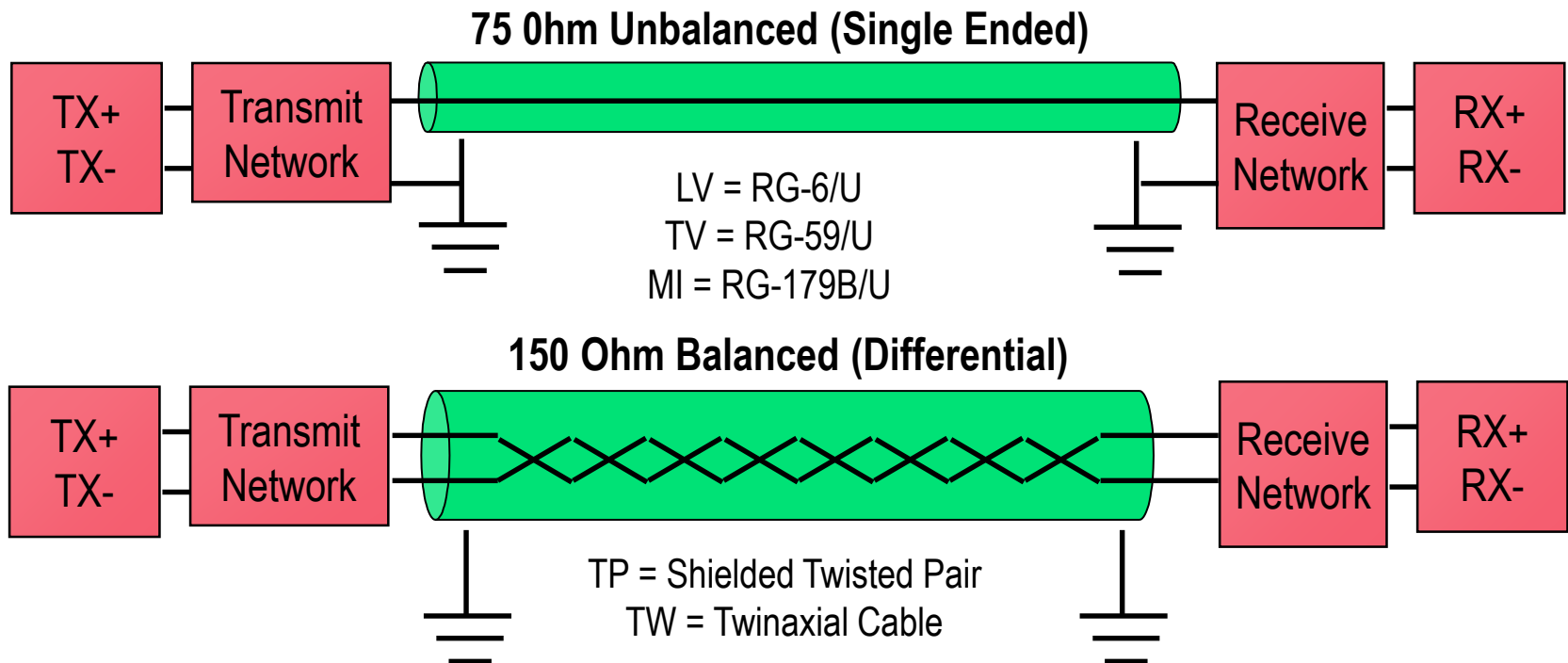


- Single mode



COPPER WIRE

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



FIBRE CHANNEL PROTOCOL

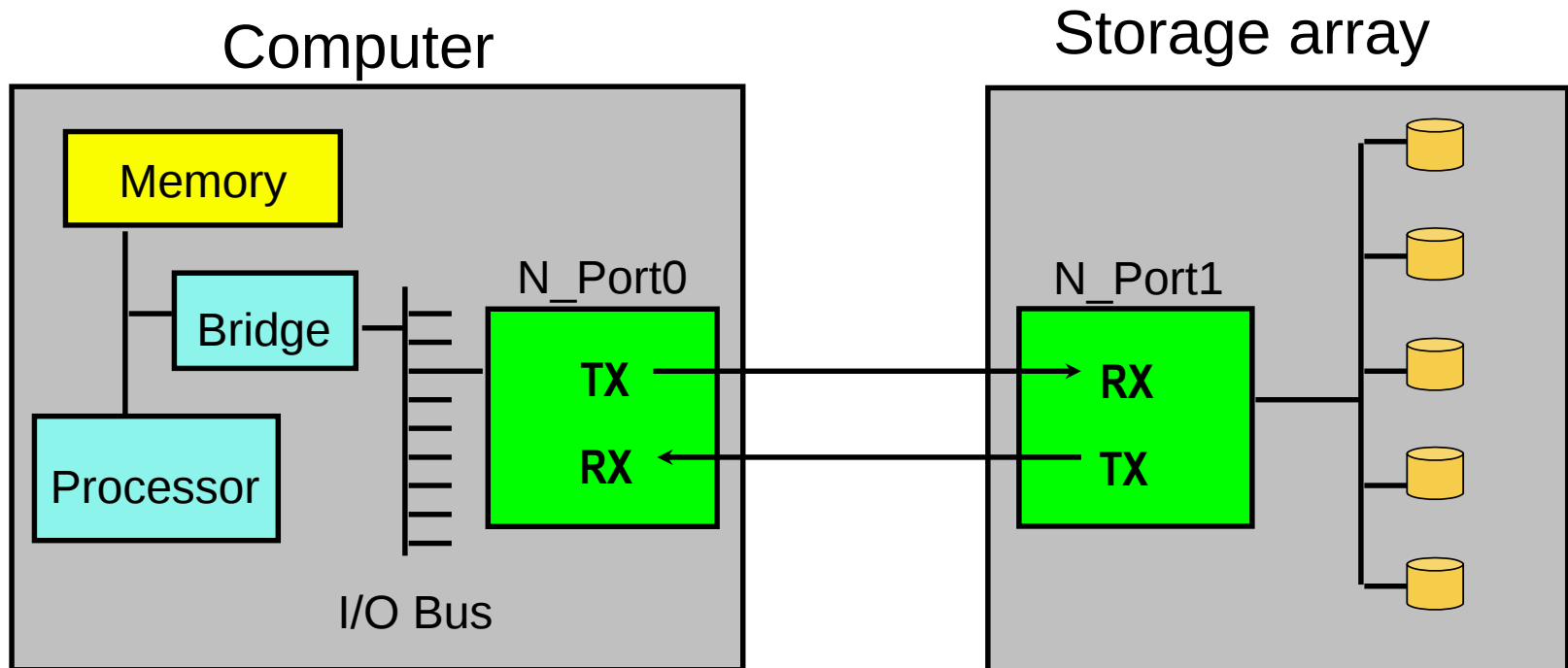
- 3 different topologies
 - Point-to-point
 - Arbitrated loop (not used nowadays)
 - Switched Fabric
- FC interconnects ports
 - Port: any entity communicating over FC, not a physical port
 - N ports (Node)
 - Disk
 - HBA (Host Bus Adapter) in computers
 - F ports
 - FC Switch



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

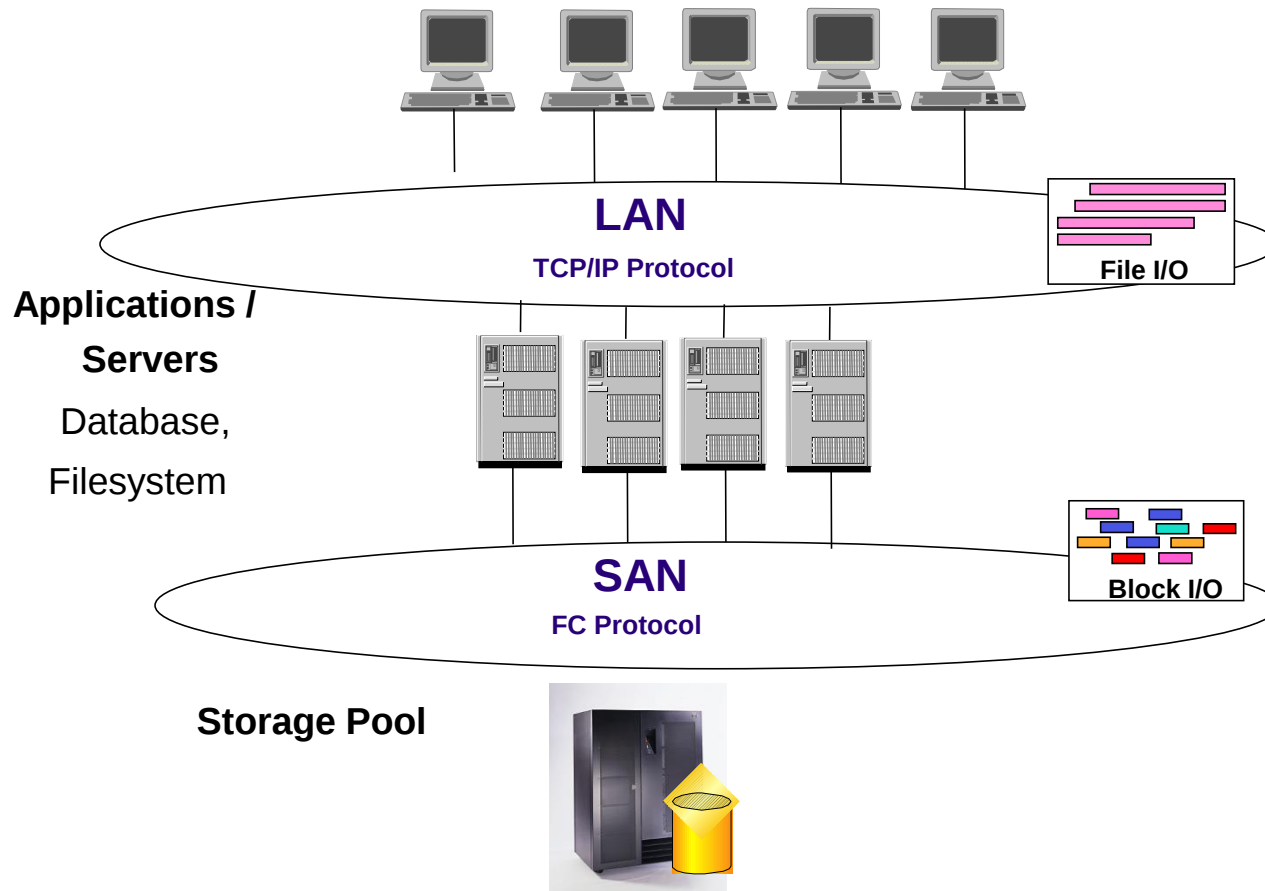


DATA TRANSMISSION SPEED

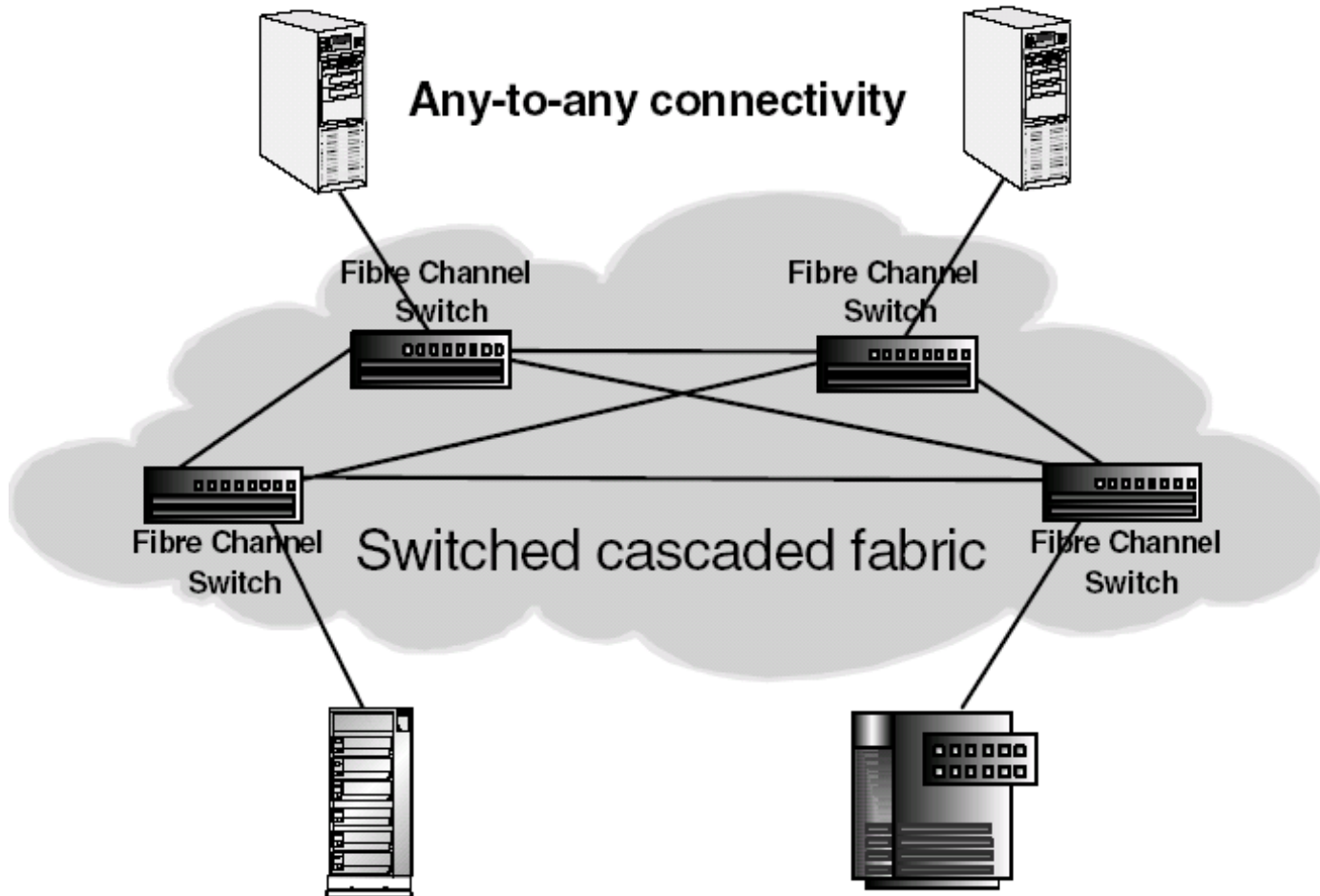
SCSI 1	12.0 Mbit/s	1.5 MB/s
Fast SCSI 2	80 Mbit/s	10 MB/s
Fast Wide SCSI 2	160 Mbit/s	20 MB/s
Ultra DMA ATA 33	264 Mbit/s	33 MB/s
Ultra Wide SCSI 40	320 Mbit/s	40 MB/s
Ultra DMA ATA 66	528 Mbit/s	66 MB/s
Ultra-2 SCSI 80	640 Mbit/s	80 MB/s
Ultra DMA ATA 100	800 Mbit/s	100 MB/s
Ultra DMA ATA 133	1064 Mbit/s	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)

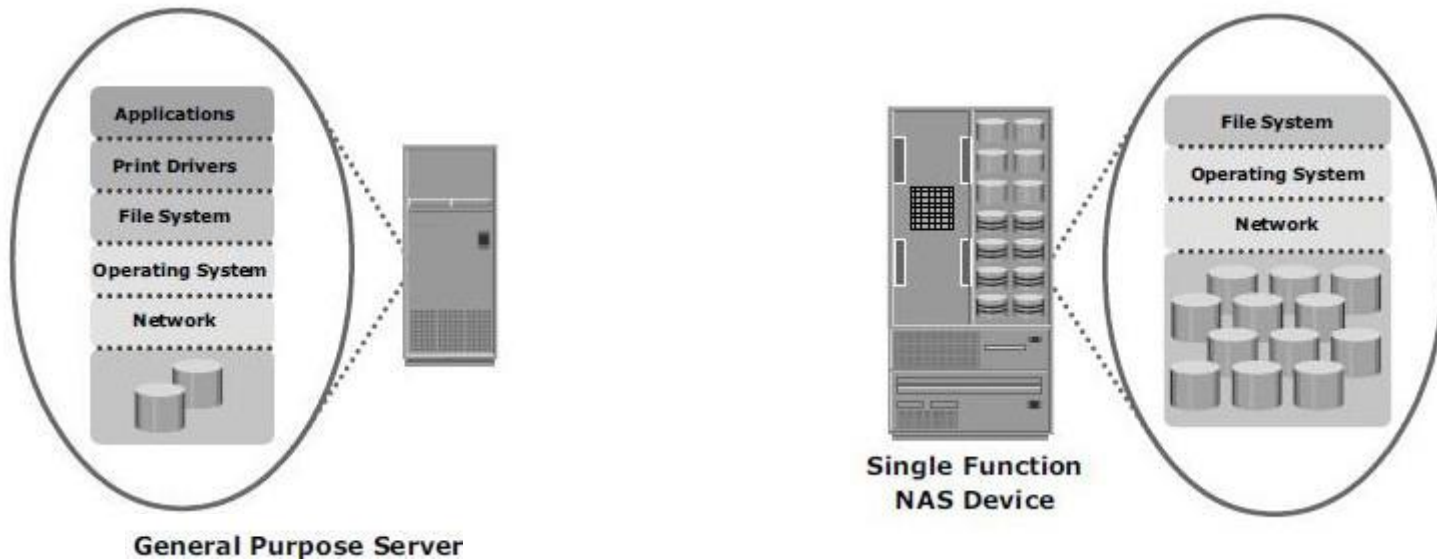


SWITCHED SAN NETWORK



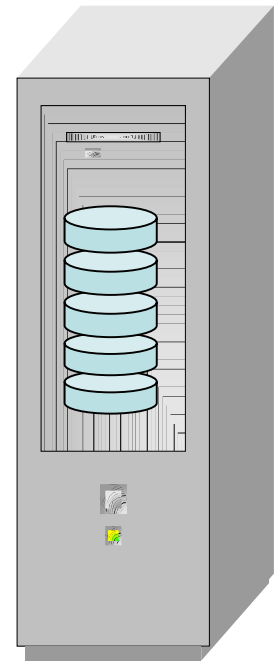
NAS – NETWORK-ATTACHED STORAGE

- 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 resistance
- File level access (Block access not supported!!)
- Easy installation
- Scalability inside the device
- Performance limitations (LAN bandwidth, 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 elements 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



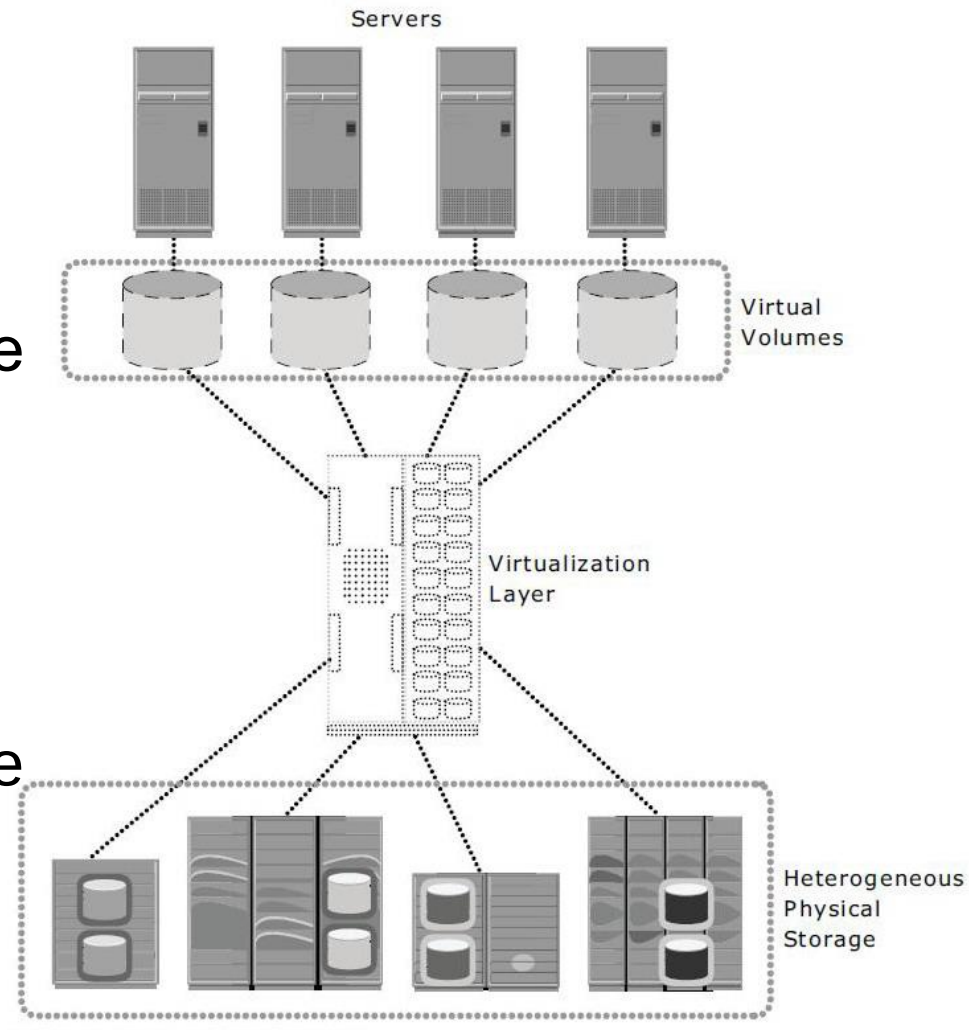
VIRTUALISATION

- **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



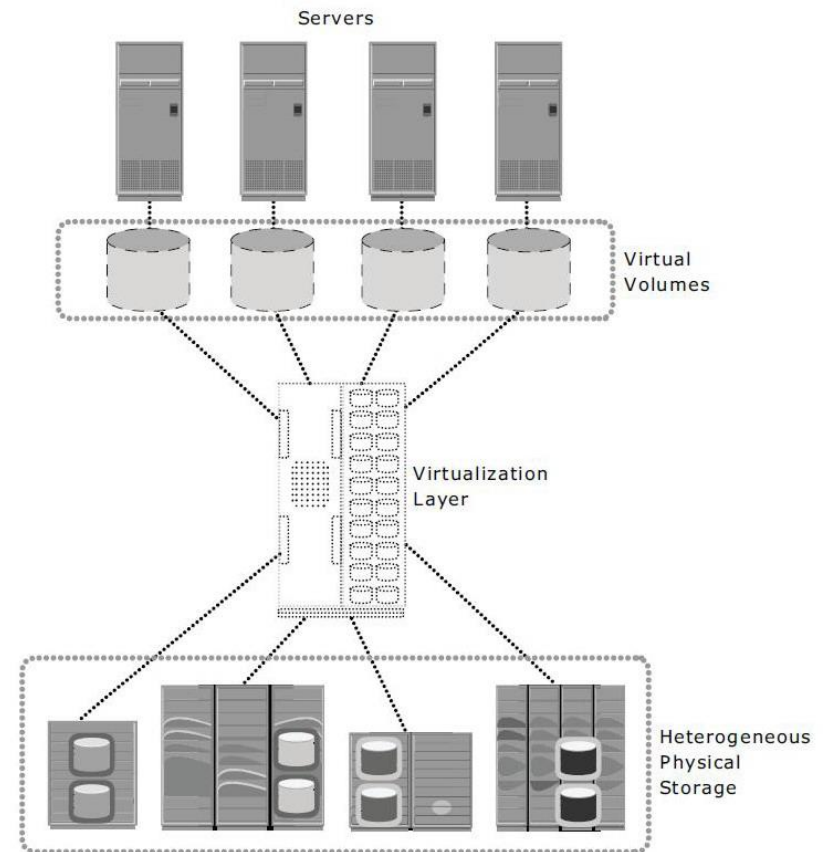
STORAGE VIRTUALISATION

- Virtualisation Appliance or Virtualisation Engine hides the differences 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



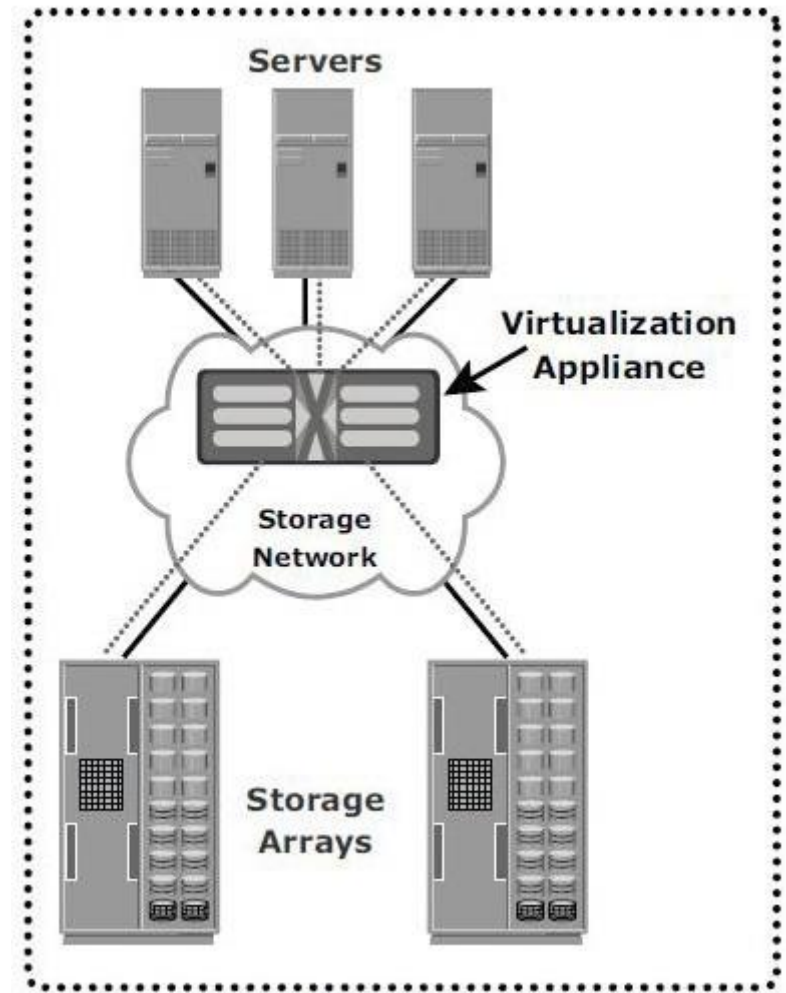
VIRTUALISATION APPLIANCE

- 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



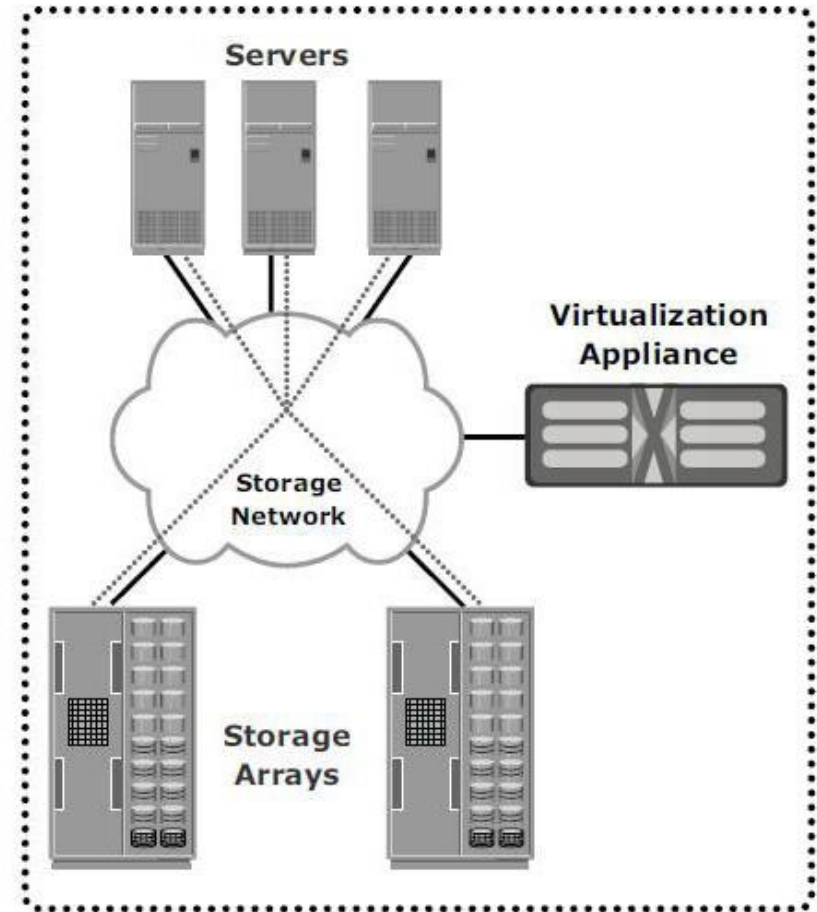
VIRTUALISATION CONFIGURATION – IN-BAND

- 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 ...

- Disks (different types, different vendors)
- Typically fix sized
- Different vendor configurations & back-up services
- Migration to new technologies is difficult

Virtual sources...

- Virtual LUNs, they seem as same type of same vendor
- Size can be modified dinamically
- Centralised management and services
- Migration without disturbing the applications



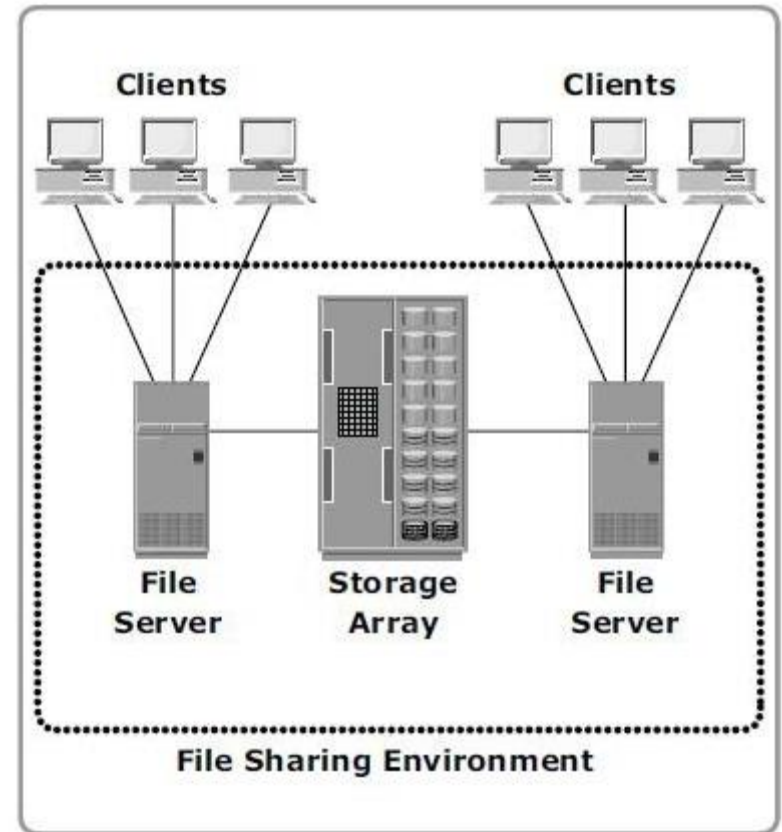
LEVELS OF VIRTUALISATION

- 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



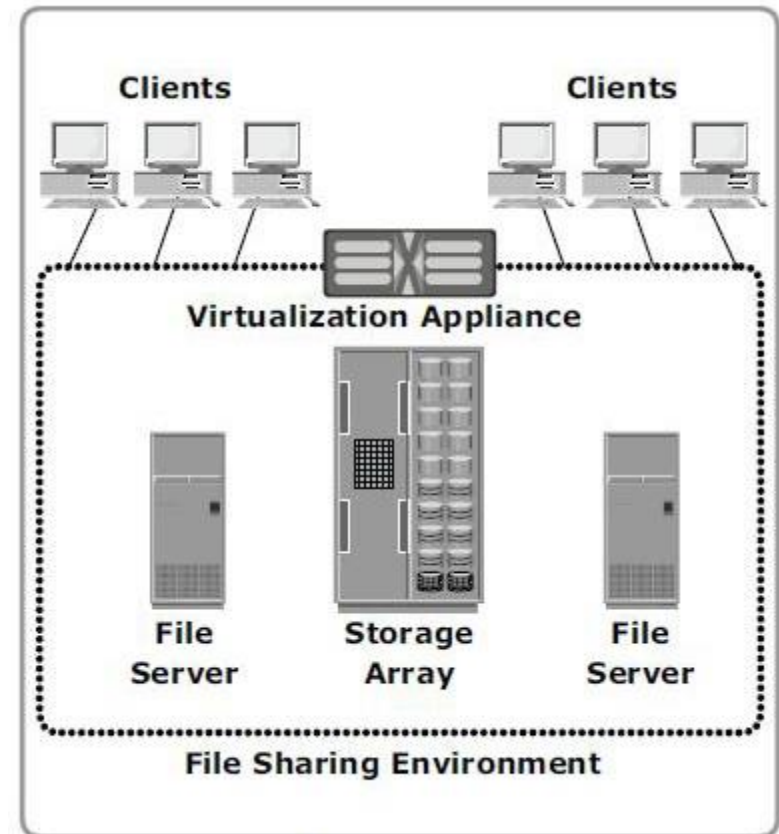
FILE LEVEL VIRTUALISATION

- 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 media + delivery boy
- 100 petabyte
 - Film archive
 - NASA satellite pictures
 - 2000 years of mp3
 - 200x the genom of all humans

