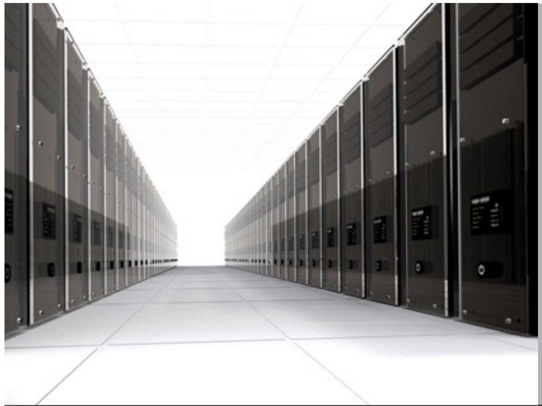


STORAGE CONCEPTS

David López
v 2.5.1
Updated spring 2022

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Magnetic vs. Optical vs. Solid State

Three basic storage technology:
- Magnetic
  - Tapes (1952-Today)
  - Hard Disk (1956-Today)
- Optical
  - Optical Disc Archive (2013 – Today)
- Solid State
  - Solid State Discs – SSD (2006 – Today)

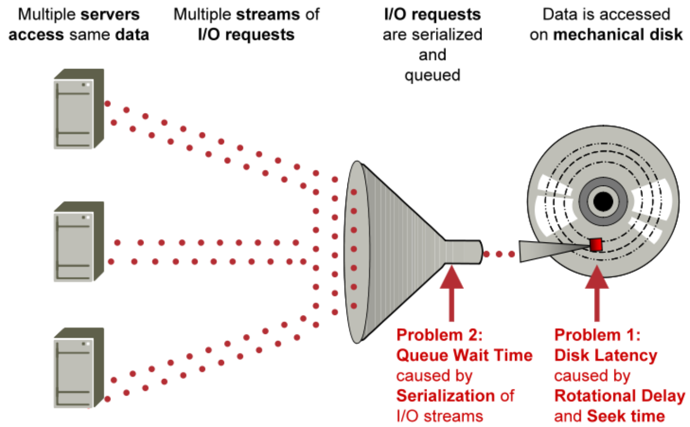
Hard Disk situation

Hard disks are “living dinosaurs”
- According to Moore’s law, the density of microelectronics doubles every 18 months
- In hard disks, this only applies to:
  - Process speed of the controller (which never was much of a problem anyway)
  - Increased speed of read/write operations because more data is packed onto each track
  - Increased capacity of the disk (that means more accesses per second)
- The problem is that it does not affect nor to the rotational speed neither to the actuators moving speed
  - And several actuators on the same rack does not work due to the high density and dilatation

BIG PROBLEM: HDD can store gigantic amounts of data, but the transactions per second are tied to the mechanical internals

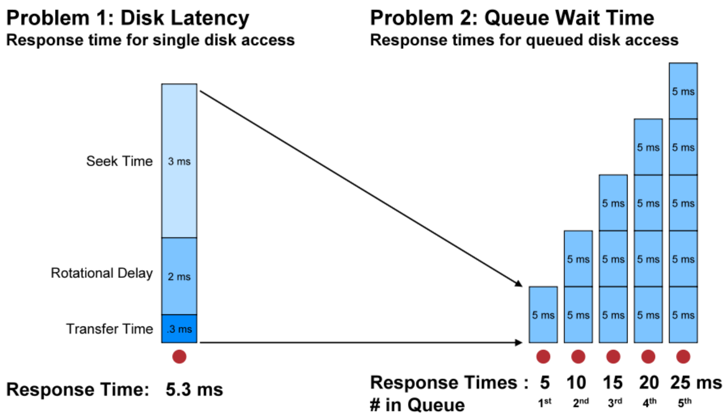
---

Hard disk problems



http://www.violin-memory.com/assets/Violin-WP-Disk-Storage-Shortfall.pdf?d=1

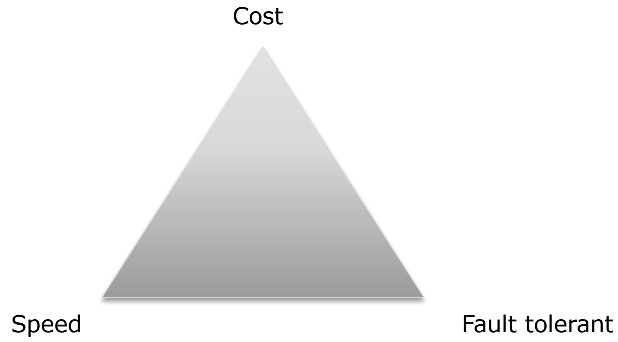
Disk latency + queue wait time



http://www.violin-memory.com/assets/Violin-WP-Disk-Storage-Shortfall.pdf?d=1

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Storage triangle



UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

LUNs and JBOD

- Divided in **LUN**s (**L**ogical **UN**its)
  - For the host computer, there are not differences between LUNs and physical disks
- Easy to work for the host computer
  - Partitions or (more often) aggregation
  - Saw as an unique disk for backup
- Example a **JBOD** (**J**ust a **B**unch **O**f **D**isks)
  - Example: three 2TB disks
  - Build a 6TB LUN
  - You can have disks of different size (not like RAID)
  - One block following the next on the same disk (not like RAID 0) "Concatenation or SPAN, not stripped"

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

JBOD

	Space Efficiency	Fault tolerance	Read Performance	Write Performance
JBOD	1	0	1	1

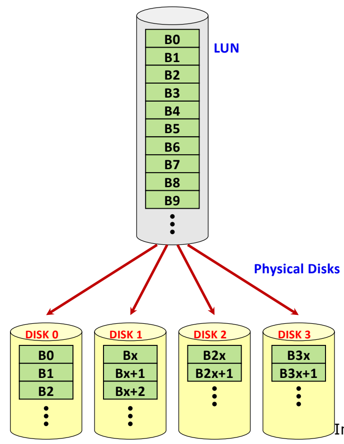
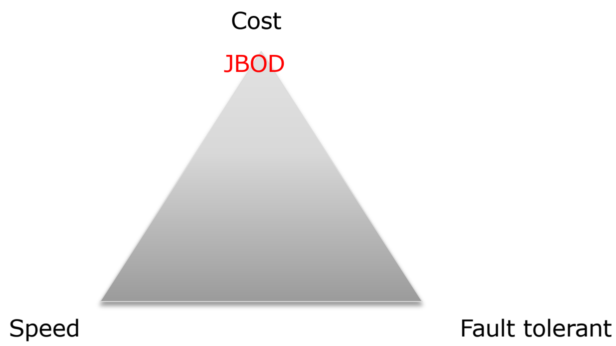


Image by Agustín Fernández (AC)

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Storage triangle



Avoiding errors: RAID

- RAID offers redundancy, BUT ALSO SPEED (at a certain cost)
- Let's calculate # of parallel R/W in
  - RAID 0
  - RAID 1
  - RAID 5
  - RAID 6
  - RAID 10, 01
  - RAID 51, 15
- Important question: WHAT ABOUT THE STRIPE SIZE?
  - 4KB-128KB?
  - In the activities we will consider 4KB but it is an interesting question

RAID 0 (stripping) & RAID 1 (mirroring)

	Space Efficiency	Fault tolerance	Read Performance	Write Performance
RAID 0	1	0	n to 1	n to 1
RAID 1	1/n	n-1	n (real)	1

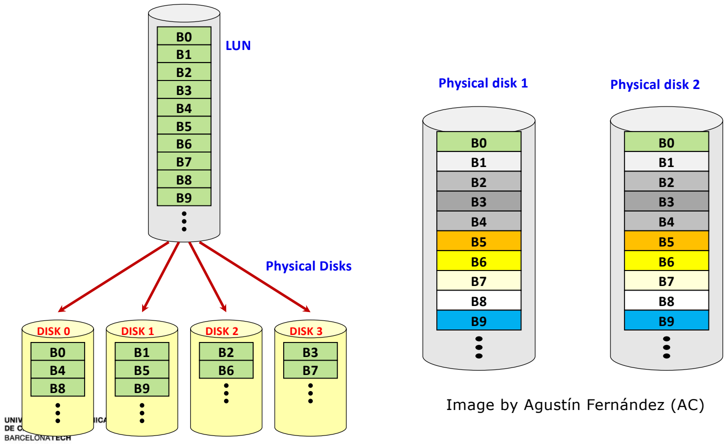
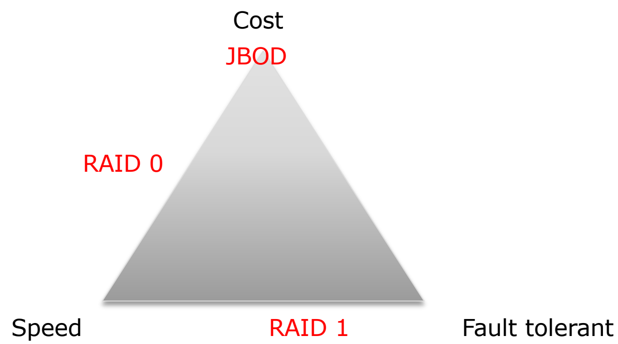
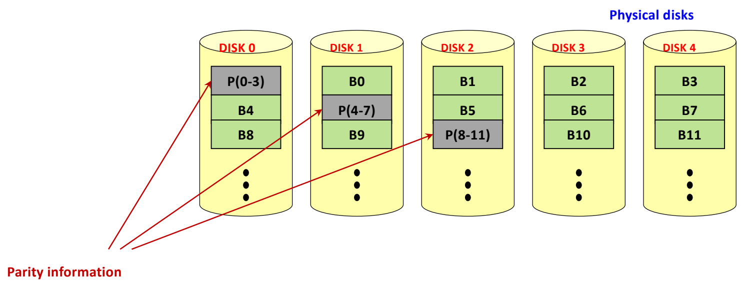


Image by Agustín Fernández (AC)

Storage triangle



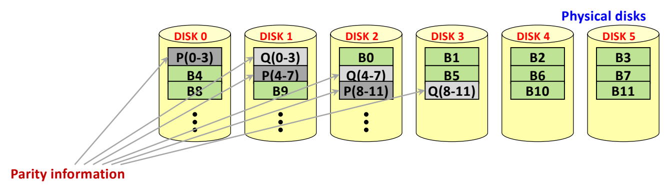
RAID 5: Block-level striping with distributed parity



	Space Efficiency	Fault tolerance	Read Performance	Write Performance
RAID 5	n-1	1	n (n/2)	(n-1) (n/2)

Image by Agustín Fernández (AC)

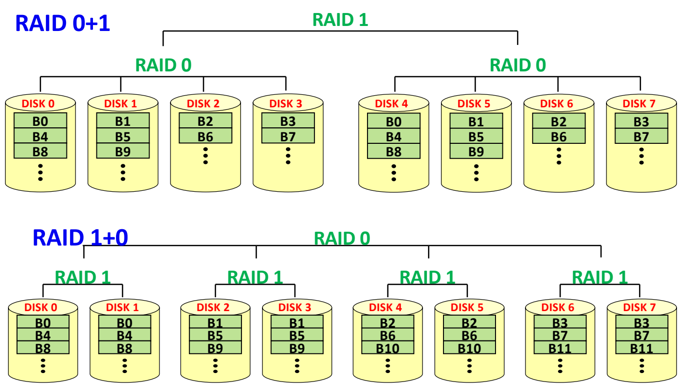
RAID 6: Block-level striping with double distributed parity



	Space Efficiency	Fault tolerance	Read Performance	Write Performance
RAID 6	n-2	2	n (n/3)	(n-2) (n/3)

Image by Agustín Fernández (AC)

RAID 10 & RAID 01



	Space Efficiency	Fault tolerance	Read Performance	Write Performance
RAID 10/01	n/mirrors	n/mirrors	n mirrors	(n/mirrors) 1

Image by Agustín Fernández (AC)

RAID 51 & 15

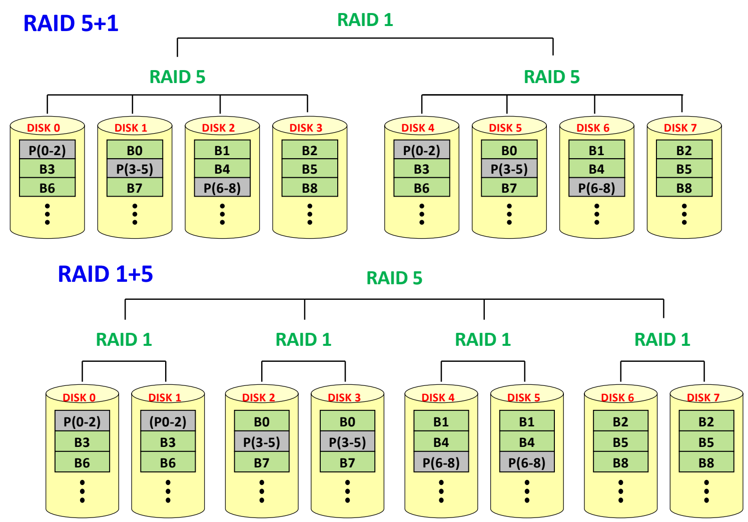
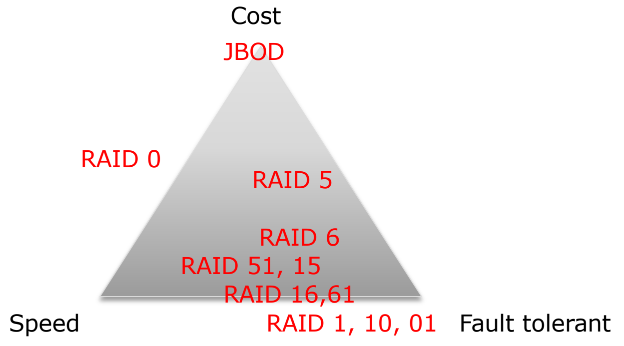


Image by Agustín Fernández (AC)

Storage triangle



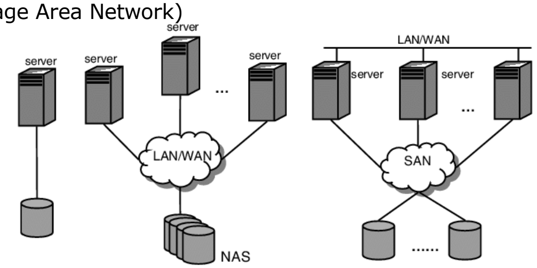
RAID, write penalty & capacity

	RAID 0	RAID 10	RAID 5	RAID 51	RAID 6	RAID 61
Operations per write	1W	2W	2R+2W	(2R+2W) x2	3R+3W	(3R+3W) x2
Write penalty	1	2	4	8	6	12
Capacity	X*C	(X/2)*C	(X-1)*C	((X-1)/2)*C	(X-2)*C	((X-2)/2*C
Minimum number of discs	2	4	3	6	4	8
Required discs (for Y Bytes)	Y/C	2*Y/C	Y/C +1	2*Y/C +1	Y/C +2	2*Y/C +2

Let's assume X discs, homogeneous, each one of capacity C

Avoiding errors: storage networks

- a) DAS (Direct Attached Storage)
- b) NAS (Network Attached Storage)
- c) SAN (Storage Area Network)



(a) DAS (b) NAS (c) SAN

Further reading:
IBM. Demystifying Storage Networking: DAS, SAN, NAS, NAS Gateways, Fibre Channel, and I SCSI. David Sacks
www-03.ibm.com/industries/ca/en/education/k12/technical/whitepapers/storagenetworking.pdf

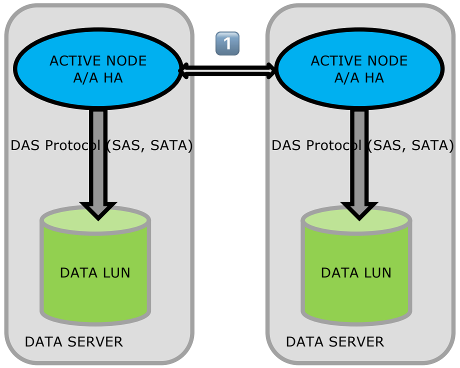
DAS (Direct Attached Storage)

The simplest form

- A single (or multiple) disk drive or tape connected to a computer
- Can have some features like RAID, partitions, …
- Can be accessed by others?
  - Yes. Not directly but through the host computer
  - There is no network device between the data storage device and the computer
- Direct connection, usually using SAS or SATA
  - SAS: Serial Attached SCSI (Small Computer System Interface)
  - SATA: Serial Advanced Technology Attached
- Low cost solution

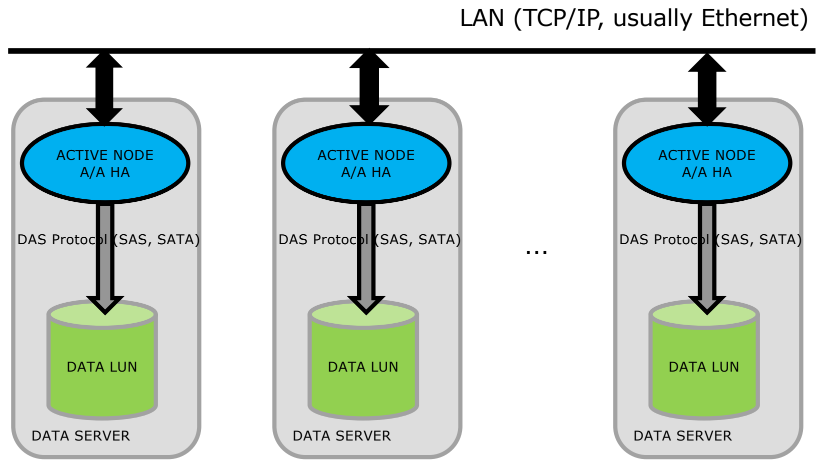
QUITE INUSUAL IN DATA CENTERS

DAS (Direct Attached Storage) idea



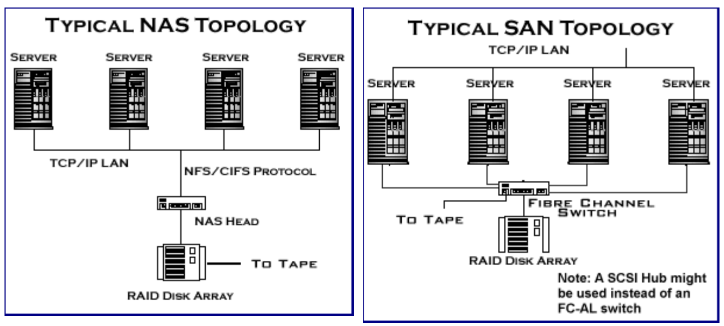
- Both nodes actives (A/A) (not Active/Passive A/P)
- When one node fails, its data are unreachable for the other nodes
- HA (High Availability) requires each piece of data replicated in other nodes (but local disks are cheap)
- Changes are not immediately visible to all nodes.
- Cluster software needed to arbitrate Read & Modify access to replicate data in order to maintain consistency (through Ethernet) 1

DAS (Direct Attached Storage)

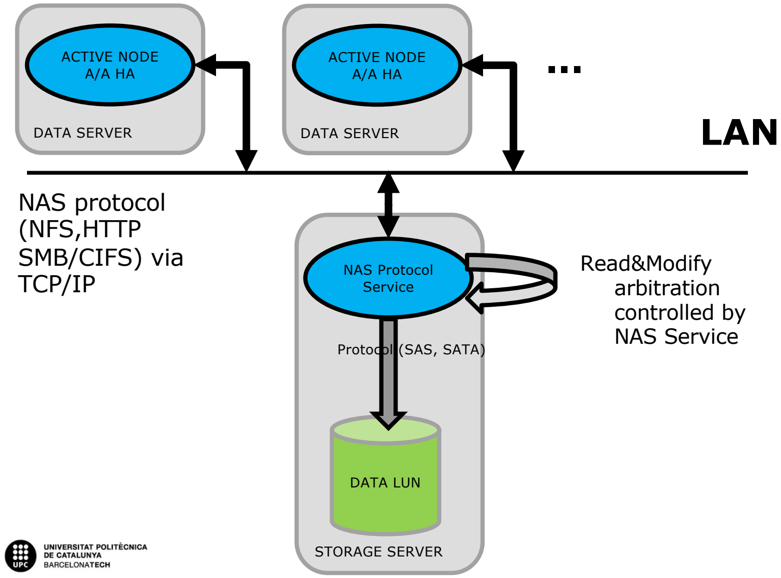


NAS and SAN

Image from NAS-SAN.com



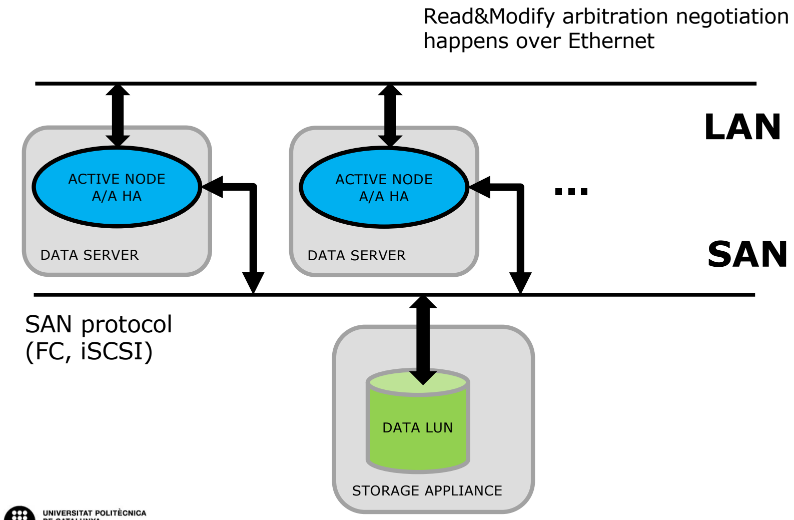
NAS (Network Attached Storage)



NAS (Network Attached Storage) details

- Attached to a TCP/IP network (usually Ethernet)
  - Typically 100 Mbps – 10 Gbps, 2-4μsec latency
- Protocol operates on files (like a network attached file)
- NAS appears to the OS as a shared folder
- NAS is LAN-dependent; if the LAN goes down so does the NAS
- Does not scale very well (in the EBH project we will ignore this)
- One weakness related with its very nature:
  - Ethernet transfer data via packets, that can be sent out-of-order (or even lost), so the file is not available until all packets has arrived
  - No problem with small files, problem with large files (video production or consumption)

SAN (Storage Area Network)



SAN (Storage Area Network) details

- Dedicated high-performance network for block-level storage
  - Typically 2-200 Gbps, <1μsec latency
- Protocol operates on blocks: multiple clients can access files at the same time with very high performance (as it was a local hard disk). Changes are visible by all nodes
- SAN is LAN-independent; if the LAN goes slow does not affect
- More complex to administrate, more expensive
- Not affected by out-or-order

NAS

- Cheaper
- Easy to manage
- Ideal for:
  - File storage and share
  - Small Databases

SAN

- High performance
- Ideal for:
  - High transaction databases
  - E-commerce
  - Video editing or broadcasting
  - If fast backup is required

IOPS (Input / Output Operations Per Second)

- Pronounced *eye-ops*
- Common performance measurement for storage devices
- There are applications to measure it
  - Iometer (Intel)
  - IOzone
  - FIO
- Not easy to define / compare
  - Mix of read / write operations
  - Sequential and random accesses
  - Data block sizes
- Typical values
  - Total IOPS (mix of R/W, Seq/RND)
  - Random read IOPS
  - Random write IOPS
  - Sequential read IOPS
  - Sequential write IOPS
- IOPS * TransferSizeInBytes = MBps

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

SSD performance

Many IOPS? Solid State Disks can offer the solution!

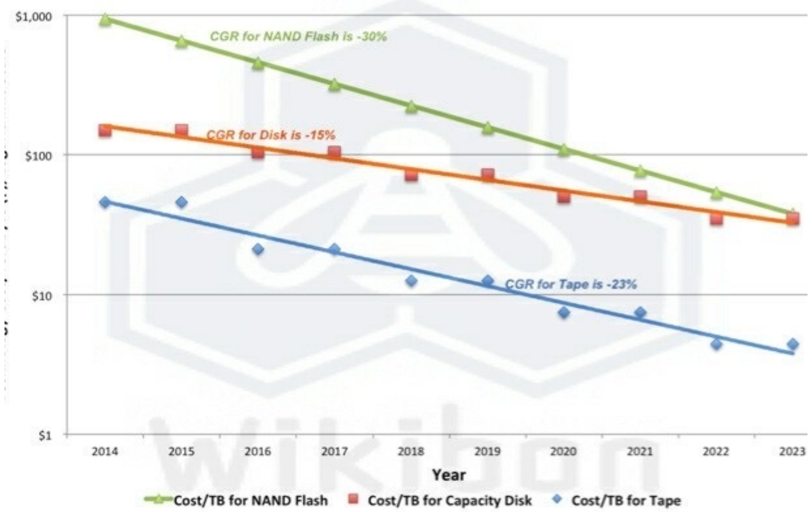
- In our project
  - HDD IOPS: 640 – 5210
  - SSD IOPS (RD/WR): 90k/10k – 540k / 205k

And the cost? Fa$t di$k$ co$t money!

- In our project
  - HDD cost: 0,029 /G (8 TB=235€) – 0,15€/G (2.4TB=360€)
  - SSD cost: 0,155 €/GB (2TB=310€) – 0,21€/GB (7,68TB=1545€)

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

SSD & HDD price forecast



UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Consumer vs Enterprise

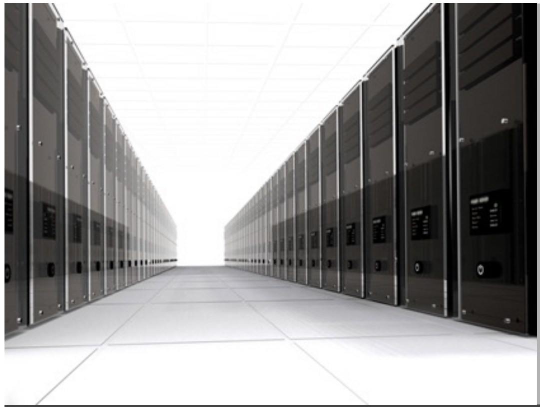
HDD

Model	Seagate Barracuda ST8000DM0004	Toshiba MG07ACA14TA	Seagate ST10000NM009G	HPE 765466-B21	HPE EG002400JWJNN
Tipus	Consumer	Enterprise	Enterprise	Enterprise	Enterprise
Capacitat (TB)	8	14	10	2	2,4
Consum (W)	6.8	7.8	9.5	7	7.1
Preu (€)	235	520	350	250	360
IOPS R/W	640	800	710	3360	5210
RPM	5400	7200	7200	10000	10000
€ / GB	0,029375	0,037142857	0,035	0,125	0,15

SSD

Model	Samsung 860 EVO	Intel Optane H10	Kingston SEDC100M	WD Gold S768T1D0D	WD Ultrastar DC SN640
Tipus	Consumer	Consumer	Enterprise	Enterprise	Enterprise
Capacitat (TB)	2	1	1,92	7,68	3,8
Consum (W)	2.2	5,8	9	12	8
Preu (€)	310	195	372	1545	750
IOPS R/W	90k / 10k	330K /250k	540K /205K	467k/ 65K	511K / 82K
Tecnologia	3D QLC NAND	3D QLC NAND	3D TLC NAND	3D TLC NAND	3D TLC NAND
€ / GB	0,155	0,195	0,19375	0,201171875	0,197368421

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH



STORAGE

David López

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH