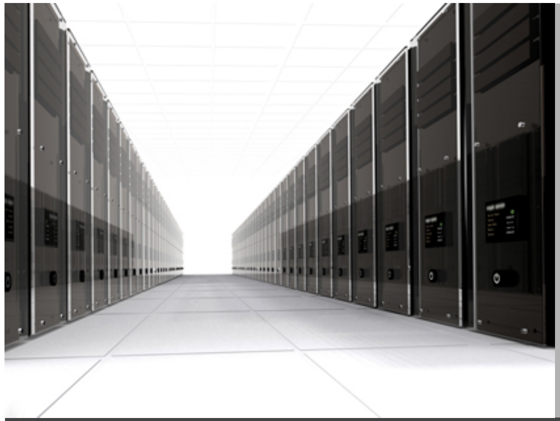


HOW STUFF WORKS

David López
V.2.5.4
Updated Spring 2021

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Magnetic vs. Optical vs. Solid State

Three basic storage technology:

- Magnetic
  - Tapes (1952-Today)
  - Hard Disk (1956-Today)
  - Removable
    - Floppy disk (1971-2000)
    - Zip, Jaz, Rev (1990's-2005)
- Optical
  - Compact Disc – CD (1984 – Today?)
  - Digital Versatile Disc – DVD (1996 – Today)
  - High Definition DVD – HDDVD (2003 – 2008)
  - BluRay Disc – BD (2003 – Today)
  - Optical Disc Archive (2013 – Today)
- Solid State
  - Flash Memory (1998 – Today)
  - Pen Drive (2000 – Today)
  - Solid State Discs – SSD (2006 – Today)

HOT, WARM AND COLD DATA

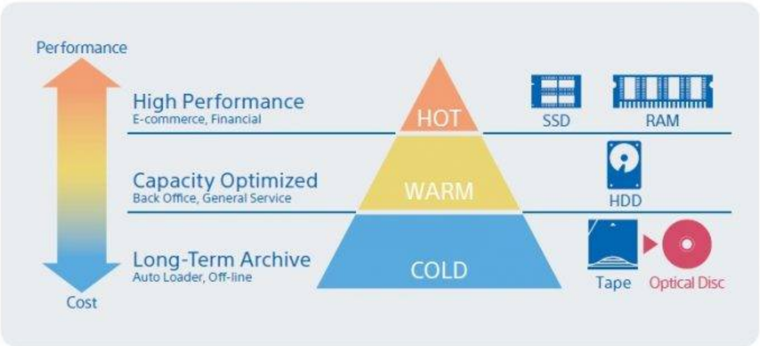


Image from Matsushita Electric Co

Magnetic Tape

Invented for audio recording in 1928 by Fritz Pfleumer
May 21, 1952 IBM develops the IBM 701 Defense Calculator
It includes the first Tape Unit for computers, the IBM 726 Tape Unit

Source: IBM

Magnetic tapes nowadays (for Data Centers)

Usually in data cartridges
Sequential, good for backup but not for random access
From small tape devices (1 tape) to tape robots
(https://www.youtube.com/watch?v=yLIE3LIPeiU)
Usually compressed data (ratio 2:1)
Biggest tape (2019): StorageTek (Oracle)

- 12TB uncompressed data
- 300 MB/s transfer rate
- 25,000 read/write cycles
- 30 years archival life

Hard Disk

September 13, 1956
IBM develops the 305 RAMAC, with a hard disk

- 3,75 MB real capacity
  - 5 million "words" of 6 bits
- 50 discs, 24" (61 cm) Ø
- Each surface has 100 tracks
- 1,200 RPM
- Data transfer: 8000 chars/sec
- Random access!
- $35,000 per year leasing cost

Source: http://ed-thelen.org

Hard Disc nowadays (for Data Centers)

Typical form factors: 2.5" and 3.5"
Rotational speeds (in rpm – revolutions per minute): 7200, 10000, 15000
DC Systems Example: iX16 iSCSI to SATA RAID Subsystem, a 3U 16-bay 3.5" discs system ($3199 without discs)

Images and price retrieved at pc-pitstop.com

Optical

Not very useful in DC … until Archival Disc appears

- Magnetic tape is cheaper for backup
- Upgrades via Internet
- Useful for long-term archive

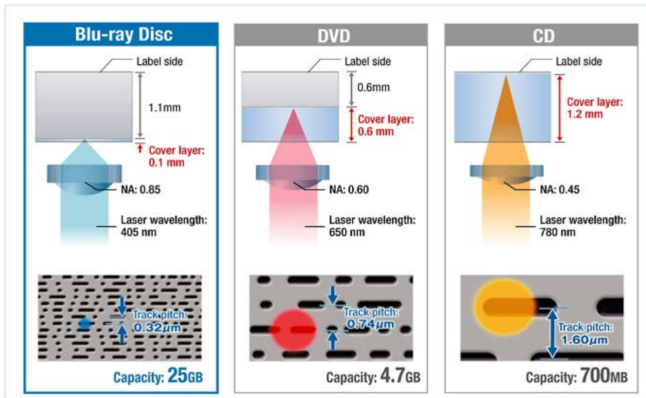


Image from Matsushita Electric Co

Archival Disc

- Archival Disc (see https://www.youtube.com/watch?v=pPbt0sARido)
- 2 sides
- more density
- more data space
- multiple heads
- 5.5 TB/ disc
- 100 + years

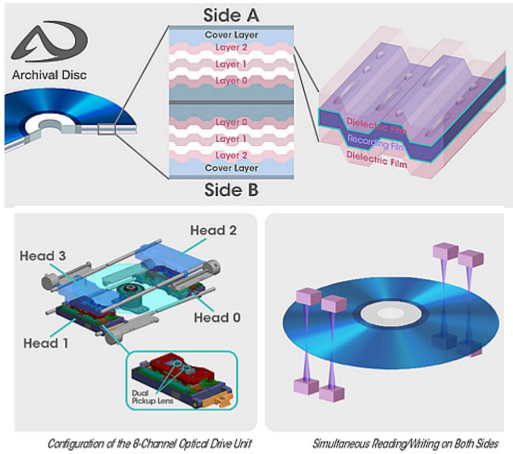


Image from cdrinfo.com

Solid State Devices

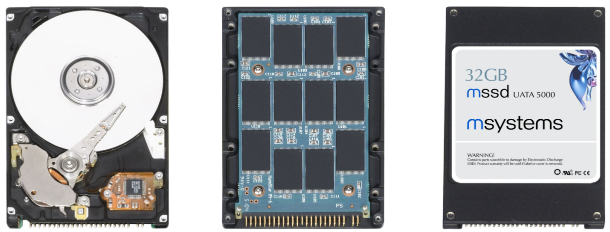
- Even that solid state technology is quite old (one radio receiver use the first solid state device in 1930!), in 1980 Dr. Fujio Masuoka from Toshiba develops the flash memory
- First appear in 1988
    - Basically for cameras
    - Quite expensive
    - SD, MicroSD, Memory Stick, SmartMedia, …

Pen drive

- A flash memory on an USB device
- First generation (USB 1.0) appeared in 2000
- Now USB2, USB3 and USB4
- Standard for (small) data transfer, kills the floppy disc (and the optical devices)

Solid State Discs (SSD)

- Based on flash memory
- 2004 first prototype discs
- 2006 first commercial discs
- Based on memory architecture, the size is in power-of-two units
    - Typical size: 128GB, 256 GB, up to 16 TB (about $15000)
- Offer clear advantages for Data Centers (the main problem is the price)



How stuff works

Magnetic devices

- Hard Disc
  - Architecture
  - Reliability: SMART
- Magnetic Tapes
  - Principles
  - Compression & Encryption
  - Viability
- Solid State Devices
  - Flash memory NAND and NOR
  - SLC, MLC, TLC. QLC
  - SSD pros and cons

Magnetic hard disk (HDD)

Hard disk at a glance

- An HDD consists of one or more rigid ("hard") rapidly rotating discs (platters) with magnetic heads arranged on a moving actuator arm to read and write data to the surfaces

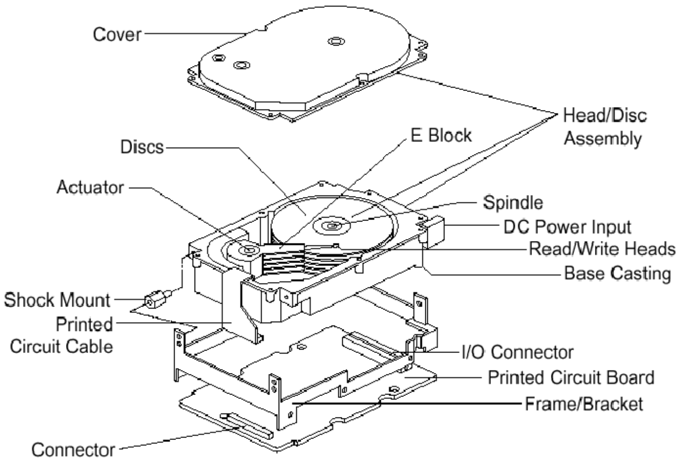


Image borrowed from Seagate

Hard Disk Architecture

- The disc is divided in cylinders, with two heads per platter (the section of cylinder in one side of the platter is called track)
- Each track / cylinder is divided into sectors
- Each sector (or block) contains **4211** bytes (raw)= **4096** bytes/ sector (**data**) + error-correcting code (100 bytes) + HCS position (15 bytes)
- Addressing by **Logical Block Addressing** (LBA) (translation via BIOS)
  - Important idea for spare area
- Not all tracks have the same number of sectors
  - They use Zoned Bit Recording

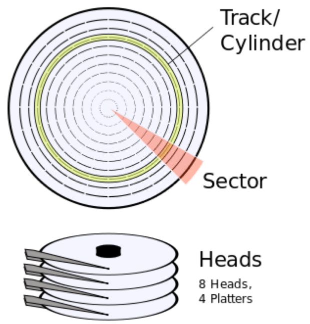


Image borrowed from Wikipedia

Hard disk information you MUST know

Transfer Rate (in MBps)
- Be careful: raw or formatted? Maximum, minimum or average?

Average access time: Average seek time + latency
- Average seek time depends on hardware: from 3ms (high end server drives) to 15ms (mobile drivers). Common desktop is around 9ms
- Latency: depends on RPM. Example: a 7200 RPM HD
  - 7200 rpm / 60 =120 revolutions per second
  - 1 revolution in 1/120= 8.33 $\times 10^{-3}$ seconds
  - ½ revolution in 8.33 $\times 10^{-3}$/2 = 4.16666$\times 10^{-3}$ seconds = 4.17ms

IN OUR CASE IOPS (**I**nput/output **O**perations **P**er **S**econd)

Power Consumption
- Heat dissipation directly tied to power consumption, and as drive age, disk failure rates increase at higher drive temperatures

Reliability

Mean Time Between Failures (MTBF)
- Well, you know, there are lies, dammed lies and statistics
- An average computer HD has about MTBF=400,000 hours (i.e. more than 45 years)
- Reality shows that a HD life is between 3 and 7 years, typically 5

SMART (Self Monitoring, Analysis and Report Technology)
- In 1992, IBM began shipping 3.5-inch hard disk drives that could actually predict their own failure
- These drives were equipped with **Predictive Failure Analysis (PFA)**, an IBM-developed technology that periodically measures selected drive attributes – things like head-to-disk flying height – and sends a warning message when a predefined threshold is exceeded
- Industry acceptance of PFA technology eventually led to becoming the industry-standard reliability prediction hard disk drives

SMART

There are two kinds of hard disk drive failures: unpredictable and predictable

Unpredictable failures happen quickly, without advance warning
- These failures can be caused by static electricity, handling damage, or thermal-related solder problems, and there is nothing that can be done to predict or avoid them

Through research and monitoring of vital functions, performance thresholds which correlate to imminent failure have be determined, and it is these types of failure that SMART attempts to predict

Image borrowed from coredatarecovery.com

Some SMART attributes

- Read Error Rate
- Throughput Performance
- Spin-Up Time
- Spin Retry Count
- Recalibration Retries
- Temperature
- Reallocated Sectors Count
- Current Pending Sector Count

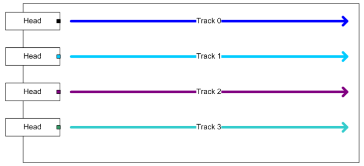
Reallocated sectors
- Each hard disk comes with a limited "pool" of empty sectors that can be used as reallocated sectors (**spare area**)
- When one sector fails, or it is predicted to fail, it’s marked as bad sector
- If failure was predicted, information migrates to one of the spare area sectors
- Due to the fact that some sectors are remapped to another area on the disk, sequential I/O on those sectors is getting randomized (becomes random I/O) with very different performance characteristics

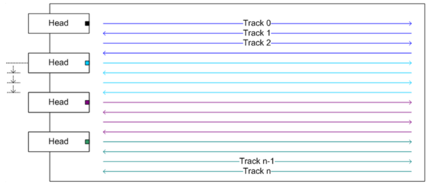
Magnetic Tape

Magnetic Tape

Recording Method:

- Linear
  - Can be linear (several heads read/write in parallel for the tape width) or linear serpentine (the tape goes forward and backward several times – also several heads)





Images from wikipedia

- Scanning: these methods write short dense tracks across the width of the tape medium, not along the length

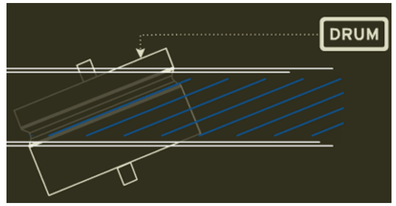


Image from tomshardware.com

Magnetic tapes: compression and encryption

- Most tape drives now include some kind of data compression
  - There are several algorithms which provide similar results: LZ (most), IDRC (Exabyte), ALDC (IBM, QIC) and DLZ1 (DLT)
  - Embedded in tape drive hardware
  - A ratio of 2:1 is typical, with some vendors claiming 2.6:1 or 3:1
- Some enterprise tape drives can encrypt data (this must be done after compression, as encrypted data cannot be compressed effectively)
- The compression algorithms used in low-end products are not the most effective known today, and better results can usually be obtained by turning off hardware compression, using software compression (and encryption if desired) instead

SSD Disks

SSD Disks

Two kinds of Flash memory: NAND and NOR

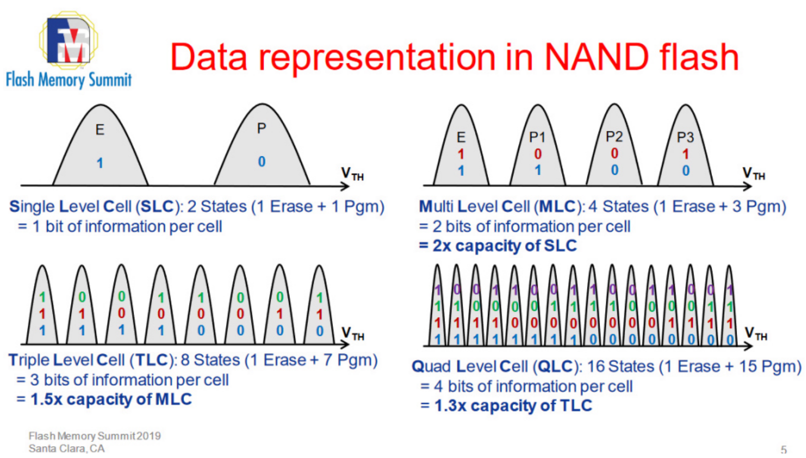
NAND Flash memory

- Block-level access (128-1024 KB)
- High write-erase cycle
- Fast erase time
- Low security (bit flipping)
- Massive storage (pen drive, **SSD disks**)

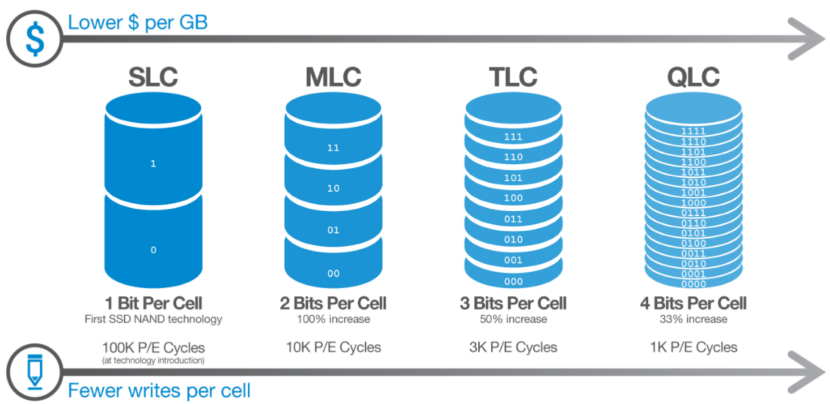
NOR Flash Memory

- Byte-level access (like a RAM), allows eXecute In Place (XIP)
- Low write-erase cycle
- Slow erase time (10 times slower than NAND)
- Highly secure
- Firmware, OS, Boot, BIOS

SSD technologies: SLC, MLC, TLC, QLC (usually NAND)



SSD technologies: SLC, MLC, TLC, QLC (usually NAND)



SSD versus HDD

Start-up time

- Faster in SSD (no spin-up)

Read and write speed

- SSD is faster ALWAYS (so, faster backups & recoveries)

Consumption and noise

- Less consumption in SSD (no rotation, no arm movement)
- HD produces much heat, so more power required for heat reduction
- Less noise in SSD (0 dB vs, 29-36 dB)

Failures

- No mechanical parts in SSD
- Non recoverable data in SSD
- SSD more vulnerable to electric / magnetic effects (shorter life)
  - We can discuss a lot about it…

SSD versus HDD

Deterministic performance

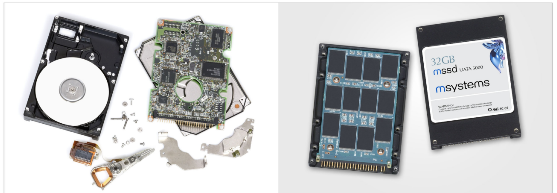
- In SSD you are able to know how much time an access can cost

Size and weight

- SSD is smaller and lighter

Storage capacity

- HD are bigger, but the problem is not the capacity, but cost (5-10 times more expensive the SSD)



STORAGE

David López