Sistemas Operativos
Capítulo 9
Dispostivos de armazenamento

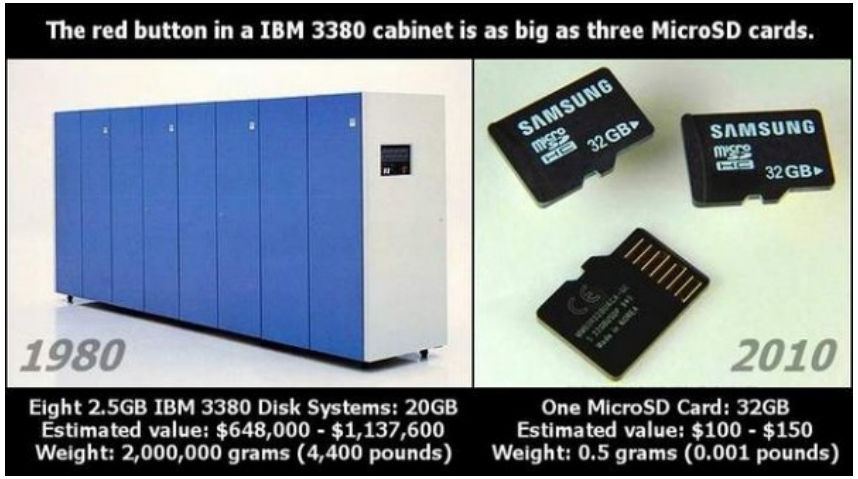
Patrício Domingues, ESTG/IPLeiria

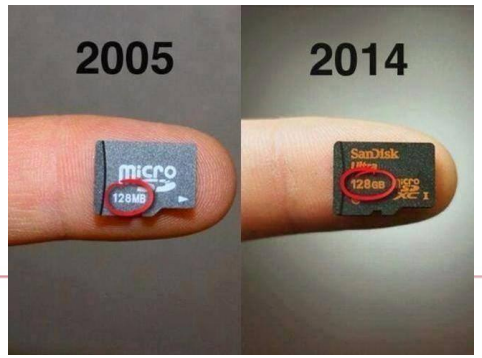
Adaptado de: CS5600 – Computer Systems – Lecture 8

STORAGE - HDD

Storage in ancient times…(#1)

- ✓1956: IBM RAMAC 305 / 5MB /
  - $50000 US dollars
- ✓1980: IBM 3380 / 20 GB
  - > $650000 US dollars

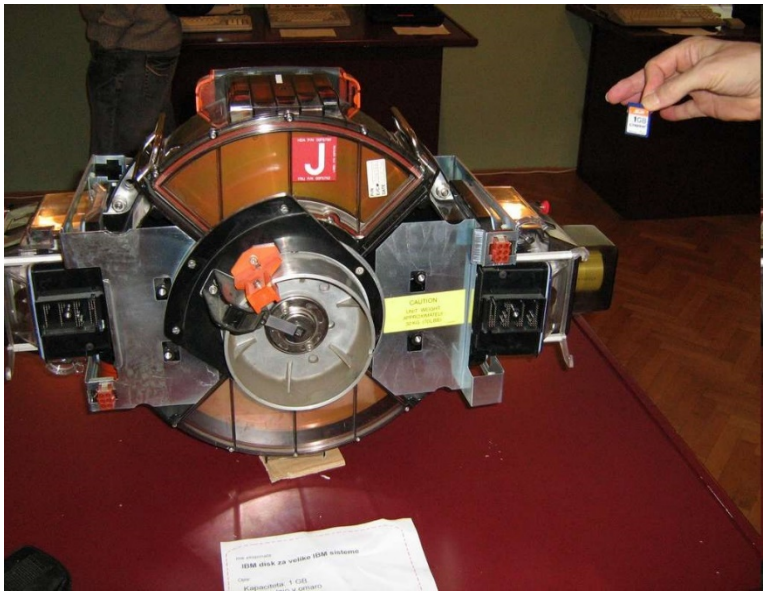




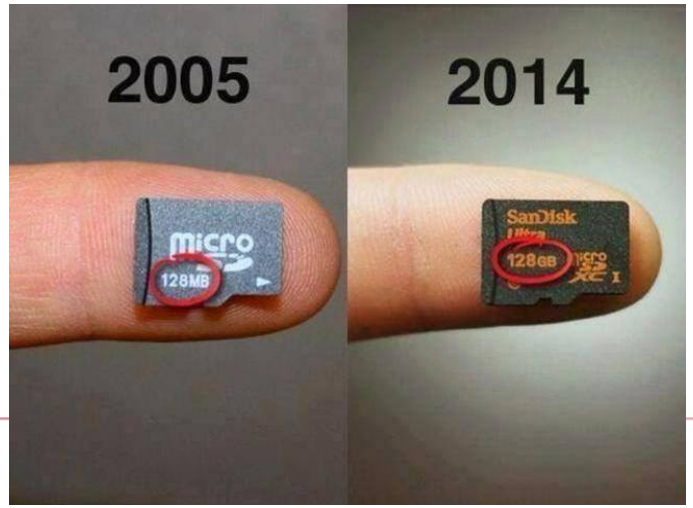
Storage in ancient times… (#2)

Storage evolution

- ✓1 GB HDD VS 1GB SD card...



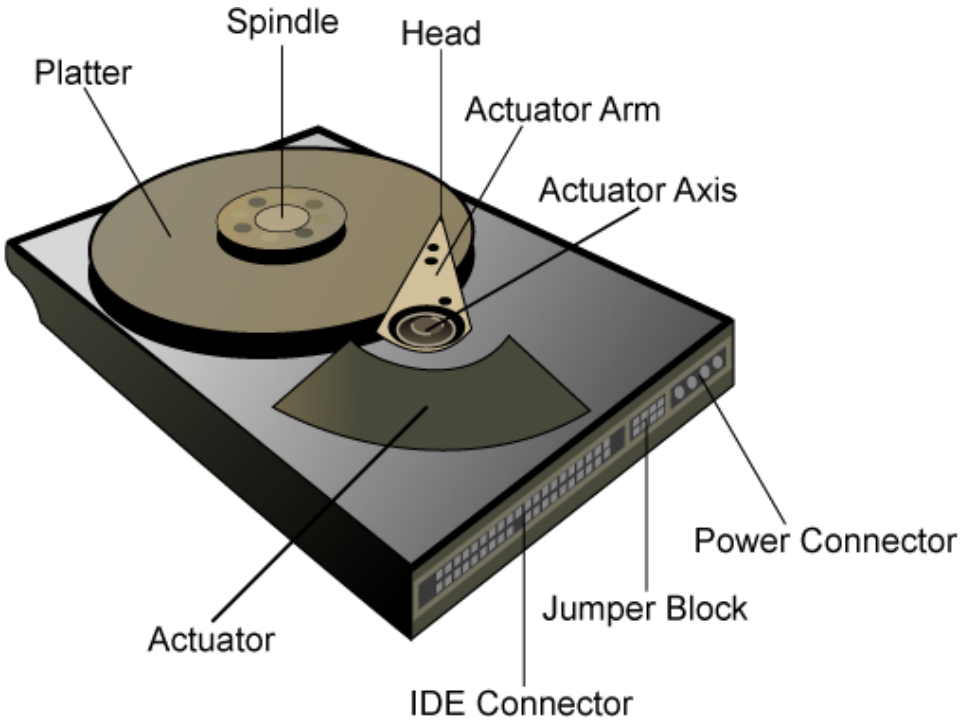
128 MB to 128 GB...

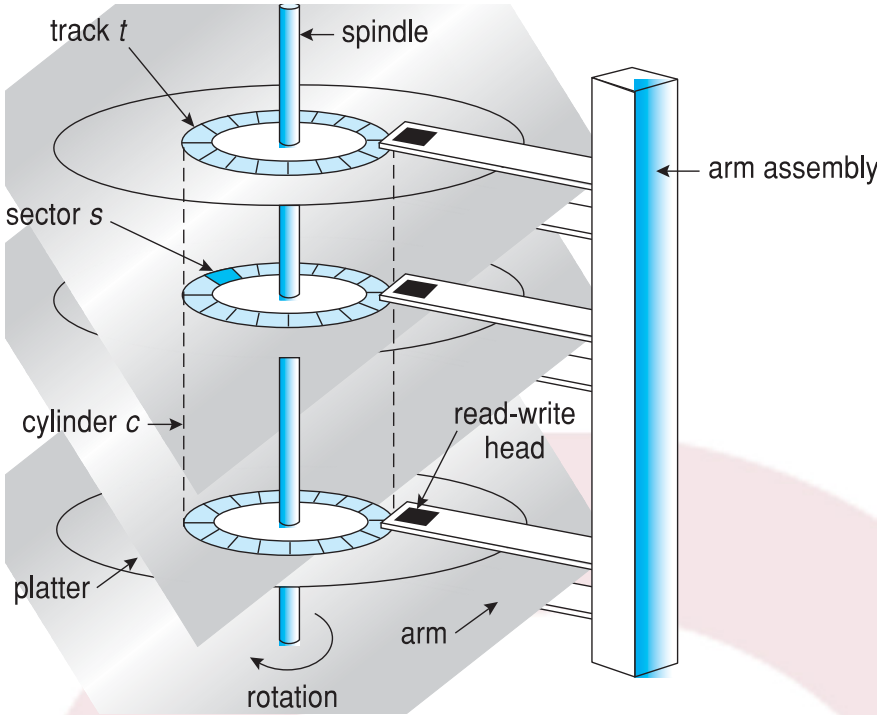


Boosting your storage is always a good idea, and today, gobbling up a lot of gigabytes doesn't have to be expensive. You can grab the 400GB SanDisk Ultra microSDXC card for just $83.98 on Amazon, down from a list price of $250 and a crazy low price for such a substantial amount of storage.

https://bit.ly/2Crq6oL

✓ *Hard Disk Drive* (HDD)

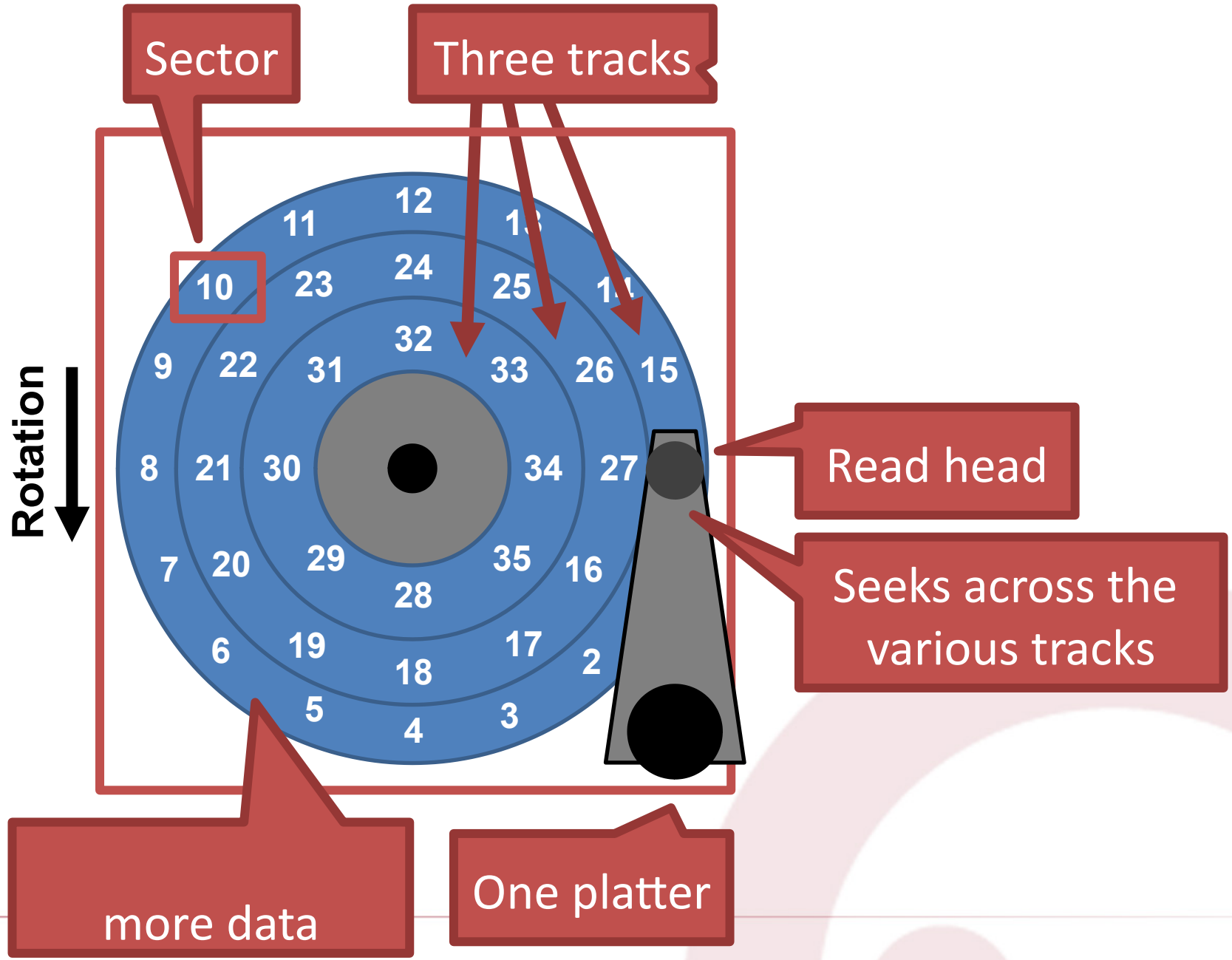




HDD: addressing and geometry

- ✓ Externally, hard drives expose a large number of sectors (blocks)
  - Typically 512 or 4096 bytes (newer disks)
  - Individual sector writes are atomic
  - Multiple sectors writes may be interrupted (torn write)
- ✓ Drive geometry
  - Sectors arranged into tracks
  - A cylinder is a particular track on multiple platters
  - Tracks arranged in concentric circles on platters
  - A disk may have multiple, double-sided platters
- ✓ Drive motor spins the platters at a constant rate
  - Measured in revolutions per minute (RPM)
  - Examples: 5400 rpm; 10000 rpm

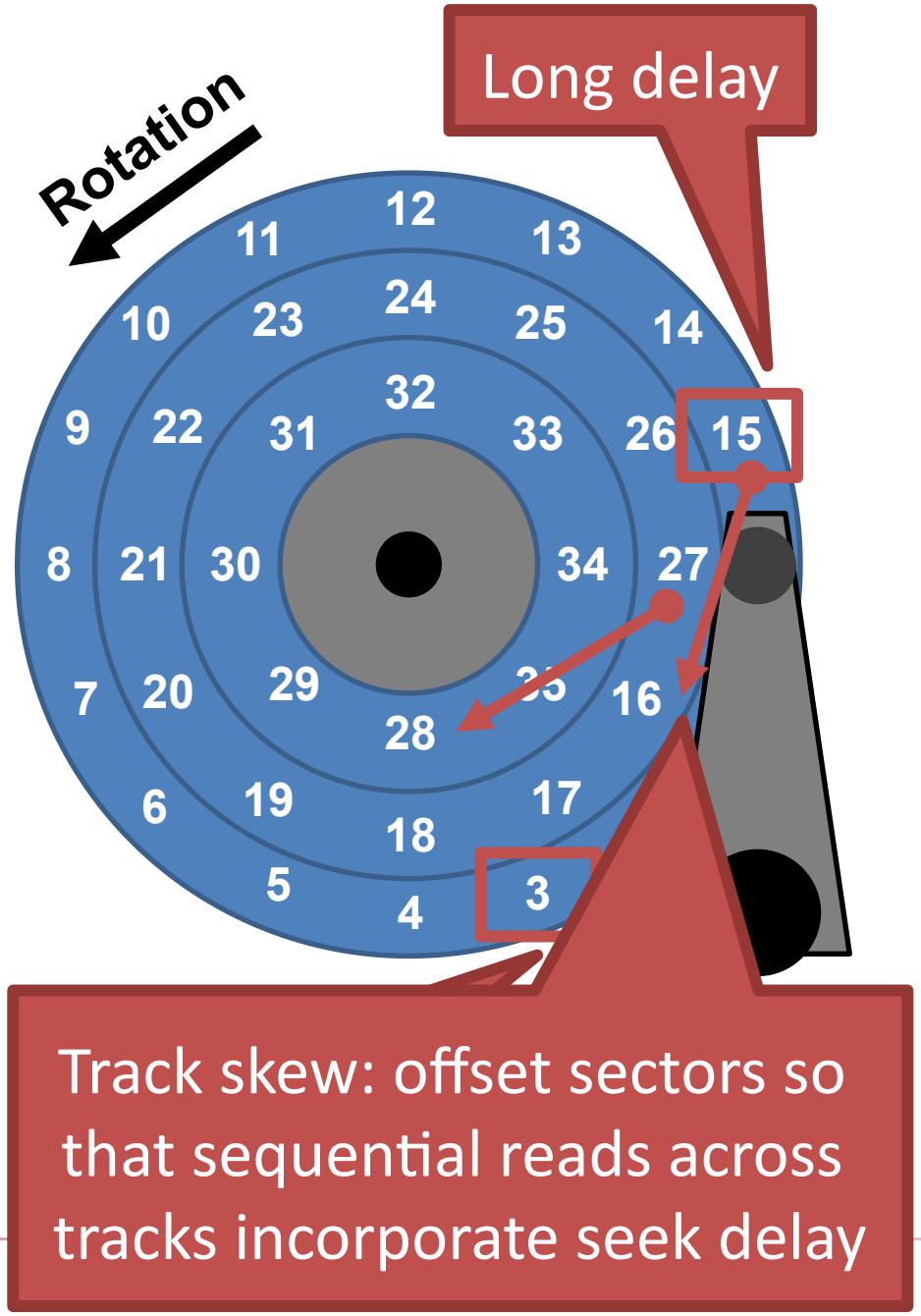
Geometry example



Common Disk Interfaces

- ✓ ST-506 → ATA → IDE → SATA (serial ATA)
  - Ancient standard
  - Commands (read/write) and addresses in cylinder/head/sector format placed in device registers
  - Recent versions support Logical Block Addresses (LBA)
- ✓ SCSI (Small Computer Systems Interface)
  - Packet based, like TCP/IP
  - Device translates LBA to internal format (e.g. c/h/s)
  - Transport independent
    - USB drives, CD/DVD/Bluray, Firewire
    - iSCSI is SCSI over TCP/IP and Ethernet

Types of Delay With Disks



Three types of delay

1. Rotational Delay
   - Time to rotate the desired sector to the read head. Related to RPM
2. Seek delay
   - Time to move the read head to a different track
3. Transfer time
   - Time to read or write bytes

Sequential vs. Random Access

Rate of I/O

$R_{I/O} = transfer\_size / T_{I/O}$

Access Type	Transfer Size		Cheetah 15K.5	Barracuda
Random	4096 B	$T_{I/O}$	6 ms	13.2 ms
		$R_{I/O}$	0.66 MB/s	0.31 MB/s
Sequential	100 MB	$T_{I/O}$	800 ms	950 ms
		$R_{I/O}$	125 MB/s	105 MB/s
Max Transfer Rate			125 MB/s	105MB/s

Random I/O results in very poor disk performance!

Caching

- ✓ Many disks incorporate caches (track buffer)
  - Small amount of RAM (8, 16, or 32 MB)
- ✓ Read caching
  - Reduces read delays due to seeking and rotation
- ✓ Write caching
  - Write back cache: drive reports that writes are complete after they have been cached
    - Possibly dangerous feature. Why?
  - Write through cache: drive reports that writes are complete after they have been written to disk
- ✓ Today, some disks include flash memory for persistent caching (hybrid drives)

Command Queuing

- ✓ Feature where a disk stores of queue of pending read/write requests
  - Called Native Command Queuing (NCQ) in SATA
- ✓ Disk may reorder items in the queue to improve performance
  - E.g. batch operations to close sectors/tracks
- ✓ Supported by SCSI and modern SATA drives
- ✓ Tagged command queuing
  - allows the host to place constraints on command re-ordering

HDD filled with helium

https://bit.ly/1qQtUXC

STORAGE - SDD

Beyond Spinning Disks

- ✓ Hard drives have been around since 1956
  - The cheapest way to store large amounts of d
  - Sizes are still increasing rapidly
- ✓ However, hard drives are typically the slowest component in most computers
  - CPU and RAM operate at GHz
  - PCI-X and Ethernet are GB/s
- ✓ Hard drives are not suitable for mobile devices
  - Fragile mechanical components can break
  - The disk motor is extremely power hungry

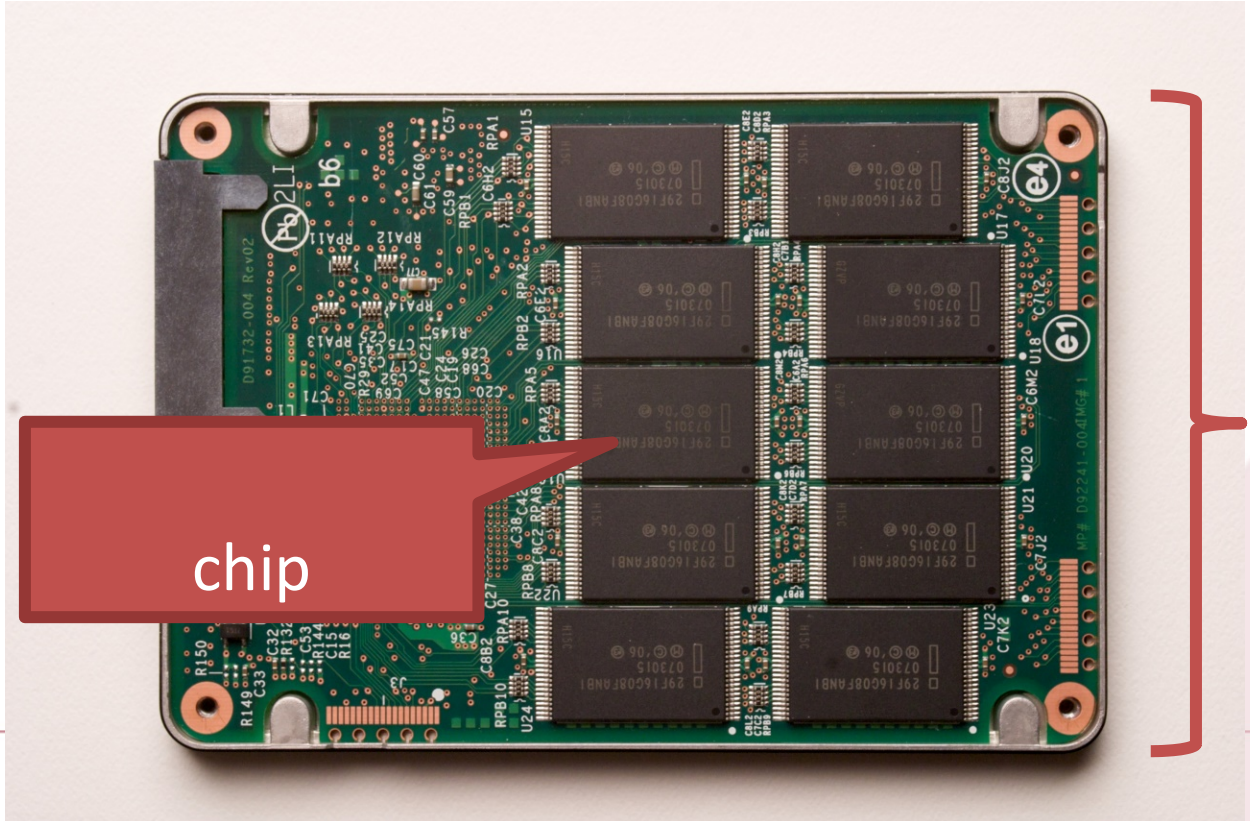
5 MB HDD (1956)

Disco Rígido 3.5" Seagate Barracuda 4TB 5400RPM 256MB SATA III

108,90 €

Solid State Drives

- ✓ NAND flash memory-based drives
  - High voltage is able to change the configuration of a floating-gate transistor
  - State of the transistor interpreted as binary data



Advantages of SSDs

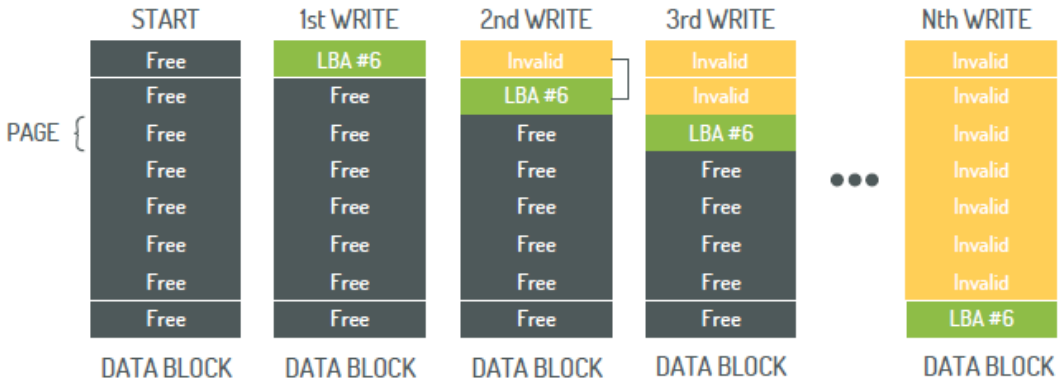
- ✓ More resilient against physical damage
  - No sensitive read head or moving parts
  - Immune to changes in temperature (kind of)
- ✓ Greatly reduced power consumption
  - No mechanical, moving parts
- ✓ Much faster than hard drives
  - >500 MB/s vs ~200 MB/s for hard drives
    - Newer NVMe SSD can reach 3500 MB/s
  - No penalty for random access
    - Each flash cell can be addressed directly
    - No need to rotate or seek
  - Extremely high throughput
    - Although each flash chip is slow, they are RAIDed

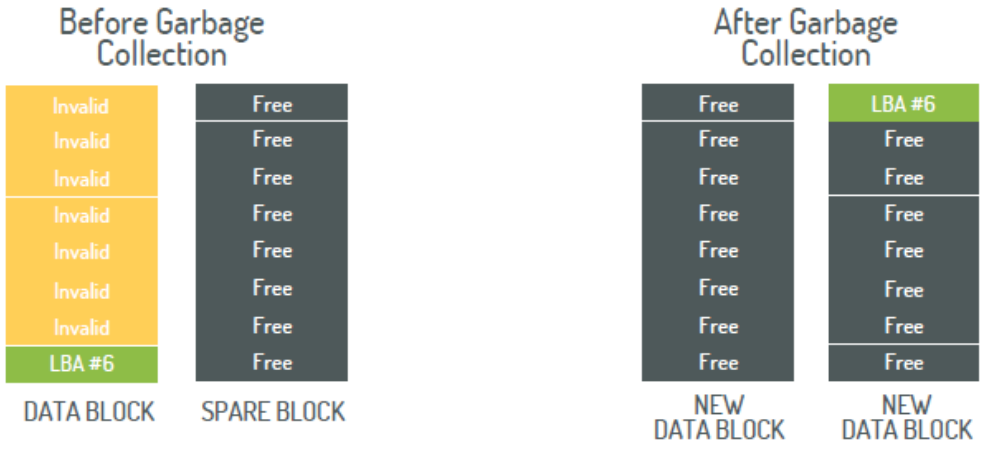
Challenges with Flash

- ✓ Flash memory is written in pages, but erased in blocks
  - Pages: 4 – 16 KB, Blocks: 128 – 256 KB
  - Thus, flash memory can become fragmented
  - Leads to the write amplification problem
- ✓ Flash memory can only be written a fixed number of times
  - Typically 3000 – 5000 cycles for MLC
  - SSDs use wear leveling to evenly distribute writes across all flash cells

Wear Leveling

- ✓ Recall: each flash cell wears out after several thousand writes
- ✓ SSDs use wear leveling to spread writes across all cells
  - Typical consumer SSDs should last at least ~5 years





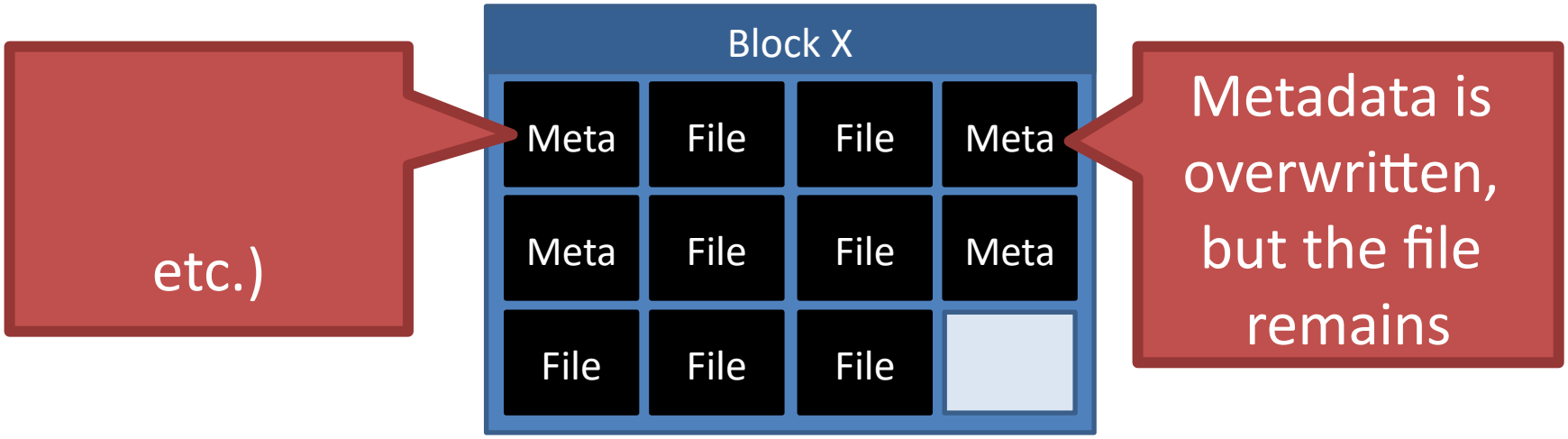
Garbage Collection

- ✓ Garbage collection (GC) is vital for the performance of SSDs
- ✓ Older SSDs had fast writes up until all pages were written once
  - Even if the drive has lots of "free space," each write is amplified, thus reducing performance
- ✓ Many SSDs over-provision to help the GC
  - 240 GB SSDs actually have 256 GB of memory
- ✓ Modern SSDs implement background GC
  - However, this doesn't always work correctly

The Ambiguity of Delete

- Goal: the SSD wants to perform background GC
  - But this assumes the SSD knows which pages are invalid
- Problem: most file systems do not actually delete data
  - On Linux, the "delete" function is unlink()
  - Removes the file meta-data, but not the file itself

Delete Example

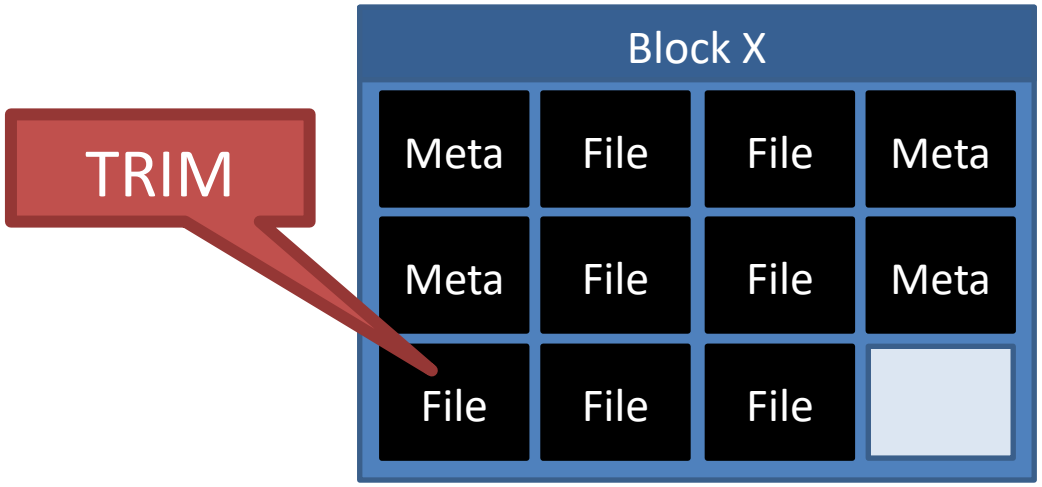


1. File is written to SSD
2. File is deleted
3. The GC executes
   - 9 pages look valid to the SSD
   - The OS knows only 2 pages are valid

- Lack of explicit delete means the GC wastes effort copying useless pages
- Hard drives are not GCed, so this was never a problem

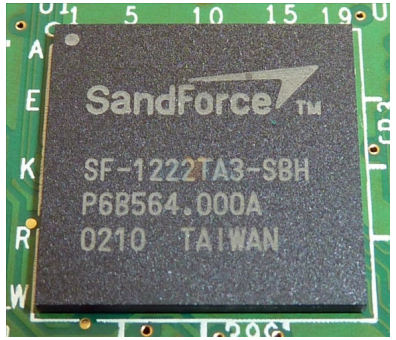
TRIM

- ✓SATA command TRIM (SCSI – UNMAP)
  - Allows the OS to tell the SSD that specific LBAs are invalid, may be GCed



- OS support for TRIM
  - Win >=7, OSX Snow Leopard, Linux >= 2.6, Android 4.3
- Must be supported by the SSD firmware

SSD Controllers



- SSDs are extremely complicated internally

✓ All operations handled by the SSD controller
  - Maps LBAs to physical pages
  - Keeps track of free pages, controls the GC
  - May implement background GC
  - Performs wear leveling via data rotation

✓ Controller performance is crucial for overall SSD performance

Flavours of NAND Flash Memory

Multi-Level Cell (MLC)

- Multiple bits per flash cell
  - For two-level: 00, 01, 10, 11
  - 2, 3, and 4-bit MLC is available
- Higher capacity and cheaper than SLC flash
- Lower throughput due to the need for error correction
- 3000 – 5000 write cycles
- Consumes more power

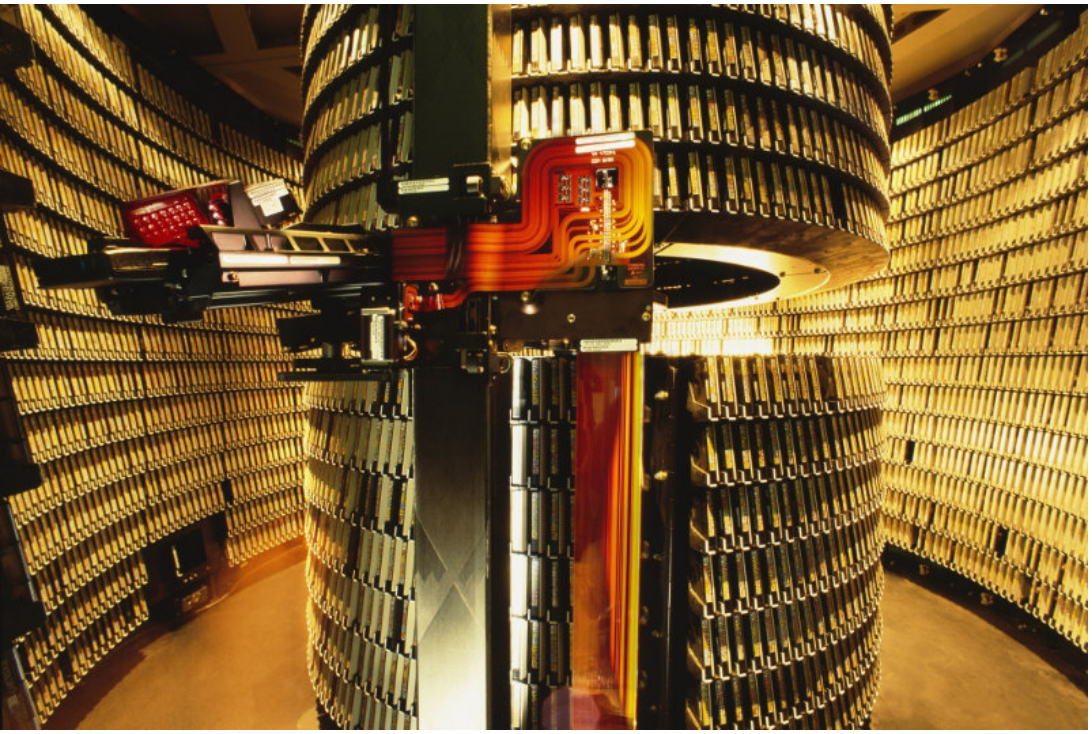
Consumer-grade drives

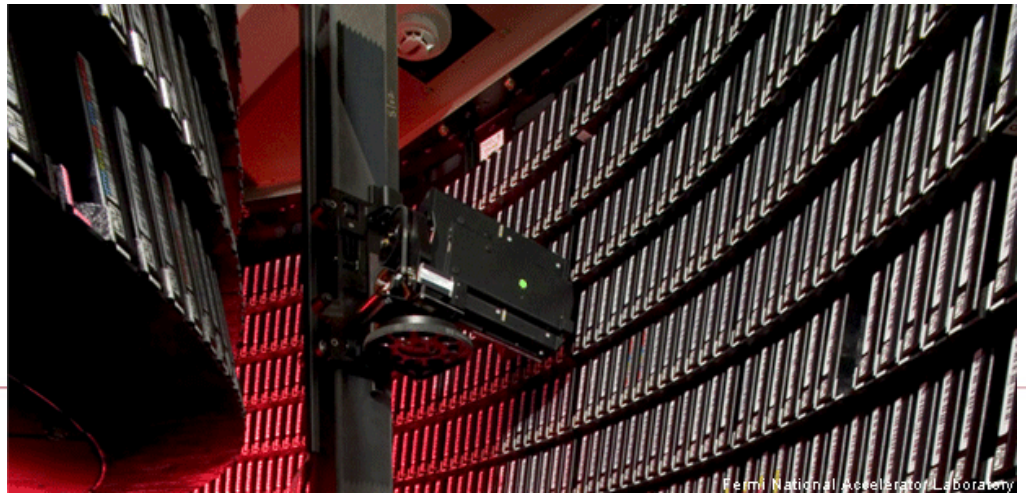
Single-Level Cell (SLC)

- One bit per flash cell
  - 0 or 1
- Lower capacity and more expensive than MLC flash
- Higher throughput than MLC
- 10000 – 100000 write cycles

Expensive, enterprise drives

✓Tape robot



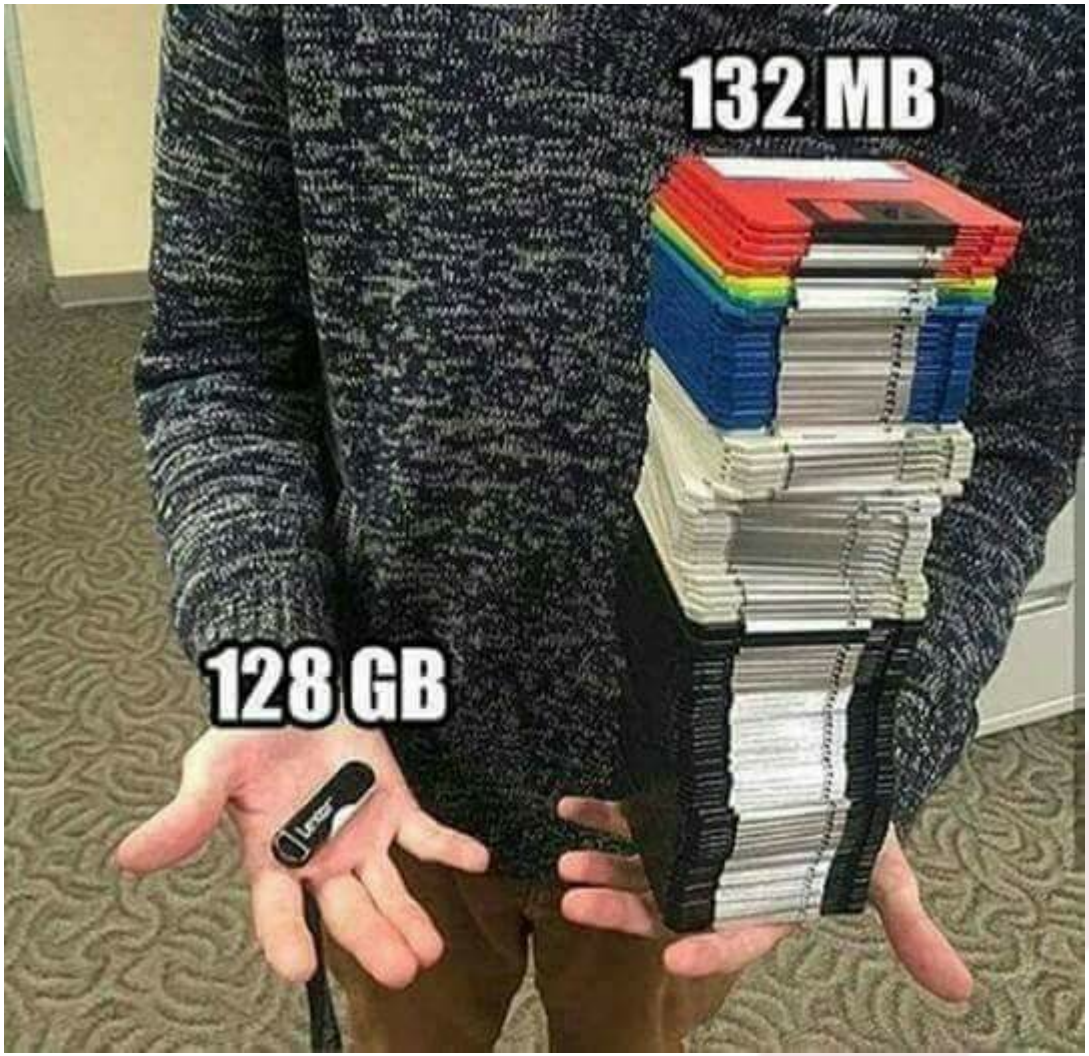


https://bit.ly/2q4cREY

The long life of floppies...

- ✓ Floppy
  - USA's nuclear arsenal still uses floppy disk technology from 1970...
    - 8'' disks (200 mm) of 237.25 KiB
    - The disks/readers will be replaced until the end of 2020
  - Floppies are still used in many ancient manufacturing devices
  - "We regularly read floppy disks from 40 years ago and they are as good as new"
    - Kevin Murrell, The National Museum of Computing at Bletchley Park
  - Source: http://www.bbc.com/news/technology-36389711

But...



Storage unit (#1)

- ✓ Measuring storage is done in byte
  - A byte is 8 bits
  - Also known as an octet
- ✓ What's a KiloByte?
  - System SI: 1 kilo = 10^3 = 1000 = 1000 bytes
    - (The same goes with KM, KV, KG, etc.)
  - But...computers are binary devices
    - Storage is measured in power of 2
    - 2^{10} = 1024
    - 1 KB = 1024 bytes

So, which is the correct unit? >>

Storage unit (#2)

- ✓ KiloByte (prefix Kilo)
  - Follows International System (SI)
  - 1 KB = 1000 bytes
- ✓ Power of 2
  - Defined by International Electrotechnical Commission (IEC) – IEC60027-2
  - 2^{10} = 1024 bytes
  - Name: KibiByte
  - 1 KiB = 2^{10} = 1024 bytes

Multiples of bytes V · T · E

Decimal			Binary				
Value	Metric		Value	JEDEC		IEC	
1000	kB	kilobyte	1024	KB	kilobyte	KiB	kibibyte
1000^2	MB	megabyte	1024^2	MB	megabyte	MiB	mebibyte
1000^3	GB	gigabyte	1024^3	GB	gigabyte	GiB	gibibyte
1000^4	TB	terabyte	1024^4	-	-	TiB	tebibyte
1000^5	PB	petabyte	1024^5	-	-	PiB	pebibyte
1000^6	EB	exabyte	1024^6	-	-	EiB	**exbibyte**
1000^7	ZB	zettabyte	1024^7	-	-	ZiB	zebibyte
1000^8	YB	yottabyte	1024^8	-	-	YiB	yobibyte

Orders of magnitude of data

http://en.wikipedia.org/

Storage unit (#3)

- ✓ Storage devices vendors use SI units
- ✓ SO and others software use the BI units (KiB, MiB, GiB, etc.) as defined by IEC
- ✓ Exemple
  - USB pen sold as “16 GB”
    - It has 14.6 GiB

When we say	but mean	we're this far off
1 kilobyte	2^{10} bytes	2.4%
1 megabyte	2^{20} bytes	4.9%
1 gigabyte	2^{30} bytes	7.4%
1 terabyte	2^{40} bytes	10.0%
1 petabyte	2^{50} bytes	12.6%
1 exabyte	2^{60} bytes	15.3%

http://cnx.org/contents/q4PmFDpc@1/Prefixes-for-binary-multiples

http://vidadesuporte.com.br/suporte-a-serie/anexo/