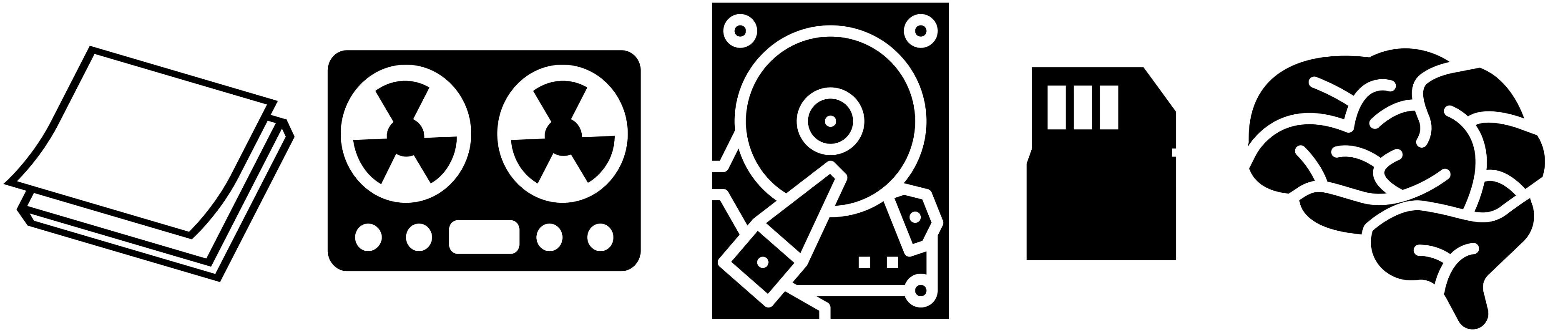
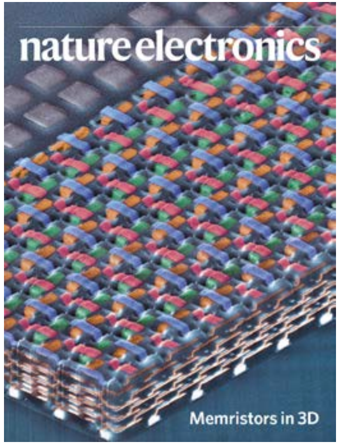
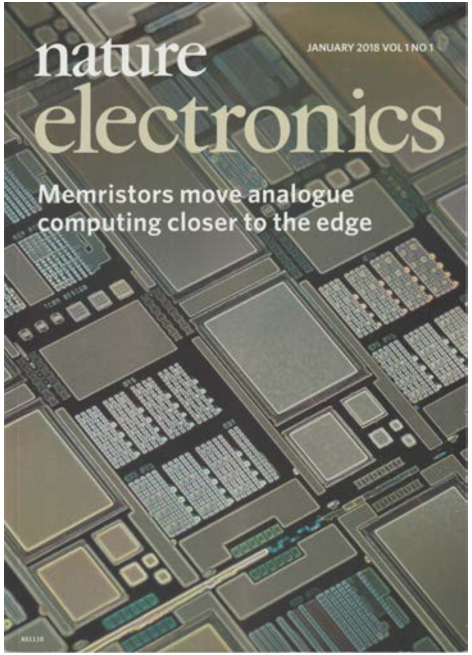
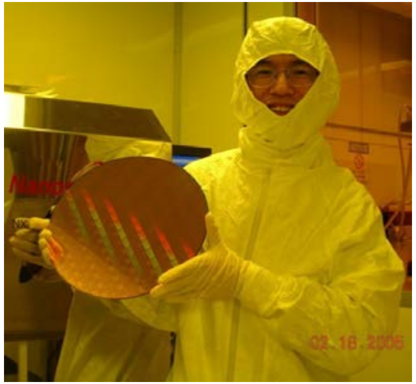
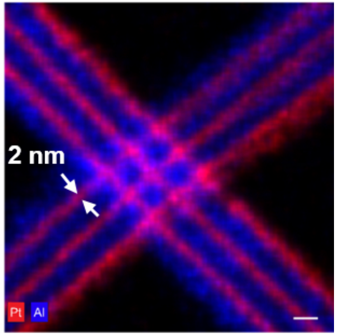


Professor Qiangfei Xia- Nanoelectronics (memristors)









2nm memristors

University of Massachusetts Amherst

Module 11 Memory and Storage

ENGIN 112 – Introduction to Electrical and Computer Engineering

Module Plan

- Fundamental Concepts
  - Some basics concepts about memory and storage
  - Early technologies
  - Sequential and Random Access
  - The role of memory and storage for computers
- Solid state Memory and Storage
  - Floating gate transistors
  - Other types of memory
  - Dynamic and Static RAM
  - Layered memory

Basics about memory & storage

- A bit is a binary digit, a term first coined by John Tukey at AT&T Bell labs in 1947.
- Tukey is also well known for developing the Fast Fourier Transform (or FFT!)
- The word "bit" is handy, because it indicates something very small, in this case, the smallest unit of information

  0 – FALSE 1 - TRUE
- A byte, of course, is 8 bits. A play on words, like "biting" something more than just a little bit. The use of the "y" helps avoid confusion with bit, and bite.
- A nibble, is 4 bits, like a tiny bit of a byte!

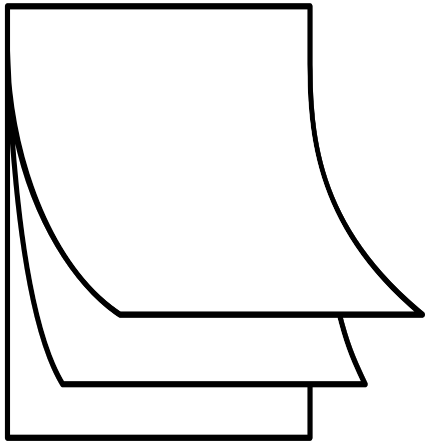
More words

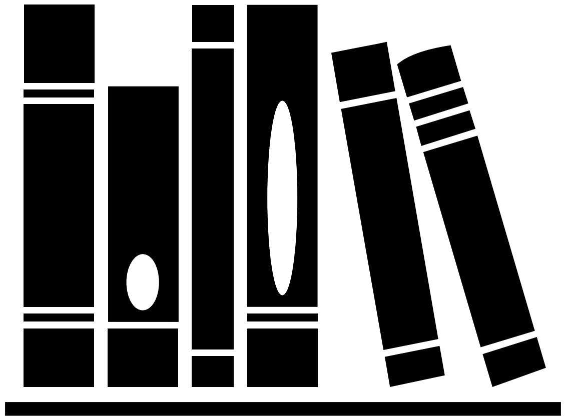
- Things get confusing when we start using metric expressions.
- A kilobyte (kB)
  - can either be 1000 bytes, or 1024 bytes.
- A Kibibyte (KiB)
  - is 1024 bytes (2^{10} bytes)
- What is a Gibibyte (GiB)?
  - (2^{30} bytes)
- What is a Gigabyte (GB)?
  - (2^{30} bytes or 1×10^9 bytes)

- Hard drive manufactures often specify numbers in Gigabytes and Terabytes. Which one do they usually mean?
  - Answer: 1 GB for a hard drive = 1×10^9 bytes
  - Hard drives are not necessarily in base-2, like many other digital systems are.

Memory versus Storage

- Storage: Save information "permanently"
- Memory: Store information "temporarily"





Methods of storing information

- Drawings and books (writing).
- Also a handy way of communicating

8000 years ago, ancient Chinese would put small pictures on tortoise shells that they would use to predict the future.

The oldest cave drawings are more than 40,000 years old

Written words for communicating

- Papyrus scrolls and hand-written books could pass on stories and communicate thoughts over many generations

Dead sea scrolls, 200 BCE

Hand written religious text, 800 AD

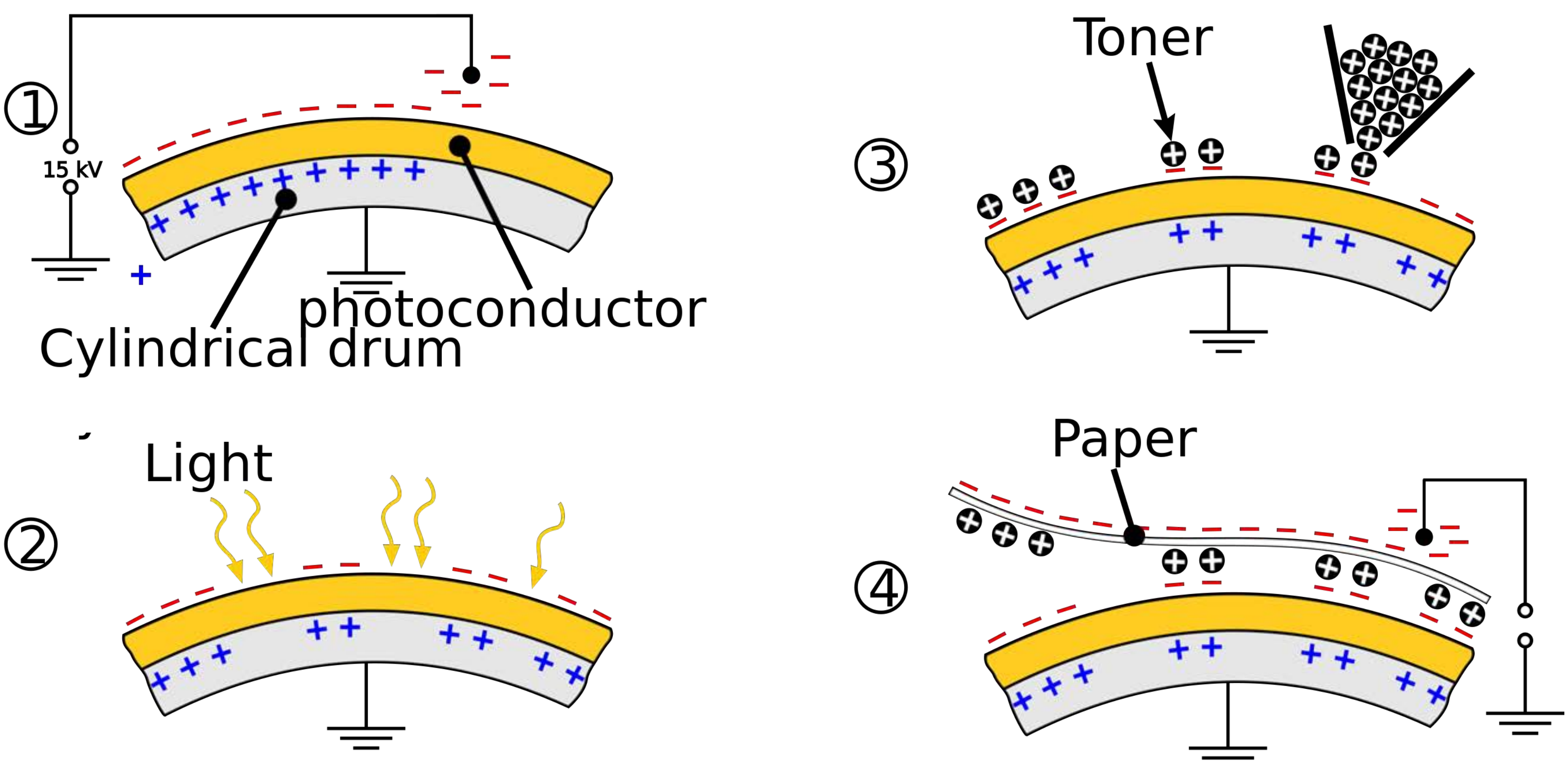
Mechanized writing

- The Guttenberg press (~1440)

Now it was possible to produce copies of the same text, many times over

Mechanized Copying

- One of the great inventions of the 20th century, the first photocopies were made in the US in the 1940’s
- A photograph is a copy of an original instant in time.
- With photocopying, we can take documents and make many replicas.



Other forms of memory & mechanized copying

A player piano

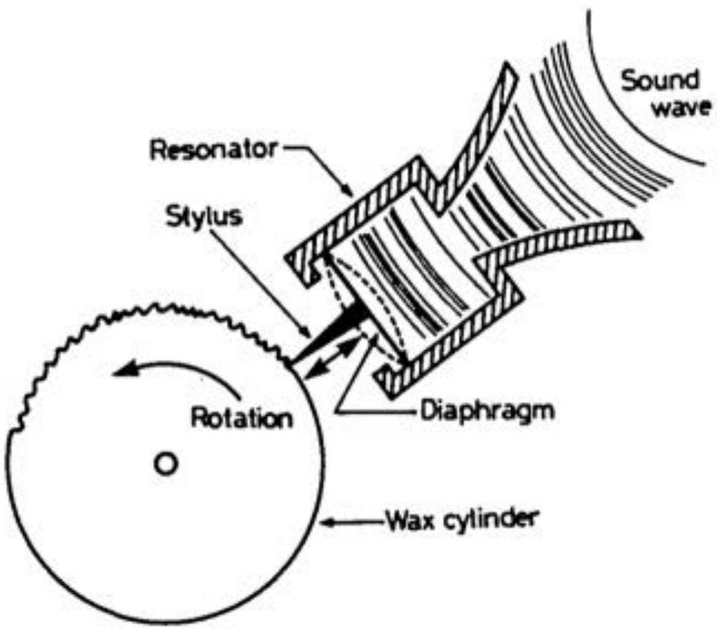
Rectangular windows cut out of a scrolling paper could be used to control small levers to replicate the mechanical levers of a piano

A phonograph

Grooves in a rotating wax cylinder are used to vibrate a membrane and create sound

What are the differences and similarities?

Early phonograph



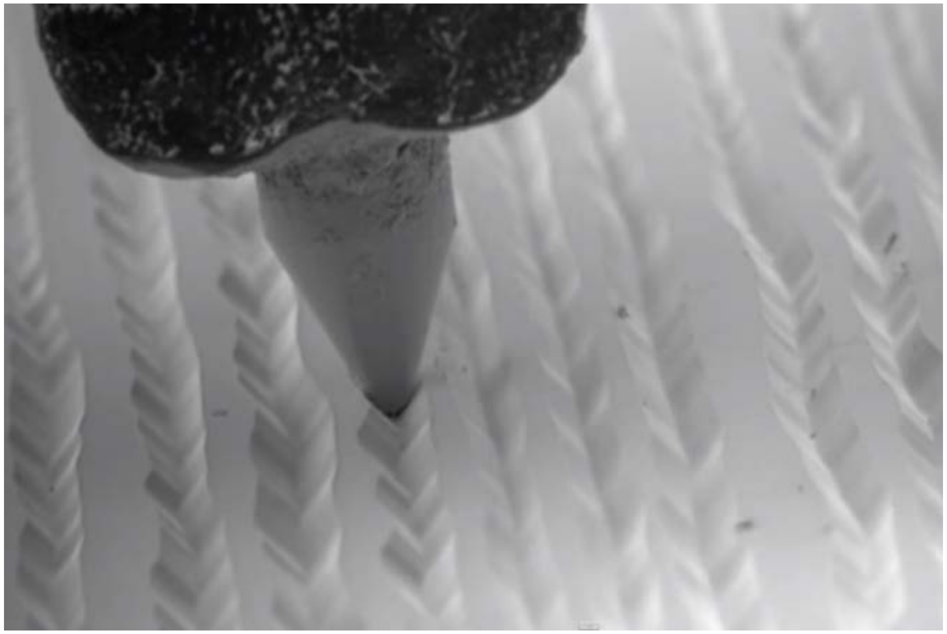
- Early audio recordings were etched and stored on these wax cylinders.
- Because they were hollow inside, they could easily break
- They also took up a lot of space for storage!

Time and technology march on

- A disk (or plate) made out of vinyl was much more durable and could easily be stacked, and take up much less space

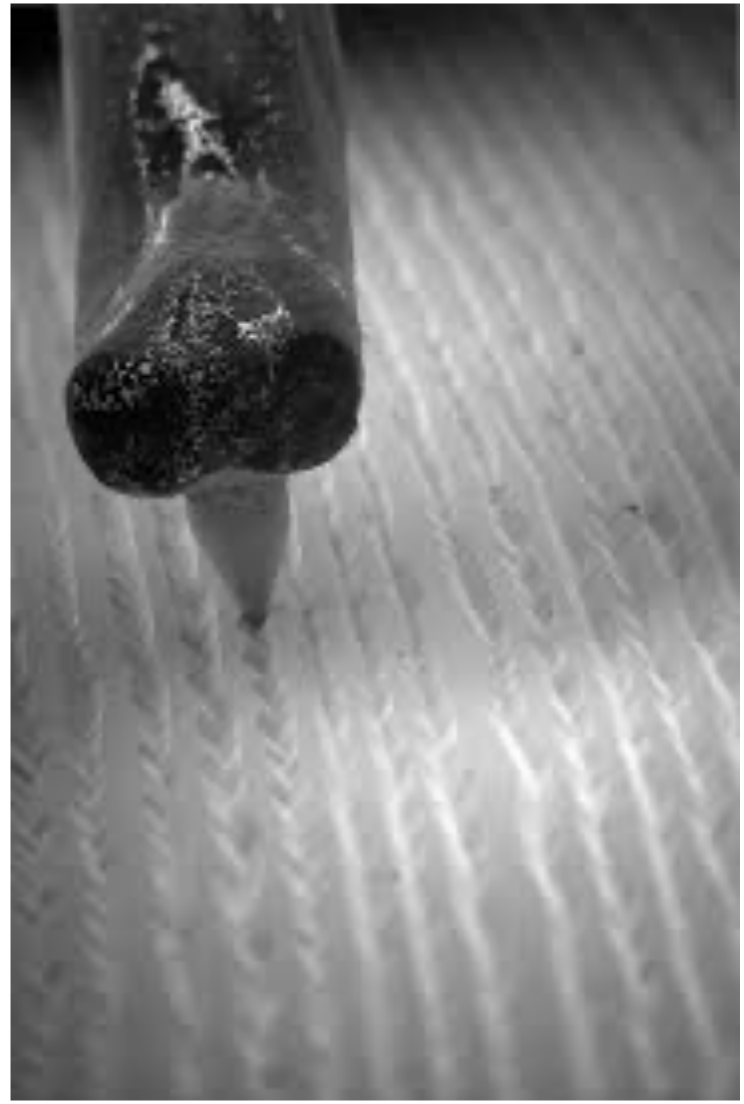
What are the dots used for on the turntable?

Closeup of a vinyl record



So far, most of the types of storage and memory that we have looked at, are mechanical in nature.

What other technologies might be used for storing?

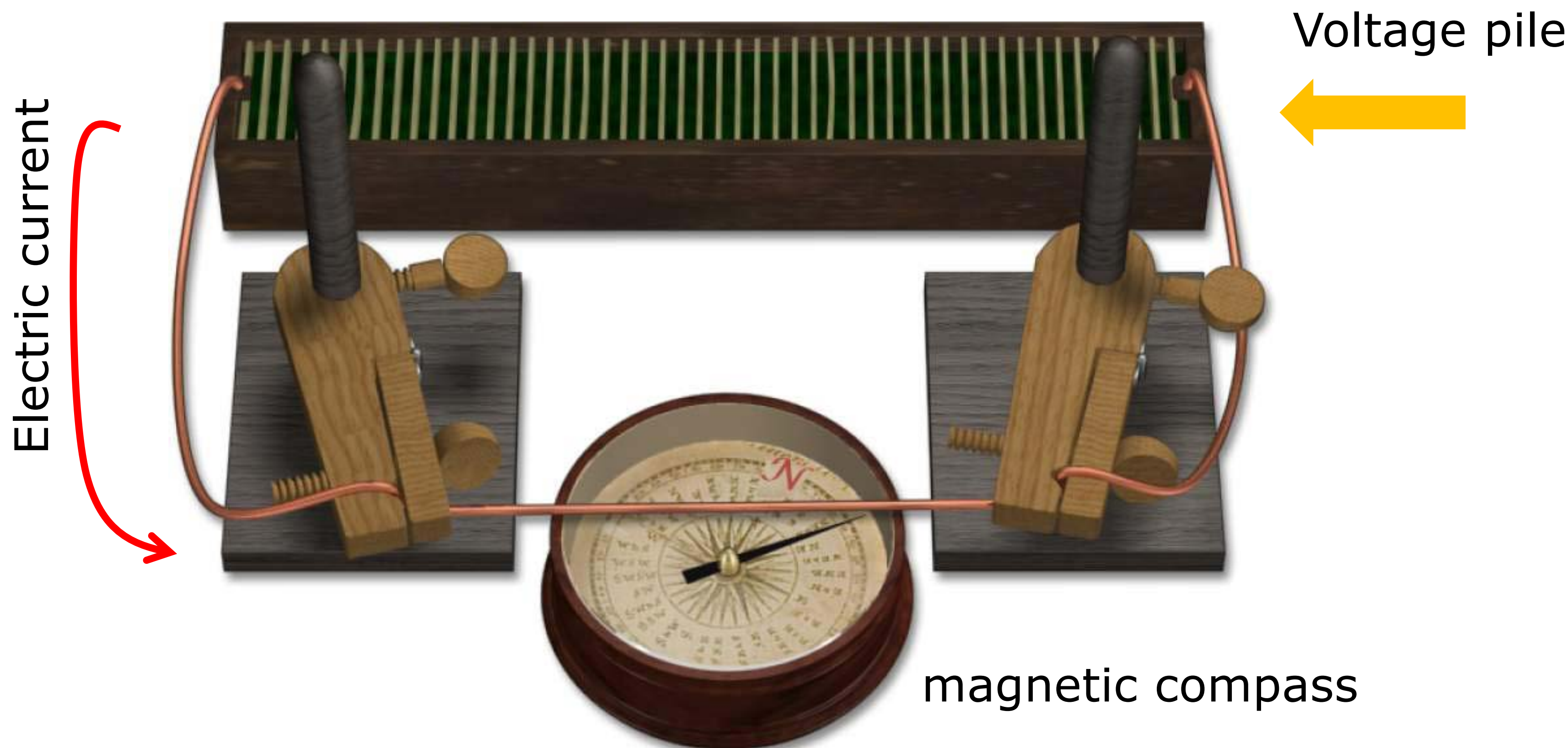


Magnetic storage

- Magnets are a form of storage
- When heated, magnetism is lost. As a magnet is cooled, it will “remember” or “store” the state of the magnetic field surrounding it.

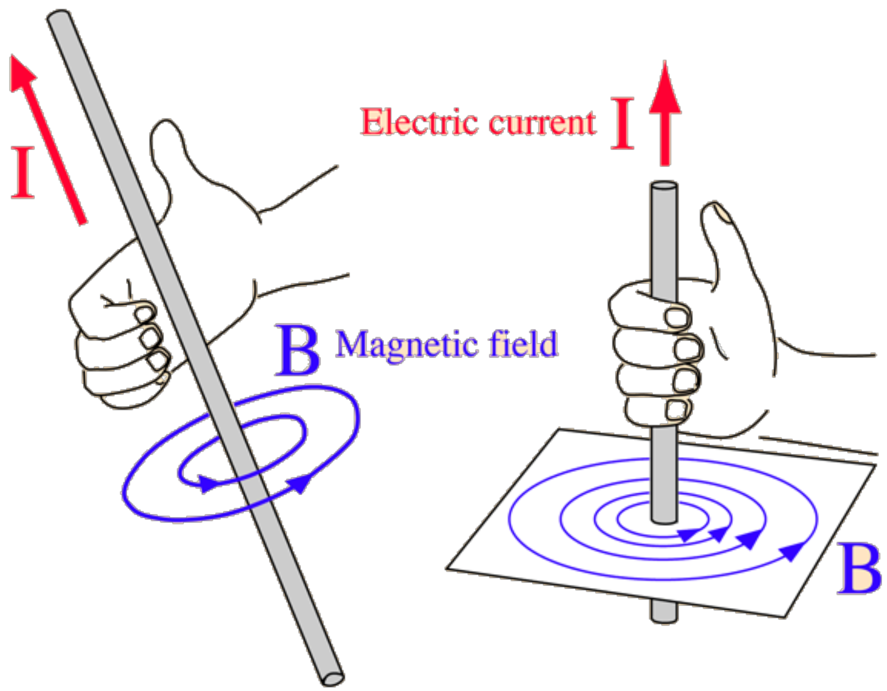
Artificial magnets

- From Oersted's work in the 1820's, he demonstrated that moving charges (a current) could deflect a compass.
- In other words, the current was generating its own magnetic field!

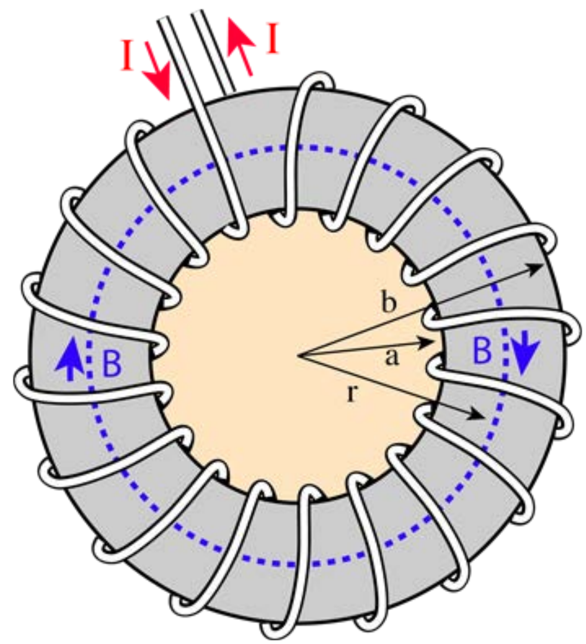


Ampere’s Law

- Magnetic field orientation around a current
- Using the right-hand rule, we can determine the direction of the magnetic field



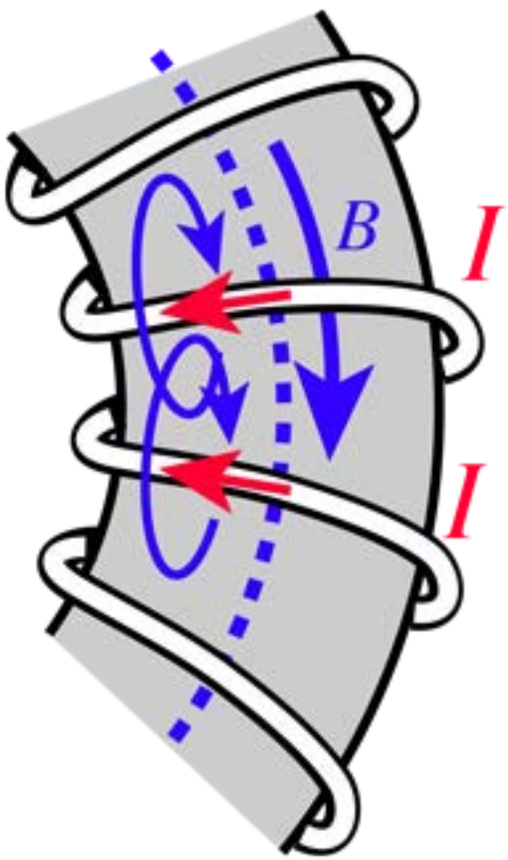
”A line current” – thumb in the direction of the current. Fingers in the direction of the magnetic field, **B**



When a wire is wrapped around a toroid, the magnetic field is circular.
This is called a magnetic core

Toroidal Current

- A closeup

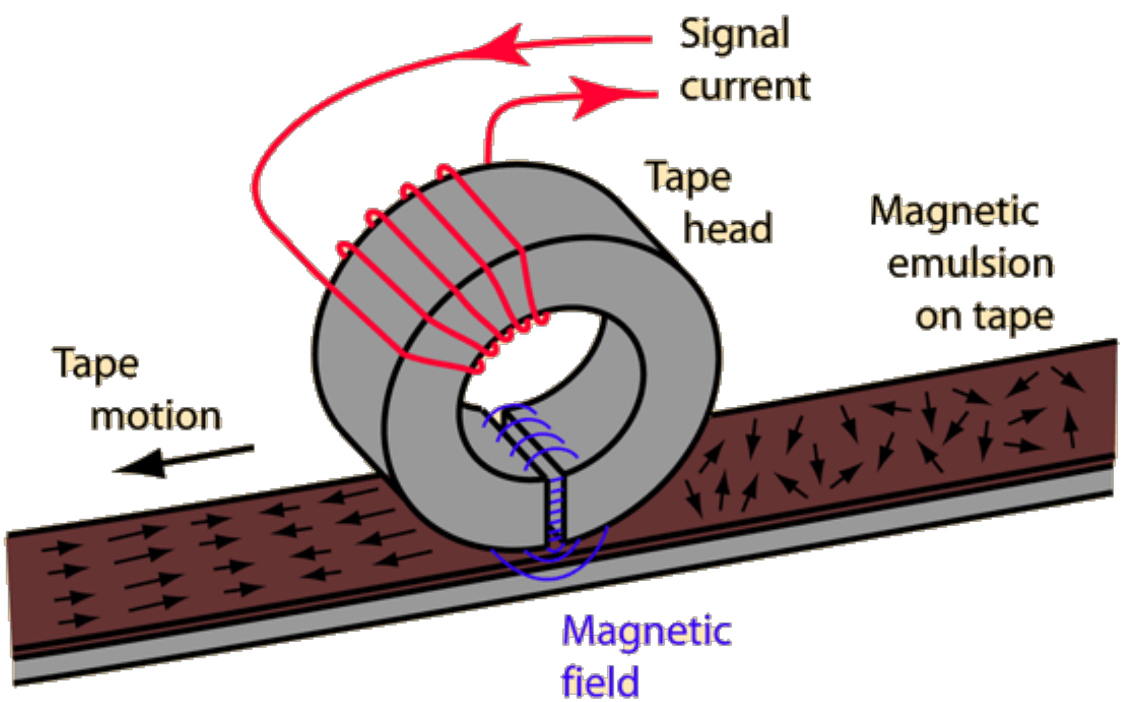


Each loop of current, contributes a small magnetic field that is held by the magnetic core

If we break the continuity of the magnetic core, the magnetic field will expand outside of the magnetic material

Artificial magnets & storage

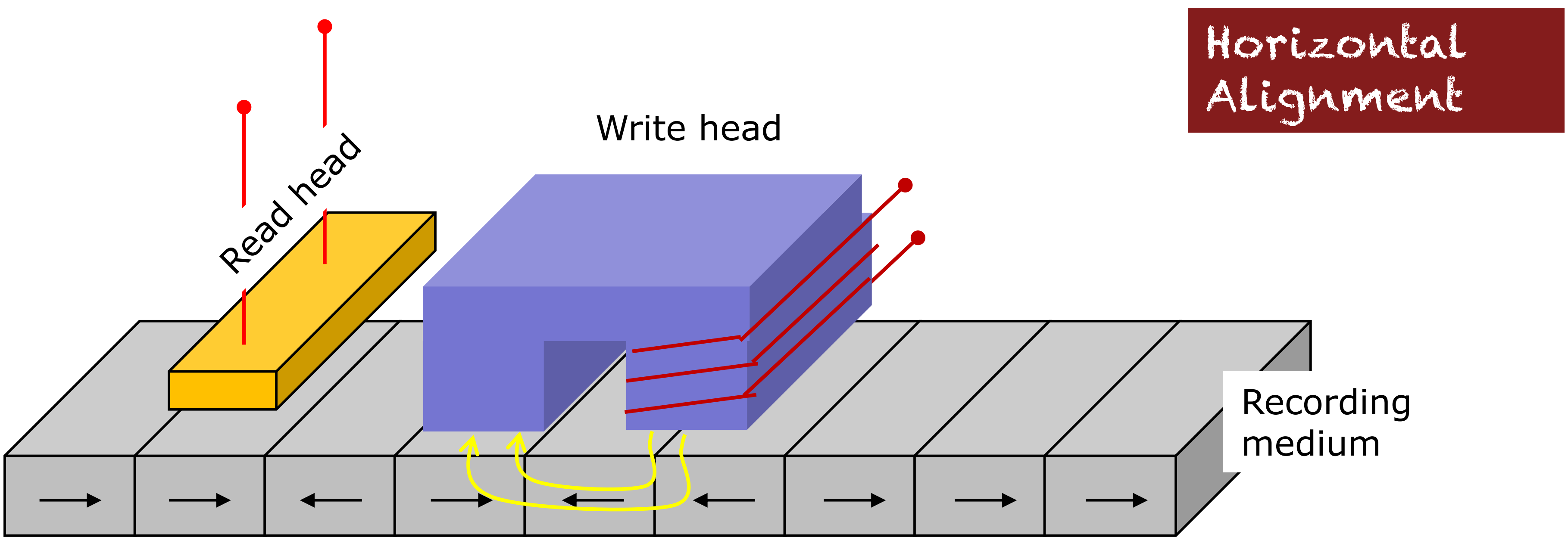
- Use of a strong magnetic field, artificially induced, and be used to align the weaker magnets bound to a surface



We can use a changing current to change the direction of magnetic fields on medium

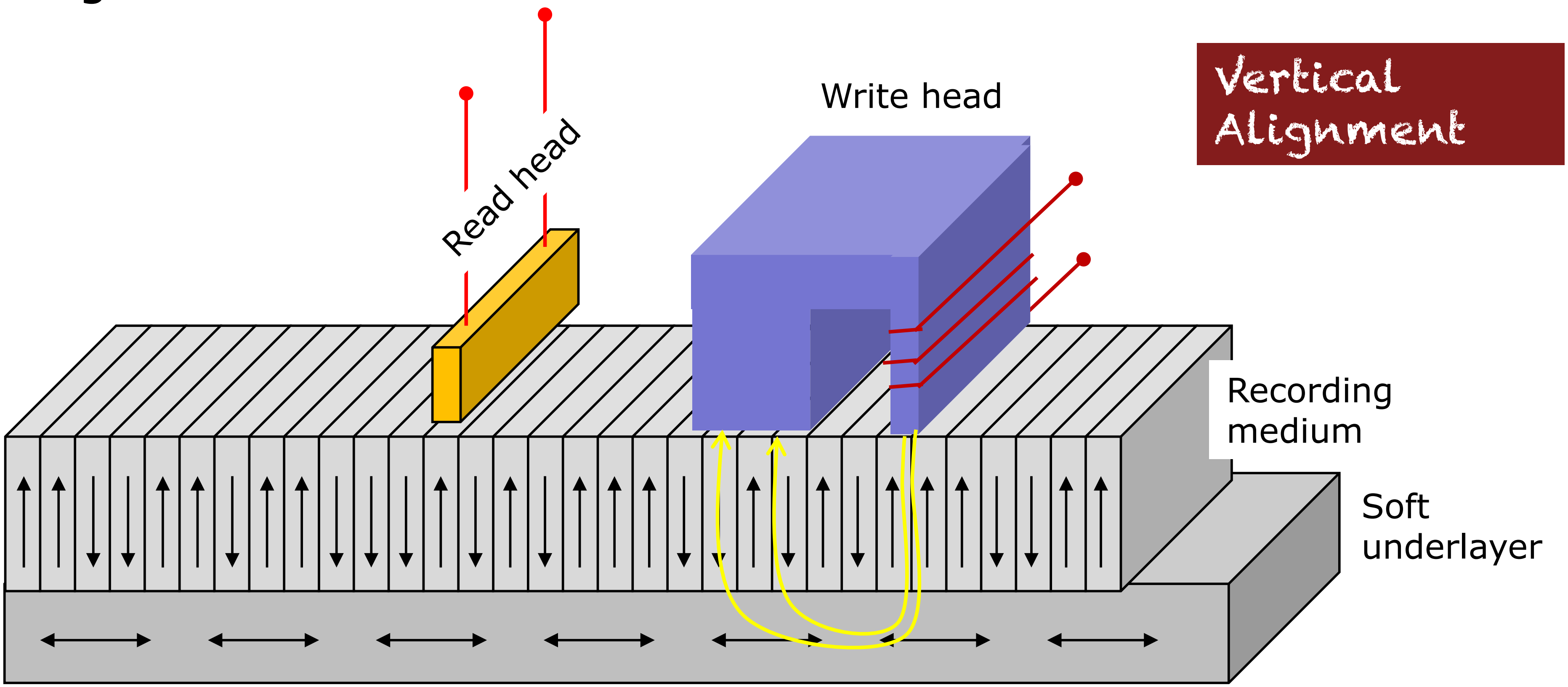
Magnetic Storage

- Alignment of magnetic fields and magnetic susceptibility can affect the density of data stored on the magnetic medium



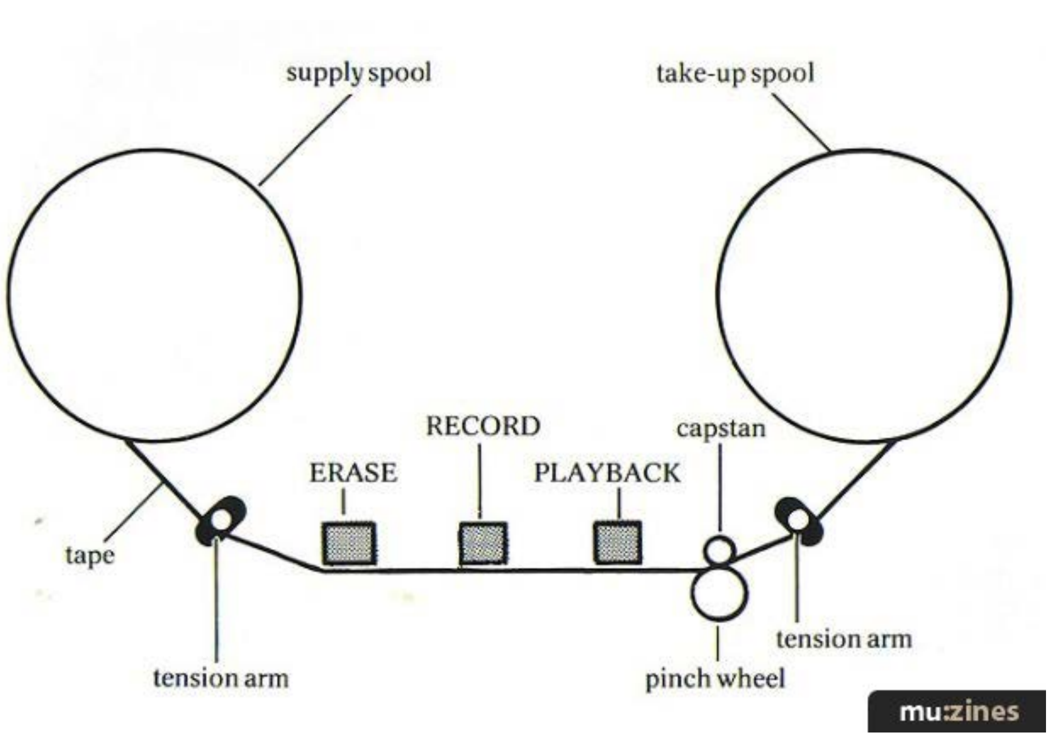
Magnetic Storage

- Increased magnetic density significantly increases memory storage.
- Physical shape of write head is used to concentrate magnetic fields
- Substrate material helps extend magnetic fields and preserve alignment



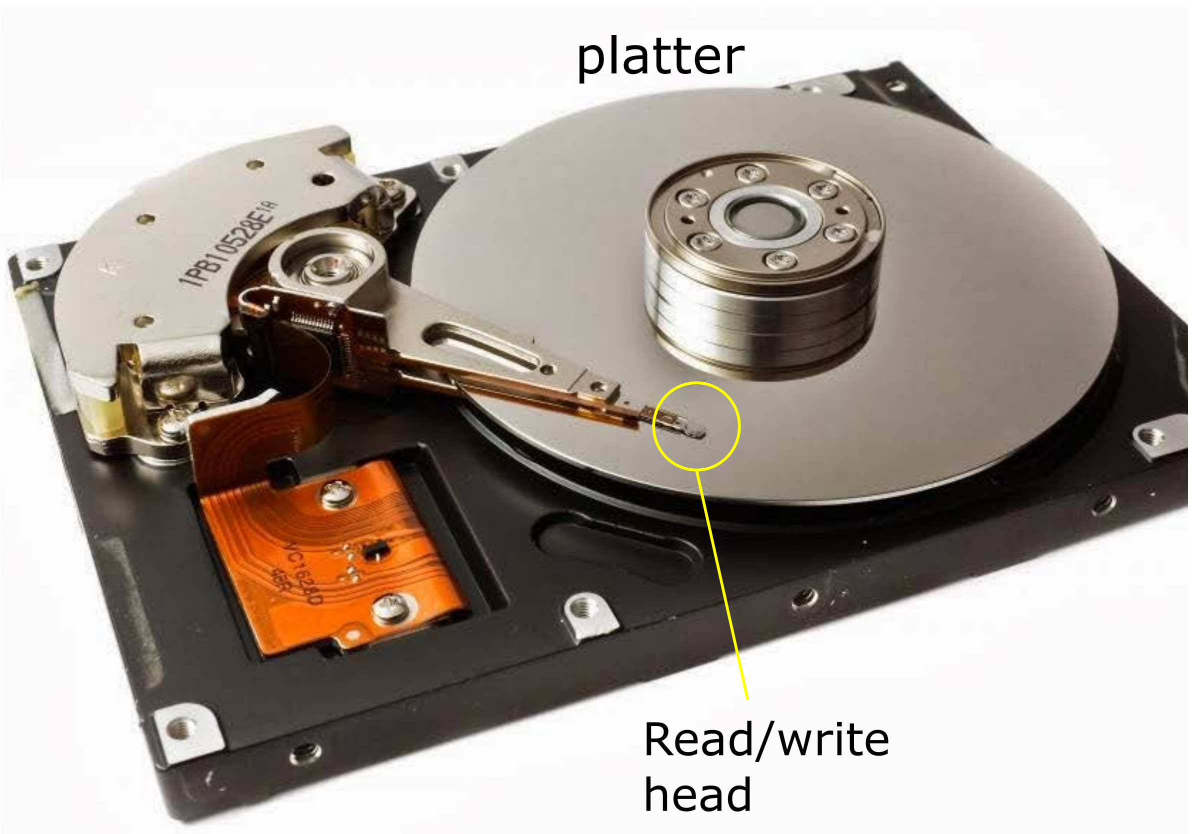
Magnetic tapes

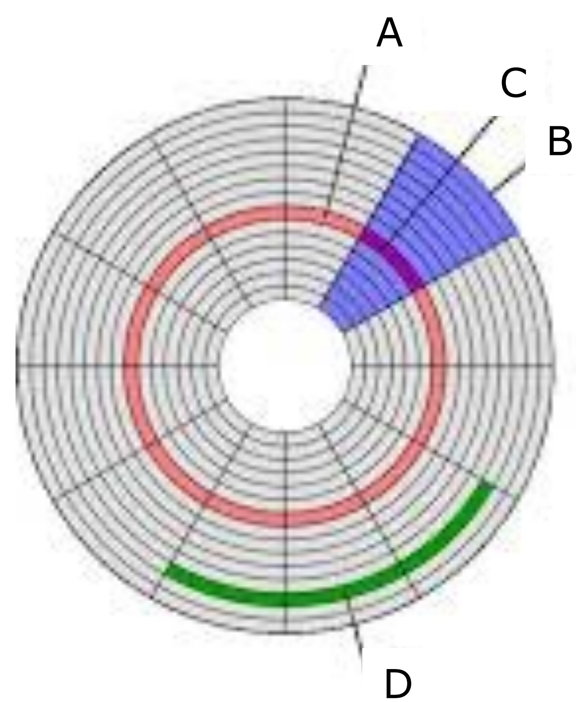
- Magnetic material is combined with a flexible plastic in order to make a magnetic tape that is used for storage
- Data is stored linearly on the tape



Magnetic disks

- Like vinyl records, data can be circumferentially around a disk.



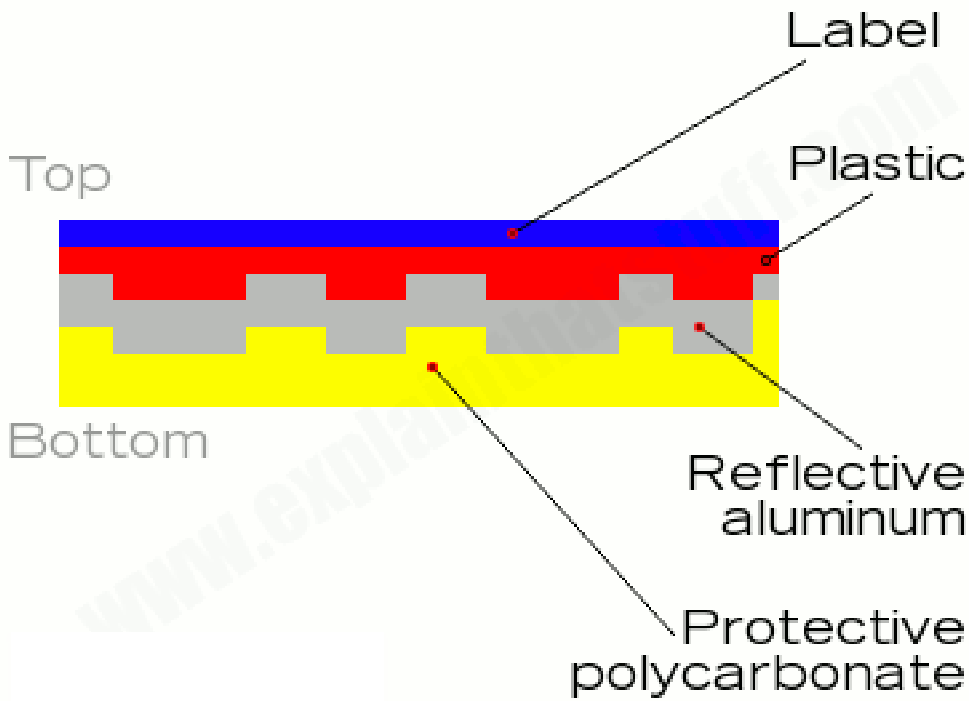


A – track
B – sector
C – sector of a track
D - cluster

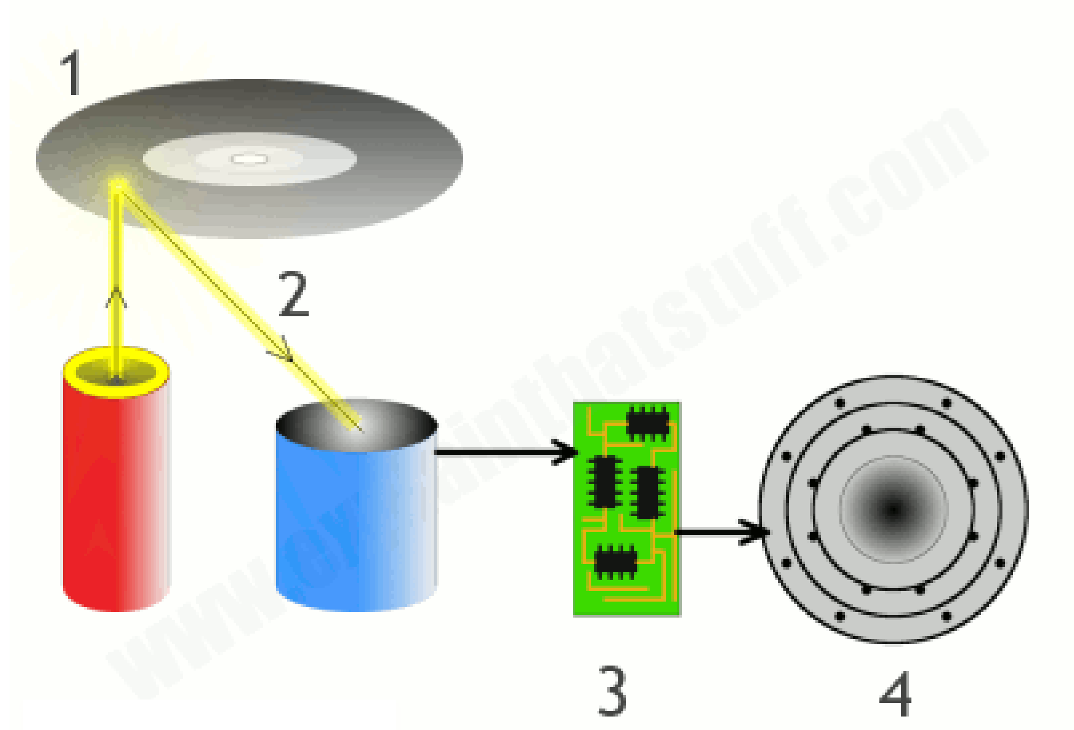

Columbia
www.explainthatstuff.com


CD mounted here
Track for laser and photocell
Photocell
Laser
Control panel
www.explainthatstuff.com
MEMORY /RECALL
REPEAT

Optical Disk Writer



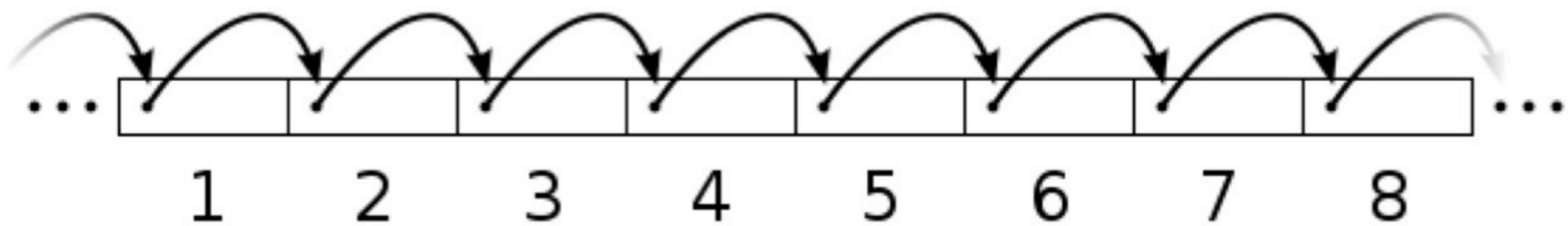
Optical Disk Reader



Sequential vs. Random Access

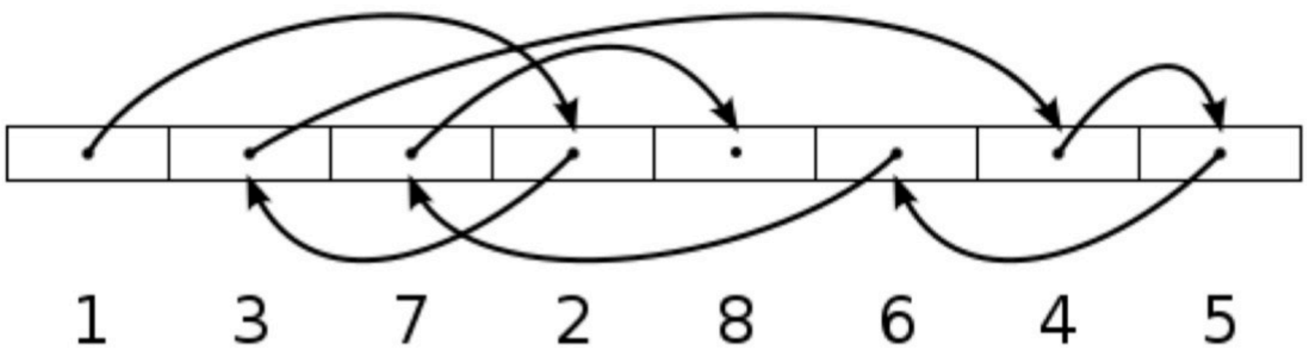
Sequential access:

- data follow, one after the other.
- data can be removed at the end of the sequence and placed back at the beginning



Random access:

- data can be moved around and used at any point

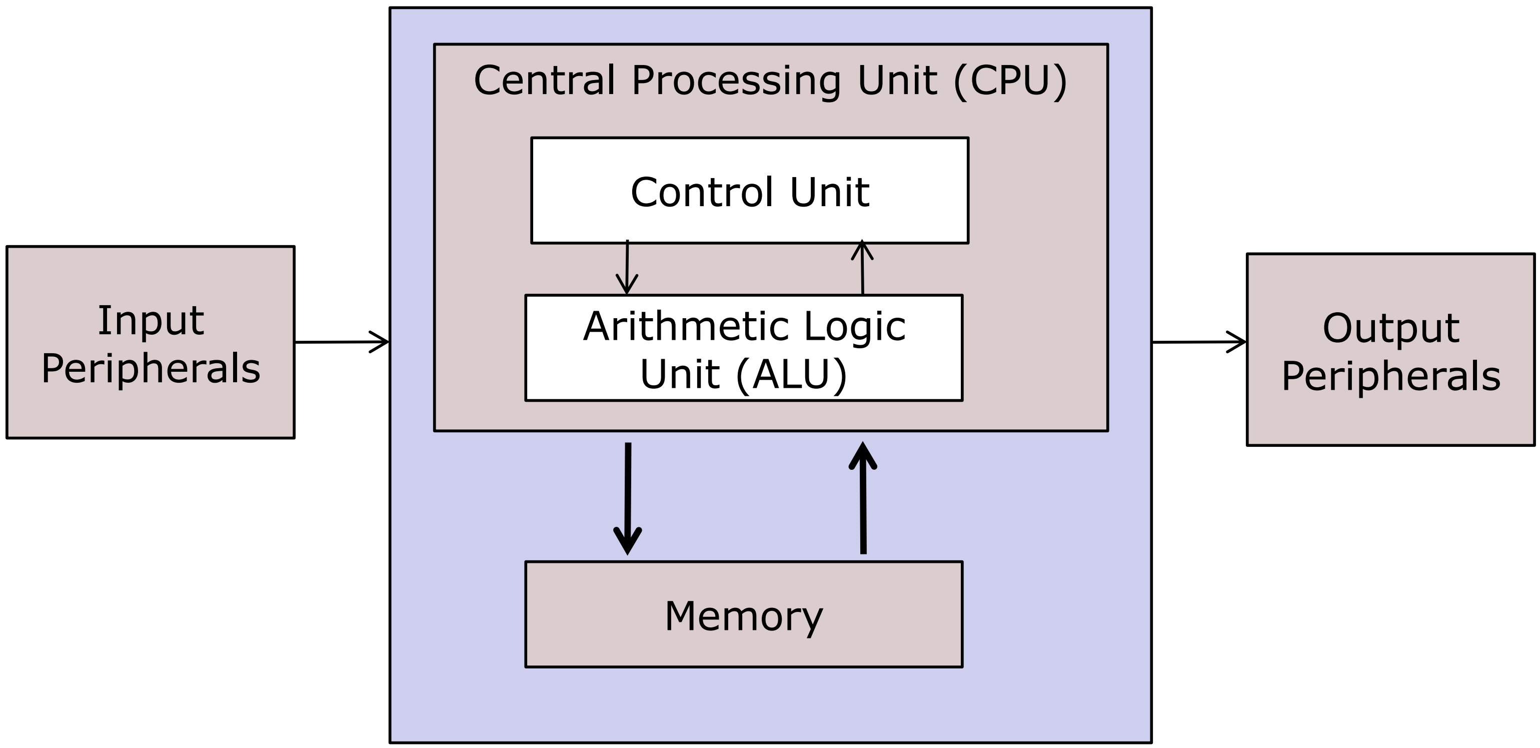


Sequential vs. Random Access

Floating Sushi Bar: Is it random or sequential?

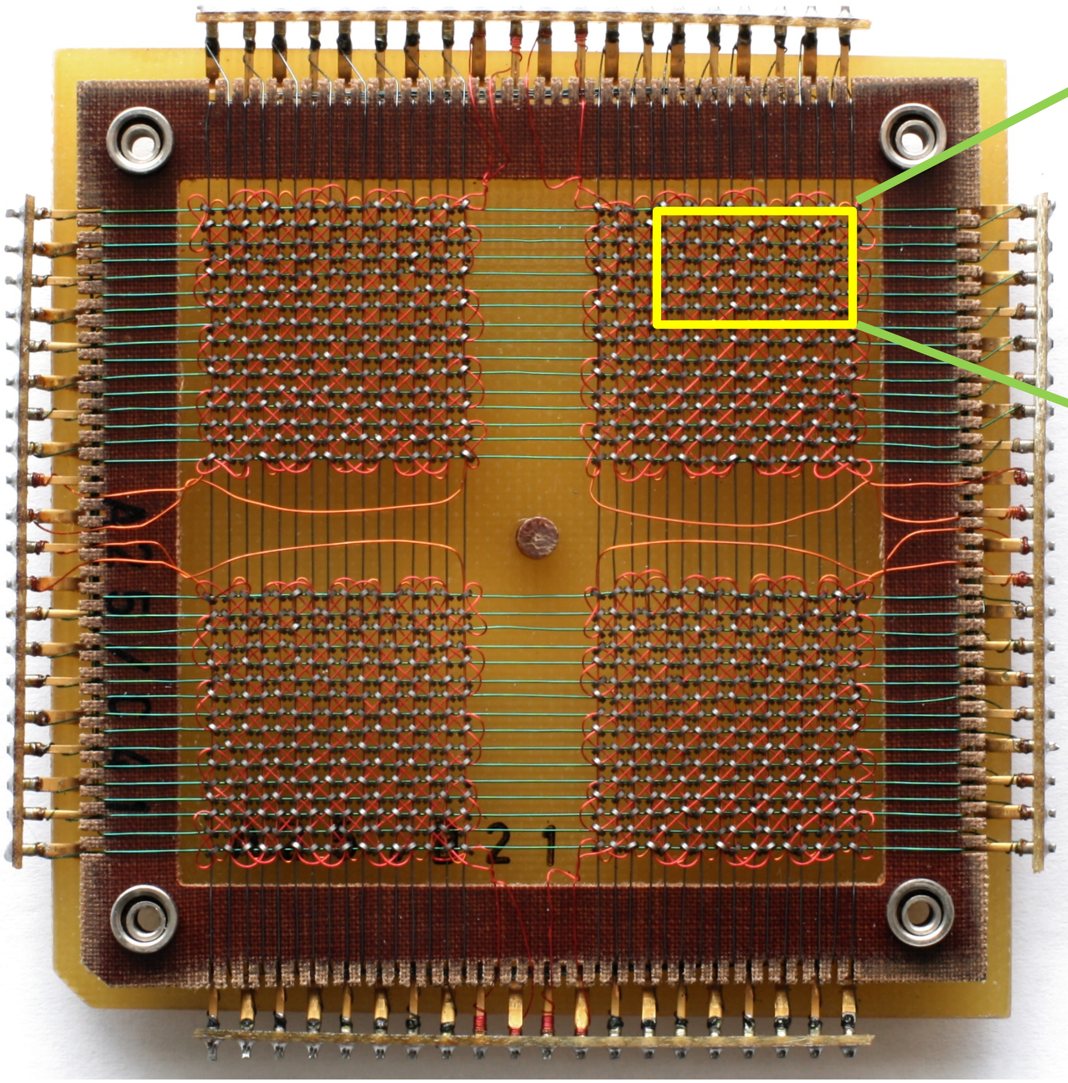
Computers & Memory

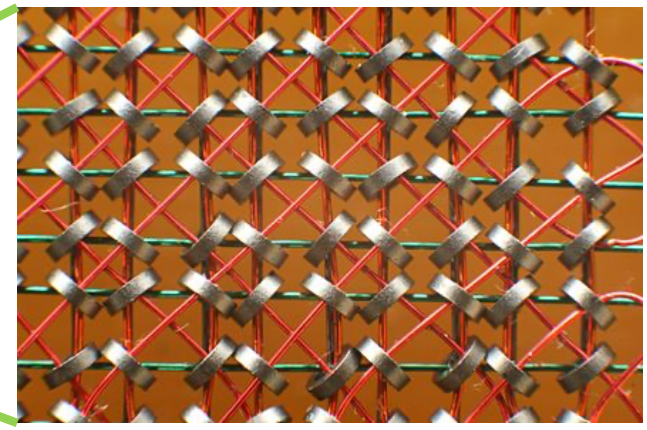
- Basic computer architecture



Magnetic Core Memory

- Early computers used magnetic cores (toroids)

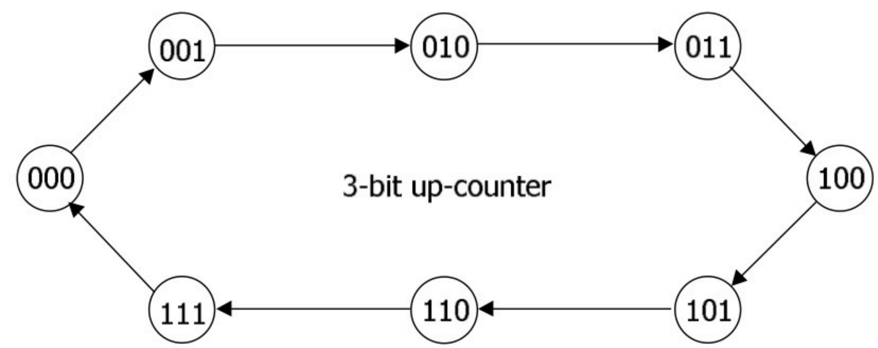




Depending on the magnitude and direction of the currents in the grid, each magnet could be either read or written to a desired state (0 or 1)

Early computers

- Early computers had a small amount of memory
- The amount of memory was small enough, that they could be considered a "finite state machine"
- A finite state machine has a fixed number of possible configurations



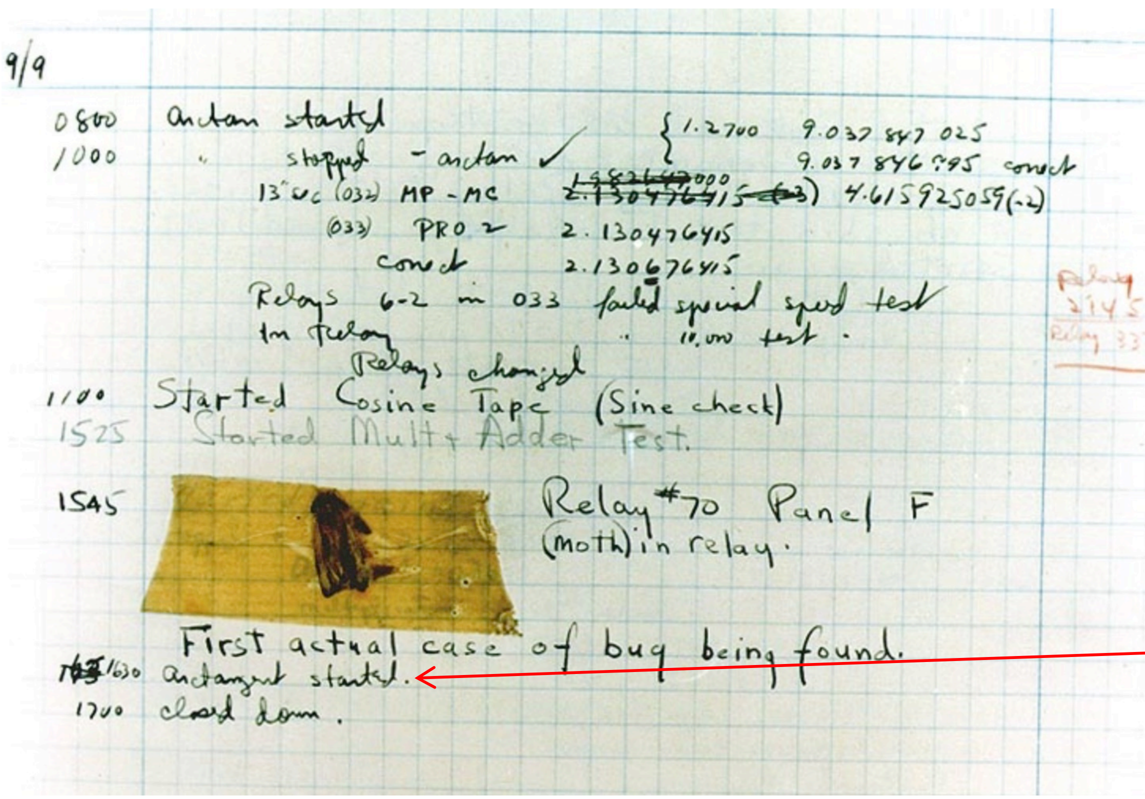
number of possible states: $2^3 = 8$ states.

In old computers, if something was wrong with a program, the entire memory could be printed out and analyzed

This was called a core dump, a term still in use today

More terminology

- "Software bug" !
- Term popularized in 1947 by Grace Hopper
- Prior to that, a "bug" was a form of monster, that caused difficulties in machine operation
- Thomas Edison used the term in talking about problems in his inventions.

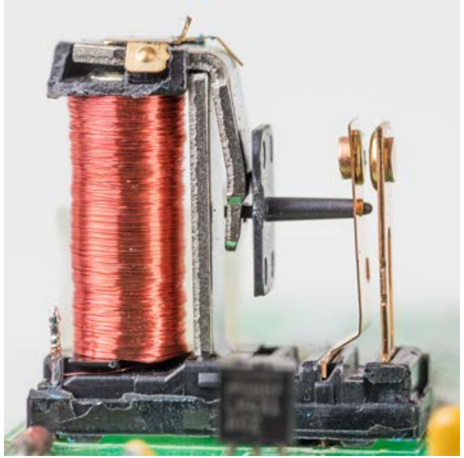


Have a look at what was being calculated, and how long it took

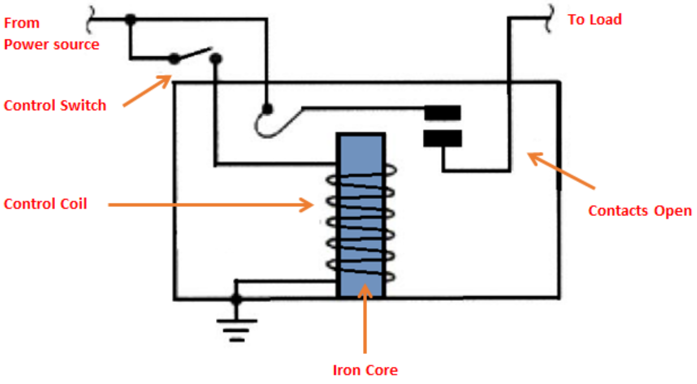
Grace Hopper

- Computer Scientist
- Navy rear admiral
- Devised theory for machine-independent computer programming (e.g., COBOL)

Relays



12VDC 30/40A
87a
87
86
30
85



From
Power source
Control Switch
Control Coil
To Load
Contacts Open
Iron Core

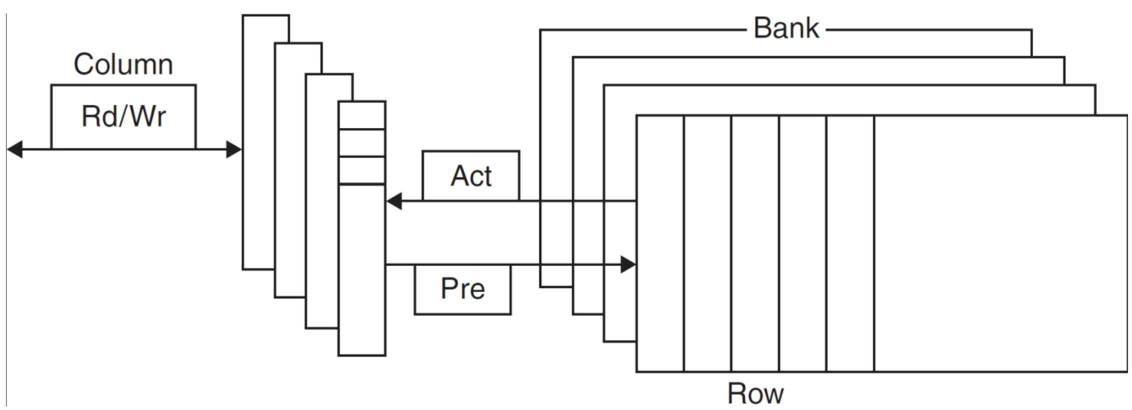
Memory Basics

- Users want large and fast memories
- Fact
  - Large memories are slow
  - Fast memories are small
- Large memories use DRAM technology: Dynamic Random Access Memory
  - High density, low power, cheap, slow
  - Dynamic: needs to be "refreshed" regularly
- DRAM access times are 50-70ns at cost of $10 to $20 per GB
  - FPM (*Fast Page Mode*)

- Fast memories use SRAM: Static Random Access Memory
  - Low density, high power, expensive, fast
  - Static: content lasts "forever" (until lose power)
- SRAM access times are 0.5 – 5ns at cost of $400 to $1,000 per GB

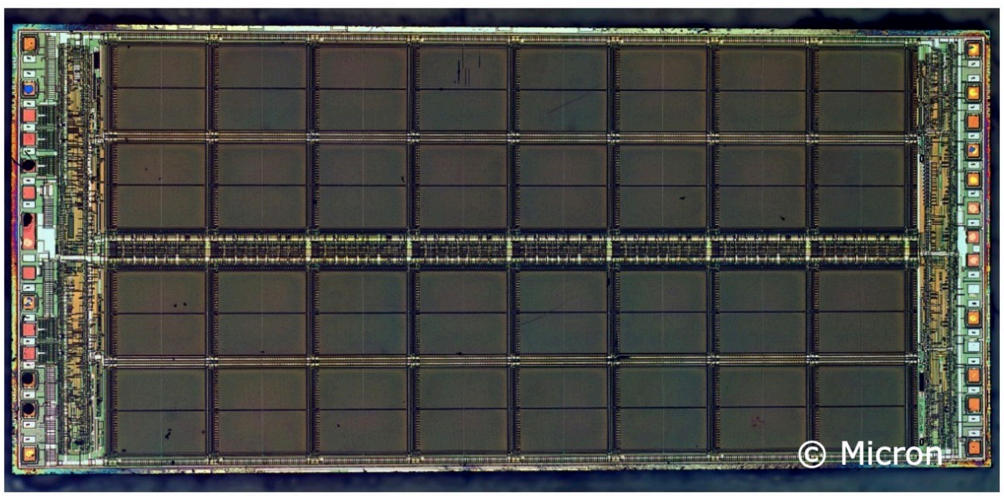
Dynamic Random Access Memory (DRAM)

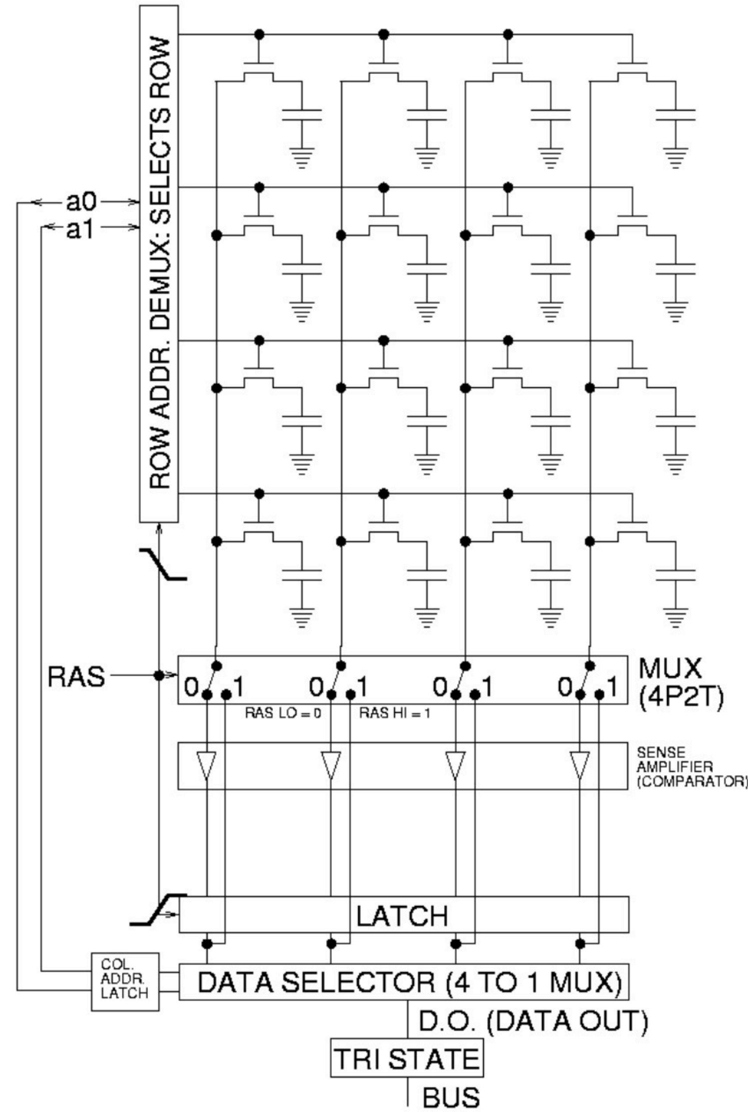
- Data stored as a charge in a capacitor
  - Single transistor used to access the charge
  - Must periodically be refreshed
    - Read contents and write back
    - Performed on a DRAM "row"
- Bits in a DRAM are organized as a rectangular array
  - DRAM accesses an entire row
  - Burst mode: supply successive words from a row with reduced latency
- Double data rate (DDR) DRAM
  - Transfer on rising and falling clock edges
- Quad data rate (QDR) DRAM
  - Separate DDR inputs and outputs



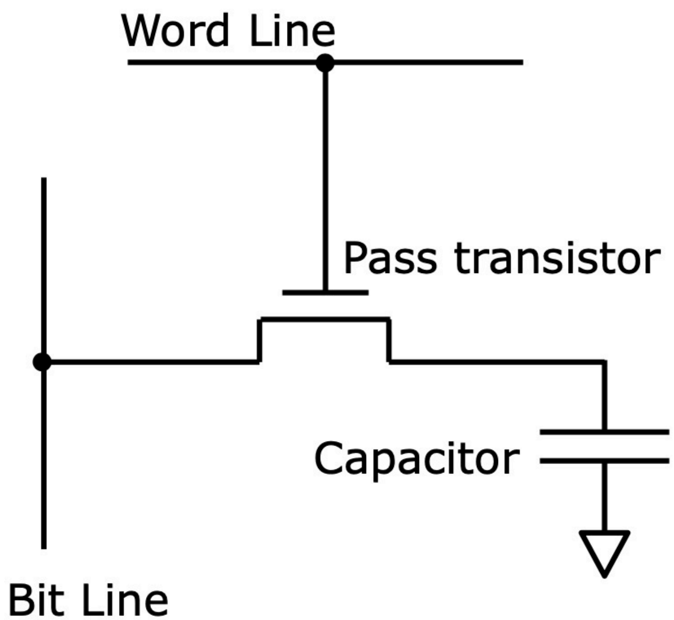
Dynamic Random Access Memory (DRAM)

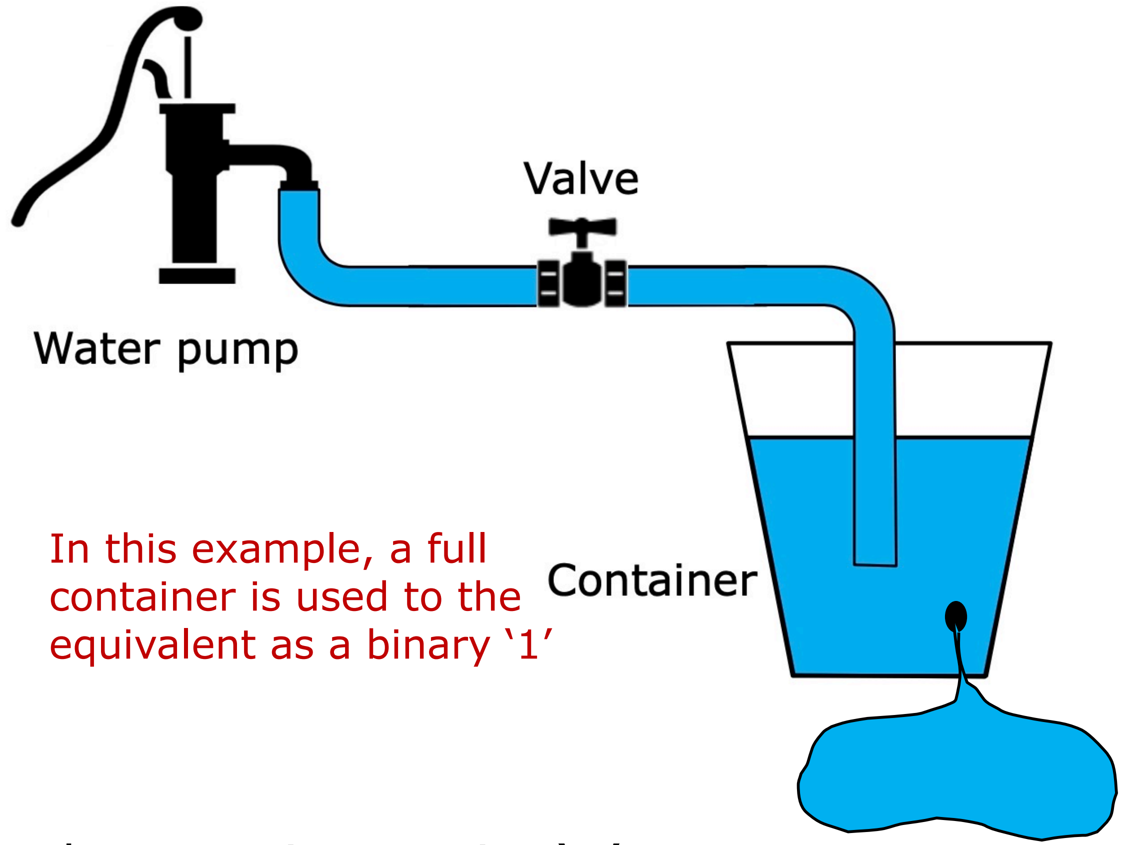
- Implementation





Dynamic Random Access Memory (DRAM)



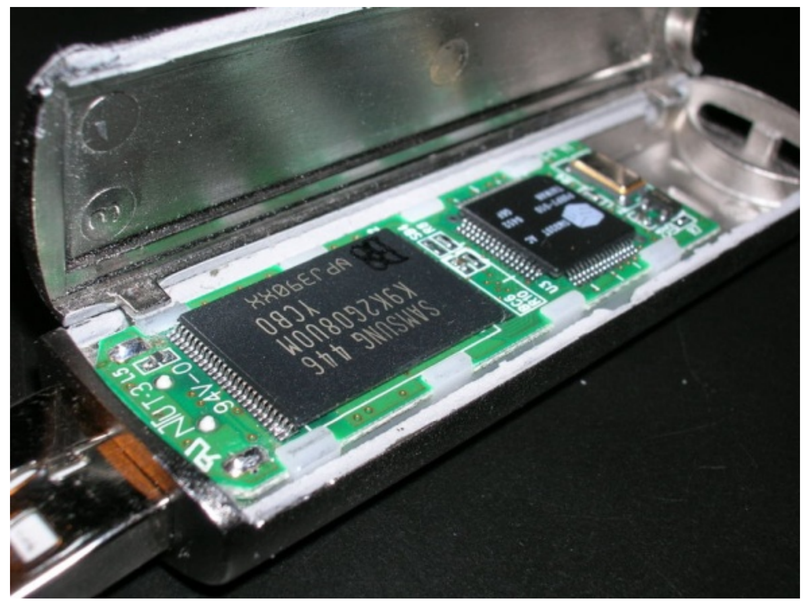


Operating Principal

- Valve opens, pump water into the container, write '1';
- Valve closes, container isolated from the pump, store '1';
- Valve opens, pump work in an opposite direction, erase '1' and write '0';
- Valve closes, container isolated from the pump, store '0'.
- The capacitive "bucket" in DRAM acts as if it has a leak, that has to be replenished with water on a periodic basis.
- This is called "volatile" memory, or memory that does not persist.

Flash Storage

- Nonvolatile semiconductor storage
  - 100× – 1000× faster than disk
  - Smaller, lower power, more robust
  - Limited number of R/W cycles
  - But more $/GB (compared to disk and DRAM)



Flash Memory

USB Drive

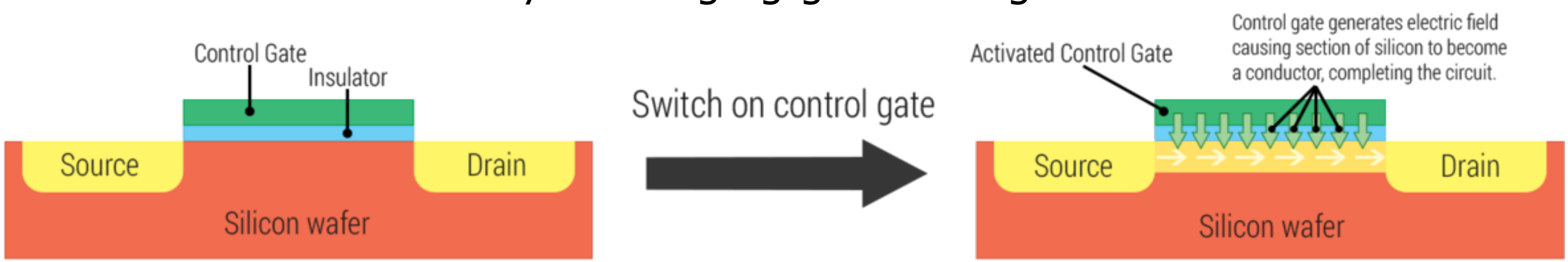
Memory Card

Solid State Drive

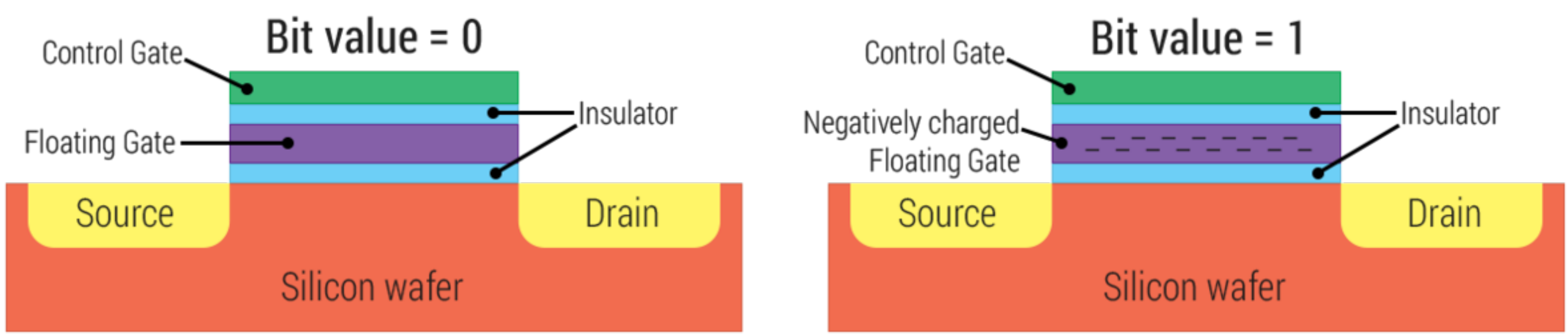
Dr. Fujio Masuoka is credited with the invention of flash memory when he worked for Toshiba in the 1980s. Masuoka's colleague, Shoji Ariizumi, coined the term flash because the process of erasing all the data from a semiconductor chip reminded him of the flash of a camera.

Flash Memory

- A normal MOSFET is a sandwich of Metal (the gate), Oxide (the insulator) and Semiconductor (the conducting channel) that can be switched on and off by a changing gate voltage

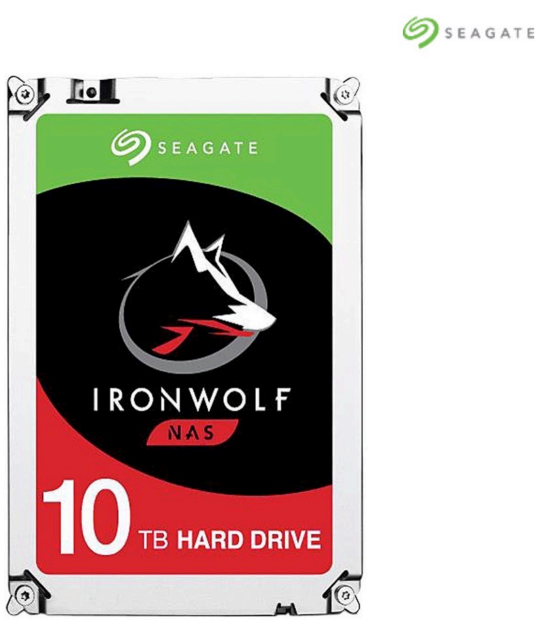


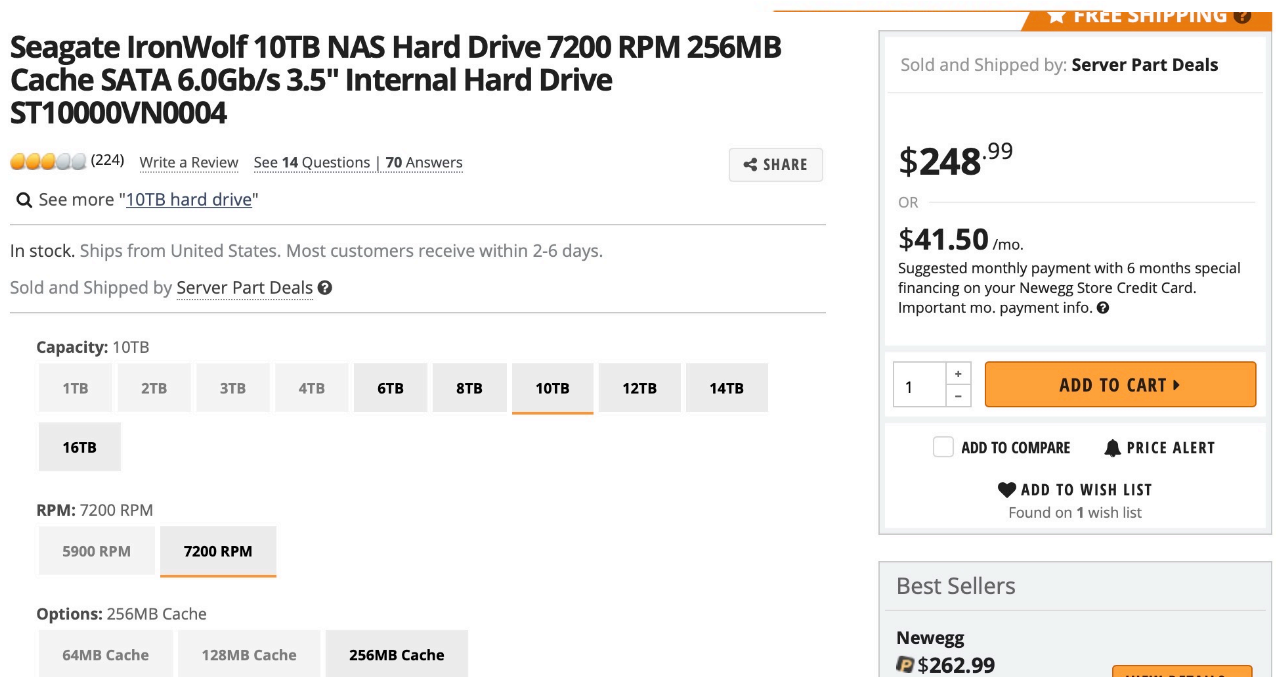
- Flash memory adds an additional insulated metal layer in between.
- Large gate voltages can be used to pull electrons from the conducting channel into the floating gate, and hence serve as a form of memory



Memory Technology

- SRAM and DRAM are: Random Access storage
  - Access time is the same for all locations (Hardware decoder used)
- For even larger and cheaper storage (than DRAM) use hard drive (Disk): Sequential Access
  - Very slow, Data accessed sequentially, access time is location dependent, considered as I/O
  - Disk access times are 2 to 20 million ns (i.e., msec) at cost of $.04 to $2.00 per GB

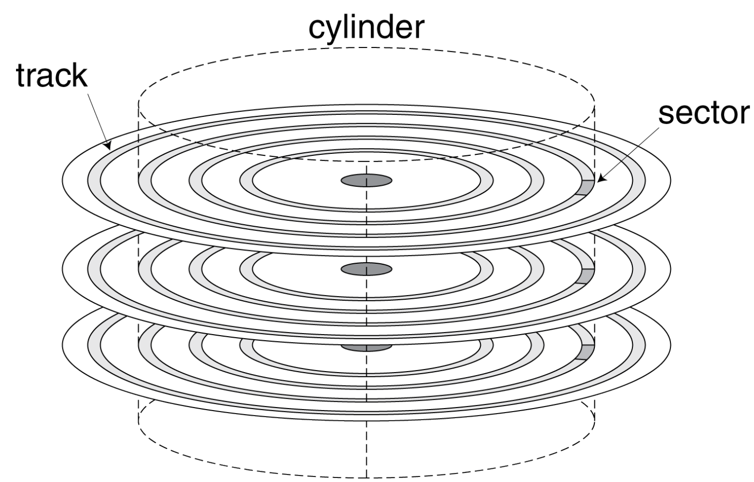




Disk Storage

- Nonvolatile, rotating magnetic storage

- Each sector records
  - Sector ID
  - Data (512 bytes or 4096 bytes)
  - Error correcting code (ECC)
    - Used to hide defects and recording errors
- Access to a sector involves
  - Queuing delay if other accesses are pending
  - Seek: move the heads
  - Rotational latency
  - Data transfer
  - Controller overhead



Disk Access Example

Example

- Given
  - 512 Byte sectors
  - 15,000 rpm (revolutions per minute)
  - 4 ms average seek time,
  - 100 MB/s transfer rate,
  - 0.2 ms controller overhead
- Average read time
  - 4 ms seek time
  - + ½ / (15,000/60) = 2 ms rotational latency
  - + 512 / 100MB/s = 0.005 ms transfer time
  - + 0.2 ms controller delay
  - = 6.2 ms
- If actual average seek time is 1 ms
  - Average read time = 3.2 ms

Disk Interface and Performance

- Manufacturers quote average seek time
  - based on all possible seeks
  - Locality and OS scheduling lead to smaller average seek times
- Smart disk controller allocate physical sectors on disk
  - Present logical sector interface to host
  - SCSI, ATA, SATA
- Disk drives include caches
  - Prefetch sectors in anticipation of access
  - Avoid seek and rotational delay

Cache

(noun) – a collection of items of the same type stored in a hidden or inaccessible place

(verb) – store away in hiding or for future use

Need for speed

- Assume CPU runs at 3GHz
- Every instruction requires 4B of instruction and at least one memory access (4B of data)
  - 3 * 8 = 24GB/sec
- Peak performance of sequential burst of transfer (*Performance for random access is much much slower due to latency*)
- **Memory bandwidth and access time is a performance bottleneck**

Interface	Width	Frequency	Bytes/Sec
4-way interleaved PC1600 (DDR200) SDRAM	4 x64bits	100 MHz DDR	6.4 GB/s
Opteron Hyper-Transport memory bus	128bits	200 MHz DDR	6.4 GB/s
Pentium 4 "800 MHz" FSB	64bits	200 MHz QDR	6.4 GB/s
PC2 6400 (DDR-II 800) SDRAM	64bits	400 MHz DDR	6.4 GB/s
PC2 5300 (DDR-II 667) SDRAM	64bits	333 MHz DDR	5.3 GB/s
Pentium 4 "533 MHz" FSB	64bits	133 MHz QDR	4.3 GB/s

FSB – Front-Side Bus; DDR – Double Data Rate; SDRAM – Synchronous DRAM

Need for Large Memory

- Small memories are fast
- So just write small programs

"640 K of memory should be enough for anybody"
-- Bill Gates, 1981

- Today's programs require large memories

- Data base applications may require Gigabytes of memory

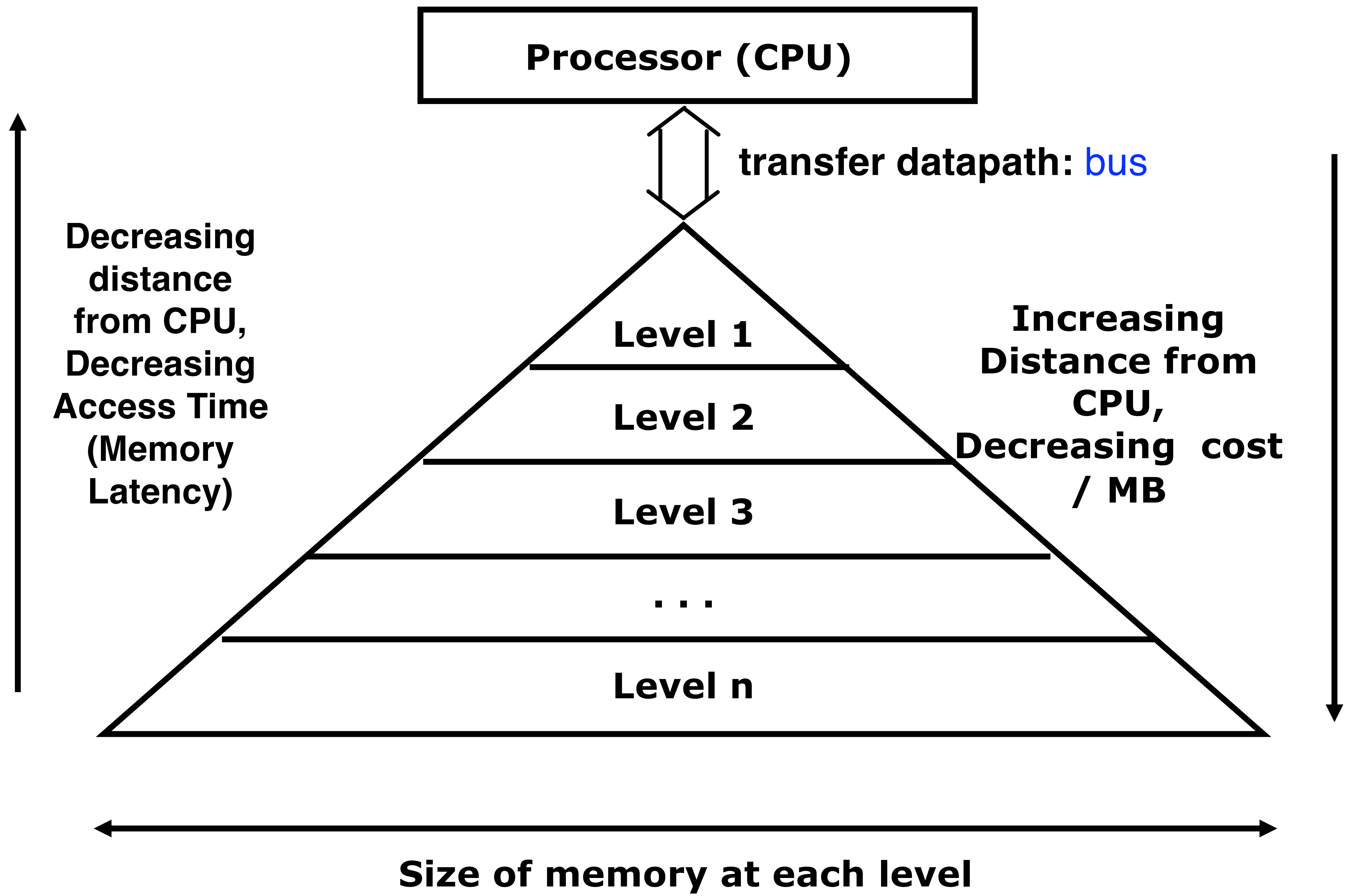
The Goal: Illusion of large, fast, cheap memory

- How do we create a memory that is large, cheap and fast (most of the time)?

- Strategy: Provide a Small, Fast Memory which holds a subset of the main memory – cache memory
  - Keep frequently-accessed locations in fast cache
  - Cache retrieves more than one word at a time
  - Sequential accesses are faster after first access

Memory Hierarchy Pyramid



Memory Hierarchy

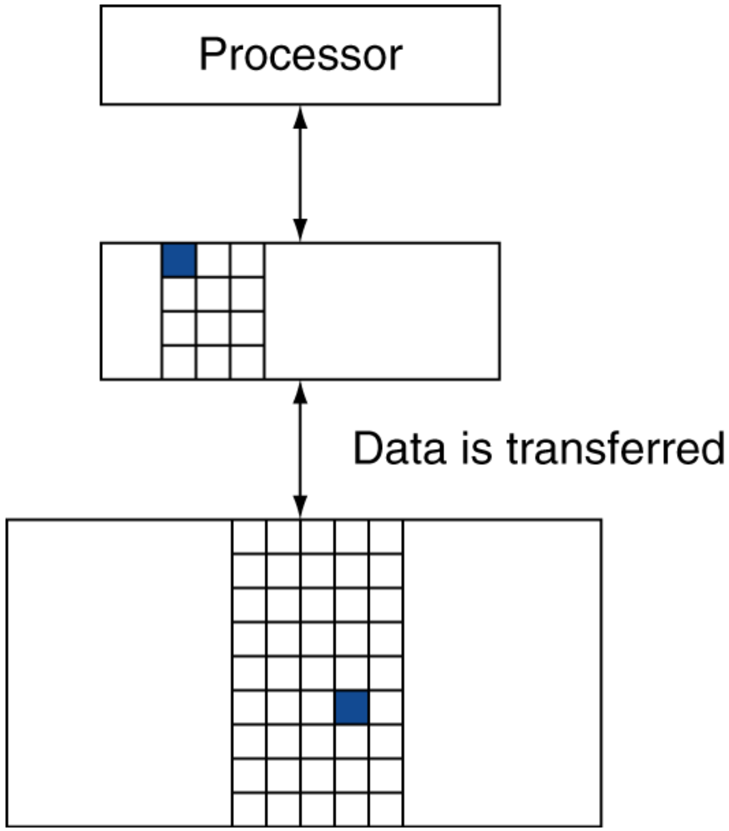
- Hierarchy of Levels
  - Uses smaller and faster memory technologies close to the processor
  - Fast access time in highest level of hierarchy
  - Cheap, slow memory furthest from processor

- The aim of memory hierarchy design is to have access time close to the highest level and size equal to the lowest level

Basic Philosophy

- Move data into ‘smaller, faster’ memory
- Operate on it
- Move it back to ‘larger, cheaper’ memory
  - How do we keep track if changed
- What if we run out of space in ‘smaller, faster’ memory?

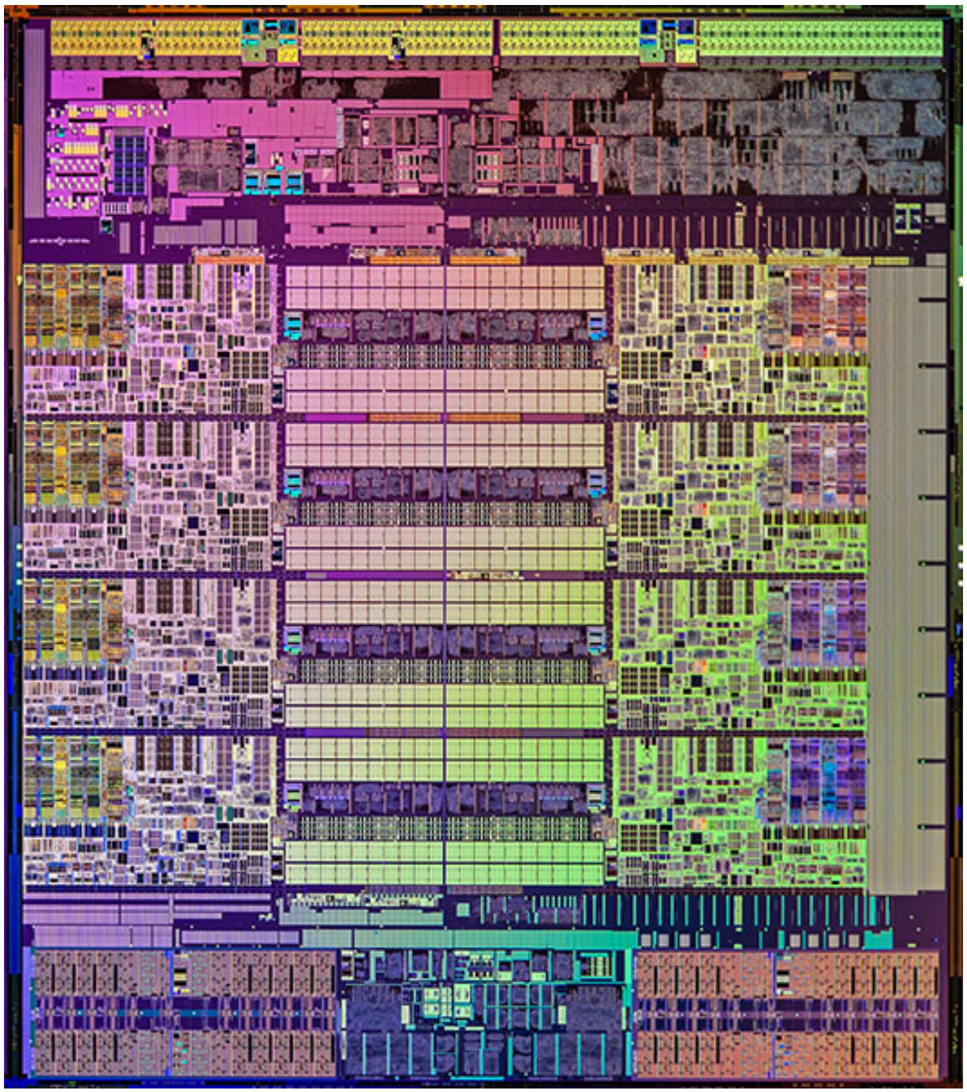
- Important Concepts: Latency, Bandwidth

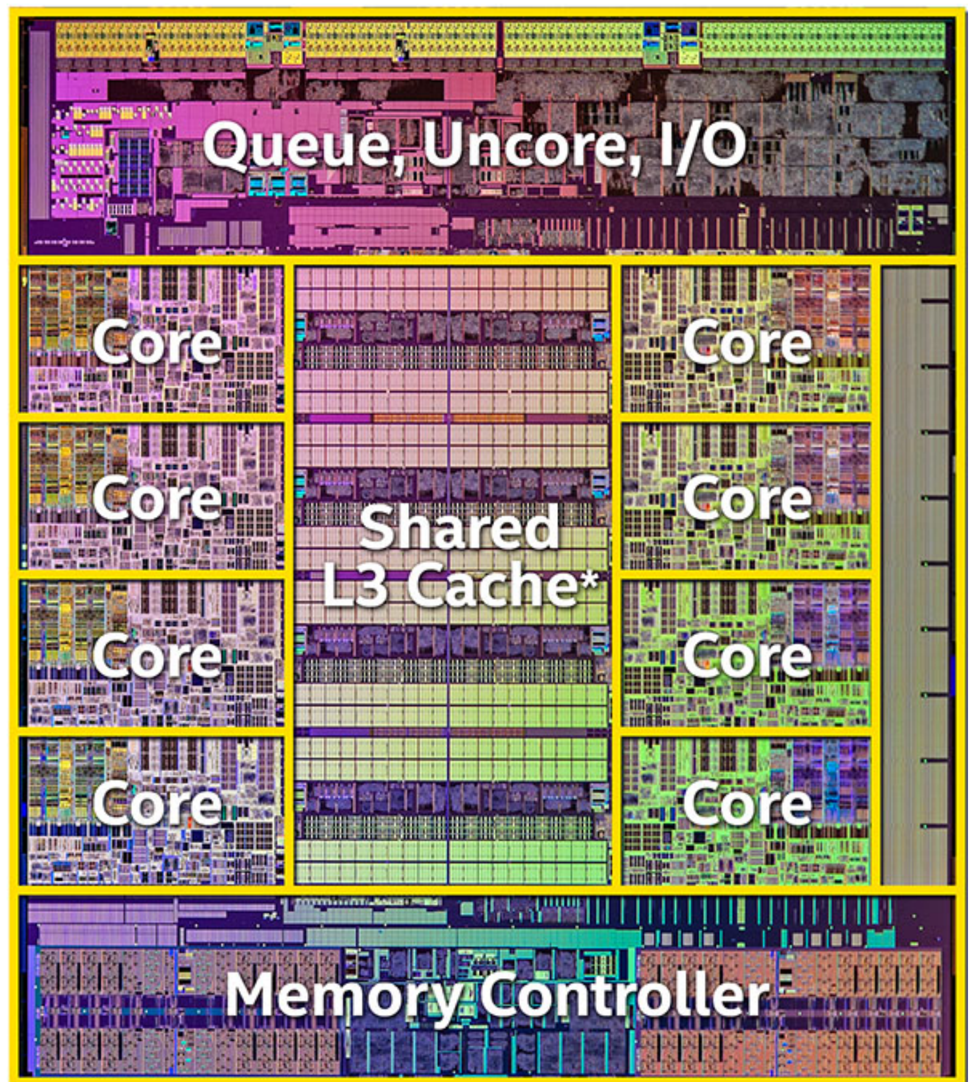


Inside of a CPU

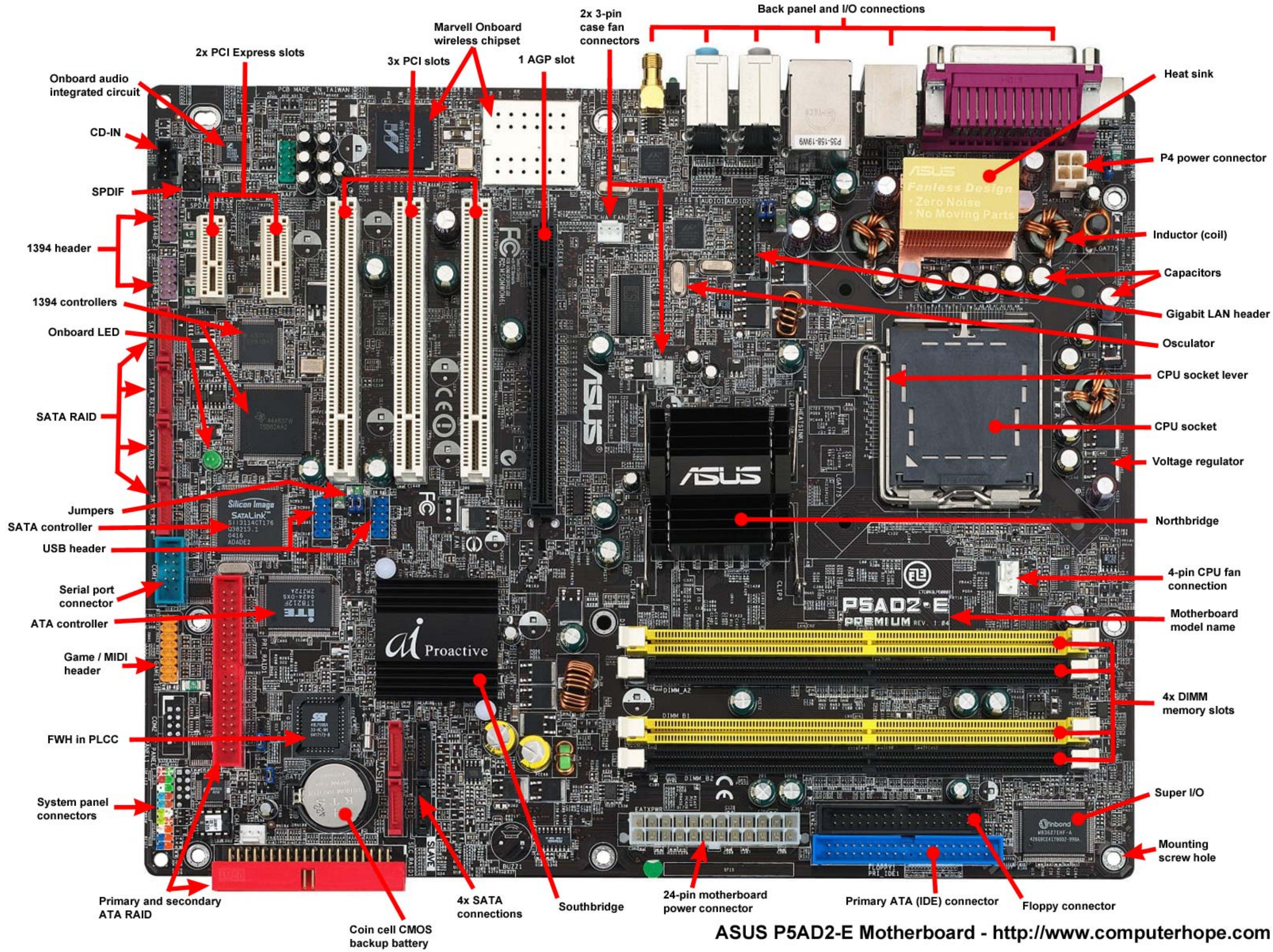
Intel 8-core i7-5960X

3 GHz, 8-core, 20 MB of cache, 140 W, 22 nm... cost: ~$1000

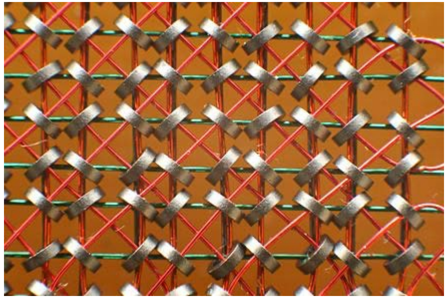




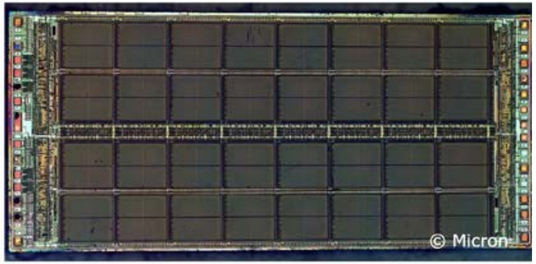
Motherboard




EDISON
HOME
PHONOGRAPH




CD mounted here
Track for laser and photocell
Photocell
Laser
Control panel
www.explainthatstuff.com


© Micron


SanDisk
ultra II
2.0GB
SanDisk
Compact Flash Card
2GB
MMC PLUS
512
Kingston