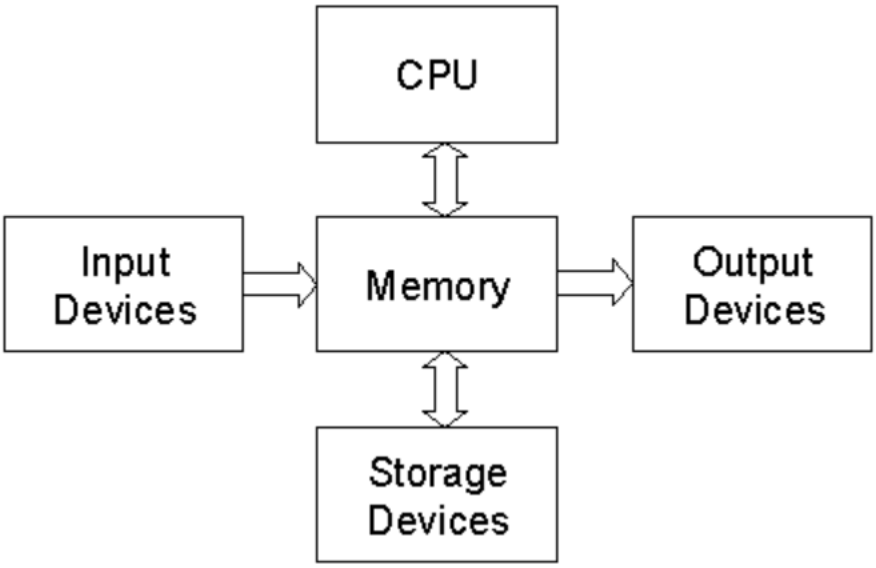
COMPUTER ARCHITECTURE

## Overview of the Organization of a Computer System



The main components of a computer

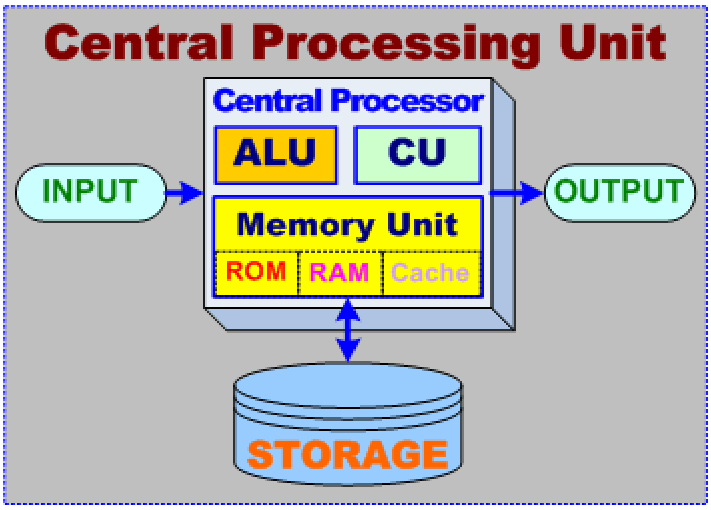
The main components of a computer are (i) the CPU, (ii) memory, (iii) storage, (iv) I/O subsystem and (v) buses. Other important components are (vi) the system clock, (vii) ROM and (viii) caches.

### 1. CPU

- Stands for Central Processing Unit (processor).
- It is the unit which executes programs i.e. sequences of instructions.
- The microprocessor is a silicon chip that contains a whole processor.

The three basic characteristics that differentiate microprocessors are:

- Instruction set: The set of instructions that the microprocessor can execute.
- Bandwidth: The number of bits processed in a single instruction.
- Clock speed: Given in gigahertz (GHz), the clock speed determines how many instructions per second the processor can execute.
  - 1 GHz = $10^9$



**Central Processing Unit**

The most important parts of a processor are:

- Control Unit (CU)
- Arithmetic-Logic Unit (ALU)
- Registers
- Cache

Control Unit

- Extracts instructions from memory and decodes them.
- It sends the necessary signals to the ALU to perform the operations needed.

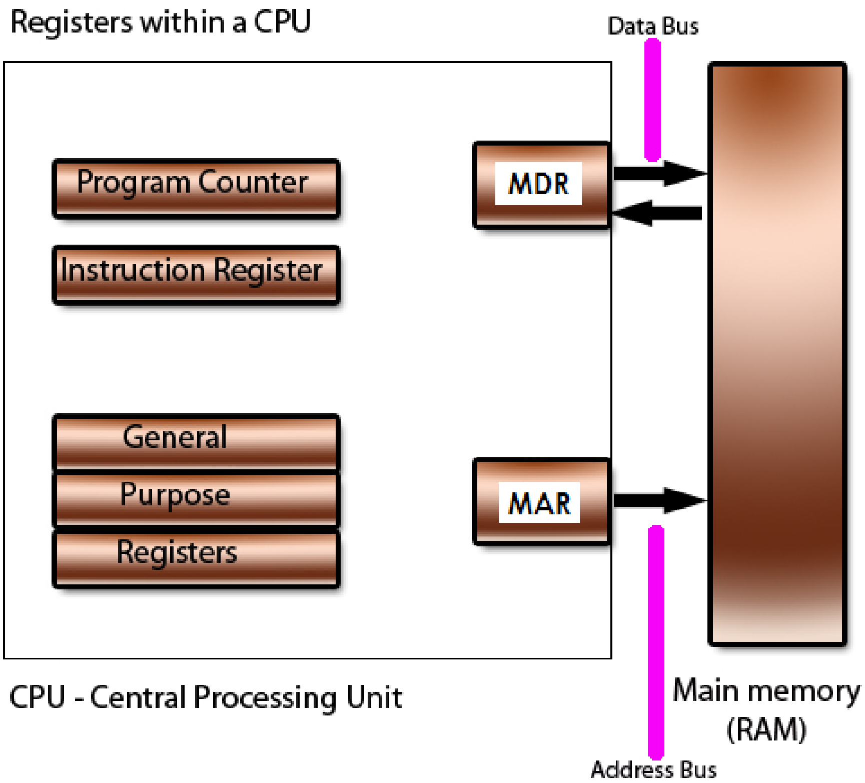
Arithmetic-Logic Unit

- The unit where all the arithmetic (+, - etc.) and logical operations (AND, OR, NOT etc.) are carried out.

Registers

- The registers are what the CPU uses for temporary storage of data.
- They are not part of any of the system memory, but are instead additional storage locations that are on the CPU itself.
- Are very fast for the CPU to use.

## Particular Registers



Particularly important registers are the following:

- CIR: Current Instruction Register (CIR): This register (also called IR) is the part of a CPU's control unit that stores the instruction currently being executed or decoded.
- PC: Program Counter (PC): The PC (also called the ‘instruction pointer’ or ‘instruction address register’) holds the address of the next instruction to be executed. In most processors, the instruction pointer is incremented automatically after fetching a program instruction.
- MAR: Memory Address Register (MAR): This register either stores the memory address from which data will be fetched to the CPU or the address to which data will be sent and stored.
- MDR: Memory Data Register (MDR): This register contains the data to be stored in the computer storage (e.g. RAM), or the data after a fetch from the computer storage. It acts like a buffer.

- Status Register: The status register (flag register) is a collection of 1-bit values which reflect the current state of the processor and the results of recent operations. Here are some examples:
    - Carry bit: set if the last arithmetic operation ended with a leftover carry bit coming off the left end of the result. This signals an overflow on unsigned numbers.
    - Parity bit: set if the low-order byte of the last data operation contained an even number of 1 bits (that is, it signals an even parity condition).
    - Zero bit: set if the last computation had a zero result. After a comparison this indicates that the values compared were equal (since their difference was zero).
    - Sign bit: set if the last computation had a negative result (a 1 in the leftmost bit).
    - Interrupt bit: when set, interrupts are enabled.
    - Overflow bit: set if the last arithmetic operation caused an overflow

## CISC and RISC

CPUs are also classified as:

- CISC (complex instruction set computer)
    - Most personal computers use a CISC architecture
    - The CPU supports as many as two hundred instructions.
    - Intel's Pentium microprocessors are CISC.
- RISC (reduced instruction set computer)
    - Macintosh computers use a RISC microprocessor.
    - One advantage of RISC computers is that they can execute their instructions very fast because the instructions are so simple.
    - Cheaper to design and produce.

## The Fetch-Decode-Execute Cycle

- Also called the Fetch-Decode-Execute cycle
- It describes how a CPU executes instructions one after the other

- The order is the following:

    1. Bring the next command from main memory.
    2. Interpret this command.
    3. Execute this command.
    4. Return to step 1.

## 2. Memory

- Also called RAM (for Random Access Memory), main memory or primary memory.
- It is a sequence of memory elements (locations) that hold content. Each location is identified by an address.
- It holds the programs currently being executed by the computer (i.e. the operating system and the applications that are 'open').
- It is volatile.

### Read and Write Memory Cycles

These are the steps in a typical Read Cycle:

1. Place the address of the location to be read on the address bus via MAR.
2. Activate the memory read control signal on the control bus.
3. Wait for the memory to retrieve the data from the addressed memory location.
4. Read the data from the data bus into MDR.
5. Drop the memory read control signal to terminate the read cycle.

A simple Pentium memory read cycle takes 3 clock cycles. Steps 1-2 and then 4-5 are done in one clock cycle each. For slower memories, wait cycles will have to be inserted.

These are the steps in a typical Write Cycle:

1. Place the address of the location to be written on the address bus via MAR.

2. Place the data to be written on the data bus via MDR.
3. Activate the memory write control signal on the control bus.
4. Wait for the memory to store the data at the addressed location.
5. Drop the memory write control signal to terminate the write cycle.

A simple Pentium memory write cycle takes 3 clock cycles. Steps 1-2 and 4-5 are done in one clock cycle each. For slower memories, wait cycles will have to be inserted.

There are two different types of RAM:

- DRAM
  - Dynamic Random Access Memory
  - Slower
  - Needs refreshing thousands of times per second
  - Access times of about 60 nanoseconds
  - Cheaper
  - Volatile
  - Smaller in size (for the same amount of memory) than SRAM (by about four times).
- SRAM
  - Static Random Access Memory
  - Faster
  - Does not need refreshing
  - Access time, about 10 nanoseconds
  - More expensive
  - Volatile

| Applications of DRAM | Applications of SRAM |
|---|---|
| • Computer RAM<br>• TVs<br>• Video cameras<br>• GPSs | • Cache (support for DRAM) in computers<br>• Digital cameras<br>• Cell phones |

## 3. Storage

- Secondary storage, also called mass storage or auxiliary storage.
- Is not volatile.
- Holds data and programs permanently until the user decides to delete them.
- If a user opens a program, this is copied from storage to RAM. It is when in the RAM that a program can be executed.
- List of secondary storage devices:
  - Hard disks:
    - Very fast.
    - Possessing very large memories.
    - Normally placed inside the computer case. Some are external.
    - Magnetic.
  - Optical disks:
    - Uses a laser to read and write data.
    - Have very large storage capacity.
    - Not as fast as hard disks.
  - Tapes:
    - Relatively inexpensive.
    - Can have very large storage capacities.
    - Serial access only.
  - Pen drives:
    - Very fast.
    - Electronic, no moving parts.
- Mass storage is measured in kilobytes (1,024 bytes), megabytes (1,024 kilobytes), gigabytes (1,024 megabytes) and terabytes (1,024 gigabytes).

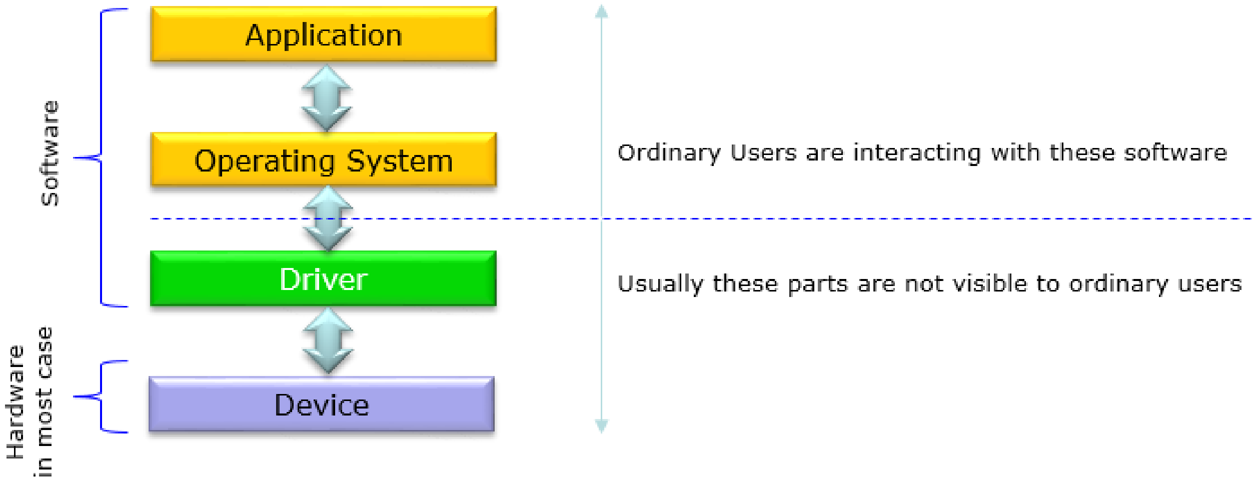
## 4. I/O Subsystem

- Controls all I/O devices
- Basic functions are:
  - Issues commands to the devices

        - Example: messages to printer which file to print.
    - Handles messages (called interrupts) from devices
        - Example: receives message to printer that it is out of paper.

## Device Driver

- A software program that controls a particular type of hardware device that is attached to a computer.
- It is an interface between the operating system and the device.
- When one buys an operating system, this already contains the most common drivers.
- Hardware that use a device driver to connect to a computer include printers, displays, CD-ROM readers, network or sound cards, computer mice or hard disks.



Software
Application
Operating System
Driver
Hardware in most case
Device
Ordinary Users are interacting with these software
Usually these parts are not visible to ordinary users

## Speed

List of devices in order of speed (speediest on top)

| |
|---|
| Register |
| L1 cache |
| L2 cache |
| RAM |
| ROM |

| |
|---|
| SSD |
| HDD |
| Optical disk |
| Magnetic tape |

I/O buffering

- A buffer is a memory that is used in the transfer of data between two components.
- An example:
    - The computer sends a document to be printed to the printer. It does so by placing the document in a buffer B.
    - The computer performs other tasks and at the same time the printer starts printing the material in the buffer.
    - This method permits the computer to do other tasks while the printer is printing.

## 5. Buses

- A bus connects computer components and transfers data between them.
- May be parallel or serial.
    - Examples of parallel bus standards:
        - Advanced technology attachment (ATA) or small computer system interface (SCSI) for printer or hard drive devices.
    - Examples of serial bus standards:
        - Universal serial bus (USB), FireWire or serial ATA.
- Frequently divided into three categories:
    - Address bus (to transfer addresses)
        - If n = width of address bus then the number of addressable memory locations is $2^n$.
    - Data bus (to transfer data)
    - Control bus (to transfer messages)

## 6. System Clock

- It generates periodic, accurately spaced signals.
- Used to regulate the operations of a processor – between two consecutive signals an operation is performed.
- It is expressed in megahertz (MHz) or gigahertz (GHz). 33 MHz means 33 million cycles per second. 4GHz means 4 billion ($10^9$) cycles per second.
- A real-time clock, also called the system clock, keeps track of the time of day and makes this data available to the software.

## 7. ROM

- Stands for Read-only memory.
- Also known as firmware.
- Not volatile.
- It is used to hold:
  - The POST program to test the computer parts like the processor, RAM and secondary storage that is run when the computer is switched on.
  - The software required to initiate bootstrapping.
  - Software drivers which interface between the operating system and hardware.
- ROMs are also used in peripheral devices such as laser printers, whose fonts are often stored in ROMs.
- ROMs are slower than RAMs.
- There are five basic ROM types:
  - ROM:
    - Contains data and programs put by the manufacturer.
    - Its contents can never be changed.
  - PROM:
    - Programmable read-only memory (PROM).
    - Blank PROM chips can be bought inexpensively and coded by anyone with a special tool.
    - They can only be programmed only once.

    - EPROM
        - Erasable PROM.
        - Can be rewritten many times.
        - The EPROM must be removed from the computer to be erased.
        - One cannot erase just part of the EPROM.
    - EEPROM
        - Electrically erasable PROM.
        - The chip does not have to be removed from the computer to be rewritten.
        - One can erase just a portion.
        - EEPROMs are changed 1 byte at a time, which makes them versatile but slow.
    - Flash memory
        - A type of EEPROM.
        - Deletes in blocks.
        - Much faster than traditional EEPROMs because it writes data in chunks, usually 512 bytes in size.
        - Used as secondary storage as USB drive (pen drive), MP3 players, digital cameras and solid-state drives.

## 8. Caches

- Fast memory
- Used to speed up the computer
- The most used programs and data are placed in the cache
- Types of memory caching (use SRAM):
    - L1 (level 1)
        - Built into the CPU itself.
        - Runs at the same clock speed as the CPU.
        - Most expensive type of cache memory.
        - It is the first place that a processor will look for data or instructions.
        - Size is about 64KB.
    - L2 (level 2)

        - Normally, also located in the CPU chip, although not as close to the core as L1 cache (sometimes located on a separate chip close to the CPU).
        - Less expensive and larger than L1 caches.
        - Size is about 256 KB per core.
    - L3 (level 3)
        - Much larger than either L1 or L2.
        - Tends to be a shared cache that is common to all the cores.
        - May be of the order of 2 MB per core.
- Disk caching
    - A disk cache is part of RAM i.e. uses DRAM.
    - It holds the most recently accessed data from the disk.
    - When a program needs to access data from the disk, it first checks the disk cache to see if the data is there.
- Cache hit refers to the occurrence when data is found in the cache.