CHAPTER 3: The Basic Architecture of a Computer

1. Historical Background

Humans initially began counting using their fingers. However, it became quickly apparent that this method was insufficient given the various things that needed to be counted or calculated. Thus, the need for calculation increased with the economic and social development of early societies (Egypt, Rome, Greece, etc.).

The activities, especially those related to commerce, compelled these early civilizations to develop the first calculating tools. The development of these calculating machines was necessitated by the limited capabilities of these machines in terms of the number of operations and the time required for each operation.

Examples (mechanical machines):

- Numeric calculating machines.
- Blaise-Pascal's calculating machine (1623-1662).
- Leibniz's machine.

Subsequently, automatic calculating machines were developed.

1.1. Vacuum Tube Calculators (First Generation Computers)

The first computer was built in 1945 based on vacuum tubes, following several studies and attempts (1941: ABC - Atanasoff Berry Computer). This machine was named:

ENIAC = Electronic Numerical Integrator And Computer. It was constructed with 18,000 vacuum tubes and consumed a significant amount of electricity. It required extensive rooms for installation and had several programming flaws.

1.2. Transistor Computers (Second Generation Computers)

In 1948, the size of computers could be minimized due to remarkable developments in electronics:

- Birth of electronic components: The transistor replaced vacuum tubes, resulting in:
  - Lower power consumption (transistors do not operate by the Joule effect).
  - Reduced physical space requirements.
  - Increased speed.

Around the 1940s, Von Neumann introduced a new approach based on the use of the memory to move closer to the modern form of computers and provide another advantage concerning time (programming).

1.3. Integrated Circuit Computers (Third Generation Computers)

In 1960, the first computer was born.

Advantages:

- Speed.
- Computer size.

The concept of miniaturization and integration emerged:

- SSI (Small Scale Integration) - "a few tens of transistors per chip."
- MSI (Medium Scale Integration) - "hundreds of transistors per chip."
- LSI (Large Scale Integration) - "1,000 transistors per chip."
- VLSI (Very Large Scale Integration) - "1,000,000 transistors per chip."

Example: IBM360 (MSI), IBM370 (LSI).

1.4. Microprocessors (Fourth Generation Computers)

After 1970 (VLSI), the era of microcomputing began, giving birth to the first microprocessors that still dominate the world of computing.

- The first microprocessor, Intel 4004, was introduced in 1971 with 2,250 transistors.
- Intel 8008 (8-bit processor).
- Intel 8086 (16-bit processor).
- Intel 80286, 80386.
- Intel 80486 (32-bit processor).
- Pentium I, II, III, 4…

2. Computer Architecture

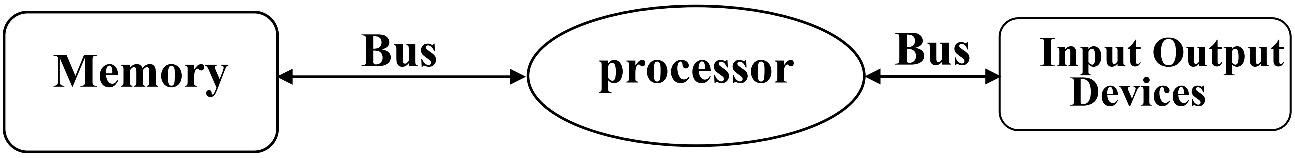
There are two fundamental types of structures, known as 'Von Neumann' and 'Harvard'

2.1. Von Neumann Model

The model on which today's computers are based is the result of research by John Von Neumann, published in 1946. Since then, the architecture of computers has remained largely unchanged.

The fundamental characteristic of the Von Neumann machine is its procedural nature. This is because any problem must be described to the computer as a sequence of operations (program).

A computer, following the Von Neumann model, consists of a memory and a processor. The memory contains both data and programs. The processor's role is to sequentially fetch instructions from the program stored in memory, interpret them, and execute them one by one until the problem is solved. The data is transformed during the execution of the program according to a succession of states, the last being the result.

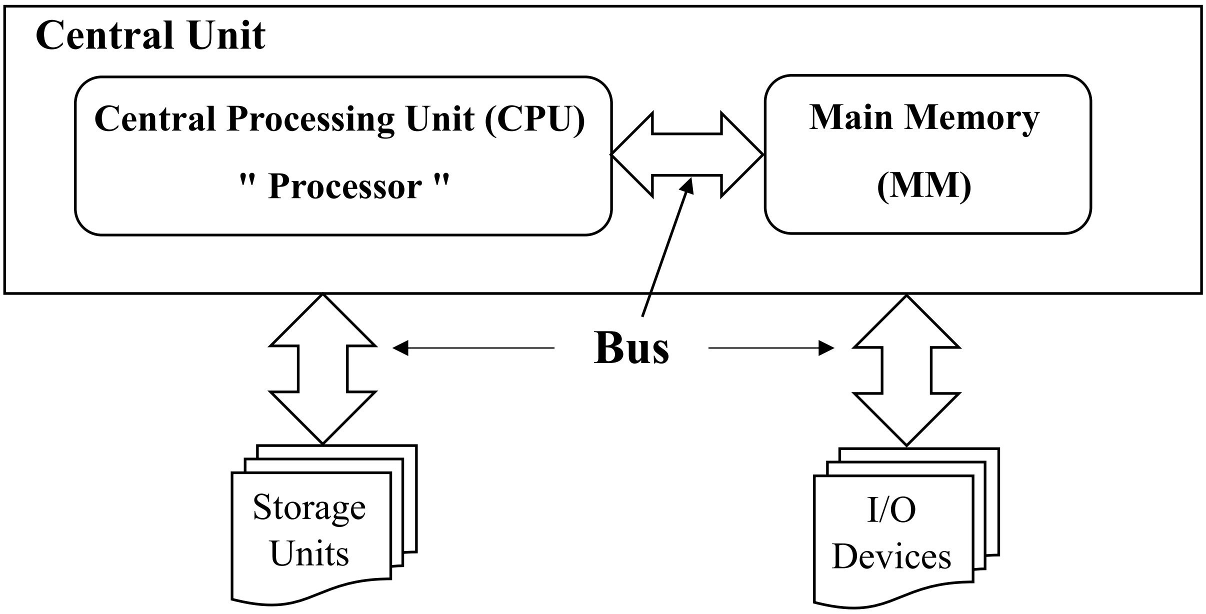


Von Neumann Model

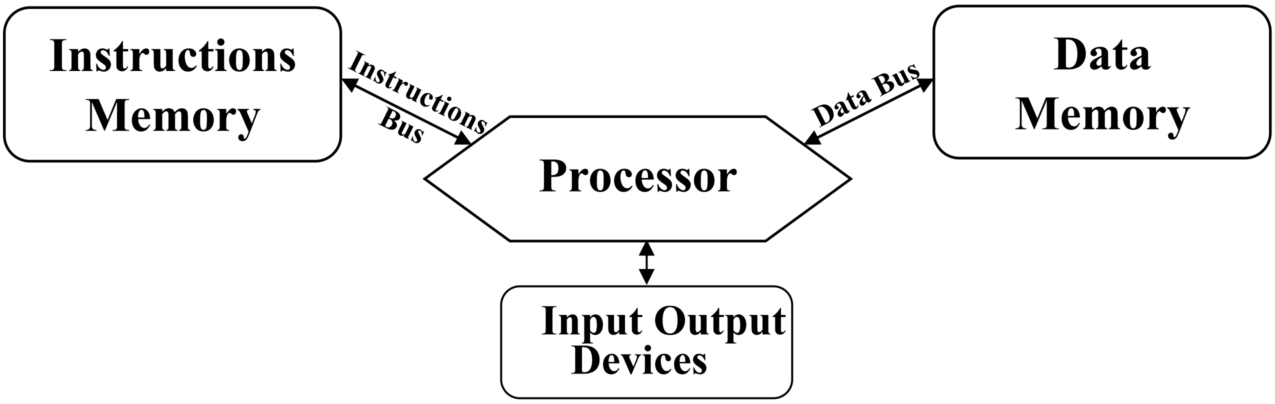
2.1.1. Logical Organization of a Von Neumann Machine

The hardware of the computer has two main units in addition to the interconnection elements:

- The central processing unit (the computer itself).
- The input/output units (all input/output and storage peripherals).
- Interconnection units (buses between the different units of the computer)



2.2. Harvard Model



Harvard Model

This structure differs from the Von Neumann architecture only in that program (instruction) and data memories are separated. Access to each of the two memories is done through a distinct path. This organization allows for the simultaneous transfer of an instruction and data, which improves performance.

2.3. The use of these structures in the real market

The architecture generally used by processors is the Von Neumann structure (examples: the Motorola 68XXX family, the Intel 80X86 family). The Harvard architecture is rather used in specialized microprocessors for real-time applications.

However, there are a few rare real-time applications with a Von Neumann structure. The reason for this is related to the higher cost of the Harvard-type structure. Indeed, it requires twice as many data and address buses.

2.4. Basic Components of a Computer

2.4.1. Memory

A memory is a device capable of:

- Recording information,
- Retaining it (memorizing),
- And restoring it (possibility to read or retrieve it later).

Memory can be in the processor (Registers), internal (Main Memory or RAM), or external (Secondary Memory)

There are different types of memories, the most used are:

2.4.1.1. Main Memory (MM or CM: Central Memory)

Memory is a device capable of storing and retrieving information (data, programs).

It is divided into fixed-size locations (cells). There are two types of memory:

- ***Random Access Memory (RAM)***: is used for both writing and reading information. It constitutes the largest part of the main memory. This memory is volatile because it retains information only as long as it is powered (it holds information in the process).
- ***Read-Only Memory (ROM)***: is only accessible for reading. Its content is often defined during manufacturing. It is used to store information that hardly ever changes.

- ***The Bit (Binary digit)***: represents the elementary unit of information (binary digit). A bit can be **0** or **1**. It is the smallest conceivable unit of information.

- Digital information is stored by distinguishing between two values (or two states) of a physical phenomenon (electric voltage or the state of a switch or magnetic core).
- A group of 8 bits is called a Byte. "1 byte = 8 bits."

- ***Memory Addresses:*** is composed of a certain number of words (cells): it is a matrix of bits.

- A memory word is a set of bits (a multiple of 8: 1 byte, 2 bytes, etc.).
- Each of these words has a number we call its address for referencing.
- If a memory has n words, the addresses will range from 0 to (n-1).
- All memory cells contain the same number of bits.
- Each word has a unique address

0	1	0	1	0	0	1	0	1
1	1	0	0	0	1	1	0	1
2	1	1	1	1	1	1	0	0
3	1	0	1	1	1	0	0	0
4	0	0	0	0	0	0	0	0
5	1	0	0	1	0	1	0	1
6	1	0	0	0	0	0	0	1
7	1	1	1	1	1	1	0	0
8	1	1	1	0	0	0	0	0
9	1	0	0	1	1	0	0	1
10	0	0	0	0	0	0	0	0

Size

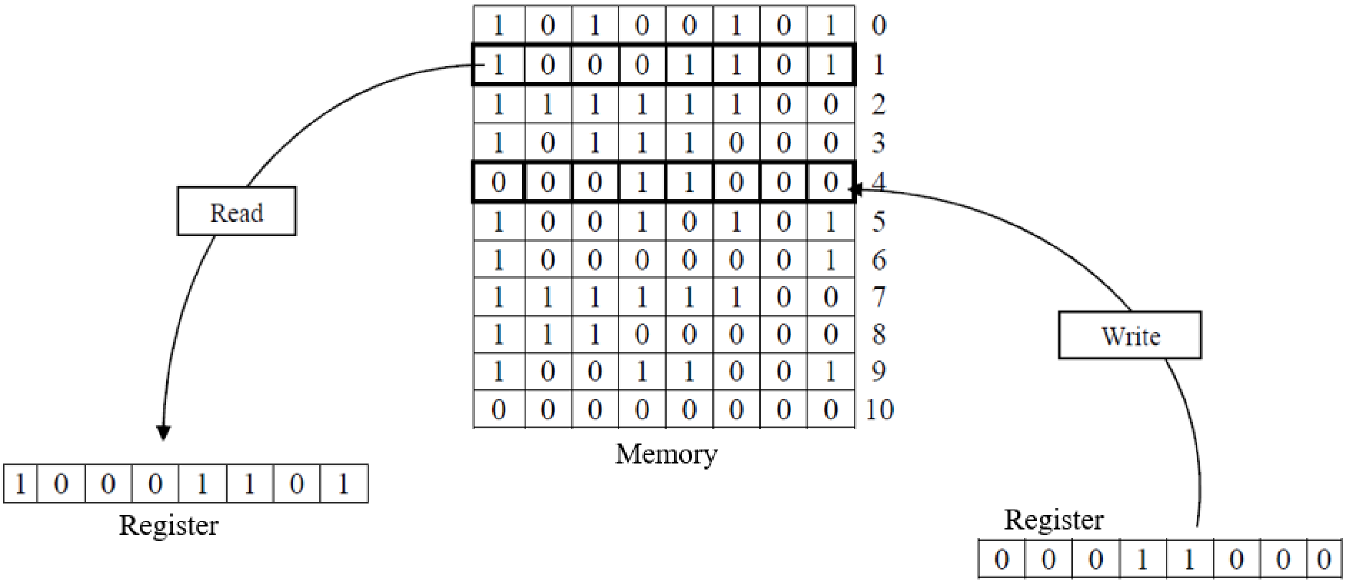
Adresse --

8-bit memory words

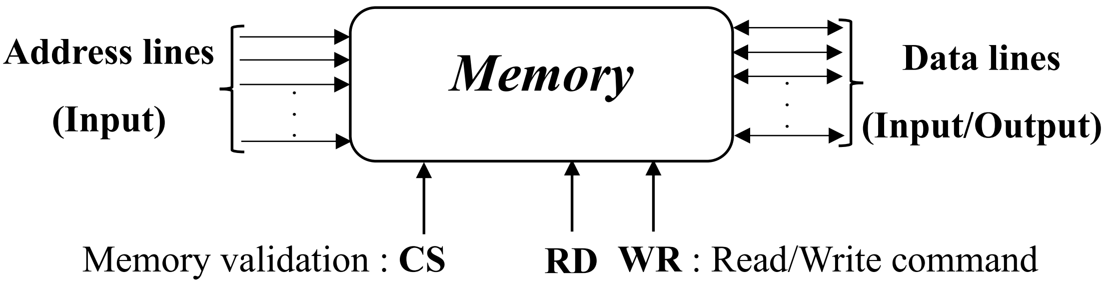
Example of 11 Bytes memory with addresses 0…10

- ***Memory Characteristics***

- ***Memory Size (Capacity):*** The amount of information that can be stored in memory is the number of words in the memory.
  - Unit of measurement: memory word.
  - Multiples of this unit are commonly used:
    - Kilobyte (KB): 1 KB = 2^{10} words = 1024 words
    - Megabyte (MB): 1 MB = 2^{20} words = 1,048,576 words
    - Gigabyte (GB): 1 GB = 2^{30} words = 1,073,741,824 words
    - Terabyte (TB): 1 TB = 2^{40} words = 1,099,511,627,776 words.
- ***Memory Word Size:*** The number of bits in each memory word (in bits or bytes).
- ***Operations on Memory (Memory Access)***
  - ***Read Access:*** This operation consists in extracting the content of a memory word or transferring the content of a memory word to an external location (register/bus).
  - ***Write Access:*** This is the transfer of content from a register or bus to a memory word.



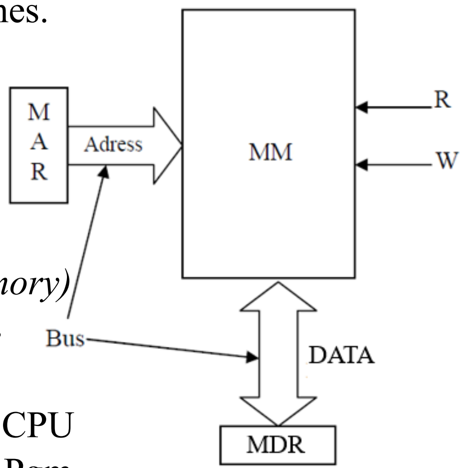
- ***Note:***
  - During a read operation, transferring the content does not mean emptying the memory word, but only to extract a copy (a duplicate) of it.
  - The read operation is not a destructive operation.
  - The write operation causes overwriting and loss of the existing information.
- ***General Operation of a Memory***



- ***Address Lines:*** *it ensures the transport od address to the memory which gives access to the memory word for an R or W operation.*
- ***Read/Write Command:*** *Selection of a read or write operation.*
- ***Data Lines:*** Routing of the data to be stored during a write operation or holding the data extracted from memory during a read operation.
- ***Memory Validation (Chip Select):*** Enabling or disabling the memory, it is a way to indicate to a particular memory chip that it should respond to read or write operations initiated by the processor or another controlling component.

- The memory works with two registers (in the processor):
    - Memory Address Register (MAR): Holds the memory address.
    - Memory Data Register (MDR) or Memory Buffer Register (MBR)
- ***The MDR Register:*** Memory Data Register, used as a receiver and holder of memory information (the content of the word in both read and write operations); it is linked to the memory's data lines.
    - ***Read MDR:*** Content of the selected memory word.
    - ***Write MDR:*** to the selected memory word.
    - ***The MAR Register:*** Contains the memory address for introducing addresses to memory before read/write operations. It is linked to the memory's address input lines.

@ → MAR.



- ***In the case of a memory read operation:***

+ Address of the target word →MAR } CPU Pgm
+ Read

→ Content of the target word → MDR } *(Memory)*

- ***In the case of a memory write operation:***

+ Data to be written (size = 1 word) →MDR } CPU Pgm
+ Memory address (target) → MAR
+ Write

→ Memory word (target) Memory }

- ***Other Memory Characteristics***
    - ***Response Time:*** The time elapsed between the R/W command and the availability of the information in:
        - ***The MDR***: in the case of a read operation.
        - ***A memory word***: in the case of a write operation.
    - ***Example:***
        - ***Dynamic Random Access Memory (DRAM):*** Access time = 70nS.
        - ***Static Random Access Memory (SRAM):*** Access time = 12nS.

2.4.1.2. Secondary Memory:

Secondary memory, also known as auxiliary storage or non-volatile memory, refers to a type of computer memory that is used for long-term data storage even when the power to the computer is turned off.

These types of memory are typically used for storing large amounts of data, such as operating system files, software applications, documents, media files, and other user data. It serves as a more permanent storage solution compared to the main memory, which is used for temporary data storage during active processing.

There are various types of secondary memory devices, including:

- **Hard Disk Drives (HDDs):** Traditional magnetic storage devices that use rotating platters to store data.
- **Solid-State Drives (SSDs):** Faster and more reliable storage devices that use flash memory technology.
- **Optical Discs:** Such as CDs, DVDs, and Blu-ray discs, which are used for storing data and media.
- **USB Flash Drives:** Portable and removable storage devices that use flash memory for data storage.
- **Memory Cards:** Used in cameras, smartphones, and other devices for data storage.

2.4.2. The Central Processing Unit (CPU)

The processor is a highly complex electronic circuit that rapidly executes each instruction, often within a few clock cycles. It acts as an interpreter of instructions. The processor is composed of three interlinked units through an internal bus:

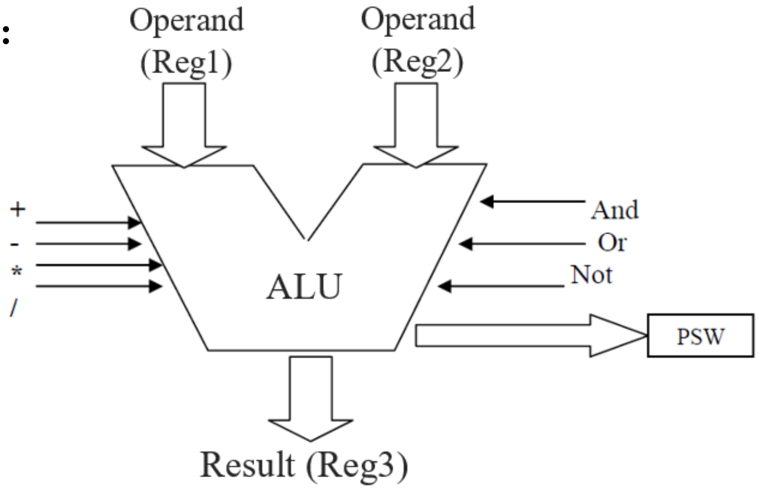
- ***Control Unit (CU)***: is a fundamental component of a processor that plays a crucial role in coordinating and managing the execution of instructions. It acts as the brain of the processor, determining the sequence in which instructions are fetched from memory, decoded, and executed.
- ***Arithmetic and Logic Unit*** **(ALU)*:*** is responsible for performing arithmetic and logic operations on data. It handles mathematical computations and logical evaluations required by the instructions executed by the processor.
- ***Memory Unit:*** Consisting of a set of registers (ACC, CO, RI, MAR, MDR, PSW, etc.) that work in conjunction with the other two units.

- ***The Arithmetic and Logic Unit (ALU)*** is a circuit that performs arithmetic operations (+, -, *, /) and logic operations (AND, OR, NAND, NOR, XOR, NOT, etc.) under the control of the control unit.

Example 1 (using 03 registers):

To perform the operation A + B:

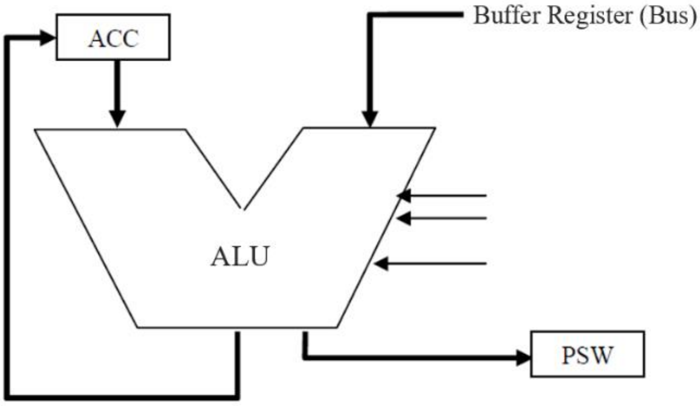
- Reg1 ← A
- Reg2 ← B } (CU: pgm)
- Addition
- Reg3 ← Result } (ALU)



Example 2 (using a single register: Accumulator):

To perform the same operation: A + B:

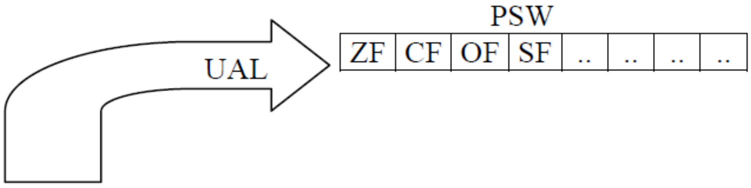
- Load A (ACC ← A) } (CU: pgm)
- Add B (ACC ← ACC + B)
- ACC ← Result } (ALU)



The registers associated with the ALU are:

- ***Accumulator Register (Acc)***: This is the most important register for the ALU. In most arithmetic and logic operations, the accumulator holds one of the operands. It can also serve as a buffer register during input/output operations. Generally, this register has the same size as a memory word. However, most machines have an extension that allows doubling its size. This extension is often called the Q or X Register. It is used in conjunction with the accumulator, for example, to hold the result of a division operation (quotient).
  Naturally, the accumulator is accessible to the programmer, as it is constantly involved during data processing.

- ***Status Register (PSW: Processor Status Word):*** Also known as the condition register, it contains indicator bits called flags, which indicate the processor's state after the execution of the last operation.



For example:

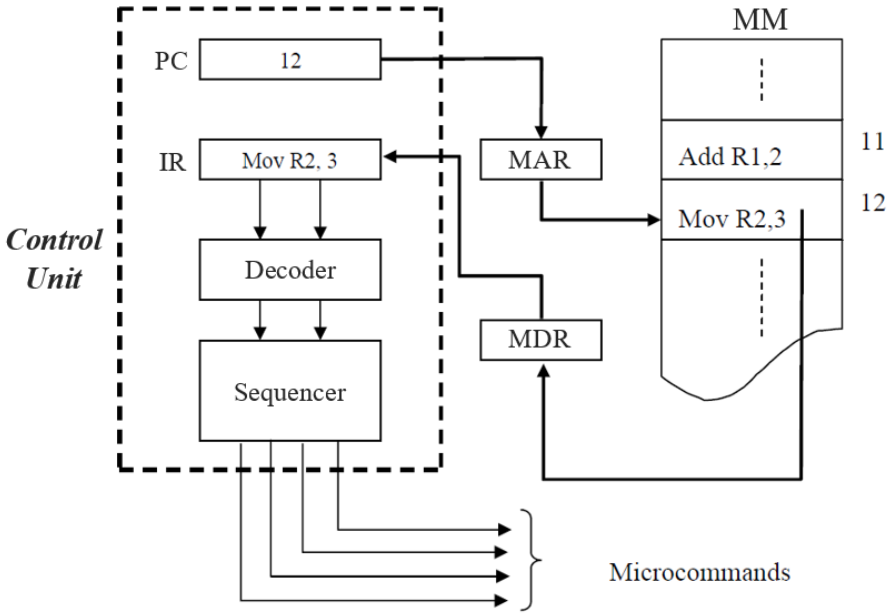
- **The Zero Flag (ZF):** indicator bit is set to 1 to indicate that the result of the last arithmetic/logic operation is zero.
- **The Overflow Flag (OF)** indicator bit is set to 1 to indicate that the result of the last arithmetic/logic operation has caused an overflow (result >> Accumulator).

- ***The control unit:*** is a set of circuits and registers that coordinate the computer's operation to correctly execute the sequence of operations specified in the program instructions. The purpose of the control unit is to:
  - Fetch instructions from main memory.
  - Analyze and decode instructions.
  - Generate ordered micro commands for other units (ALU, Accumulator, Memory, etc.) to execute the currently analyzed instruction at specific time points, facilitated by a set of registers and circuits.
- ***Note:*** During execution, the control unit must take into consideration the response time of each component (unit).

The control unit consists of:

- ***The Program Counter (PC)***: This is a register that always holds the memory address of the next instruction to be executed. It is automatically incremented (PC ← PC + 1).
- ***The Instruction Register (IR)***: This is a register designed to receive and hold the instruction during its interpretation.
- ***The Decoder:*** This is a circuit used to analyze instructions in order to prepare for their execution.

- ***The sequencer:*** is a circuit enabling the generation of micro-commands - signals - (addition, Read (R), Write (W), open BUS, etc.) to different units (ALU, Bus, Accumulator, I/O, Memory Control, etc.) as required by the operation (based on the operation's nature). These operations correspond to the interpretation of the instruction to be executed within the CPU.



Notes:

- A microprocessor is a processor in which all components are integrated onto the same chip (integrated circuit), aiming to reduce manufacturing costs and increase processing speed.
- Microcomputers are computers equipped with microprocessors (such as personal computers or PCs).
- Additional components have been added to the microprocessor to enhance its performances, such as cache memory, floating-point units, arithmetic coprocessors, pipelines, etc.

2.4.3. Input/Output (I/O) Devices: These are units that facilitate the exchange of information between the computer and the external world (users, programmers, etc.). There are three types of peripherals:

- ***Input devices:*** These units allow the introduction of information (text, numbers, images, audio, etc.) to the central processing unit (CPU) of the computer. Examples include keyboards, mice, scanners, microphones, networks, sensors, etc.

- ***Output devices:*** These units allow information to be sent from the central unit (main memory or processor registers) to the external world (users, for example). Examples include screens, printers, speakers, actuators, etc.
- ***Storage devices:*** Also known as secondary memory, these devices enable the storage and retrieval of information, similar to RAM, but on secondary media that can retain information permanently even without electrical power. Examples include punched cards, magnetic tapes, cassettes, floppy disks, hard drives, CD-ROMs, DVD-ROMs, and more.

2.4.4. Buses (Interconnection Units): The bus is a collection of communication lines that facilitate communication among various components of the computer, including the processor, memory, and I/O units.

The bus lines are categorized into three types:

- ***Address Bus:*** This is the set of lines intended to carry out the transport of addresses from the processor to the main memory.
- ***Data Bus:*** They transport data between the processor and the main memory or I/O units.
- ***Control Bus:*** These lines include commands necessary for proper communication between different elements (Read, Write, Add, Interrupt, etc.).

Note: There are three types of buses based on the elements being interconnected:

- ***Internal bus (local):*** This type of bus connects the internal elements of the processor (ALU, Control Unit, Accumulator, etc.).
- ***System bus (external):*** This bus connects the processor with the main memory and I/O units.
- ***Expansion bus (I/O bus):*** This bus connects peripheral controllers (specialized processors for managing peripherals) to the central unit. Examples include ISA and PCI buses.