-

UNIT 2.
ARCHITECTURE AND COMPONENTS

Computer Systems
CFGS DAM

Revisado por: Francisco Lifante Rivera
f.lifanterivera@edu.gva.es

2023/2024

Versión:231010.1855

Licencia

Nomenclatura

A lo largo de este tema se utilizarán distintos símbolos para distinguir elementos importantes dentro del contenido. Estos símbolos son:

Importante

Atención

Interesante

ÍNDICE DE CONTENIDO

UD02. Architecture and Components

1. Historical evolution

1.1 Before digital computers

Before the computers even existed, there were other machines that tried to solve the complexity of calculus. One of them was the Abacus, invented around 500B.C., created to perform complex arithmetical operations in a simple and fast way. After it other machines emerged, like Pascaline by Blaise Pascal (XVII century) or Bagage Machine by Charles Bagage (XIX century).

Also, other scientists and mathematicians set the basis for computer sciencie. Among others, binary code (Leibniz), bool algebra (George Bool), turing machine (Alan Turing, who set the math basis for computing), etc. There were so many contributors to the field and without their work, probably we couldn't be reading this digital document.

1.2 Digital computers

First digital computers arise during II World War, used mainly for miliitary purposes: de-encryption of communications, calculations of long shots, etc. Those first computers main characteristics were that they were really hard to reprogram, because they were usually hardware programmed through wires:

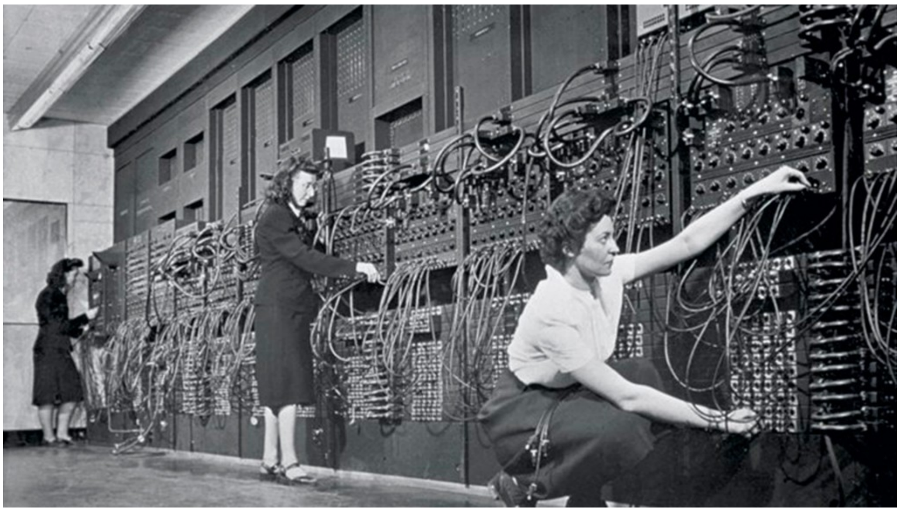

Figure 1. Eniac programmers

Also another feature of the first digital computers was the usage of vacuum tubes. They were big, unreliable and energy consuming so they were an obstacle. In 1947 the first transistor was invented, reduzing size, increasing the reliabilty of computers and their need of energy. Two years later the Integrated Circuits were invented, to create chips that could be miniaturized.

But it wasn't until the 1970s were the personal computers weren't born. Until then, only gigantic institutional computers were built. 1970s saw the appereance of the IBM Pcs, the IBM compatibles (with an interesting story about them portrayed at the tv show Halt Catch and Fire), Apple II, Spectrum… which changed the whole way computers were used.

1990s and, specially, 2000s saw the interconnection of all personal computers through internet. Nowadays we have computers everyhwhere, shaped like an smartphone, smartwatch, tablets… Even cars, washing machines and many other home appliances can fit the definition of a computer system.

1.3 Definition of a computer system

What's a computer system? Basically a multi-purpose machine that can solve many different problems through computing.

They are composed of two parts: hardware (physical elements) and software (the instructions that tell the hardware how to solve the specific problem). Some authors also include the user and the data, since more and more the software uses the inputs the user gives them and the previous data to take choices.

2. Computer Architectures

Computer architectures define the abstract model where computer system is defined based on its components, their “tasks” and the interconnection between them.

2.1 Von Neumann architecture

Von Neumann architecture is the most implemented architecture. That doesn’t mean that it’s the only one existing, but due to its importance we are going to study it. Von Neumann’s architecture defines the main elements of a computer as:

- CPU (Central Process Unit)
- Memory Unit
- Buses
- I/O (Input/Output to external devices)

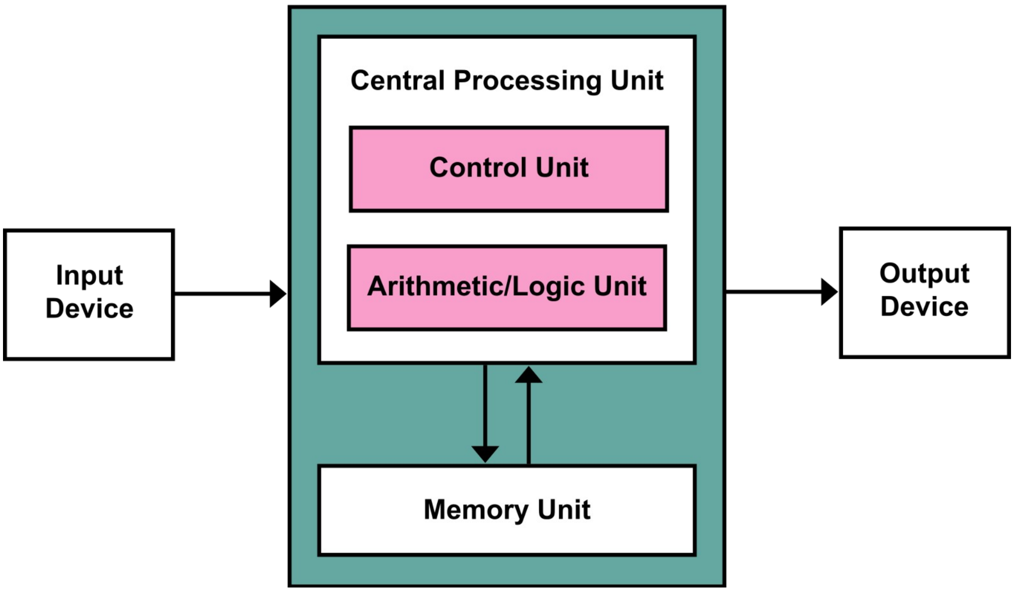

Figure 2. Von Neumann Architecture
By Kapooht - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=25789639

2.2 CPU (Central Process Unit)

The CPU is the core of the computer. CPU read instructions from memory and execute them, doing the required operations. It is composed of:

- **Arithmetic-Logic Unit (~~ULA)~~:** it is a unit designed to perform arithmetic operations (addition, subtraction, multiplication,...) and logic operations (not, and, or, ...). It's the worker.
- **Control unit (~~UC)~~:** it decides what to do in each moment. It's the boss. Its components are:
  - **Decoder**: its function is to decode instructions and prepare all the signals to execute the program properly. For instance an instruction can be ADD (which adds up two numbers that are in memory). When it comes to executing, the decoder generates a lot of small

signals such as reading the first number of the memory, taking it to the ULA, reading the second number, taking it to the ULA, perfom the operation (adding) and saving the result in memory. In this case an instruction has generated six signals[1]

  - **Clock**: is the system that sets the pace at which things are done. It's like the conductor's baton. I allows all system components to operate perfectly together.

- **Registers**: are little and fast memories. They are used as temporal storage when an instruction is being executed or to store temporally results of operations. Most important registers are:
  - **Program counter (PC)**: an incremental counter that stores the memory address of the next instruction that is going to be executed.
  - **Memory address register (MAR)**: it stores the address of a block of memory for reading from or writing to.
  - **Memory data register (MDR)**: a two-way register that stores data fetched from memory (and ready for the CPU to process) or data waiting to be stored in memory.
  - **Instruction register (IR)**: a temporally register to store the instruction that has just been fetched from memory.
  - **General purpose registers**: those registers are used to store information temporally, for example to store operands of and addition or to store the result of an arithmetic-logic operation.
  - **Flags**: These are very simple records. They can only be in two states. on or off. They are used to indicate the status of an operation. For example, flag Z is set to 1 if the operation has resulted in 0., N is set to 1 if the result is negative.

To measure the speed of a CPU we use frequency units that indicate the number of times a thing repeats itself per unit of time. In geneal we use the Hz (or its multiples) that indicates the number of operations that can be performed in a second.

2.3 Memory Unit

The memory unit is an internal memory that contains the instructions to execute and the data or programs to use. It consist of several cells that store 0 or 1 (1 bit).

A set of cells of a determined size is called word. Computers usually use words of 32 bits or 64 bits. Each word has an address associated with it. When you read information of the memory unit, you have to read the entire selected word.

1 Here, the process es simplified. Actually many more signals are generated

2.4 I/O (Input/Output external devices)

The I/O devices are the devices used to obtain information (input devices) or to provide information (output devices). They are usually called peripheral devices. A mouse or a keyboard are examples of input devices and a printer or a screen are examples of output devices. Also there are devices that are input and output at the same time (a hard disk, a touch screen, internal components like a sound card, graphic card, expansion port,…).

2.5 Buses

A bus is a set of wires that connect components of the computer.

A bus is defined by its boundary (number of bits that can be sent in one operation). Usually it is 32 bits, 64 bits,128 bits….

> The bus speed can be measured in the number of operations that can be performed in a time unit. Usually it is measured in Hz (operations that can be done in one second).

- Serial bus: buses send information by one wire each time. They are very simple and require a simple hardware to be managed. For this reason, in spite of using only one wire, they obtain very good performance. For example, SATA and USB are serial buses.
- Parallel bus: buses send information parallel, by several wires. They are complex to process and require complicated hardware. Usually it obtains less performance than serial bus. For example, the old com port or the old printer port were parallel buses.

2.6 Instructions

An instruction is the data that CPU interprets to perform an operation.

When we talk about a CPU of 16 bits, 32-bit, 64-bits, … we are talking about the number of bits than an instruction can use.

Usually an instruction is divided on two fields:

- Operation code (OP_CODE): is the code that indicates the CPU the type of operation to perform.
- Operation data (OPERAND): is the data associated with the operation that is going to be performed. There can be more than one.

The addressing mode is used to find the data stored in memory through the field/s specified in the instruction, called operand. The common addressing modes are:

- Immediate, the operand (data) is given explicitly in the instruction.
- Absolute or direct, the address where the operand is stored is given explicitly in the instruction.
- Indirect, the operand’s effective address is stored in the location whose address is given explicitly in the instruction.

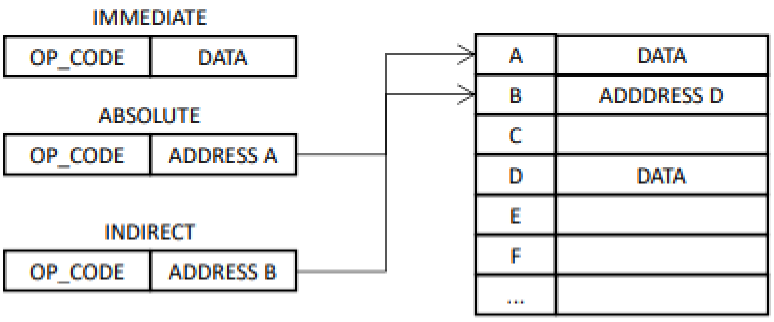

Figure 3. Addressing modes

2.6.1 Instruction Cycle

The instruction cycle is the steps that CPU takes to obtain the next instruction and perform it. Usually, those steps are:

1. Fetch the instruction: the next instruction is fetched from the memory address that is stored in the Program Counter (PC) and it is stored in the Instruction Register (IR). At the end of this operation, the Program Counter (PC) points to the next instruction (that will be read at the next iteration of the cycle).
2. Decode the instruction: during this cycle the encoded instruction present in the Instruction Register (IR) is interpreted by the decoder of the CPU.
3. Execute the instruction: The control unit of the CPU passes the decoded information as a sequence of control signals to perform the actions required by the instruction. (Example: reading values from registers, passing them to the ALU, writing the result back to a register).
4. Store: The result generated by the operation is stored in the main memory, in a register or sent to an output device

The Program Counter (PC) may be updated to a different address from which the next instruction will be fetched. (It is the way to jump from one place of the code to another). When the cycle ends, it is repeated again.

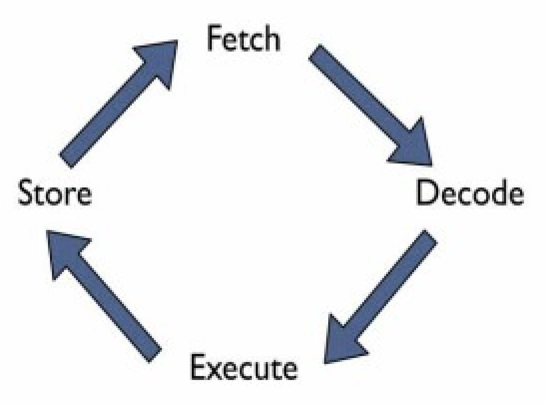

Figure 4. Instruction Cycle

2.6.2 Instruction sets

Each CPU has their own instructions that can perform, based on what's been designed in their integrated circuits. The instructions can be:

- Arithmetic-logic operations: perform arithmetic-logic operations.
- Data handling and memory operations: read and store date from/to memory and from/to input/output devices.
- Control flow operations: operations that change the flow of the program modifying the Program Counter (PC). Example can be conditional operations, loops…

2.6.2.1 Example of instruction set

In this point we are going to describe a fictional set of instructions of a fictional computer to understand what a set of instructions is and how an instruction is codified.

In our fictional computer:

- Each instruction is codified using 8 bits. The first 2 bits are used to store the operation and the next 6 bits are used to store information of the related operation.
- We have a memory with 2^6 addresses, each one with one word of 8 bits (1 byte)
- Our CPU has 3 temporally registers R1, R2 and R3. They can store 8 bit.
- In our memory, address 000000 contains the word 00001001 (9 in ~~two's complement~~) and
- 000001 contains the word 00001000 (8 in ~~two's complement~~).

In our instruction set, there are 4 instructions:

- Instruction 0: in binary 00. It indicates that we can read 2 numbers from temporally CPU
- Registers R1 and R2 and add them. The result of the addition will be stored on R3. In this instruction only 2 first bits are used, the next 6 bits are ignored.
- Instruction 1: in binary 01. It uses a memory address. The content of that memory address is stored in R1.
- Instruction 2: in binary 10. It uses a memory address. The content of that memory address is stored in R2.
- Instruction 3: in binary 11. It shows on screen the value of the temporally register R3. In this instruction only 2 first bits are used, the next 6 bits are ignored.

For example, if we want to codify an instruction for reading the word on the memory address 000010 and store it in R2, we use the following instruction:

10 000010

Where the first 2 bits indicate the operation (10) and the next 6 bits indicate the memory address to read.

The next code will show in screen a number 17.

01 000000	Load in R1 content of address 000000.

10 000001	Load in R2 content of address 000001.
00 000000	Perform the addition R1 + R2 and saves it in R3.
11 000000	Show the result (R3) on screen.

2.6.2.2 RISC and CISC

RISC and CISC are the most popular approach to design computers. They determine how is the instruction set of the CPU.

- CISC: Complex Instruction Set Computer (CISC) has the next characteristics
  - High set of instructions (can do many things)
  - Variable size instructions
  - Complex instructions, with one instruction doing complex operations or several things
- RISC: Reduced Instruction Set Computer (RISC) has the next characteristics
  - Reduced set of instructions
  - Fixed size instructions
  - Only load/store instructions can access to memory

Usually, RISC computers are an advantage to CISC computers. Mainly, because simpler instructions take less time consuming the CPU clock cycles, leaving space for other instructions to be performed, improving the overall performance of the CPU.

3. Components

3.1 Motherboard

The motherboard, also known as MoBo, is a printed circuit board where the computer components are connected. It tends to give more Importance to other elements such as the microprocessor or memory, but the motherboard is critical: what and how many components can be installed in the computer depends on it.

> At first, each MoBo could hold a specific model of processor (CPU) but from any manufacturer (Intel, AMD, Cyrix..). But since 2004, although the processors were compatible in functionalities, manufacturers were differentiating their models in the pinout, so the motherboard has to be specific to each type of processor ever since.

It has installed a series of integrated circuits, including the chipset which serves as a hub between the microprocessor, random access memory (RAM), expansion slots and other devices. All the components of the motherboard are connected with "paths" whose name is buses.

Besides, the motherboard includes a firmware (software stored in a read-only memory like EEPROM, flash[2]...) called BIOS, which lets you perform basic functions such as testing devices, video and keyboard operations, recognition devices and operating system load.

As the motherboard it's the main component where any other component has to connect, many standards have been developed. That way, there amny manufacturers, brands, kinds of components but the standards make the interoperability possible.

3.1.1 Form factor

The form factor of a motherboard defines:

- The shape of the motherboard: square or rectangular.
- Their **exact** physical dimensions: width and length.
- The position of the anchors. That is, the coordinates where the screws are positioned to anchor it to the casing.
- The areas where certain components are located. Specifically, the expansion slots and the rear connectors (keyboard, mouse, USB, etc.)
- The electrical connections of the power supply, that is, the number of power supply wires that the motherboard requires, their voltages and function.

Following we can see some form factors:

2 Types of memories will be studied later

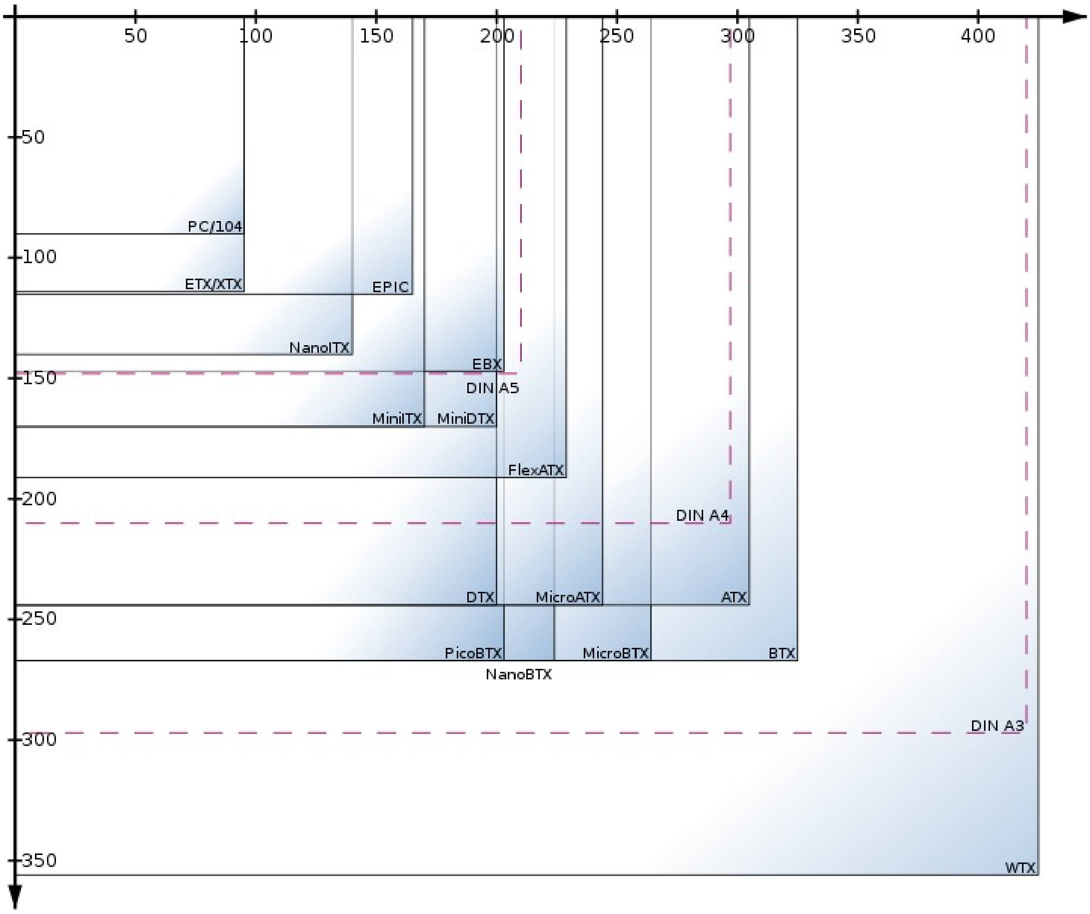

Figure 5. Instruction Cycle

There are many form factors, ranging from the ones for traditional personal computers shaped in the traditional box case, from laptops, smartphones… But the thing is that all of them specify the same features of the mother board.

3.1.2 Connectors

3.1.2.1 Internal

Internal connectors are all the connectors that are in the motherboard and are not shown externally, outside the case. Some of them are:

- **Power Supply:** also called MOLEX connector.
- **Control pins:** they connect the case buttons and connectors with the motherboard. Power on/off button, reset button, HD led, front USBs, front audio connectors…
- **SATA:** transfers data between the mother board and storage devices, like hard drives
- **RAM Slots:** where the RAM modules are going to be inserted

3.1.2.2 External

External connectors are the ones shown outside the motherboard. Some of them are:

- **PS/2:** to connect keyboard and mouse. Nowadays they are usually connected through USB.

- **USB**: Universal Serial Bus. The most standard connector nowadays. Tons of devices can be connected to our computers through USB.
- **VGA/HDMI:** graphic connectors
- **Jack:** audio connectors

3.1.3 BIOS

The BIOS (Basic Input Output System) is the basic computer software for controlling the hardware elements.

> A common mistake is to think that the BIOS is a hardware component. It is software. In fact, is not a single program. It s a set of programs that perform different tasks.

One of the most important programs is POST. It runs when you turn on the computer and locates and recognizes all devices needed to load the operating system into RAM. It manages at least the computer keyboard, and provides a basic output in the form of sounds trough the motherboard speaker when there is an error, such as a device failure or an error in the connection. These error messages are used by technicians to find solutions when assemble or repair the computer.

Formerly, it was stored in a chip ROM (Read Only Memory), a memory that can only be read, and not deleted if it lacks the power. So, if you wanted to change the program you had to change the whole chip which, obviously it was a very complicated task. Today, it is installed on a Flash or EPROM chip memory, which allows updating, although in a complicated and delicate way. All models used (ROM, EPROM or Flash) are slower than conventional RAM, therefore the execution of BIOS programs usually slow.

Another well-known program of the BIOS is the SETUP. This program provides a graphical interface that allows access to another memory, CMOS, where part of the system settings are saved. CMOS is not a ROM, so it needs power to save the data. For that, the motherboard has a battery. If the battery runs out, all settings are lost and we need to change the battery and configure the system again.

3.1.4 UEFI

Although BIOS is an effective system, it has lost in efficiency over the years, so the industry needed to evolve it. That is why in recent years the BIOS has been replaced by UEFI.

UEFi's main idea is to replace the BIOS by adding new features such as a setup with a much more modern graphical interface, a secure boot system, a higher boot speed or support for hard disks over 2TB.

3.1.5 Expansion slots

The expansion slots are connectors to insert cards into the motherboard and connect devices to the bus to which they are connected. For instance, we can connect a graphic, a sound or a TV card…

The most common expansion slot is PCIe. It has 2 versions: 1.0 and 2.0. Useful for all kinds of devices, including graphics. It is made up of 1, 4, 8, 16 or 32 lanes of data between the motherboard and cards. The number of lanes is shown with x1, x2, x4, ... x32. For every lan can reach 250 MB/s

(1.0) or 500 MB/s (2.0). The standard for graphics cards is x16, which supports 4 GB/s to 8 GB/s.

3.1.6 Socket

The socket is the place of the motherboard where the processor is inserted. Each generation of processors have their sockets. Each is compatible only with the same family.

Today all the CPU sockets use the ZIF (Zero Insertion Force) mechanism. With this mechanism, the processor is inserted and removed without apply any pressure on him. Lifting the lever near the socket, the microprocessor is released, being extremely simple remove it.

There are basically two types of sockets, LGA (Land grid array) which consists of a socket with pins to place the processor on, and PGA (pin grid array) where the pins are on the processor and it is inserted with the appropriate holes to place it.

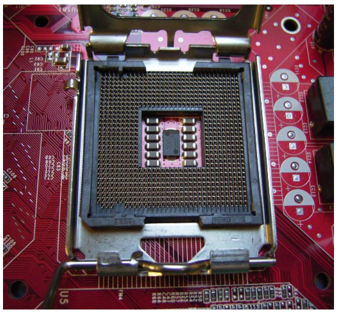

Figure 6. LGA Socket

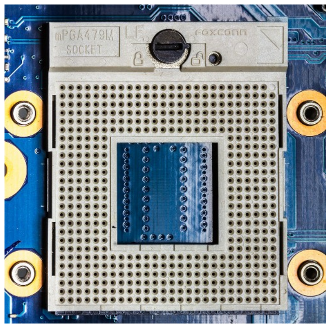

Figure 7. PGA Socket

3.1.7 Chipset

Chipset function is to control how the CPU interacts with any other component. Usually the chipset is composed of two chips:

- Northbridge: controls how the CPU interacts with the memory and the video
- Southbridge: controls how the CPU interacts with everything else, which means storage, buses, ethernet…

As the chipset is the intermediary between the CPU and everything else, it has taken a huge importance in matters of performance. If not chosen wisely, it can become a big bottleneck to access a powerful CPU.

3.2 Processors

We could say that the processor is the brain of the computer. It is a chip that has within it thousands or millions of transistors[3] . They usually have a shape of a square or rectangle and black can be found on an element called socket.

3.2.1 Internal architecture

Over time and with the advancement of technology, the scale of integration of components is

3 4 A transistor is a very simple but important electronic device. Its task is similar to a valve. It has three connectors. If we put electrical current in one of them the transistor, allows the current pass through the other connectors. Otherwise, the transistor will not allow pass.

increased (the number of transistors included in the processor is increasing, that is, they are becoming smaller).

Today, there are two kinds of processors: single and multi core. Multi core are way more expanded, but single core are easier to explain and are still used in many devices that don't require too much cmputing capacity.

3.2.1.1 Single core

The implementation of abstract von neumann architecture (see CPU (Central Process Unit)) for CPUs in single core CPUs is as follows:

- Control Unit: Responsible for seeking the instructions in the main memory and decode them to send control commands to other components.
- ALU. Arithmetic Logic Unit: responsible for performing arithmetic and logical operations.
- FPU. Float Point Unit: also called math coprocessor. It is responsible for performing floatingpoint operations. It usually performed addition and multiplication operations, although someone perform higher functions such as exponents.
- L1 cache or Level 1 cache: is a volatile memory built into the processor core operating at the same speed as this. Its function is to store the most common data for faster location.
- L2 cache or Level 2 cache: is a volatile memory built into the processor, although not directly in the heart of this. The purpose is the same as L1, but slower although bigger
- BSB (Back-Side Bus): is the connection between the microprocessor and its L2 cache.
- FSB Front-Side Bus): is the data bus used as principal in some of Intel microprocessors (in AMD processor is called HyperTransport).

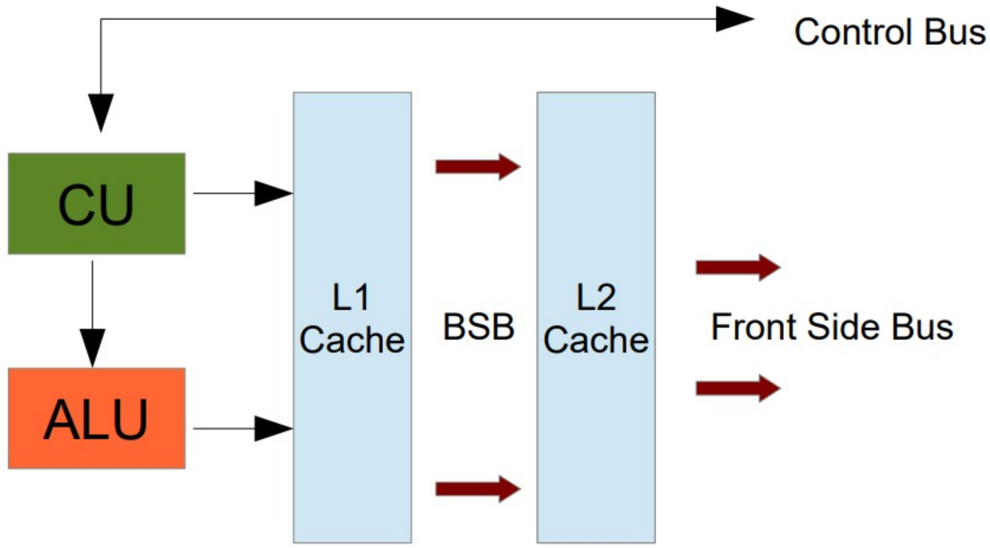

Figure 8. Single core diagram

3.2.1.2 Multi core

The current microprocessor manufacturing technology is reaching its limit. Each time smaller and faster processors are manufactured and that causes an increase in temperatures of the processors.

Also our current knowledge of physics dictates that we are getting to sizes were only quantic physics matter, where our current designs won't work.

These problems make it difficult to increase processor performance, so for future enhancements, it was necessary to find another way other than the increased speed.

The idea was increase the number of core. In a single core processor only one instruction can be processed in an instant, but if we have n cores the number of instructions to be processed in an instant will be n. Based on this kind of work, the parallel processing, the manufacturers began to build multicore processors.

The general struct is similar to single core processors, but we have to add:

- Integrated memory controller: the memory controller for faster access to RAM
- Bus transport high speed.
- Another L3 cache, shared among the cores.

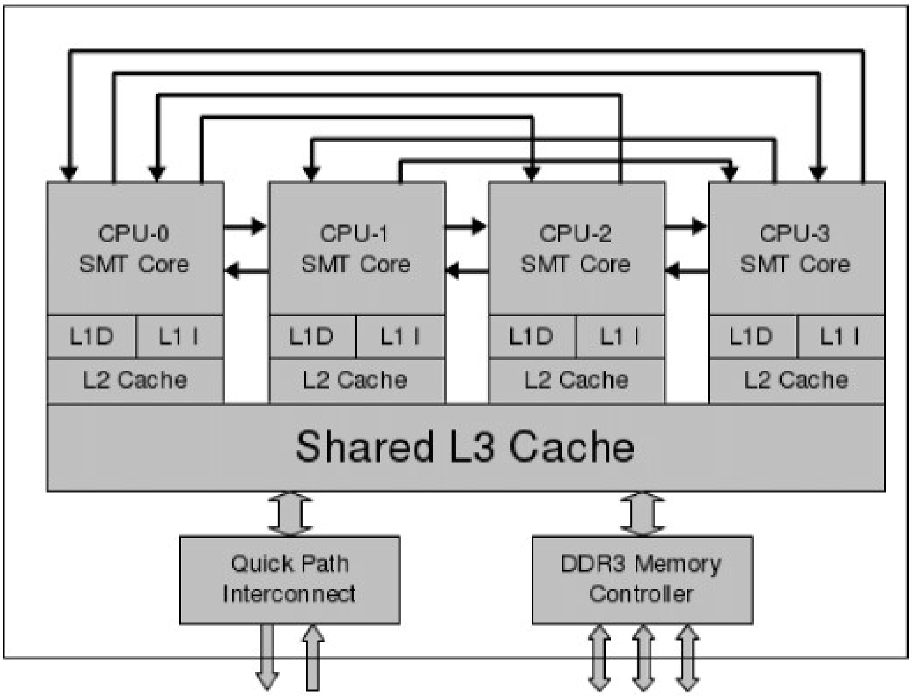

Figure 9. Multi core diagram

3.2.2 Features

3.2.2.1 Frequency

Processors are devices that perform a big number of tasks in stages, so they need a way so that all the components work at the same rate, with the same rhythm. This is achieved with a clock signal whose frequency is called frequency microprocessor.

The frequency of a processor is measured in millions or billions of hertz[4] , ie, the commonly used MHz or GHz units.

4 A hertz measures the number of times that something occurs in a second

The frequency has never been a very good parameter to assess the real speed (the one received by a user) of a processor. It should only be used to compare processors of the same family: two Pentium IV, two i7... which are identical in all other respects. This is because each processor can perform too much different work in a single cycle.

3.2.3 Bus speed

Processors have a data bus to communicate with other system elements. This bus is connected to the Northbridge, which communicates with memory or some expansion buses like PCIe. This bus is called FSB or Front Side Bus in Intel microprocessors and HyperTransport in AMD. The transmission will be better if the flow rate is high.

- The first parameter affecting the flow rate is the width of the bus, that is, the number of bits that can be transmitted in each clock pulse.
- The second parameter is the clock speed at which the bus operates. Because of the difficulty to manufacture electronic components operating at processor speeds (MHz or GHz), the bus speed is usually lower than the processor (66MHz, 133MHz, 266MHz …).

Modern processors use buses with multiple use of the clock signal, so that for each clock pulse, it sends various data groups of bit and not just one. So we talk about a bus speed in MHz equivalent (which value is greater than real MHz) , because this MHz are not the real, but the user feels that the speed is greater. For example, in an Intel Core 2 that uses the clock signal to send 4 bit, its FSB is called FSB800 (although the bus runs at 200MHz physical), or FSB1333 (although it works 333 physical Mhz).

3.2.4 Cache

A cache is a fast memory that is used to store a copy of the data most likely required, then accelerating performance by reducing the number of times to access the RAM (which is slower). There are three kinds of processor cache:

- L1 cache: always internal, built into the kernel itself. it always runs at the same speed than processor. It is usually divided into two parts: data and instructions. Today is usually around 512 KB- 1MB.
- L2 and L3 cache: they are connected the BSB (faster than the FSB). They can be included in the micro core inside the encapsulated or external.

When the processor is looking for a piece of data, it look first in the L1, if it find it, it use it. otherwise it goes through levels to reach RAM.

In the nomenclatures, when 64KB + 64KB appears, it indicates that is 64 KB for data and the same size for instructions. If it appears 2 x 4MB, it indicates that are 4 MB per core (if two cores).

3.2.5 32 – 64 bits

Although, many years ago, some parts of the processors were 64-bit or more, most records were still 32 bits, allowing access to only 4 GB of RAM (2^{32}).

In the last years, AMD (first) and Intel (after) supplemented the x86 32-bit technology with 64-bit option, creating the x86-64 technology. In this technology, all 32-bits and 64-bits programs are supported. The addressing memory has increased, but not using the 64-bit but, for now, 44. That allows to address 2^{44} = 16TB of memory.

3.3 Current Intel and AMD Processors

3.3.1 Intel

The current generation of Intel processors is based on the Alder Lake architecture. This architecture introduces a hybrid approach with high-performance cores (P-cores) and high-efficiency cores (E-cores), striking a balance between performance and energy efficiency. Alder Lake processors are built on a 10-nanometer (10nm) manufacturing process, providing improvements in power consumption and transistor density.

In terms of performance, Intel Alder Lake processors have proven to be highly competitive, delivering excellent performance in both single-threaded tasks and multi-threaded workloads.

3.3.2 AMD

The current generation of AMD processors is based on the Zen 3 architecture. These processors are built using a 7-nanometer (7nm) manufacturing process, making them highly efficient in terms of power and computational power. The Zen 3 architecture brings significant improvements in instruction efficiency and cache latency, resulting in outstanding performance across a wide range of applications.

AMD processors based on Zen 3 have proven to be leaders in multi-threaded performance, making them ideal for tasks that demand high processing power, such as video editing and rendering.

3.3.3 Sockets

3.3.3.1 Intel

The Alder Lake generation from Intel uses the LGA 1700 socket. This socket is specific to the Alder Lake architecture and features a higher number of pins compared to previous generations. It's important to ensure that the motherboard is compatible with this socket before purchasing an Alder Lake processor.

3.3.3.2 AMD

Processors based on AMD's Zen 3 architecture use the AM4 socket for desktop platforms, while server platforms use the SP3 socket. However, AMD has announced that their next generation of processors, based on the Zen 4 architecture, will require a new socket, the AM5.

3.4 Memory

3.4.1 Computer memory

RAM memory has two main characteristics:

- RAM is used to store data temporarily, that is, until the computer shuts down or is restarted.
- Its access is random (in fact, RAM means Random Access Memory), that is, we can be accessed anywhere at any time.

We can imagine the memory like a big grid. To identify each cell we use the address of the cell, or, in other words, the memory address. In each cell, the memory can save one bit. To work in a easier way, the computer puts in groups of 2n , with n>=3 n<=7. Each group is called word. The next figure shows an example of a memory of 4Kb (or 512B) with a word length of 8 bits:

Address	b0	b1	b2	b3	b4	b5	b6	b7
000-007	0	0	0	0	0	0	0	0
008-00F	0	0	0	0	0	0	0	0
010-017	0	0	0	0	0	0	0	0
018-01F	0	0	0	0	0	0	0	0
020-027	0	0	0	0	0	0	0	0
...								
FF0-FF7	0	0	0	0	0	0	0	0
FF8-FFF	0	0	0	0	0	0	0	0

Figure 10. 512B memory

3.4.2 Memory Bus

It is the path to the data. It is divided in two:

- Address Bus: It is in charge of sending the memory addresses. As discussed above, depending on its width, you can address a maximum amount of memory.
- Data Bus: the part that transmits data and instructions itself. Current processors use 64-bit buses, that is, can be transmitted 8 bytes in each clock cycle..

This bus has a speed, which is measure in Hz. The higher the speed of the bus more data can be sent. For instance, for a 32bits bus to 100MHz, the flow rate will be:

32*100*106 b/s = 32*108 b/s = 4*108 B = 4*108/220 MB/s= 381,47 MB/s

3.4.3 Features

3.4.3.1 *Clock speed*

Clock speed in memories specifies the frequency its bus supports. Or, at least, their equivalent speed. For instance, PC100 means that is capable of supporting a 100MHz bus. But a DDR333 memory means that it can work at 166MHz but can send data twice per clock cycle.

3.4.3.2 Bandwidth

The goal is to get a memory that transmits the maximum amount of data. To represent this value we use the bandwidth, one of the fundamental parameters on the performance of a memory.

This is a mix between the clock speed and the bus width, and express the flow rate that the memory can achieve. Its value can be calculated by:

(real clock speed x use of cycle[5] x bus width) / 8 = bytes/s.

For example, DDR-667 memory operates on 333MHz physical bus but with double use. Working on a 64-bit bus which is capable of transmitting 5333MB/s

3.4.3.3 Number of channels

There is one more factor to consider: the number of channels. It really is not just a factor of memory, also of the MoBo in which you are connecting the memory. The memory controller can allow access to two channels simultaneously (or even three or four), which makes multiplying the effective bandwidth. This technologies are called dual, triple or quad channel.

To use Dual Channel, Triple Channel or Quad Channel requirements are usually: to have a good and identical (two to two in Dual, three to three in Triple or four to four in Quad Channel) memory modules that were DDR, DDR2, DDR3 (in the case of Triple Channel only DDR3) and install them in special slots memory following the instructions of MoBo's manufacturer.

In this case the way to calculate the bandwidth is:

(number of channels x real clock speed x use of cycle x bus width) / 8 = bytes/s

3.4.4 Parity modules and ECC modules

When the computer is used for critical tasks, we must ensure the integrity of the data memory or, at least, know if there have been failures to read or write.

Methods to achieve this include the application of a mathematical algorithm to the data and compare the result with the data to know if there is an error. There are two methods:

- Parity generates a bit that indicates whether an even or odd number of zeros (or ones) exists. Before sending the byte, the system check the byte and the parity bit to see if there are any changes. This has the problem that if failures occur in pairs, could pass as OK.
- ECC: error correction code. It allow detect and fix small errors (its functionality is beyond the list of topic for this course)

Both of them require more bits per byte, so visually there are more memory chips in each module. Moreover, error checking takes time, so the performance is lower.

It is rare to find a server that does not use ECC. The characteristics registered and ECC are independent, although in practice they almost always go together.

3.5 Hard Disk

Any PC can start and run without hard disk, because it is a secondary storage, but today, this

5 How often it sends information per cycle?

configuration does computers least useful. For 50 years hard drives (from now HDD) have been based on magnetic technology, although in recent years have appeared new hard drives in solid state (SDD) based on flash memory, with the same external appearance that hard disk, but faster, quieter and more efficient.

A hard disk does not belong to a Von Neumann Machine, so it is a peripheral

3.5.1 Physical format

In modern personal computers, physical size is 3,5” and 1” height. Laptops world usually is 2,5” or 1,8”. Same as SSD, which we usually found them sized 2,5” or 1,8”. There are many adapters to use small-format disks in larger bays, as well as housings for connection to external ports.

When we talk about 2.5'' we refer to the size of the disk itself, not the case.

3.5.2 Capacity

Usually we measure the capacity of hard drives in GB or TB.

In the advertisement world, 1GB means 10^9 bytes, not 2^9 bytes

3.5.3 Internal transfer speed

This parameter is one of the most important of a hard disk. Refers to the maximum data flow (measured in MB / s) can be read or written within a device at a particular time. Normally it is lower than the speed of the interface (external speed).

Two different internal speeds are taken into account:

- Maximum speed internal transfer: for a 7200 rpm disk can be by 175 MB/s. It is a theoretical and ideally given to read information value.
- Sustained speed: indicates performance when reading data in appreciable amounts. It is more important than the maximum speed internal transfer and always inferior. For a hard disk as before it would reach 125 MB/s.

In the SSD, reading operations are faster than writing. In the HDD are almost the same, though a bit faster reading. If a manufacturer specifies only one, will be reading.

Overall, they not reach these values in practice, but they serve to compare the performance of hard disks. These speeds are heavily influenced by the mechanical part of the disk, so that SSD’s, that do not include this part, is usually higher.

Hard disks are filled from the outside to the center, so that the outside, where the internal speed is maximum, will be the first fill area. In other words, a hard disk becomes slower as it fills.

3.5.4 External transfer speed

It is the most knowing value, since it is much higher than the internal speed. This is the rate at which the output interface transmits data between the hard disk and the rest of the PC. In a SATA disk can be up to 600MB/s or a SCSI 320MB/s.

It is not the most important parameter because with a mechanical hard disk is almost impossible to saturate (with an SSD might be). To be important in computer performance (and still little), you have to have a good buffer and good software that optimizes access.

3.5.5 SATA Connector

Transmitting simultaneously relatively large groups of bits, using a relatively large number of wires, it is very useful when we are working with low speeds (that is, few MHz). But at high speeds interferences in the signal begin to show, just because so many connectors in parallel. The solution is usually to minimize the number of cables, isolate and greatly increase the clock speed to compensate. This is the reason why technology moves from a PATA (Parallel ATA, older connectors) to **SATA** (Serial ATA).

3.5.6 Magnetic Hard Disks

Most used type of hard disk until recent years was the magnetic hard disk. It works taking advantage of the capacity of certain materials to permanently store a given magnetic state that is imposed from outside.

Their main advantages are:

- The storage is permanent.
- It can be altered any time.
- Its price is very low with respect their capacity (although this is changing slowly).

The disadvantages are:

- They are quite delicate. The recorded data may be affected by high and low temperatures, humidity…
- They can be affected by shock
- They can be affected magnetic fields.

3.5.6.1 How magnetic disks work

The hard disk is a magnetic and mechanical device with moving parts and therefore more sensitive than other systems such as memory. Data is stored in magnetic form on the surface of a series of disks, called platters. These platters are rigid and rotate about a common axis at high speed.

To read and write data heads are used. These heads are placed on the platters surface without touching it. The heads are attached to arms which, in turn, are attached to an own axis located next to the platters, so that the heads can slide radially on the surface of the platters.

The data is distributed over the platters in thousands of concentric circles called tracks. Each of these tracks is divided into hundreds of adjacent parts of the same capacity called sectors. This capacity is usually 512 Bytes (although nowadays, in modern and high capacity disks, this value is

4096 Bytes) and correspond to the smallest unit of information that can be read and recorded on the disk.

Finally there are the cylinders. Since the arms of the heads move in unison, on disks with various platters, all heads always be located on the same track number in each. The cylinder is the set of tracks that can be positioned heads in a moment.

To find the information inside the hard disk, storage protocols have used the cylinder-head-sector to locate a data system for a long time. All of them, cylinders, sectors and the faces of the platters (heads) are identified by a number. Knowing these 3 numbers is essential to locate the information. Currently there are disks that can be configured in LBA mode (Logical Block Address) where all disk sectors without consecutively numbered from 0 to n-1.

3.5.6.2 Rotation speed

It is the angular velocity of the rotating platters, measured in revolutions per minute (rpm) and is a key parameter disk performance. Hard drives 3.5'' usually have 7200 rpm. Only low-end or very high capacity has speeds of 5400rpm. In smaller hard drives, the speed is reduced. Thus, in 2.5 '' it is 5400rpm and as 1.8'' 4200rpm.

3.5.6.3 Size of buffer or cache

The buffer is a memory that performs the function of data storage between a "fast" device (disk controller) and a slow (the mechanical part of a hard disk).

This buffer also serves as a disk cache. Thus, when reading data from the processor to the disk requested, may these are already located in the buffer. In this case the data will be read much faster.

The larger the buffer, the more likely that a data was saved on it. Today the tendency is to have 16MB, although there are still 8MB disk and it can be found easily with some 32MB and 64MB.

3.5.6.4 SMART technology

Nowadays, a lot of hard disks support SMART technology. This is technology of self-monitoring, analysis and reporting, through which they can come to alert the user to foreseeable problems.

A disk with SMART measures thousands of variables (flight altitude, position of head temperatures, etc.) and compares these values with a number of nominal values and, if the trend seems to lead to a failure, it alerts the user.

3.5.7 SSD Hard Disks

3.5.7.1 Features of SSDs

1. Flash Storage Technology: SSDs use flash memory to store data permanently. This means they have no moving parts, making them more durable and less prone to mechanical failures compared to traditional hard drives.

2. SATA and PCIe Interfaces: SSDs can be connected through a SATA (Serial Advanced Technology Attachment) or PCIe (Peripheral Component Interconnect Express) interface. SATA SSDs are more common and offer significantly better performance than conventional hard drives. PCIe SSDs, especially those using the NVMe protocol, offer even higher performance.

3. Variety of Form Factors: SSDs are available in a variety of forms, including 2.5-inch models that resemble traditional hard drives, as well as M.2 and PCIe formats that are more compact.

3.5.7.2 Advantages of SSDs

1. High Performance: SSDs offer much higher read and write speeds than traditional hard drives. Newer models, especially those using the PCIe interface and NVMe protocol, can achieve exceptionally fast transfer speeds.

2. Faster Boot and Application Load Times: Due to their high read speeds, operating systems and applications start up faster, and loading times are significantly reduced.

3. Lower Power Consumption: SSDs consume less power than traditional hard drives, which can increase battery life in laptops and reduce electricity costs in desktop systems.

4. Greater Durability and Reliability: Since they have no moving parts, SSDs are more resistant to shocks and vibrations, and they have a longer lifespan compared to hard drives.

3.5.7.3 Disadvantages of SSDs

1. Cost: While prices have come down in recent years, SSDs are still more expensive per gigabyte than traditional hard drives. However, their superior performance and durability can justify this additional cost.

2. Limited Capacity: High-capacity SSDs can be costly, which can be a significant consideration if you need large amounts of storage.

3. Wear and Degradation Over Time: While SSDs are very durable, they have a limit on write cycles and can experience performance degradation over time, especially if used for intensive write tasks.

In summary, SSDs offer significantly improved performance compared to traditional hard drives, making them an excellent choice for enhancing the speed and efficiency of any computing system. Despite their higher cost per gigabyte, their durability and speed make them a valuable investment for many users.

3.5.8 M.2 Drives

3.5.8.1 Features of M.2 Drives

1. Compact Form Factor: M.2 drives are very small and thin, making them ideal for systems with limited space for storage components.

2. M.2 Interface: These drives use an M.2 interface to connect directly to the computer's motherboard. The M.2 interface can support various communication protocols, including SATA and PCIe.

3. High Data Transfer Speeds: M.2 drives can achieve very high read and write speeds, especially when connected through the PCIe interface. This means they can load and transfer data much faster than traditional hard drives or even some older SSDs.

4. Diversity of Communication Protocols: M.2 drives can use various communication protocols, such as SATA, PCIe, and NVMe. NVMe (Non-Volatile Memory Express) is a communication protocol specially designed to make the most of SSD performance.

3.5.8.2 Advantages of M.2 Drives

1. High Performance: Due to their ability to leverage the PCIe interface and NVMe protocol, M.2 drives are much faster than conventional SATA SSDs.

2. Space Saving: Being small and able to connect directly to the motherboard, they free up space inside the computer case.

3. Lower Power Consumption: Overall, M.2 drives tend to consume less power than traditional hard drives and some SSDs, which can improve system energy efficiency.

3.5.8.3 Disadvantages of M.2 Drives

1. Potential Overheating Issues: Due to their small size and lack of space for heat sinks, some M.2 drives may experience overheating issues under intensive workloads. However, this can depend on the specific design of the drive and your system's configuration.

2. Compatibility and Limited Slots: Not all motherboards have M.2 slots, and those that do may not be compatible with all types of M.2 (for example, some only support M.2 SATA, while others support PCIe and NVMe).

3. Cost: While prices have come down over time, M.2 drives tend to be more expensive than traditional SATA SSDs. However, their superior performance can justify this additional cost for those seeking high performance.

3.6 Graphic Cards

In most computers, we represent images. The component which is responsible for interpreting the graphic data that come to the microprocessor and transform them into a signal that can understand the monitor is what is usually called graphic card or graphic adapter.

Today we can find this adapter:

- Integrated into the chipset: frequent office PCs, small and portable PC. In general, any system that does not require a graphical power.
- Integrated into the CPU: some processors have included the adapter. This provides advantages such as lower consumption, lower latency and lower price, but without a great 3D performance.
- Integrated motherboard. Not to be confused with the integrated into the chipset. This is a separate chip. They have the disadvantage that power is often limited. Usually a widely it used on computers that require little graphics requirements such as servers.
- Expansion cards: the usual graphics cards. the reason for its use is to connect more monitors or to increase performance.

> Current graphics cards have a higher complexity of the rest of the PC. For example, a Core i5 with 4 cores has 750 million transistors, but a graphics chip can have between 1500 and 2000 million transistors and all of them in a dedicated way.

3.6.1 GPU

GPU (Graphics Processing Unit) is a processor such as CPU, but specifically dedicated to graphics processing; its task is to reduce the workload of the central processor, being optimized for floating point computing, which is the type of data that is handled more 3D operations.

GPUs are able to perform certain operations denominated graphics primitives, as the drawing of certain basic shapes: rectangles, triangles... as well as the pre-processing and post-processing for high quality images like antialiasing, accelerating the screen representation of the images much more than if these were made via software.

That's because their architecture and, specially, their instruction set are designed taking that in mind. For instance, pixel state or vertex calculations are usually included in their instruction set.

3.6.2 Video Memory

Video memory is memory dedicated solely to the use of the graphics card, the place where information of the calculations performed by the graphics chip is stored. Formerly the only function was to serve video memory frame buffer (storage space for the image before moving on to the monitor).

Today in graphics cards with 3D functions, things change. Memory is used to store data from multiple additional functions, such as the depth of the image points (ZBuffer), textures, images to apply antialiasing...

It's a special type of memory called VRAM (video RAM) and they also have another type of DDR, which is called GDDR (graphics DDR). Nowadays (2023) the most normal is found between 2GB and 16GB of GDDR5/6 memory.

> If the dedicated memory on the graphics card it is over, RAM memory is used as video memory, making the 3D performance much lower.

3.7 Screens

Screens are the deveice where we show visually the information the computer (in any of its shapes) returns. There are many type of screens, based on the device it's being used by (computer, mobile...), its technology (LCD, plasma…) but all of them share the next features that define them.

3.7.1.1 Size

Usually measured in inches, it refers to the diagonal of the screen. Take in mind that, for curve screens, the size will include the permiter of the curve where the diagonal transits.

3.7.1.2 Resolution

The screen resolution is the number of points that form the image that will be displayed in the monitor expressed as "horizontal dots x vertical dots" for example 1280x1024. The greater the resolution will be the detail that will be on display. However, the chosen resolution should be consistent with the size and type of the screen.

3.7.1.3 Aspect ratio of current monitors

The aspect ratio of a monitor refers to the proportion between its width and height. It defines the physical shape of the screen and can affect how content appears on it. Here's an explanation of the most common aspect ratios for current monitors:

1. Aspect Ratio 4:3: This was one of the most common aspect ratios in the early days of computer monitors. In a 4:3 screen, the width is 4 units and the height is 3 units. This means the screen is relatively square. While these monitors are less common today, they can still be found in some older devices.

2. Aspect Ratio 16:9: This is the most common aspect ratio in modern monitors. It's used in the majority of televisions and computer monitors. Here, the width is 16 units and the height is 9 units. This produces a wider screen compared to its height, which is ideal for viewing high-definition and widescreen content.

3. Aspect Ratio 16:10: This aspect ratio was popular in computer monitors some years ago but has declined in popularity in favor of 16:9. In a 16:10 screen, the width is 16 units and the height is 10 units. This provides a bit more height compared to a 16:9, which can be beneficial for tasks that require more vertical space, such as programming or document editing.

4. Ultra-Wide Aspect Ratio (21:9): This aspect ratio is significantly wider than the standard 16:9. With a 21:9 ratio, the screen is great for multitasking and for viewing cinematic content. Ultra-wide monitors are popular among gamers and professionals who require ample screen space for multiple windows or applications.

5. Ultra-Extended Aspect Ratio (32:9, 32:10): These are extremely wide aspect ratios found in super ultra-wide monitors. With a 32:9 or 32:10 ratio, these monitors offer a highly immersive viewing experience and are ideal for gaming and professional tasks that require extensive screen real estate.

Remember, the aspect ratio is just one of the specifications to consider when choosing a monitor. Other factors like resolution, refresh rate, and panel type are also important depending on your specific needs.

3.7.1.4 Pixel density

Pixel density refers the number of pixels per inch. Although for computer screens this feature is important, it’s extremely important in mobile devices.