Theme 2. Computer architecture

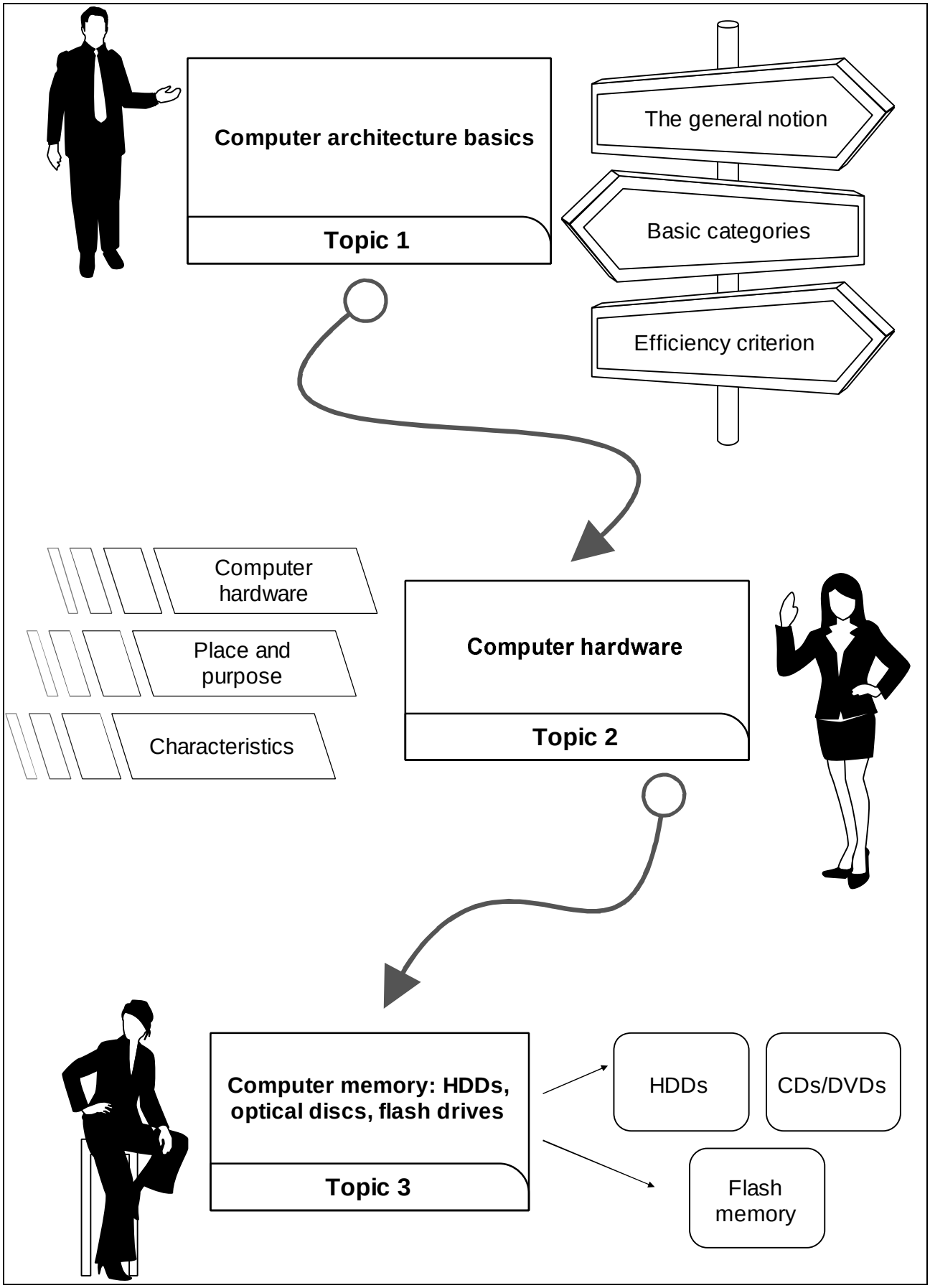


Fig. 2.1. **System elements of the theme**

g Introduction

The purpose of this theme is to give a clear understanding of the structure of computer architecture as a field of scientific research. A thorough consideration is given to the analysis of factors that determine the efficiency of computer architecture operation. A close attention is given to the study of the constituent parts of computer hardware.

Clue notions of the theme include: computer architecture; computer hardware; latency; throughput; clock speed; on-board devices; peripheral devices; computer memory.

This theme covers the following **topics**: What is computer architecture? What are the categories that are included in the general structure of computer architecture? What are the main criteria for assessing computer architecture efficiency? What devices are included in computer hardware?

As a result of the study of the material represented in this theme students will acquire the following **competencies**:

they would know the notion and structure of computer architecture;

they would learn retrospective peculiarities in the development process of computer architecture;

they would become aware of the basic components hat form computer hardware and understand the place and interconnection of each part inside the computer chassis.

2.1. Computer architecture basics

& ***Computer architecture*** *is the conceptual design and fundamental operational structure of a computer system. It is a blueprint and functional description of requirements (especially speeds and interconnections) and design implementations for the various parts of a computer — focusing largely on the way by which the central processing unit (CPU) performs internally and accesses addresses in memory.*

The term also covers the design of system software, such as the operating system (the program that controls the computer), as well as referring to the combination of hardware and basic software that links the machines on a computer network**. Computer architecture** refers to an entire structure and to the details needed to make it functional.

Thus, computer architecture covers computer systems,

microprocessors, circuits, and system programs. Typically the term does not refer to application programs, such as spreadsheets or word processing, which are required to perform a task but not to make the system run.

Computer architecture may also be defined as the science and art of selecting and interconnecting hardware components to create computers that meet functional, performance and cost goals.

Computer architecture comprises at least **three main subcategories** (fig. 2.2):

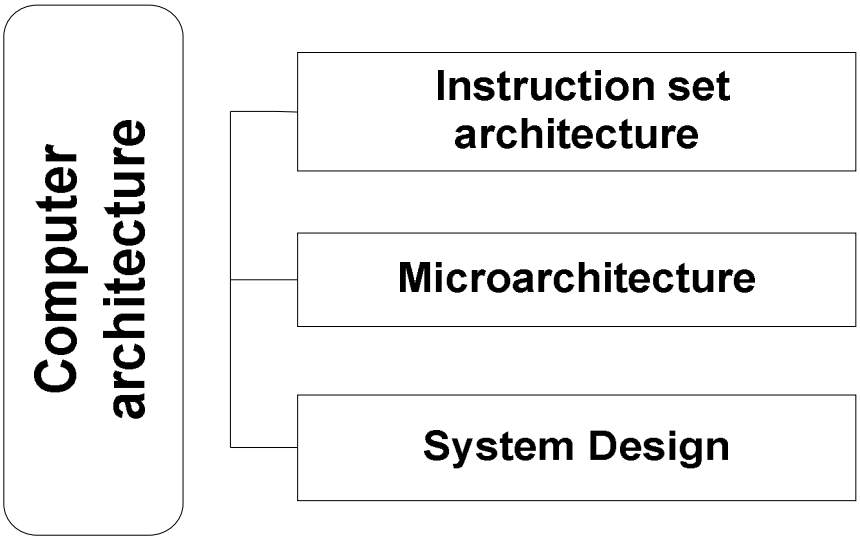


Fig. 2.2. **The structure of computer architecture**

Instruction set architecture, or ISA, is the abstract image of a computing system that is seen by a machine language (or assembly language) programmer, including the instruction set, memory address modes, processor registers, and address and data formats.

Microarchitecture, also known as Computer organization is a lower level, more concrete and detailed, description of the system that involves how the constituent parts of the system are interconnected and how they interoperate in order to implement the ISA. The size of a computer's cache for instance, is an organizational issue that generally has nothing to do with the ISA.

System Design which includes all of the other hardware components within a computing system such as:

1) system interconnects such as computer buses and switches;
2) memory controllers and hierarchies;
3) CPU off-load mechanisms such as direct memory access;
4) issues like multi-processing.

Once both ISA and microarchitecture have been specified, the actual device needs to be designed into **hardware**. This design process is called **implementation**. Implementation is usually not considered architectural definition, but rather hardware design engineering.

Implementation can be further broken down into three (not fully distinct) pieces:

Logic Implementation – design of blocks defined in the microarchitecture at (primarily) the register-transfer and gate levels.

Circuit Implementation – transistor-level design of basic elements (gates, multiplexers, latches etc) as well as of some larger blocks (ALUs, caches etc) that may be implemented at this level, or even (partly) at the physical level, for performance reasons.

Physical Implementation – physical circuits are drawn out, the different circuit components are placed in a chip floor-plan or on a board and the wires connecting them are routed.

The exact form of a computer system depends on the constraints and goals for which it was optimized. Computer architectures usually trade off standards, cost, memory capacity, latency and throughput. Sometimes other considerations, such as features, size, weight, reliability, expandability and power consumption are factors as well.

The most common scheme carefully chooses the bottleneck that most reduces the computer's speed. Ideally, the cost is allocated proportionally to assure that the data rate is nearly the same for all parts of the computer, with the most costly part being the slowest. This is how skillful commercial integrators optimize personal computers.

Main criteria for computer architecture efficiency.

1. Performance.

Computer performance is often described in terms of **clock speed** (usually in MHz or GHz). This refers to the cycles per second of the main clock of the CPU. However, this metric is somewhat misleading, as a machine with a higher clock rate may not necessarily have higher performance. As a result manufacturers have moved away from clock speed as a measure of performance.

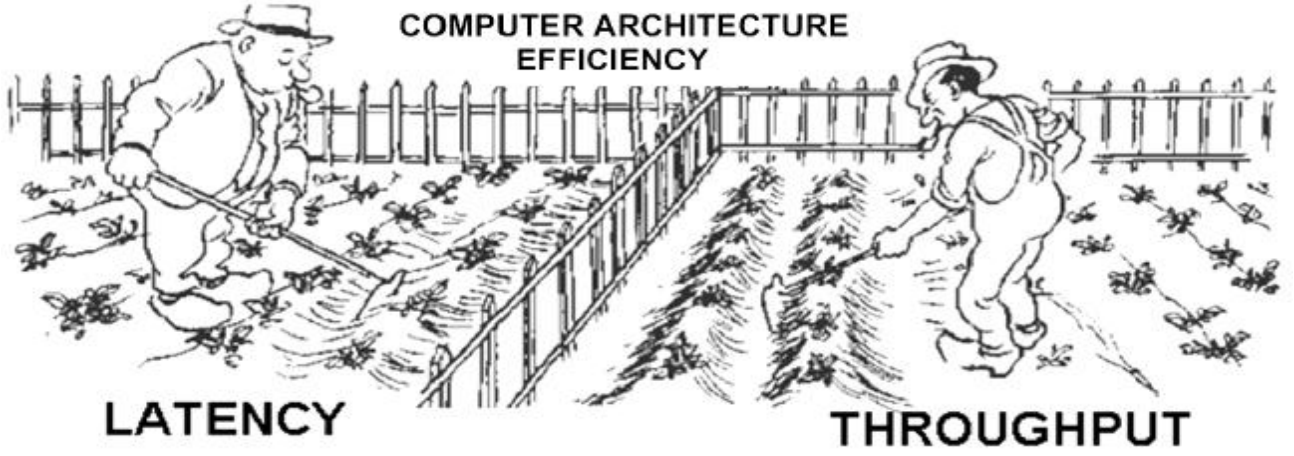
Computer performance can also be measured with the amount of cache a processor has. If the speed, MHz or GHz, were to be a car then the cache is like a traffic light. No matter how fast the car goes, it still will be stopped by a red traffic light. The higher the speed, and the greater the cache, the faster

a processor runs. Modern CPUs can execute multiple instructions per clock cycle, which dramatically speeds up a program. Other factors influence speed, such as the mix of functional units, bus speeds, available memory, and the type and order of instructions in the programs being run.

There are two main types of speed, latency and throughput.

& ***Latency*** *is the time between the start of a process and its completion.*

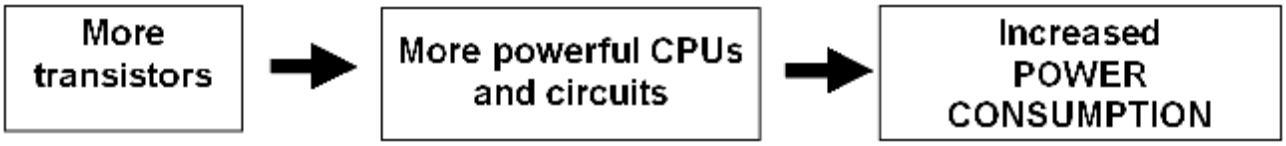
& ***Throughput*** *is the amount of work done per unit time.*



Computer performance development is believed to submit to Moore's Law. In 1965 semiconductor pioneer Gordon Moore predicted that the number of transistors contained on a computer chip would double every year. This is now known as Moore's Law, and it has proven to be somewhat accurate. The number of transistors and the computational speed of microprocessors currently doubles approximately every 18 months. Components continue to shrink in size and are becoming faster, cheaper, and more versatile.

2. Power consumption

Power consumption is another design criterion that factors in the design of modern computers. Power efficiency can often be traded for performance or cost benefits. With the increasing power density of modern circuits as the number of transistors per chip scales, power efficiency has increased in importance.



Recent processor designs such as the Intel Core 2 put more emphasis on increasing power efficiency. Also, in the world of embedded computing, power efficiency has long been and remains the primary design goal next to performance.

powerful hardware leads to encreased power consumption

The term **“architecture”** in computer literature can be traced to the work of Lyle R. Johnson and Frederick P. Brooks, Jr., members in 1959 of the Machine Organization department in IBM’s main research center.

Johnson had occasion to write a proprietary research communication about Stretch, an IBM-developed supercomputer for Los Alamos Scientific Laboratory; in attempting to characterize his chosen level of detail for discussing the luxuriously embellished computer, he noted that his description of formats, instruction types, hardware parameters, and speed enhancements aimed at the level of “system architecture” – a term that seemed more useful than “machine organization.”

Subsequently Brooks, one of the Stretch designers, started Chapter 2 of a book (Planning a Computer System: Project Stretch, ed. W. Buchholz, 1962) by writing, “Computer architecture, like other architecture, is the art of determining the needs of the user of a structure and then designing to meet those needs as effectively as possible within economic and technological constraints.” Brooks went on to play a major role in the development of the IBM System/360 line of computers, where “architecture” gained currency as a noun with the definition “what the user needs to know.” Later the computer world would employ the term in many less-explicit ways.

The first mention of the term **architecture** in the referred computer literature is in a 1964 article describing the IBM System/360. The article defines architecture as the set of “attributes of a system as seen by the

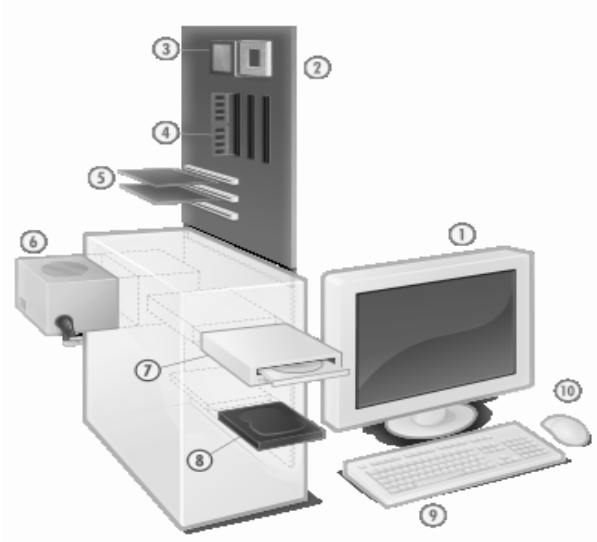
programmer, i.e., the conceptual structure and functional behavior, as distinct from the organization of the data flow and controls, the logical design, and the physical implementation.” In the definition, the programmer perspective of the computer’s functional behavior is key. The conceptual structure part of an architecture description makes the functional behavior comprehensible, and extrapolatable to a range of use cases. Only later on did ‘internals’ such as “the way by which the CPU performs internally and accesses addresses in memory,” mentioned above, slip into the definition of computer architecture.

2.2. Computer hardware

Once both ISA (instruction set architecture) and microarchitecture has been specified, the actual device needs to be designed into **hardware**. A personal computer is made up of computer hardware, multiple physical components onto which can be loaded into a multitude of software that perform the functions of the computer.

& ***Computer hardware*** *is the equipment involved in the function of a computer. Computer hardware consists of the components that can be physically handled.*

The function of these components is typically divided into three main categories: input, output, and storage. Though a PC comes in many different form factors, a typical personal computer consists of a **case or chassis** in a tower shape (desktop) and the following parts (fig. 2.3):



1. Monitor. 2. Motherboard. 3. CPU. 4. RAM Memory. 5. Expansion card. 6. Power Supply. 7. CD-rom. 8. Hard Disk. 9. Keyboard. 10. Mouse

Fig. 2.3. **Hardware of a personal computer [10]**

Typical PC hardware consists of the following elements (fig. 2.4):

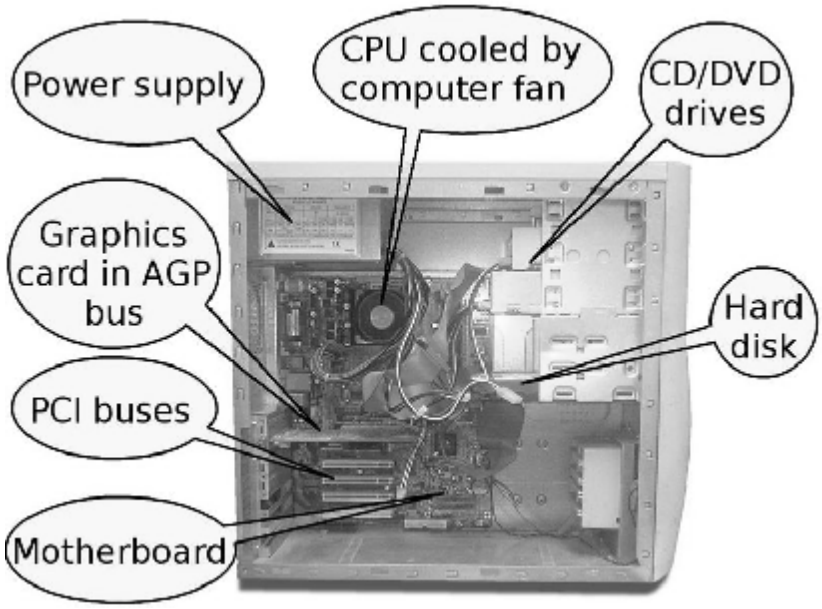


Fig. 2.4. **Typical PC hardware elements**

Motherboard – the motherboard is sometimes alternatively known as the mainboard, system board, or, on Apple computers, the logic board. It is also sometimes casually shortened to mobo. It is the "body" or mainframe of the computer, through which all other components interface (fig. 2.5).

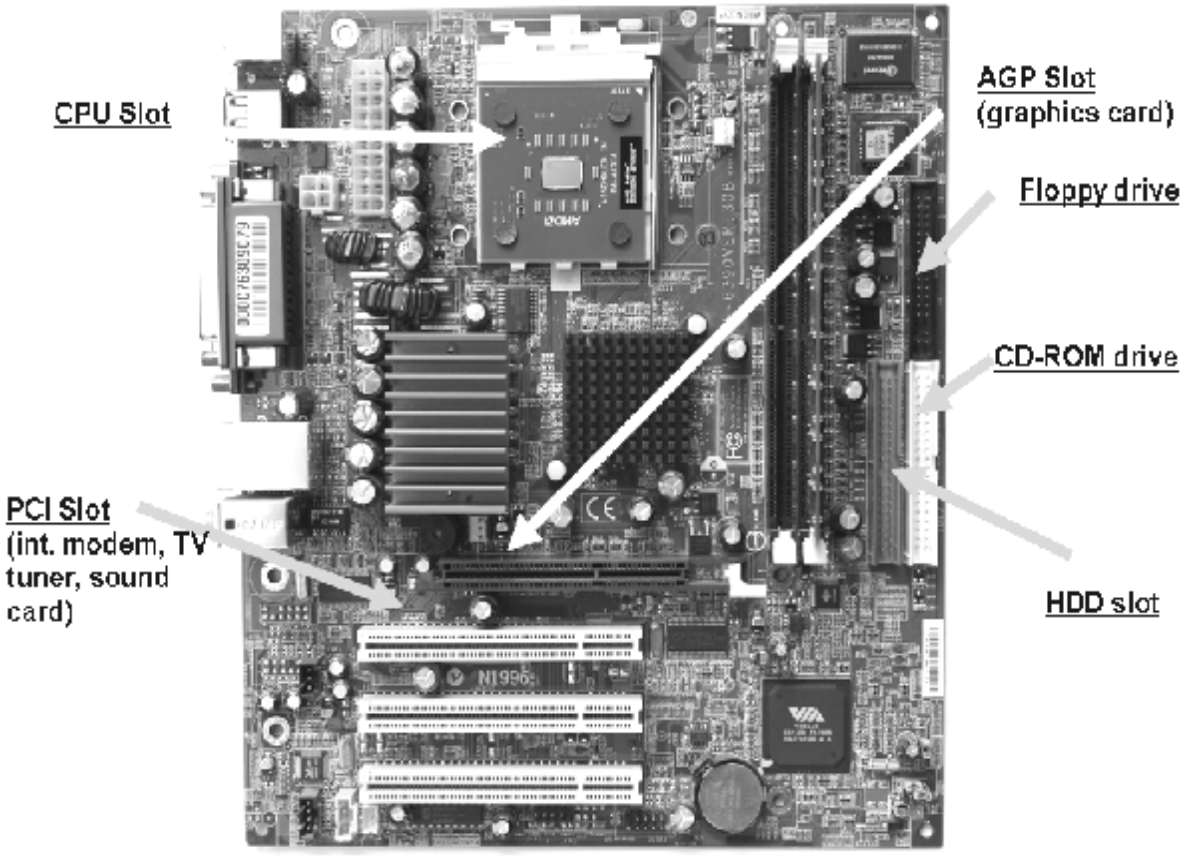


Fig. 2.5. **Computer motherboard**

Central processing unit (CPU) – performs most of the calculations which enable a computer to function, sometimes referred to as the "backbone or brain" of the computer.

Computer fan – is used to lower the temperature of the computer. A fan is almost always attached to the CPU, and the computer case will generally have several fans to maintain a constant airflow. Liquid cooling can also be used to cool a computer, though it focuses more on individual parts rather than the overall temperature inside the chassis.

Random Access Memory (RAM) – is a form of computer data storage. Today it takes the form of integrated circuits that allows the stored data to be accessed in any order (i.e., at random). The word random thus refers to the fact that any piece of data can be returned in a constant time, regardless of its physical location and whether or not it is related to the previous piece of data. It is the physical memory of the computer in the form of DRAM memory modules. RAM is a fast-access memory that is cleared when the computer is powered-down. RAM attaches directly to the motherboard, and is used to store programs that are currently running.

Internal storage (HDD – hard disk drive) – hardware that keeps data inside the computer for later use and remains persistent even when the computer has no power.

Video display controller – produces the output for the visual display unit. This will either be built into the motherboard or attached in its own separate slot (PCI or AGP), in the form of a graphics card. Graphics card is an expansion card whose function is to generate and output images to a display. Some video cards offer added functions, such as video capture, TV tuner adapter, MPEG-2 and MPEG-4 decoding, FireWire, light pen, TV output, or the ability to connect multiple monitors.

Power supply – is the component that supplies power to a computer. It usually includes a case control, and (usually) a cooling fan. Most common types of power supplies are AT and ATX (Advanced Technology Extended). On newer ATX power supplies, the switch goes to the motherboard, allowing other hardware or software to turn the system on or off.

Buses – in computer architecture, a bus is a subsystem that transfers data between computer components inside a computer or between computers. Unlike a point-to-point connection, a bus can logically connect several peripherals over the same set of wires. Each bus defines its set of connectors to physically plug devices, cards or cables together.

Early computer buses were literally parallel electrical buses with multiple connections, but the term is now used for any physical arrangement that provides the same logical functionality as a parallel electrical bus. Modern computer buses can use both parallel and bit-serial connections, and can be wired in either a multidrop (electrical parallel) or daisy chain topology, or connected by switched hubs, as in the case of USB. PC buses fall into two groups (fig. 2.5):

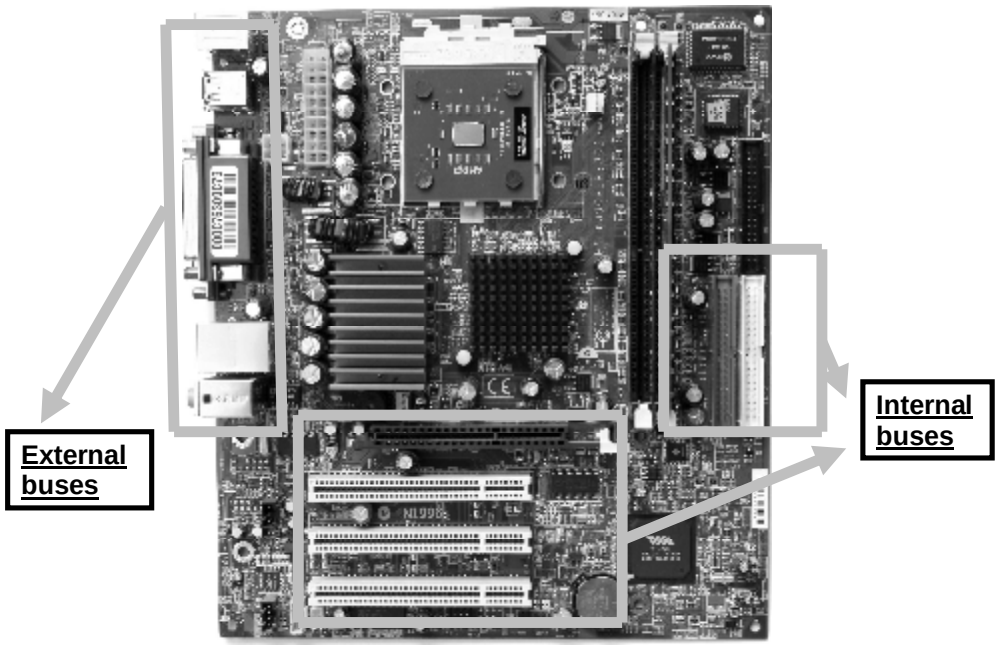


Fig. 2.6. **Position of computer buses on the motherboard**

Internal Buses – Connections to various internal components. For example, PCI, PCI-E, USB, AGP, IDE etc.

External Buses – used to connect to external peripherals, such as printers and input devices. These ports may also be based upon expansion cards, attached to the internal buses.

Sound card – enables the computer to output sound to audio devices, as well as accept input from a microphone. Most modern computers have sound cards built-in to the motherboard, though it is common for a user to install a separate sound card as an upgrade.

Removable media devices include:

CD, CD-ROM Drive, DVD (digital versatile disc), DVD-ROM Drive, Blu-ray CD, HD DVD CD – optical storages, which store information in deformities on the surface of a circular discs and read this information by illuminating the surface with a laser diode and observing the reflection. Optical disc storages are non-volatile. The deformities may be permanent (read only media),

formed once (write once media) or reversible (recordable or read/write media).

Floppy disk – an outdated storage device consisting of a thin disk of a flexible magnetic storage medium.

USB flash drive – a flash memory data storage device integrated with a USB interface, typically small, lightweight, removable, and rewritable.

Tape drive – a device that reads and writes data on a magnetic tape, used for long term storage.

Other peripherals. In addition, hardware devices can include external components of a computer system. The following are either standard or very common. Peripheral devices include various input and output devices, usually external to the computer system.

Input devices:

Text input devices.

Keyboard – a device to input text and characters by depressing buttons (referred to as keys), similar to a typewriter. The most common English-language key layout is the QWERTY layout. The QWERTY design is based on a layout created by Christopher Latham Sholes in 1873 for the Sholes and Glidden typewriter and sold to Remington in the same year, when it first appeared in typewriters (fig. 2.7).

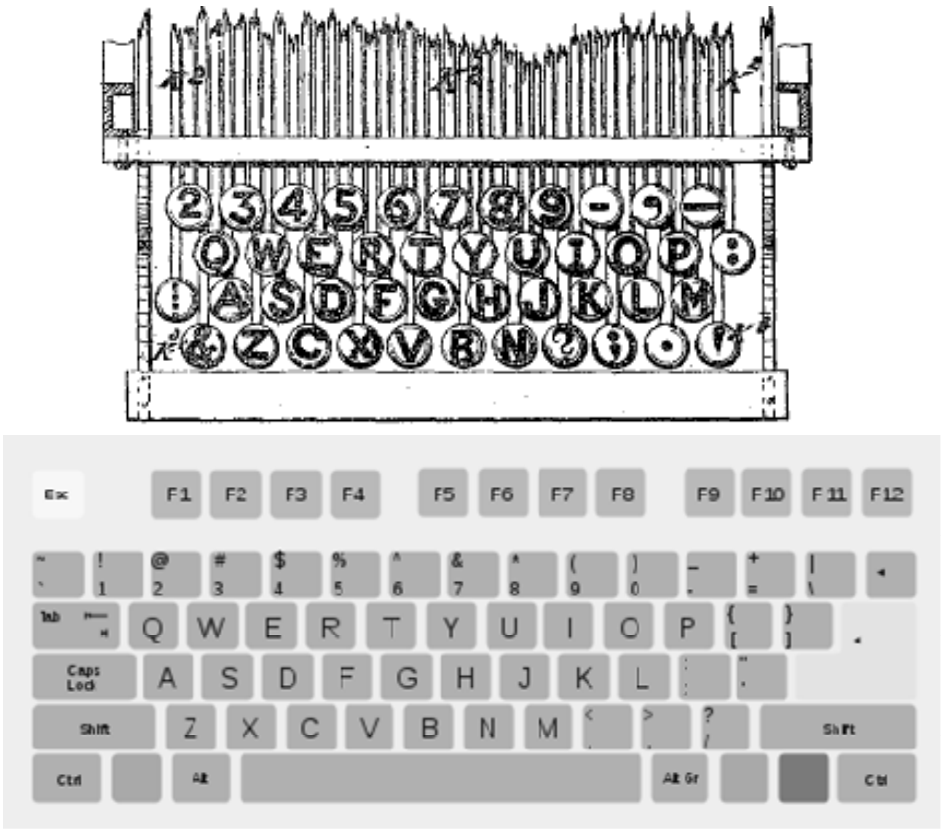


Fig. 2.7. **Typewriter and PC QWERTY layouts [11]**

Pointing devices.

Mouse – a pointing device that detects two dimensional motion relative to its supporting surface.

Trackball – a pointing device consisting of an exposed protruding ball housed in a socket that detects rotation about two axes (fig. 2.8).

Fig. 2.8. **A trackball**

Fig. 2.9. **A gamepad**

Gaming devices.

Joystick – a general control device that consists of a handheld stick that pivots around one end, to detect angles in two or three dimensions.

Gamepad – a general handheld game controller that relies on the digits (especially thumbs) to provide input (fig. 2.9).

Game controller – a specific type of controller specialized for certain gaming purposes.

Image, Video input devices.

Image scanner – a device that provides input by analyzing images, printed text, handwriting, or an object.

Webcam – a low resolution video camera used to provide visual input that can be easily transferred over the internet.

Audio input devices.

Microphone – an acoustic sensor that provides input by converting sound into electrical signals.

Output devices:

Image, Video output devices.

Printer – a peripheral which produces a hard copy (that is a permanent human-readable text and/or graphics) of documents stored in electronic form, usually on physical print media such as paper or transparencies

Monitor – a piece of electrical equipment which displays images generated from the video output of devices such as a PC graphics card, without producing a permanent record.

Audio output devices.

Speakers – are external sound producing devices, commonly equipped with a low-power internal amplifier. The standard audio connection is a 3.5mm (1/8 inch) stereo jack plug often colour-coded lime green (following the PC 99 standard) for computer sound cards.

Headset – a pair of small loudspeakers, or less commonly a single speaker, with a way of holding them close to a user's ears and a means of connecting them to a signal source such as an audio amplifier, radio or CD player. They are also known as earphones, earbuds, stereophones, headsets or, informally cans.

2.3. Computer memory: HDDs, optical discs, flash drives

There are two types of memory which are at a disposal of a personal computer: random access memory (RAM) and permanent memory (hard disc drives). Random access memory was given a close attention in the previous section, so now we'll focus on PC's permanent memory that is represented by hard disc drives (HDD).

& ***A hard disk drive (HDD),*** *commonly referred to as a hard drive, hard disk, or fixed disk drive, is a non-volatile storage device which stores digitally encoded data on rapidly rotating platters with magnetic surfaces.*

Strictly speaking, "drive" refers to a device distinct from its medium, such as a tape drive and its tape, or a floppy disk drive and its floppy disk. Early HDDs had removable media; however, an HDD today is typically a sealed unit (except for a filtered vent hole to equalize air pressure) with fixed media.

HDDs record data by **magnetizing ferromagnetic material** directionally, to represent either a 0 or a 1 binary digit. They read the data back by detecting the magnetization of the material. A typical HDD design consists of a spindle which holds one or more flat circular disks called platters, onto which the data are recorded (fig. 2.10). The platters are made from a non-magnetic material, usually aluminum alloy or glass, and are coated with a thin layer of magnetic material. Older disks used iron oxide as the magnetic material, but current disks use a cobalt-based alloy.

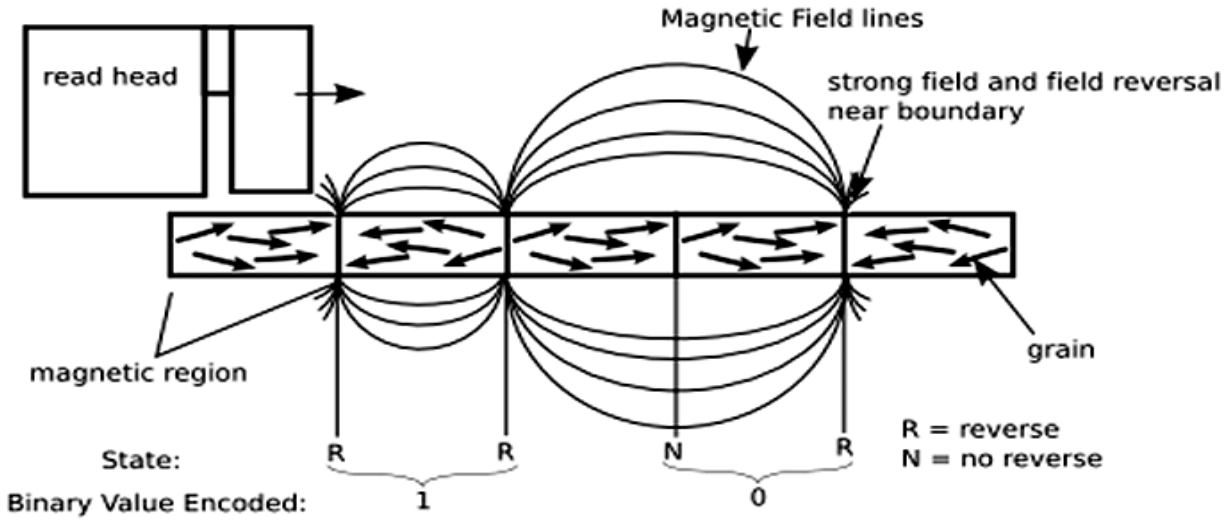


Fig. 2.10. **The structure of HDD surface [15]**

Initially the regions were oriented horizontally, but beginning about 2005, the orientation was changed to perpendicular (2.11).

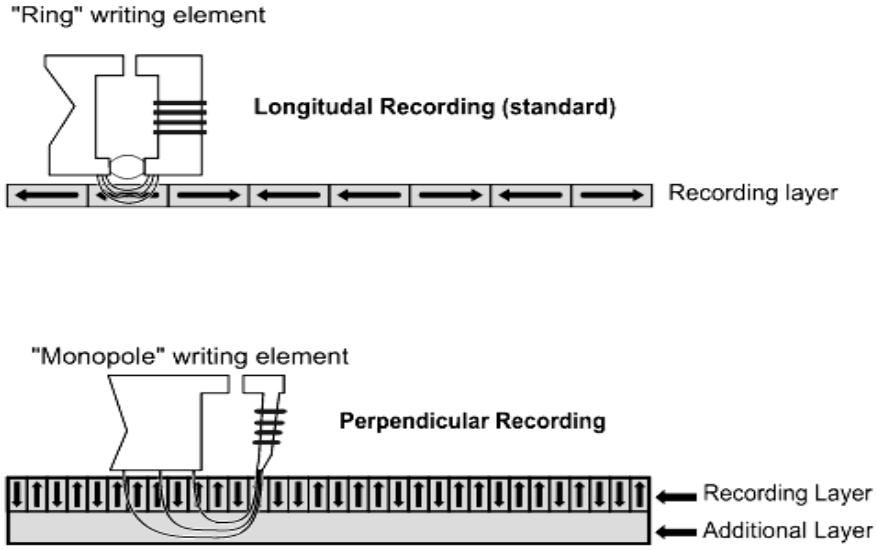


Fig. 2.11. **The direction of recording layers [15]**

The platters are spun at very high speeds. Information is written to a platter as it rotates past devices called read-and-write heads that operate very close (tens of nanometers in new drives) over the magnetic surface. The read-and-write head is used to detect and modify the magnetization of the material immediately under it. There is one head for each magnetic platter surface on the spindle, mounted on a common arm. An actuator arm (or access arm) moves the heads on an arc (roughly radially) across the platters as they spin, allowing each head to access almost the entire surface of the platter as it spins. The arm is moved using a voice coil actuator or (in older designs) a stepper motor.

HD heads are kept from contacting the platter surface by the air that is extremely close to the platter; that air moves at, or close to, the platter speed. The record and playback head are mounted on a block called a slider, and the surface next to the platter is shaped to keep it just barely out of contact. It's a type of air bearing (fig. 2.12).

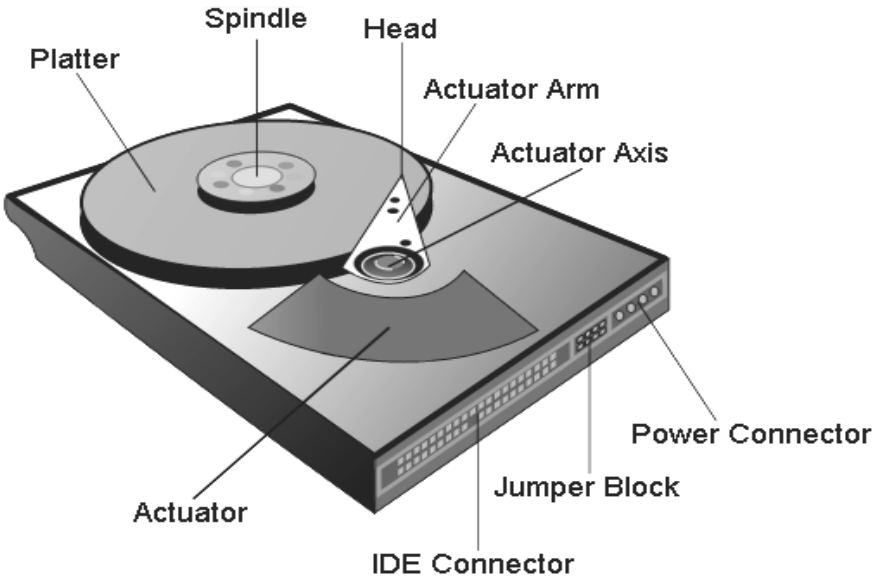


Fig. 2.12. **Computer hard drive's structural parts [15]**

The magnetic surface of each platter is conceptually divided into many small sub-micrometre-sized magnetic regions, each of which is used to encode a single binary unit of information. In today's HDDs, each of these magnetic regions is composed of a few hundred magnetic grains. Each magnetic region forms a magnetic dipole which generates a highly localized magnetic field nearby. The write head magnetizes a region by generating a strong local magnetic field.

& ***Flash memory*** *is a non-volatile computer storage chip that can be electrically erased and reprogrammed.*

Since flash memory is non-volatile, no power is needed to maintain the information stored in the chip. It is primarily used in memory cards, USB flash drives, MP3 players and solid-state drives for general storage and transfer of data between computers and other digital products.

Flash memory was invented by Dr. Fujio Masuoka while working for Toshiba circa 1980. Dr. Masuoka presented the invention at the IEEE 1984 International Electron Devices Meeting (IEDM) held in San Francisco, California. According to Toshiba, the name "flash" was suggested by Dr. Ma-

suoka's colleague, Mr. Shoji Ariizumi, because the erasure process of the memory contents reminded him of the flash of a camera.

First flash memory devices worked only with large block sizes. *That meant that in order to erase a piece of data the whole data had to be erased***.** Modern flash memory is erasable and rewritable in small blocks, typically bytes.

Singaporean Trek Technology and IBM began selling the first USB flash drives commercially in 2000. The model under the brand name "ThumbDrive", and "DiskOnKey" became available on December 15, 2000, and had a storage capacity of 8 MB, more than five times the capacity of the then-common floppy disks (fig. 2.13).

Fig. 2.13. **First USB flash drive [18]**

Storage capacities in 2011 can be as large as 512 GB with steady improvements in size and price per capacity expected. Some allow 1 million write or erase cycles and have a 10-year data retention cycle.

Most commercially available flash products are guaranteed to withstand around 100,000 program-erase (P/E) cycles, before the wear begins to deteriorate the integrity of the storage. Flash drives can be defragmented, but this brings little advantage as there is no mechanical head that moves from fragment to fragment. *Defragmenting shortens the life of the drive by making many unnecessary writes.* There also exist viruses that “kill” or shorten the operation period of a flash drive by making a lot of unnecessary read-write operations. Flash memory stores information in an array of memory cells made from floating-gate transistors. In traditional devices, each cell stores only one bit of information (fig.2.14).

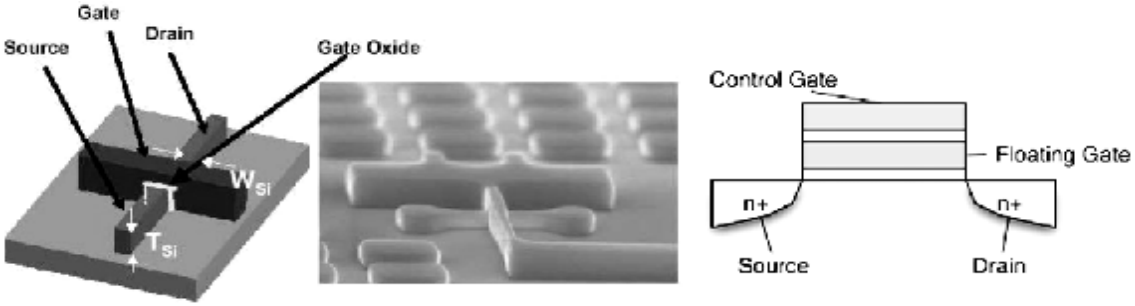


Fig. 2.14. **Flash memory structure**

A flash cell can be programmed, or set to a binary "0" value, by the following procedure:

an elevated on-voltage (typically >5 V) is applied to the CG;

the channel is now turned on, so electrons can flow from the source to the drain;

the source-drain current is sufficiently high to cause some high energy electrons to jump through the insulating layer onto the FG, via a process called hot-electron injection (fig. 2.15).

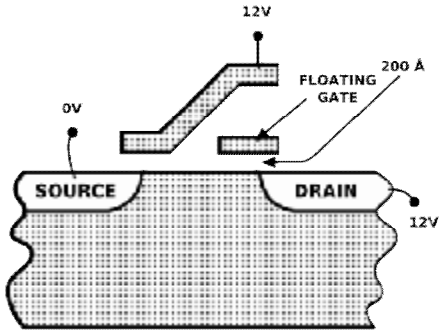


Fig. 2.15. **Flash memory operation principles**

In order to “erase” a flash cell a large voltage *of the opposite polarity* is applied between the CG and source, pulling the electrons off the FG.

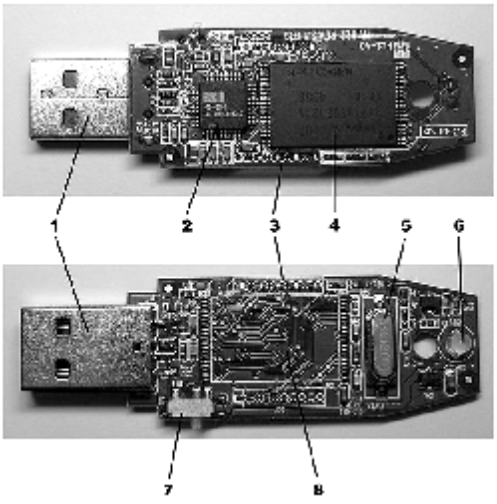
There are typically four parts to a flash drive:

USB connector – provides a physical interface to the host computer;

USB mass storage controller – implements the USB host controller. The controller contains a small microcontroller with a small amount of on-chip ROM and RAM;

flash memory chip – stores data;

crystal oscillator – produces the device's main 12 MHz clock signal and controls the device's data output through a phase-locked loop (fig. 2.16).



1	USB connector
2	USB mass storage controller
3	Test points
4	Flash memory chip
5	Crystal oscillator
6	LED
7	Write-protect switch (Optional)
8	Space for second flash memory chip

Fig. 2.16. **Typical parts of a flash drive**

& **Optical read only memory** is represented by compact optical discs (also known as a CDs or DVDs) that are used to store digital data.

It was originally developed to store sound recordings exclusively, but later it also allowed the preservation of other types of data. Standard CDs have a diameter of 120 mm and can hold up to 80 minutes of uncompressed audio (700 MB of data for common CDs or 4.7 Gb for DVDs).

Sony first publicly demonstrated an optical digital audio disc in September 1976. The first test CD was presented at a conference in Hannover, Germany, by the Polydor Pressing Operations plant in 1982. The disc contained a recording of Richard Strauss's Eine Alpensinfonie (in English language, An Alpine Symphony).

The size and length of the first CD had to be enough to hold the 9th Symphony of Beethoven – the favorite composition of the head of Sony Corporation. Later it became a standard.

CD data are stored as a series of tiny indentations known as **"pits"**, encoded in a spiral track molded into the top of the polycarbonate layer. The areas between pits are known as **"lands".** Each pit is approximately 100 nm deep by 500 nm wide, and varies from 850 nm to 3.5 nm in length. The distance between the tracks, the **pitch**, is 1.6 nm (fig. 2.17).

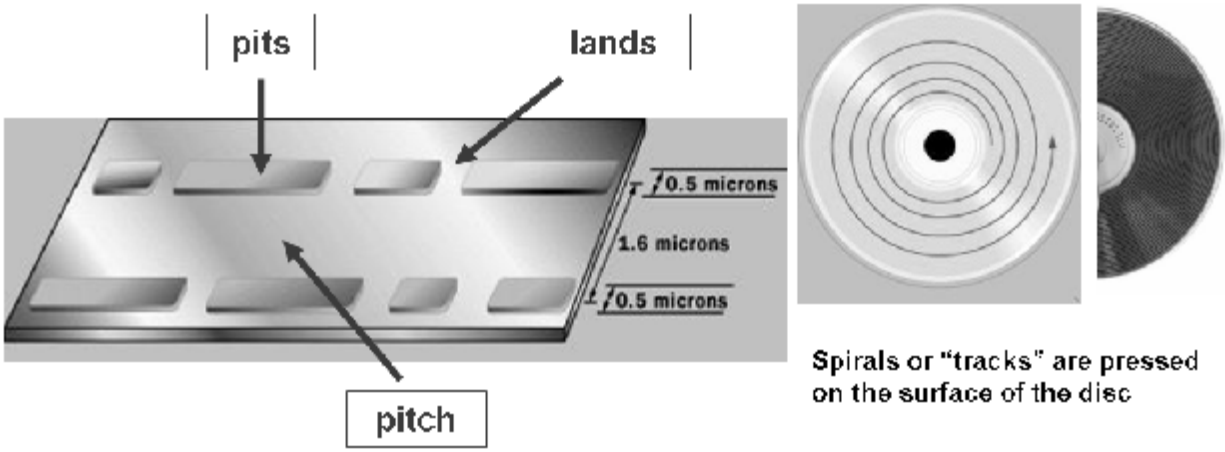


Fig. 2.17. **Optical disc structure**

A typical CD consists of a set of layers (fig.2.18):

A. A polycarbonate disc layer has the data encoded by using bumps.

B. A shiny layer reflects the laser.

C. A layer of lacquer helps keep the shiny layer shiny.

D. Artwork is screen printed on the top of the disc.

E. A laser beam reads the CD and is reflected back to a sensor, which converts it into electronic data

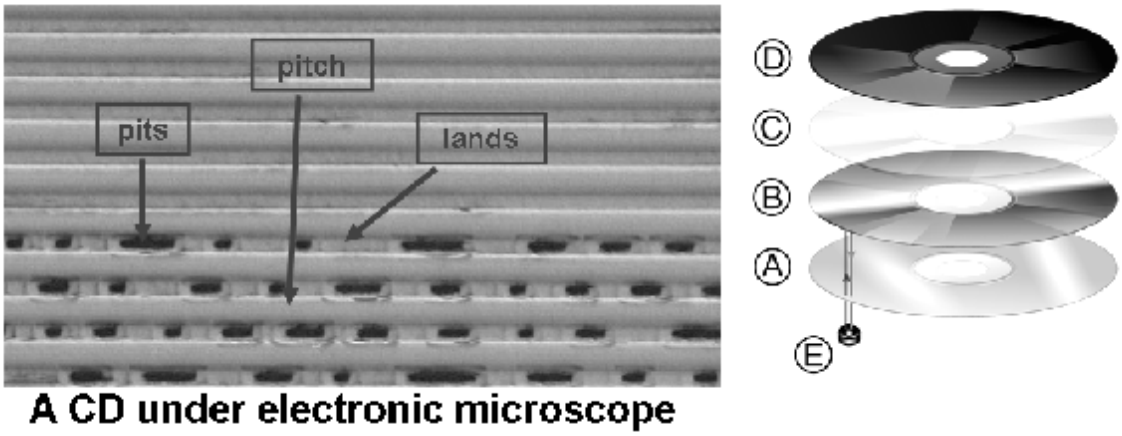


Fig. 2.18. **Optical disc layers [16]**

Optical discs can be recorded in **Disc At Once**, **Track At Once**, **Session at Once** (i.e. multiple burning sessions for one disc), or packet writing modes. Many disc manufacturers extend a recordable disc to leave a small margin of extra groove at the outer edge. This lead-out was originally intended to provide tolerance for the read head of an audio CD player should it overseek, by providing a padding of up to 90 seconds of silent digital audio. It can be used to "overburn" a disc.

& **Overburning** is the process of recording data past the normal size limit of a compact disc.

An optical disc is designed to support one of three recording types: read-only (e.g.: CD and CD-ROM), recordable (write-once, e.g. CD-R), or re-recordable (rewritable, e.g. CD-RW). Write-once optical discs commonly have an organic dye recording layer between the substrate and the reflective layer. Rewritable discs typically contain an alloy recording layer composed of a phase change material that can restore its original state.

s Questions

1. What is computer hardware? What elements does it include?
2. How does magnetic memory work?
3. What is the operational principle of flash memory devices?
4. What are the basic elements of a compact disc?
5. What is overburning?

@Tests

1. Computer architecture:	a) includes application software b) includes system software c) doesn't deal with software d) includes user software
2. Check out pieces of software that are part of computer architecture:	a) drivers b) scripts c) low level instructions d) office applications e) web browsers f) simulation games
3. The abstract image of a computer system is:	a) microarchitecture b) instruction set architecture c) system design
4. The most costly part of a computer system is usually:	a) the fastest b) the slowest c) the same speed as other parts
5. A computer system's performance is usually measured in:	a) MHz b) GHz c) clock-speed d) all points true e) no right answer available
6. The time between the start of a process and its completion is called:	a) latency b) throughput c) clock-speed d) frequency
7. The work done per unit of time is called:	a) latency b) throughput c) multitude d) frequency
8. Moore's law states that:	a) the number of transistor on a chip doubles every year b) the number of computers doubles every year c) computational abilities of computers double every 18 months d) the cost of computers reduces every 12 months e) the number of I-net users doubles every half a year