Essential Computing (BPM-119)

Computer Hardware

Memory and storage

The Memory:

Like the memory in the human brain, computer memory stores data. Also like human memory, there is both short-term (temporary) and long-term (permanent) memory storage. Some memory is dynamic, in that it stores data only until the computer is turned off. Dynamic memory (also called volatile memory) must be constantly refreshed. Static memory (also called non-volatile memory) retains whatever you put in it indefinitely.

Another way to classify memory is whether it can be overwritten with new data or not. Random Access Memory (RAM) can be rewritten freely; Read-Only Memory (ROM) cannot (at least not in the same way that RAM can; see the following Note). All ROM is static, but RAM can be either dynamic (more common) or static.

How Computers Use Memory

Computers use different types of memory in various ways. Here are some of the most common memory uses:

System memory: The main memory in a computer system, such as the RAM installed on the PC's motherboard, is dynamic RAM (DRAM). System memory is installed in slots on the motherboard, perpendicular to the motherboard itself.

Component memory: Many components have a small amount of memory built in for their own use. For example, a printer might have RAM that holds the information about the page it is printing. Component memory is typically DRAM unless the component's function requires it to be otherwise because that's the least expensive kind.

ROM-BIOS: A motherboard has an EEPROM (electrically erasable programmable read-only memory) chip that contains the low-level startup instructions for the hardware. To prevent corruption that would prevent the system from starting up, this chip is not rewritable except with a special utility program.

Caches: The caches in a CPU are a type of static RAM (SRAM).

USB flash drives, memory cards, and solid-state hard drives: These devices use a type of EEPROM to store data. This type of EEPROM is also called flash memory; its enhanced technology allows data to be written and rewritten to it multiple times.

Understanding System Memory

Each computer has a certain amount of system memory, also called RAM or main memory. It is dynamic RAM (DRAM), so it loses its contents if it is not constantly being electrically refreshed. System memory forms a workspace that the operating system uses. When we open an application, that application is placed in system memory, and when we open a data file, it is placed there too. When we make changes to a data file, those changes wait in memory until we write them to a more permanent storage area by saving your work. The more system memory we have, the more applications and files we can have open at once.

Physically, system memory comes mounted on small rectangular circuit boards called **dual inline memory module (DIMM)**. These circuit boards fit into memory slots on the motherboard. There are different types and speeds of DIMMs, so we must match the memory to the motherboard's requirements. Figure below shows two different types of DIMMs.

© iStockphoto.com/aguirre_mar, © iStockphoto.com/mkos83

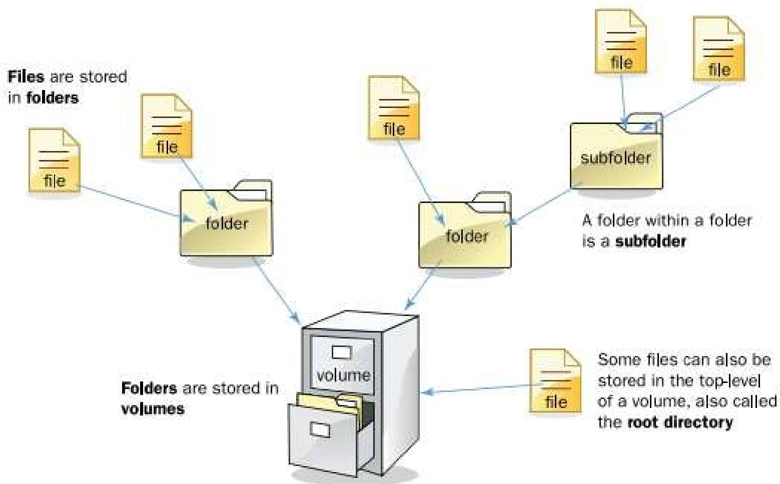
Some DIMMs have a metal plate covering the chips for better heat dissipation.

For notebook computers, small outline DIMMs (SO-DIMMs) are used instead; they do the same thing as DIMMs but are a different size and shape.

RAM is known as **primary storage** because it's where the data must be in order to interact with the CPU. For data to be safely and permanently stored, however, it must be placed in **secondary storage**. Secondary storage devices are nonvolatile, so they don't lose their contents when the computer's power is turned off. Secondary storage can include hard disk drives, solid state drives, CDs, DVDs, and network and cloud-based storage.

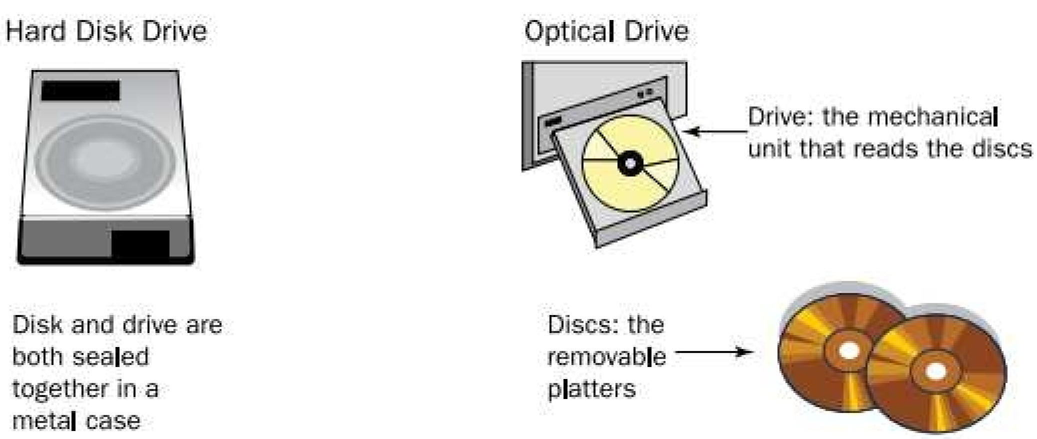
Data Storage Basics

A **file** is a named collection of bits that work together to represent a single object, such as an executable program, a spreadsheet, a picture, or a system file. Folders are logical organizing units for files. For example, a Windows-based computer has a Windows folder containing the system files needed to run Windows, a Users folder containing the data files and settings for each user account, and a Program Files folder containing the files needed for each installed application. A folder within a folder is called a subfolder; each of the aforementioned folders contains multiple subfolders. Files and folders are stored on volumes. Each volume has a letter followed by a colon, such as C: or D:. Figure below summarizes this structure.



Files are stored in folders, which are stored on volumes.

Some people use the terms disk (or disc), drive, and volume interchange-ably, but they have different meanings. The term disk refers to a platter or set of platters on which data is stored. Examples include a hard disk drive (HDD) and a DVD. When referring to a CD, DVD, or Blu-ray, it is customary to spell the term disc. The term drive refers to the mechanical components that read and write the data on a disk. In some cases, such as with an HDD, the drive and its platters are physically inseparable, so the terms disk and drive have come to be synonymous; hard disk and hard drive refer to the same thing. When the disc is removable from the drive, however, as with a CD or DVD, the terms are separate. See Figure below:.



In hard disk drives, the disk and drive are one unit; in optical drives, the disc and drive are separate.

Hard Disk Drives

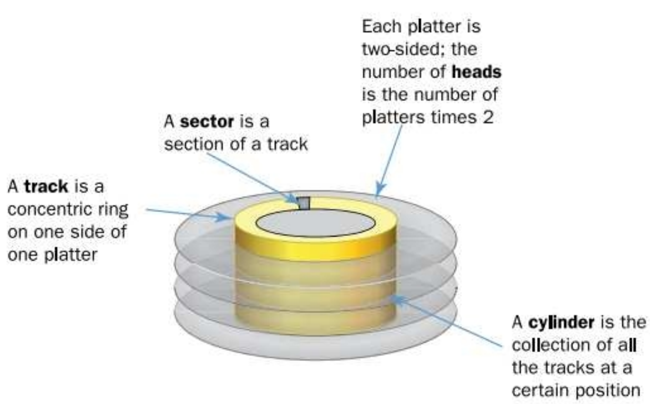
A hard disk drive (HDD) is the most popular type of secondary storage for personal computers. Although newer technologies are emerging, HDDs remain the standard because of their high capacity and low cost.

An HDD consists of a stack of metal platters (usually four to six) that arecoated with iron dust. These platters spin on a common spindle inside a sealed metal casing. A set of read/write heads inside the HDD casing reads and writes data on the platters; there are heads on each side of each platter. The arms move in and out to reach different spots on the platters, and the disks move past the heads as they spin. The most common platter sizes are 3.5” in diameter on HDDs for desktop systems and 2.5” on HDDs for notebooks. Figure below shows the inside of a hard disk drive case.

© iStockphoto.com/rhambley

The inside of a hard disk drive.

A precise system exists for describing the physical locations on an HDD. The platters are divided into concentric circles, like the rings on a tree trunk cross-section, called tracks. Each platter’s first track is track 0, the next one track 1, and so on. The track number tells the drive where to position the actuator arm that controls the position of the read/write heads. All the heads move together, so if one head is reading track 1 on one side of one platter, all the other heads are also reading track 1 on their side of their platter too. For this reason, tracks are not referred to individually, but as cylinders. A cylinder is the group of all the tracks at a single arm position. Figure below illustrates the relationship between these parts.



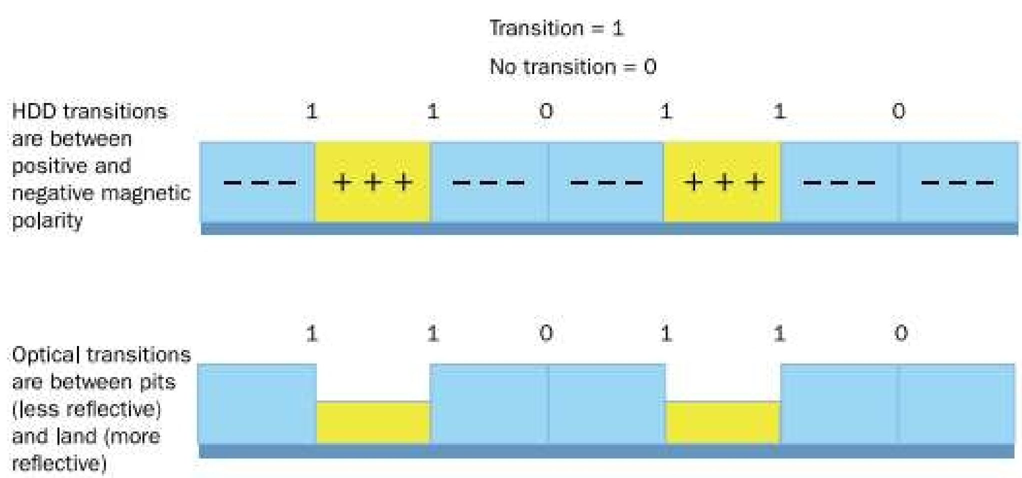
A hard disk drive is organized into tracks, cylinders, heads, and sectors.

Each track is divided into segments called sectors. A sector holds exactly 512 bytes. Because there are so many sectors, and because almost every file is much larger than 512 bytes, sectors are grouped together into clusters, and the drive's controller addresses clusters rather than individual sectors. The number of sectors per cluster is determined by the drive's size and formatting, but most modern HDDs have 32 sectors per cluster.

An HDD can be internal or external, but most are internal, installed inside the system unit. An internal HDD connects to the motherboard using either a parallel ATA (PATA[ATA: Advanced Technology Attachment]) or serial ATA (SATA) cable. An external HDD connects to a port on the outside of the system unit, and depending on the model may use a USB port, a FireWire (IEEE 1394a) port, or an external SATA (eSATA) port. You can buy enclosures that will convert an internal HDD to an external one.

Optical Drives

An **optical drive** uses a light beam and sensor to read the data. The surface of a blank optical disc is shiny and reflects light strongly. When data is written to an optical disc, certain areas are burned with a laser so they are less reflective. The shiny areas are called land, and the less-shiny areas are called pits. Recall from the preceding section that on an HDD, transitions between positive and negative magnetic polarity indicate a 1 bit, and lack of transition indicates a 0 bit. On an optical disc, transitions between areas of greater and lesser reflectivity indicate a 1, and a consistent level of reflectivity indicates a 0. See Figure below.



Magnetic and optical storage both look for transitions in the state of the disc surface.

There are several types of optical drives and discs. The oldest and most basic type is a **compact disc (CD)**, which holds up to 900MB of data. CDs are used for small amounts of data and also for audio recordings. A **digital versatile disc (DVD)** can store up to 4.7GB per disc (single-sided, single layer). DVDs can also be double-sided, with recordings on both sides rather than recording on one and a label on the other. DVDs can also be dual-layer, where the top layer is semitransparent and read using a laser and sensor at a different angle than the lower layer. A double-sided, dual-layer DVD can hold up to 17GB of data. DVDs are used to distribute large applications, large amounts of data, and standard-definition movies. **Blu-ray discs** (BD) can store up to 128GB in up to four layers. They are used to distribute even larger amounts of data, or high-definition movies.

CD	DVD	BD
Compact disc (CD) is an optical disc used for storing music and data, holding up to 900 MB.	Digital Versatile Disc (DVD) An optical disc used for storing standard-definition movies and data, holding up to 4.17 GB per side per layer.	Blu-ray disc (BD) An optical disc used for storing high-definition movies and data, holding up to 25 GB per layer.

Solid-State Drives

A **solid-state drive** (SSD) uses a type of EEPROM to store data in tiny transistors. Electrically Erasable Programmable Read-only Memory (EEPROM) is a type of memory that stores its contents permanently; it is nonvolatile. Because it is electrically erasable, the computer can erase what's written there and rewrite it. The type of EEPROM used in solid-state drives can be erased and rewritten in small blocks, making it suitable for use as a storage device.

Solid-state drives are silent because they don't have any moving parts, and the access time is very fast because there are no read/write heads that have to move anywhere to get to the data. SSDs are more expensive than HDDs, though; you get much less capacity for the money. For this reason, SSHDs (Solid-State Hybrid Drives) are found mostly in high-end desktop and notebook computers. Solid-state storage is common in tablets and smartphones, where high-capacity storage is not needed and being lightweight is a primary concern.

Processing Data

The CPU:

Every computer has at least one processor, also called a central processing unit (CPU). The CPU contains millions of tiny transistors and pathways.

The CPU is the "brain" of the computer, the place where data is manipulated. In large computer systems, such as supercomputers and mainframes, processing tasks may be handled by multiple processing chips. (Some powerful computer systems use hundreds or even thousands of separate processing units.) In the average microcomputer, the entire CPU is a single unit, called a microprocessor.

Every CPU includes the following components:

The **Control Unit** manages the flow of data through the CPU. It directs data to and from the other components within the CPU.

The **Arithmetic Logic Unit (ALU)** component does the actual processing. It receives data and instructions and delivers a result.

Because all computer data is stored as numbers, much of the processing that takes place involves comparing numbers or carrying out mathematical operations. In addition to establishing ordered sequences and changing those sequences, the computer can perform two types of operations: arithmetic operations and logical operations. Arithmetic operations include addition, subtraction, multiplication, and division. Logical operations include comparisons, such as determining whether one number is equal to, greater than, or less than another number. Also, every logical operation has an opposite. For example, in addition to "equal to" there is "not equal to." Table below shows the symbols for all the arithmetic and logical operations.

Operations Performed by the Arithmetic Logic Unit

Arithmetic Operations		Logical Operations	
+	add	=, ≠	equal to, not equal to
−	subtract	>, ≯	greater than, not greater than
×	multiply	<, ≮	less than, not less than
÷	divide	≥, ≱	greater than or equal to, not greater than or equal to
∧	raise by a power	≤, ≰	less than or equal to, not less than or equal to

Many instructions carried out by the control unit involve simply moving data from one place to another—from memory to storage, from memory to the printer, and so forth. When the control unit encounters an instruction that involves arithmetic or logic, however, it passes that instruction to the second component of the CPU, the arithmetic logic unit, or ALU. The ALU actually performs the arithmetic and logical operations described earlier.

The ALU includes a group of **registers**—high-speed memory locations built directly into the CPU that are used to hold the data currently being processed. Registers are holding areas for both data and instructions. We can think of the register as a scratchpad. The ALU will use the register to hold the data currently being used for a calculation. For example, the control unit might load two numbers from memory into the registers in the ALU. Then it might tell the ALU to divide the two numbers (an arithmetic operation) or to see whether the numbers are equal (a logical operation). The answer to this calculation will be stored in another register before being sent out of the CPU. That means there are many different registers, each with its own special purpose.

The physical composition of CPU:

Physically, a CPU is a very small and thin sheet of **semiconductor** material (usually silicon) with a complex array of tiny transistors and buses stamped into it with a die. Semiconductor material is used for CPUs because it does not affect the flow of electricity one way or another: The semiconductor neither conducts nor impedes the electrical flow. The components inside the CPU are so small that you need a microscope to see them. The CPU is fragile, so it is protected by a sturdy ceramic and metal casing; some models are mounted on miniature square circuit boards. The underside of this casing contains either tiny pins or tiny metal dots, precisely spaced, that align with connectors on the CPU socket on the motherboard. Each pin or dot is a separate communication line, carrying a different piece of data. Figure below shows both sides of a CPU designed for a desktop PC, for example. CPUs come in different sizes and have to match the socket type on the corresponding motherboard.

Most modern CPUs have multiple **cores**, so they can complete multiple tasks simultaneously, as if they physically were more than one CPU. A core consists of a separate set of the essential processor components (control unit, ALU, and registers). Most models of the Intel i7 processor have four cores, for example. All the CPU's cores are located on the same chip.

The system clock and cache memory:

Every microcomputer has a **system clock**. Like most modern wristwatches, the clock is driven by a quartz crystal. When electricity is applied, the molecules in the crystal vibrate millions of times per second, a rate chat never changes. The speed of the vibrations is determined by the thickness of the crystal. The computer uses the vibrations of the quartz in the system clock to time its processing operations. In theory, The CPU is capable of executing a function with every tick of the system clock (called a clock cycle). However, in practice, the CPU sometimes is idle because there is a delay between the request for data to be retrieved from memory and its delivery. A delay caused by waiting for another component to deliver data is called **latency**.

To help minimize latency, CPUs have **caches**. A cache is a small amount of very fast memory located near (or within) the CPU. Data that the CPU has recently used, or is predicted to need soon, is placed in the cache for temporary holding. That way, if the CPU calls for the data, it's more readily available and there is less delay.

Since the late 1980s, most PC CPUs have had cache memory built into them. This CPU-resident cache is often called [evel-1 (L1) cache. Today, many CPUs have as much as 256 KB built in. However, some really powerful CPUs are now taking it close to 1MB.

To add even more speed to modern CPUs, an additional cache is added to CPUs. This cache is called Level-2 (L2) cache. This cache used to be found on the motherboard. However, Intel and AMD found that placing the L2 cache on the CPU greatly increased CPU response. L2 (Level 2) cache is slower than L1 cache, but bigger in size. Its size typically varies between 256KB to 8MB, although the newer, powerful CPUs tend to go past that.

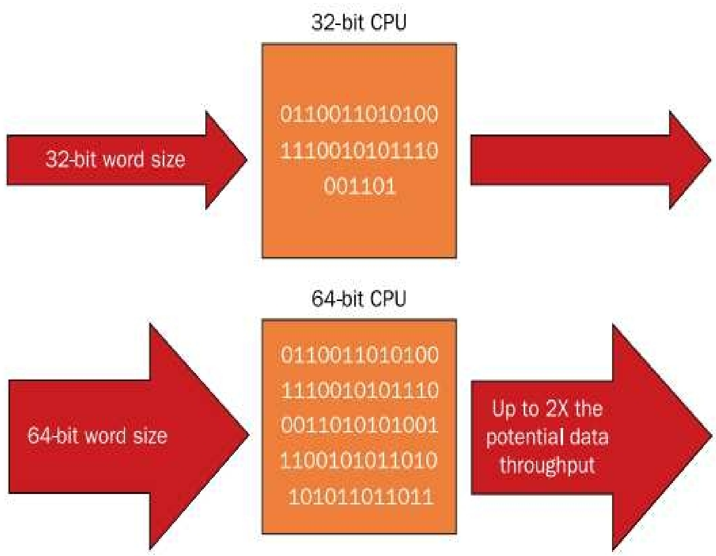
In addition to the cache memory built into the CPU, cache is also added to the motherboard. This motherboard-resident cache is now called Level-4 (13) cache. L3 (Level 3) cache is the largest cache memory unit, and also the slowest one. It is found on very-high-end computers. It is not necessary for a computer to have L3 cache. It can range between 4MB to upwards of 50MB.

L1, L2, and L3 all speed up the CPU, although in different ways. L1 cache holds instructions that have recently run. L2 cache holds potential upcoming instructions. L3 holds many of the possible instructions. In all cases, the cache memory is faster for the CPU to access, resulting in a quicker program execution.

CPU performance factors:

The most obvious performance factor for a CPU is its maximum speed, measured in billions of hertz, or gigahertz (GHz). Although the actual speed is dictated by the motherboard's system clock, each CPU has an advertised maximum speed, which is the highest speed at which the CPU's manufacturer assures that it will reliably function.

A CPU's word size also makes a difference. The word size is the number of bits that the CPU (or a single core of the CPU, if multi-core) can accept as input simultaneously. It is actually equal to the size of the registers. Most desktop and notebook PCs have 64-bit CPUs, but some tablets, netbooks, and smartphones have 32-bit CPUs.



A 64-bit CPU has twice the word size of a 32-bit CPU, so the 64-bit CPU can theoretically input and output date at twice the rate. In actual use, many other factors also play a part, so the performance boost is less than 2x.

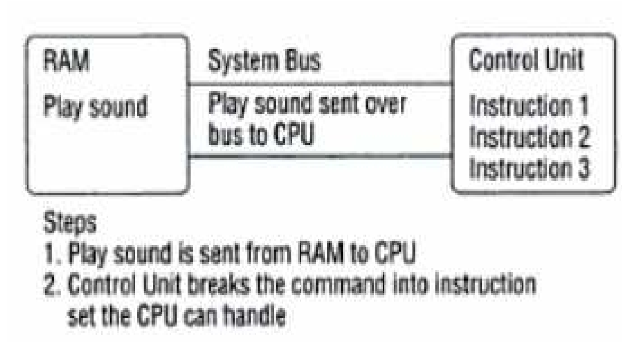
Machine Cycle:

Each time the CPU executes an instruction, it takes a series of steps. The completed series of steps is called a machine cycle. A machine cycle itself can be broken down into two smaller cycles: the instruction cycle and the execution cycle.

At the beginning of the **machine cycle** (that is, during the instruction cycle), the CPU takes two steps:

1. Fetching: Before the CPU can execute an instruction, the control unit must retrieve (or fetch) a command or data from the computer's memory.

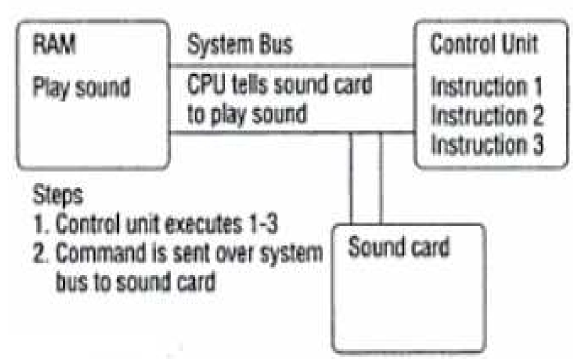
2. Decoding: Before a command can be executed, the control unit must break down (or decode) the command into instructions that correspond to those in the CPU's instruction set. Figure below shows how the CPU plays a sound:



At this point, the CPU is ready to begin the execution cycle:

1. Executing: When the command is executed, the CPU carries out the instructions in order by converting them into microcode.

2. Storing: The CPU may be required to store the results of an instruction in memory (but this condition is not always required). Figure below shows the result of the sound being played.



Microprocessor manufacturers/designers:

For two decades after the birth of the personal computer, the biggest player in the PC CPU market was Intel Corporation. This dominance began to change in 1998 when several leading computer makers began offering lower-priced systems using chips made by AMD and other chip manufacturers. Initially, these microprocessors offered less performance and a lower price. That situation has changed, however, as AMD made rapid advances in its products' capabilities. Today, Intel and AMD chips compete head to head, not only in performance, but also in price.

Some microprocessor manufacturers/designers:

- Intel.
- AMD.
- Qualcomm.
- NVIDIA.
- IBM.
- Samsung.
- Motorola.
- Hewlett-Packard (hp)