1.4 Under the covers

Now that we have looked below your program to uncover the underlying software, let's open the covers of your computer to learn about the underlying hardware. The underlying hardware in any computer performs the same basic functions: inputting data, outputting data, processing data, and storing data. How these functions are performed is the primary topic of this book, and subsequent chapters deal with different parts of these four tasks.

When we come to an important point in this book, a point so significant that we hope you will remember it forever, we emphasize it by identifying it as a *Big Picture* item. We have about a dozen Big Pictures in this book, the first being the five components of a computer that perform the tasks of inputting, outputting, processing, and storing data.

Two key components of computers are *input devices*, such as the microphone, and *output devices*, such as the speaker. As the names suggest, input feeds the computer, and output is the result of computation sent to the user. Some devices, such as wireless networks, provide both input and output to the computer.

Input device: A mechanism through which the computer is fed information, such as a keyboard.

Output device: A mechanism that conveys the result of a computation to a user, such as a display, or to another computer.

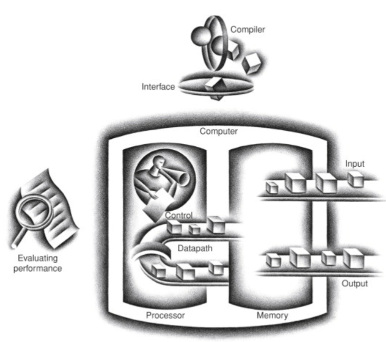
COD Chapters 5 (Large and Fast: Exploiting Memory Hierarchy) and 6 (Parallel Processor from Client to Cloud) describe input/output (I/O) devices in more detail, but let's take an introductory tour through the computer hardware, starting with the external I/O devices.

The Big Picture

The five classic components of a computer are input, output, memory, datapath, and control, with the last two sometimes combined and called the processor. The figure below shows the standard organization of a computer. This organization is independent of hardware technology: you can place every piece of every computer, past and present, into one of these five categories.

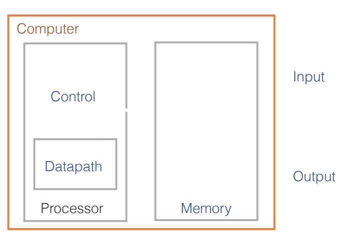
Figure 1.4.1: The organization of a computer, showing the five classic components (COD Figure 1.5).

The processor gets instructions and data from memory. Input writes data to memory, and output reads data from memory. Control sends the signals that determine the operations of the datapath, memory, input, and output.



PARTICIPATION ACTIVITY 1.4.1: The organization of a computer, showing the five classic components (COD Figure 1.5).

Start 2x speed



PARTICIPATION ACTIVITY 1.4.2: The five components of a computer.

Control Output Memory Input Datapath

Writes data to memory. Ex: Keyboard.

Reads data from memory. Ex: Display.

Stores instructions and data.

Sends signals that determine the operation of the other components.

Performs computations.

Reset

Through the looking glass

Source: zyBooks

The most fascinating I/O device is probably the graphics display. Most personal mobile devices use *liquid crystal displays (LCDs)* to get a thin, low-power display. The LCD is not the source of light; instead, it controls the transmission of light. A typical LCD includes rod-shaped molecules in a liquid that form a twisting helix that bends light entering the display, from either a light source behind the display or less often from reflected light. The rods straighten out when a current is applied and no longer bend the light. Since the liquid crystal material is between two screens polarized at 90 degrees, the light cannot pass through unless it is bent. Today, most LCD displays use an *active matrix* that has a tiny transistor switch at each pixel to control current precisely and make sharper images. A red-green-blue mask associated with each dot on the display determines the intensity of the three-color components in the final image; in a color active matrix LCD, there are three transistor switches at each point.

Liquid crystal display: A display technology using a thin layer of liquid polymers that can be used to transmit or block light according to whether a charge is applied.

Active matrix display: A liquid crystal display using a transistor to control the transmission of light at each individual pixel.

The image is composed of a matrix of picture elements, or *pixels*, which can be represented as a matrix of bits, called a ***bit map***. Depending on the size of the screen and the resolution, the display matrix in a typical tablet ranges in size from 1024 × 768 to 2048 × 1536. A color display might use 8 bits for each of the three colors (red, blue, and green), for 24 bits per pixel, permitting millions of different colors to be displayed.

Pixel: The smallest individual picture element. Screens are composed of hundreds of thousands to millions of pixels, organized in a matrix.

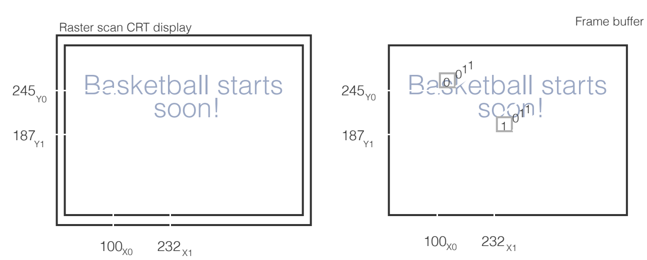
The computer hardware support for graphics consists mainly of a ***raster refresh buffer***, or ***frame buffer***, to store the bit map. The image to be represented onscreen is stored in the frame buffer, and the bit pattern per pixel is read out to the graphics display at the refresh rate. The animation below shows a frame buffer with a simplified design of just 4 bits per pixel.

The goal of the bit map is to represent faithfully what is on the screen. The challenges in graphics systems arise because the human eye is very good at detecting even subtle changes on the screen.

PARTICIPATION ACTIVITY

1.4.3: Each coordinate in the frame buffer on the right determines the color of the corresponding coordinate for the raster scan CRT display on the left (COD Figure 1.6).

Start 2x speed



> "Through computer displays I have landed an airplane on the deck of a moving carrier, observed a nuclear particle hit a potential well, flown in a rocket at nearly the speed of light and watched a computer reveal its innermost workings.
> *Ivan Sutherland, the "father" of computer graphics, Scientific American, 1984*

Touchscreen

While PCs also use LCD displays, the tablets and smartphones of the post-PC era have replaced the keyboard and mouse with touch-sensitive displays, which has the wonderful user interface advantage of users pointing directly at what they are interested in rather than indirectly with a mouse.

While there are a variety of ways to implement a touch screen, many tablets today use capacitive sensing. Since people are electrical conductors, if an insulator like glass is covered with a transparent conductor, touching distorts the electrostatic field of the screen, which results in a change in capacitance. This technology can allow multiple touches simultaneously, which recognizes gestures that can lead to attractive user interfaces.

PARTICIPATION ACTIVITY

1.4.4: Displays.

1) A liquid crystal display works by having the liquid crystal generate different colors of light.

- True
- False

2) A typical computer display is made up of hundreds of pixels.

- True
- False

3) An active matrix display is an LCD that uses a transistor to control whether light passes for each pixel.

- True
- False

4) Each pixel of a display typically involves 24 bits, with 8 bits for each of red, blue, and green.

- True
- False

5) A frame buffer stores the pixel values for a display.

- True
- False

6) A touchscreen combines a display for output with a touch-sensitive screen for input.

- True
- False

Opening the box

The figure below shows the contents of the Apple iPad 2 tablet computer. Unsurprisingly, of the five classic components of the computer, I/O dominates this reading device. The list of I/O devices includes a capacitive multitouch LCD display, front-facing camera, rear-facing camera, microphone, headphone jack, speakers, accelerometer, gyroscope, Wi-Fi network, and Bluetooth network. The datapath, control, and memory are a tiny portion of the components.

Figure 1.4.2: Components of the Apple iPad 2 A1395 (COD Figure 1.7).

The metal back of the iPad (with the reversed Apple logo in the middle) is in the center. At the top is the capacitive multitouch screen and LCD display. To the far right is the 3.8V, 25 watt-hour, polymer battery, which consists of three Li-ion cell cases and offers 10 hours of battery life. To the far left is the metal frame that attaches the LCD to the back of the iPad. The small components surrounding the metal back in the center are what we think of as the computer; they are often L-shaped to fit compactly inside the case next to the battery. The next figure shows a close-up of the L-shaped board to the lower left of the metal case, which is the logic printed circuit board that contains the processor and the memory. The tiny rectangle below the logic board contains a chip that provides wireless communication: Wi-Fi, Bluetooth, and FM tuner. It fits into a small slot in the lower left corner of the logic board. Near the upper left corner of the case is another L-shaped component, which is a front-facing camera assembly that includes the camera, headphone jack, and microphone. Near the right upper corner of the case is the board containing the volume control and silent/screen rotation lock button along with a gyroscope and accelerometer. These last two chips combine to allow the iPad to recognize six-axis motion. The tiny rectangle next to it is the rear-facing camera. Near the bottom right of the case is the L-shaped speaker assembly. The cable at the bottom is the connector between the logic board and the camera/volume control board. The board between the cable and the speaker assembly is the controller for the capacitive touchscreen. (Courtesy iFixit, www.ifixit.com)

The small rectangles in the figure below contain the devices that drive our advancing technology, called *integrated circuits* and nicknamed *chips*. The A5 package seen in the middle of the figure below contains two ARM processors that operate at a clock rate of 1 GHz. The processor is the active part of the computer, following the instructions of a program to the letter. It adds numbers, tests numbers, signals I/O devices to activate, and so on. Occasionally, people call the processor the *CPU*, for the more bureaucratic-sounding *central processor unit*.

Integrated circuit: Also called a ***chip***. A device combining dozens to millions of transistors.

Central processor unit (***CPU***): Also called ***processor***. The active part of the computer, which contains the datapath and control and which adds numbers, tests numbers, signals I/O devices to activate, and so on.

Figure 1.4.3: Logic board of Apple iPad 2 (COD Figure 1.8).

The logic board of Apple iPad 2 in the previous figure. The photo highlights five integrated circuits. The large integrated circuit in the middle is the Apple A5 chip, which contains dual ARM processor cores that run at 1 GHz as well as 512 MB of main memory inside the package. The next figure shows a photograph of the processor chip inside the A5 package. The similar-sized chip to the left is the 32GB flash memory chip for non-volatile storage. There is an empty space between the two chips where a second flash chip can be installed to double storage capacity of the iPad. The chips to the right of the A5 include power controller and I/O controller chips. (Courtesy iFixit, www.ifixit.com)

Descending even lower into the hardware, the figure below reveals details of a microprocessor. The processor logically comprises two main components: datapath and control, the respective brawn and brain of the processor. The *datapath* performs the arithmetic operations, and *control* tells the datapath, memory, and I/O devices what to do according to the wishes of the instructions of the program. COD Chapter 4 (The Processor) explains the datapath and control for a higher-performance design.

Datapath: The component of the processor that performs arithmetic operations.

Control: The component of the processor that commands the datapath, memory, and I/O devices according to the instructions of the program.

The A5 package in the figure above also includes two memory chips, each with 2 gibibits of capacity, thereby supplying 512MiB. The *memory* is where the programs are kept when they are running; it also contains the data needed by the running programs. The memory is built from DRAM chips. *DRAM* stands for *dynamic random access memory*. Multiple DRAMs are used together to contain the instructions and data of a program. In contrast to sequential access memories, such as magnetic tapes, the RAM portion of the term DRAM means that memory accesses take basically the same amount of time no matter what portion of the memory is read.

Memory: The storage area in which programs are kept when they are running and that contains the data needed by the running programs.

Dynamic random access memory (***DRAM***): Memory built as an integrated circuit; it provides random access to any location. Access times are 50 nanoseconds and cost per gigabyte in 2012 was \$5 to \$10.

Figure 1.4.4: Processor integrated circuit (COD Figure 1.9).

The processor integrated circuit inside the A5 package. The size of chip is 12.1 by 10.1 mm, and it was manufactured originally in a 45-nm process (see COD Section 1.5 (Technologies for building processors and memory)). It has two identical ARM processors or cores in the middle left of the chip and a PowerVR graphical processor unit (GPU) with four datapaths in the upper left quadrant. To the left and bottom side of the ARM cores are interfaces to main memory (DRAM). (Courtesy Chipworks, www.chipworks.com).

PARTICIPATION ACTIVITY 1.4.5: Processors, chips, and the iPad 2.

1) An integrated circuit is often called a _____.
- chip
- CPU

2) The CPU chip physically occupies _____ of the size of the iPad 2.
- most
- a small fraction

3) The iPad 2 consists of how many chips?
- 1
- 2
- 5

4) The A5 package has a chip containing ____ ARM processors.
- 1
- 2
- 5

5) A CPU is also known as ________.
- a datapath
- control
- a processor

Descending into the depths of any component of the hardware reveals insights into the computer. Inside the processor is another type of memory—cache memory.

Cache memory consists of a small, fast memory that acts as a buffer for the DRAM memory. (The nontechnical definition of *cache* is a safe place for hiding things.) Cache is built using a different memory technology, *static random access memory (SRAM)*. SRAM is faster but less dense, and hence more expensive, than DRAM (see COD Chapter 5 (Large and Fast: Exploiting Memory Hierarchy)). SRAM and DRAM are two layers of the **memory hierarchy**.

HIERARCHY

Cache memory: A small, fast memory that acts as a buffer for a slower, larger memory.

Static random access memory (SRAM): Also memory built as an integrated circuit, but faster and less dense than DRAM.

PARTICIPATION ACTIVITY 1.4.6: RAM and cache.

DRAM SRAM Cache

Large memory where most data is stored.

A faster memory technology than DRAM, but using more area to store a bit.

A small memory that keeps a copy of data from larger memory.

Reset

As mentioned above, one of the great ideas to improve design is abstraction. One of the most important **abstractions** is the interface between the hardware and the lowest-level software. Because of its importance, it is given a special name: the *instruction set architecture*, or simply *architecture*, of a computer. The instruction set architecture includes anything programmers need to know to make a binary machine language program work correctly, including instructions, I/O devices, and so on. Typically, the operating system will encapsulate the details of doing I/O, allocating memory, and other low-level system functions so that application programmers do not need to worry about such details. The combination of the basic instruction set and the operating system interface provided for application programmers is called the *application binary interface (ABI)*.

Instruction set architecture: Also called ***architecture***. An abstract interface between the hardware and the lowest-level software that encompasses all the information necessary to write a machine language program that will run correctly, including instructions, registers, memory access, I/O, and so on.

Application binary interface (ABI): The user portion of the instruction set plus the operating system interfaces used by application programmers. It defines a standard for binary portability across computers.

An instruction set architecture allows computer designers to talk about functions independently from the hardware that performs them. For example, we can talk about the functions of a digital clock (keeping time, displaying the time, setting the alarm) separately from the clock hardware (quartz crystal, LED displays, plastic buttons). Computer designers distinguish architecture from an implementation of an architecture along the same lines: an *implementation* is hardware that obeys the architecture abstraction. These ideas bring us to another Big Picture.

Implementation: Hardware that obeys the architecture abstraction.

The Big Picture

Both hardware and software consist of hierarchical layers using abstraction, with each lower layer hiding details from the level above. One key interface between the levels of abstraction is the *instruction set architecture*—the interface between the hardware and low-level software. This abstract interface enables many *implementations* of varying cost and performance to run identical software.

PARTICIPATION ACTIVITY 1.4.7: Instruction set architecture.

1) An instruction set architecture enables a machine language program to run on different hardware implementations.
   - True
   - False

2) While different hardware implementations may run the same program, designers strive to keep the performance of new hardware implementations the same as older implementations.
   - True
   - False

A safe place for data

Thus far, we have seen how to input data, compute using the data, and display data. If we were to lose power to the computer, however, everything would be lost because the memory inside the computer is *volatile*—that is, when it loses power, it forgets. In contrast, a DVD disk doesn't forget the movie when you turn off the power to the DVD player, and is therefore a *nonvolatile memory* technology.

Volatile memory: Storage, such as DRAM, that retains data only if it is receiving power.

Nonvolatile memory: A form of memory that retains data even in the absence of a power source and that is used to store programs between runs. A DVD disk is nonvolatile.

To distinguish between the volatile memory used to hold data and programs while they are running and this nonvolatile memory used to store data and programs between runs, the term *main memory* or *primary memory* is used for the former, and *secondary memory* for the latter. Secondary memory forms the next lower layer of the **memory hierarchy**. DRAMs have dominated main memory since 1975, but *magnetic disks* dominated secondary memory starting even earlier. Because of their size and form factor, personal Mobile Devices use *flash memory*, a nonvolatile semiconductor memory, instead of disks. A previous figure (COD Figure 1.8) shows the chip containing the flash memory of the iPad 2. While slower than DRAM, it is much cheaper than DRAM in addition to being nonvolatile. Although costing more per bit than disks, it is smaller, it comes in much smaller capacities, it is more rugged, and it is more power efficient than disks. Hence, flash memory is the standard secondary memory for PMDs (personal mobile devices). Alas, unlike disks and DRAM, flash memory bits wear out after 100,000 to 1,000,000 writes. Thus, file systems must keep track of the number of writes and have a strategy to avoid wearing out storage, such as by moving popular data. COD Chapter 5 (Large and Fast: Exploiting Memory Hierarchy) describes disks and flash memory in more detail.

HIERARCHY

Main memory: Also called ***primary memory***. Memory used to hold programs while they are running; typically consists of DRAM in today's computers.

Secondary memory: Nonvolatile memory used to store programs and data between runs; typically consists of flash memory in PMDs and magnetic disks in servers.

Magnetic disk: Also called hard disk. A form of nonvolatile secondary memory composed of rotating platters coated with a magnetic recording material. Because they are rotating mechanical devices, access times are about 5 to 20 milliseconds and cost per gigabyte in 2012 was $0.05 to $0.10.

Flash memory: A nonvolatile semiconductor memory. It is cheaper and slower than DRAM but more expensive per bit and faster than magnetic disks. Access times are about 5 to 50 microseconds and cost per gigabyte in 2012 was $0.75 to $1.00.

PARTICIPATION ACTIVITY 1.4.8: Memories.

Main memory Secondary memory Flash memory Volatile memory

Nonvolatile memory Magnetic disk

Memory layer used to hold programs and data while programs are running.

Memory layer used to store programs and data between runs.

A form of memory that retains data only if the memory is receiving power.

A form of memory that retains data in the absence of a power.

A nonvolatile semiconductor memory often used as secondary memory for personal mobile devices.

A form of nonvolatile secondary memory composed of rotating platters coated with a magnetic recording material

Reset

Communicating with other computers

We've explained how we can input, compute, display, and save data, but there is still one missing item found in today's computers: computer networks. Just as the processor shown in a previous figure (COD Figure 1.5) is connected to memory and I/O devices, ***networks*** interconnect whole computers, allowing computer users to extend the power of computing by including communication. Networks have become so popular that they are the backbone of current computer systems; a new personal mobile device or server without a network interface would be ridiculed. Networked computers have several major advantages:

- *Communication*: Information is exchanged between computers at high speeds.
- *Resource sharing*: Rather than each computer having its own I/O devices, computers on the network can share I/O devices.
- *Nonlocal access*: By connecting computers over long distances, users need not be near the computer they are using.

Networks vary in length and performance, with the cost of communication increasing according to both the speed of communication and the distance that information travels. Perhaps the most popular type of network is *Ethernet*. It can be up to a kilometer long and transfer at up to 40 gigabits per second. Its length and speed make Ethernet useful to connect computers on the same floor of a building; hence, it is an example of what is generically called a *local area network*. Local area networks are interconnected with switches that can also provide routing services and security. *Wide area networks* cross continents and are the backbone of the Internet, which supports the web. They are typically based on optical fibers and are leased from telecommunication companies.

Local area network (***LAN***): A network designed to carry data within a geographically confined area, typically within a single building.

Wide area network (***WAN***): A network extended over hundreds of kilometers that can span a continent.

Networks have changed the face of computing in the last 30 years, both by becoming much more ubiquitous and by making dramatic increases in performance. In the 1970s, very few individuals had access to electronic mail, the Internet and web did not exist, and physically mailing magnetic tapes was the primary way to transfer large amounts of data between two locations. Local area networks were almost nonexistent, and the few existing wide area networks had limited capacity and restricted access.

As networking technology improved, it became considerably cheaper and had a significantly higher capacity. For example, the first standardized local area network technology, developed about 30 years ago, was a version of Ethernet that had a maximum capacity (also called bandwidth) of 10 million bits per second, typically shared by tens of, if not a hundred, computers. Today, local area network technology offers a capacity of from 1 to 40 gigabits per second, usually shared by at most a few computers. Optical communications technology has allowed similar growth in the capacity of wide area networks, from hundreds of kilobits to gigabits and from hundreds of computers connected to a worldwide network to millions of computers connected. This dramatic rise in deployment of networking combined with increases in capacity have made network technology central to the information revolution of the last 30 years.

For the last decade another innovation in networking is reshaping the way computers communicate. Wireless technology is widespread, which enabled the post-PC era. The ability to make a radio in the same low-cost semiconductor technology (CMOS) used for memory and microprocessors enabled a significant improvement in price, leading to an explosion in deployment. Currently available wireless technologies, called by the IEEE standard name 802.11, allow for transmission rates from 1 to nearly 100 million bits per second. Wireless technology is quite a bit different from wire-based networks, since all users in an immediate area share the airwaves.

PARTICIPATION ACTIVITY 1.4.9: Networks.

1) Network connectivity is a key part of modern computing devices like personal computers, tablets, and smartphones.
   - True
   - False

2) Ethernet is an example of a wide area network.
   - True
   - False

3) Wireless networks continue to become more common.
   - True
   - False

PARTICIPATION ACTIVITY 1.4.10: Check yourself: DRAM memory, flash memory, and disk storage characteristics.

1) _____ memory is volatile.
   - DRAM
   - Flash
   - Disk

2) _____ memory is the most expensive per GB.
   - DRAM
   - Flash
   - Disk

3) _____ memory has the longest access time.
   - DRAM
   - Flash
   - Disk