


ch 001

An Introduction to Hardware, Software, and the Internet

- 1.1 Prologue 2
- 1.2 Hardware Overview 5
 - 1.2.1 The CPU 5
 - 1.2.2 Memory 6
 - 1.2.3 Secondary Storage Devices 7
 - 1.2.4 Input and Output Devices 7
- 1.3 Software Overview 8
- 1.4 What Do Programmers Do? 10
- 1.5 Representation of Information in Computer Memory 14
 - 1.4.1 Numbers 15
 - 1.4.2 Characters 19
- 1.6 The Internet 21
- 1.7 Summary 23
 - Exercises 

1.1 Prologue

Have you ever opened that gray or black (or purple) box sitting on or beside your desk, with tangled wires snaking out of it in all directions? If you do, you will find a mostly empty box, with a power supply, a printed circuit board on the bottom connected by ribbon cables to a small disk drive and a few other devices neatly tucked away in their bays. And that's all. But it brings you an entire world.

The most important piece of a typical computer is the *Central Processing Unit* or *CPU* [1, 2, 3]. In a personal computer, the CPU is a microprocessor made from a tiny chip of silicon, sometimes as small as half an inch square. Immensely precise manufacturing processes etch a huge number of semiconductor devices, called *transistors*, into the silicon wafer. Each transistor is a microscopic digital switch and together they control, with perfect precision, billions of signals — little spikes of electricity — that arise and disappear every second. The size of the spikes doesn't matter, only their presence or absence. The transistors in the CPU recognize only two states of a signal, “on” or “off,” “high” or “low,” “1” or “0,” “true” or “false.” This is called *digital electronics* (as opposed to *analog electronics* where the actual amplitudes of signals carry information).

The transistors on a chip combine to form logical devices called *gates*. Gates implement *Boolean* operations (named after the British mathematician George Boole, 1815-1864 [1] who studied the properties of logical relations). For example, an *AND* gate takes two inputs and combines them into one output signal. The output is set to “true” if both the first and the second input are “true,” and to “false” otherwise (Figure 1-1-a). In an *OR* gate, the output is set to “true” if either the first or the second (or both) inputs are true (Figure 1-1-b). A *NOT* gate takes one input and sets the output to its opposite (Figure 1-1-c). Note the special shapes used to denote each type of gate.

These three basic types of gates can be combined to make other Boolean operations and logical circuits. Figure 1-2, for example, shows how you can combine AND, OR, and NOT gates to make an *XOR* (“*eXclusive OR*”) operation. This operation sets the output to “true” if exactly one of its two inputs is “true.” In the late 1940s, John von Neumann, [1] a great mathematician and one of the founding fathers of computer technology, showed that all arithmetic operations can be reduced to AND, OR, and NOT logical operations.

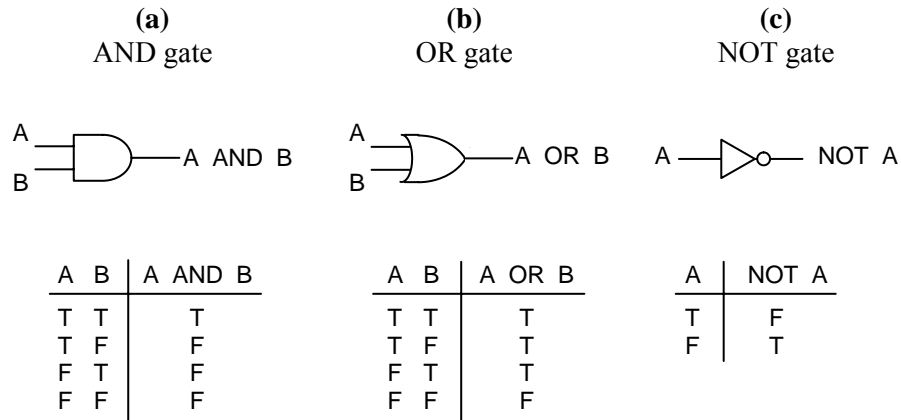


Figure 1-1. AND, OR, and NOT gates

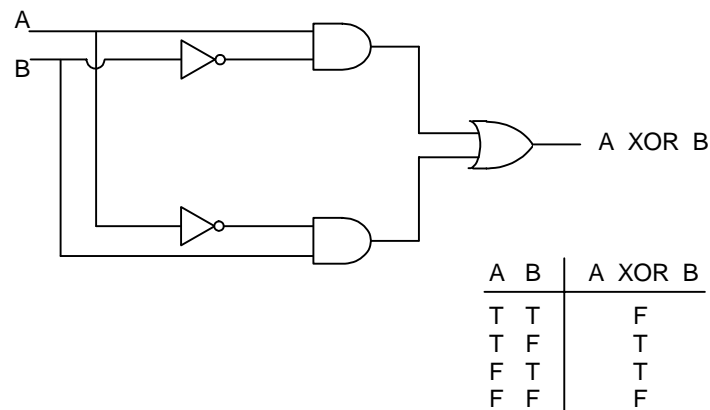


Figure 1-2. XOR circuit made of AND, OR, and NOT gates

The microprocessor is protected by a small ceramic case mounted on a *PC board* (*Printed Circuit board*) called the *motherboard* [1, 2]. Also on the motherboard are memory chips. The computer memory is a uniform pool of storage units called *bits*. A bit stores the smallest possible unit of information: “on” or “off,” “1” or “0.” For practical reasons, bits are grouped into groups of eight, called *bytes*.

One byte is eight bits.

There is no other structure to memory: the same memory is used to store numbers and letters and sounds and images and programs. All these things must be encoded, one way or another, in sequences of 0s and 1s. A typical personal computer made in the year 2005 had 256 “megs” (megabytes; 1 megabyte is 1,048,576 bytes) to 1 “gig” (gigabyte; 1 gigabyte is 1,024 megabytes) of *RAM* (Random-Access Memory) packed in a few *SIMMs* (Single In-Line Memory Modules).

The CPU interprets and carries out computer programs, or sequences of instructions stored in the memory. The CPU fetches the next instruction, interprets its operation code, and performs the appropriate operation. There are instructions for arithmetic and logical operations, for copying bytes from one location to another, and for changing the order of execution of instructions. The instructions are executed in sequence unless a particular instruction tells the CPU to “jump” to another place in the program. Conditional branching instructions tell the CPU to continue with the next instruction or jump to another place depending on the result of the previous operation.

Besides the CPU, a general-purpose computer system also includes *peripheral devices* [1] that provide input and output and secondary mass storage. In a notebook (laptop) computer, the “peripheral” devices are no longer quite so peripheral: a keyboard, a display, a hard drive, a CD-ROM or DVD drive, a fax/modem, a touch pad, a microphone, and speakers are all built into one portable unit.

CPU, memory, peripherals — all of this is called *hardware*. It is a lot of power concentrated in a small device. But to make it useful, to bring life into it, you need programs, *software*. Computer programs are also miracles of engineering, but of a different kind: *software engineering*. They are not cast in iron, nor even silicon, but in intangible texts that can be analyzed, modified, translated from one computer language into another, copied into various media, transmitted over networks, or lost forever. Software is to a computer as tunes are to a band: the best musicians will be silent if they don’t have music to play.

Take this amazing device with its software and connect it to the *Internet*, a network of millions of computers of all kinds connected to each other via communication lines of all kinds and running programs of all kinds, and you end up with a whole new world. Welcome to cyberspace!

In the rest of this chapter we will briefly discuss:

- The main hardware components: CPU, memory, peripheral devices
- What software is
- How numbers and characters are represented in computer memory
- What the Internet is

1.2 Hardware Overview

1.2.1 The CPU

What a CPU can do is defined by its instruction set and internal *registers* [1]. The registers are built-in memory cells that hold operands, memory addresses, and intermediate results. Some of the registers are accessible to the programmer. The instruction set includes instructions for loading CPU registers from memory and storing their values into memory, for logical and arithmetic operations, and for altering the sequence of operations.

Every computer has an internal “clock” that generates electrical pulses at a fixed frequency. All CPU operations and their component steps are synchronized with the clock’s pulses; their duration is measured in *clock cycles*. The CPU’s speed depends on the frequency of the clock. The Intel 8088 microprocessor in the original IBM Personal Computer (1981), for example, ran at 4.77 MHz (megahertz, or million pulses per second). Twenty-five years and a dozen computer generations later, CPU chips are pushing into the 3 GHz range (1 GHz, gigahertz, is equal to 1,000 MHz) [1].

A microprocessor CPU connects to memory and to other devices through a set of parallel lines (wires printed on the motherboard) controlled by digital electronics, called a *bus* [1]. A CPU may use a separate *address bus* for specifying memory addresses and a *data bus* for reading and writing memory values. Besides the internal clock speed, the computer’s overall performance depends on the speed of the bus transfers and the width of the bus. The Mac 512k computer [1], introduced in September of 1984, had the Motorola MC68000 8-MHz CPU and a 16-bit bus running at 8 MHz, so it could carry 16 bits of data concurrently from memory to the CPU roughly at the same speed as the CPU could handle the data. The PowerMac G5 released by Apple Computer Inc. in 2005 has a 64-bit bus running at around 1 GHz.

1.2.2 Memory

Each byte of memory has a unique address that can be used to fetch the value stored in the byte or write a new value into it. A CPU does not have to read or write memory bytes sequentially: bytes can be accessed in any arbitrary order. This is why computer memory is called *random-access memory* or *RAM*. This is similar to a CD where you can choose any track to play, as opposed to a tape that has to be played in sequence. The size of RAM is measured in *kilobytes* (KB, or simply K), *megabytes* (MB), or *gigabytes* (GB).

A kilobyte is $2^{10} = 1024$ bytes; a megabyte is $2^{20} = 1,048,576$ bytes; a gigabyte is equal to $2^{30} = 1,073,741,824$ bytes.

(Powers of 2 have a special significance in computer technology for a reason that will become clear shortly.)

In the early days, designers of personal computers thought 64K of RAM would suffice for the foreseeable future. An additional hardware mechanism, the *segment registers*, had to be added to the later versions of Intel's microprocessors to access a larger memory space, up to one megabyte, while maintaining compatibility with the old programs. But the one megabyte limit very quickly proved inadequate too. A 32-bit memory address bus allows programs to directly address four *gigabytes* (GB) of memory. This should suffice for a while, but no one knows how long.

A small part of the computer memory is permanent non-erasable memory, known as *read-only memory* or *ROM*. A personal computer's ROM contains, among other things, the initialization code that *boots up* the *operating system* (that is, loads into memory the *boot record* or initialization code from the disk and passes control to it).

An executable program has to be loaded from a hard disk or a CD-ROM (or from the Internet) into RAM before it can run.

ROM solves the “first-program” dilemma — some program must already be running to load any other program into memory. The *operating system* is a program that has the job of loading and executing other programs. In a personal computer, ROM also contains the computer configuration program and hardware diagnostic programs that check various computer components. The ROM BIOS (Basic Input Output System) contains programs for controlling the keyboard, display, disk drives, and other devices. A special small memory — EPROM (Erasable Programmable ROM) — preserves system configuration data when the power is off.

1.2.3 Secondary Storage Devices

A computer's RAM has only limited space, and its contents are wiped out when the power is turned off. All the programs and data in a computer system have to be stored in secondary mass storage. The auxiliary storage devices include hard disks, floppy disk drives, CD-ROM drives, memory sticks, and other devices. A hard disk can hold from 10 to hundreds of gigabytes and even terabytes.

■ **1 terabyte is 1024 gigabytes.**

A CD-ROM can hold more than 700 MB (or up to 2 or 3 GB with data compression). Access to data on these devices is much slower than access to RAM.

The operating system software organizes the data in secondary storage into *files*. A file may contain a set of related data, a program, a document, an image, and so on; it has a unique name. The operating system maintains a *directory* of file names, locations, sizes, dates and times of the last updates, and other attributes.

■ **Thus a “file” is not a hardware but rather a software concept.**

When a program is running, it can read and write data directly to and from files stored on secondary storage devices.

1.2.4 Input and Output Devices

A personal computer receives user input through the keyboard and displays the output on the computer display (also called the *monitor*). In many programs the input is echoed on the screen as you type, creating the illusion that the keyboard is directly connected to the monitor. In fact these are two entirely different devices connected only through the CPU and the currently running program. The keyboard sends to the program digital codes that represent the pressed keys. The program captures these codes and takes appropriate actions, which may include displaying corresponding characters on the screen.

The screen is controlled by a *video adapter* and displays the contents of special video memory in the adapter, called *VRAM*. VRAM is addressable by the CPU and may contain codes, colors and attributes of characters when running in text modes, or colors or intensities of individual *pixels* (“picture elements”) in graphics modes. The original IBM PC ran mostly in the text mode. You don't see the text mode much any more — everything is in graphics now, except diagnostic and setup programs.

A *mainframe* computer [1, 2] (a very large multi-user computer) may have hundreds of terminals attached to it. The terminals send keystrokes and receive commands and character codes from the computer through digital transmission lines.

Printers, plotters, digitizing tablets, scanners, and other devices receive commands and data from the computer in digital form and may send data or control codes back to the computer according to a specific communications protocol.

Network adapters and modems are essential for getting your computer connected to other computers. Network adapters and cables are used to connect several computers into a LAN (Local Area Network), which in turn may be connected to the Internet. Modems transmit digital information through telephone or cable lines or a wireless connection [1, 2, 3]. A sending modem encodes bits of information into a sequence of electrical pulses and a receiving modem decodes the signal back into digital form.

Special data acquisition devices equipped with *A/D (analog-to-digital)* converters [1] allow computers to convert an electrical signal into digital form by frequently sampling the amplitude of the signal and storing the digitized values in memory. *D/A (digital-to-analog)* converters perform the reverse transformation: they generate electrical currents from the digitized amplitudes stored in the computer. These devices allow the computer to receive data from all kinds of instruments and to serve as a universal control device in industrial applications and scientific experiments.

Input and output devices are connected to the computer via hardware *interface* modules that implement specific data transfer protocols. In a personal computer, the interfaces may be built into the motherboard or take the form of special adapter cards that plug into special sockets on the motherboard, called extension slots. Devices connected to the computer are usually controlled by special programs called *drivers* that handle all the details and peculiarities of the device and the data transfer protocol. Today's personal computers are equipped with several USB (Universal Serial Bus) ports, used for connecting virtually all peripheral devices: mice, printers, digital cameras, memory sticks, and so on. A high-speed USB interface can handle a data transmission rate of up to 480 megabits per second.

1.3 Software Overview

The term *software* refers to computer programs; it is also used as an adjective to refer to tasks or functions implemented through programs, as in “software interface,” “software fonts,” and so on. The line between hardware and software is not always clear. In the modern world, microprocessors are embedded in many objects, from microwave ovens and DVD players to cars and satellites. Their programs are

developed using simulation tools on normal computers; when a program is finalized, it is permanently “burned” into ROMs. Such programs are referred to as *firmware*.

A modern computer not only runs individual programs but also maintains a “software environment.” This environment involves several functional layers (Figure 1-3). The bottom layer in this environment comprises BIOS, device drivers, interrupt handlers, etc. — programs that directly support hardware devices and functions. The next layer is the *operating system*, a software program that provides computer access services to users and standard support functions to other programs. The top layer is software applications (word processors, Internet browsers, business, industrial, or scientific applications, games, etc.).

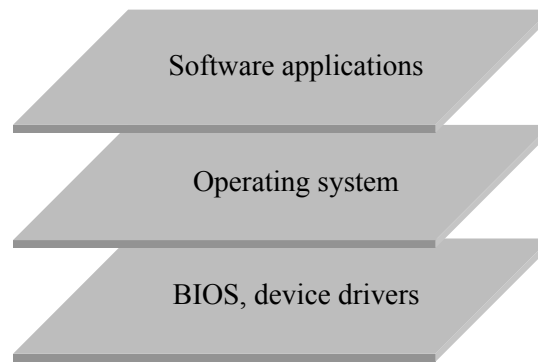


Figure 1-3. Software functional layers

The operating system loads programs into RAM from secondary storage and runs them. On a mainframe [1, 2], the operating system allows multiple users to work on the computer at once through *time sharing*. In such a multi-user system, one user may be slowly editing a file or entering data on a terminal using only a small fraction of the available CPU time. At the same time another program may be doing “number crunching.” A multi-user operating system allocates “time slices” to each program and automatically switches between them. The operating system prioritizes the jobs and swaps segments of programs in and out of memory as needed. A personal computer assumes one user, but contemporary users enjoy *multi-tasking* operating systems that let them keep several programs active concurrently (for example, a word processor, an e-mail program, and a music player).

The operating system also establishes and maintains a file system in secondary storage. Files are organized into a branching structure of directories and subdirectories. The operating system provides commands and utility programs for navigating through the directory tree.

Part of the operating system is a set of *routines* (sets of instructions, callable from other programs) that provide standard service functions to programs. These include functions for creating, reading, and writing files. The operating system *shell* provides a set of user commands, including commands for displaying, copying, deleting, and printing files, executing programs, and so on. Modern operating systems use *GUI* (*Graphical User Interface*, pronounced “goeey”), where commands can be entered by selecting items in menus or by clicking a mouse on an icon that represents a command or an object graphically.

The top layer of software consists of *application* programs that make computers useful to people. In the old days, application programs mostly performed calculations or processed files. If interaction with the user was required, it was accomplished via an unsophisticated text dialog: a computer would display questions (or *prompts*) in plain text and let the user enter responses, then print out the resulting numbers, tables, or reports. Such applications are called *console applications*. In a modern computing environment, users expect programs to have GUI. Menus, buttons, and icons select different functions in the program; dialog boxes and text-edit fields are used to set options and enter data or text. The shift from console applications to GUI applications has brought changes in the way programs are designed and written and brought about new software development tools and methodologies, such as *OOP* (*Object-Oriented Programming*), one of the main subjects of this book.

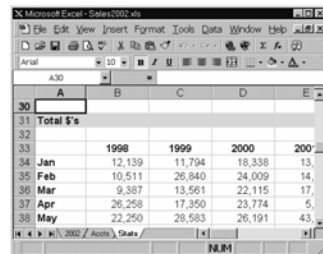
1.4 What do Programmers Do?

Don’t they write computer software? In brief, yes, of course. But this simple answer no longer explains what the computer programming (or, to sound more lofty, “software engineering”) profession has become. Software professionals create intricate and immensely complex invisible and intangible worlds. These worlds have their laws, their architectures, their aesthetics. These worlds also have “windows” into the “real” world. So many different skills are involved that computer programming has become one of the most versatile and demanding professions. We are not talking about specific technical skills, listed by cryptic acronyms in a typical employment ad for a software engineer. They are important, of course, but they are changing so rapidly that many programmers just learn them on the job. What we are talking about are the core skills that define the profession.

The most important skill is, probably, the ability to absorb new technical information all the time and to put it to work. In no other technical field are things changing so fast. Some of the changes are driven by new hardware and the new capabilities it provides; others by user demands, and some simply by fashion.

A programmer also needs to be an architect. Computers manipulate data and follow specific instructions written by a programmer. To make a computer perform useful tasks, a programmer must represent both the data and the instructions in coherent structures and procedures. Consider, for example, a word processor application. Before you proceed to write code for it, you need to have an abstract model of a “document.” You consider: a document may have text, drawings, images, links, and so on. The user has to be able to edit a document, cut and paste pieces, and so on. Each of these elements and procedures must be reflected in the structure of your program. Sound software design makes it easier to implement and test the program, modify it when necessary, share the work among a team of programmers, and reuse pieces of the software in future projects.

Programmers are not building cities, bridges, or cathedrals, of course, and you won’t find pieces of computer software exhibited in a museum. Very few people in the world will be able to examine the software structures that a programmer creates. But if you could enter and walk around software systems, you would find all kinds of things, from rather elegant structures to huge monstrosities, often barely standing, with pieces held together by the software equivalent of string and duct tape. A small change in one place upsets the whole structure; more strings and tape and props become necessary to prevent the contraption from collapsing. Perhaps these specimens do belong in a museum after all! For better or worse, software remains mostly invisible, hidden behind thousands of lines of code. It is mysterious, sometimes even full of surprises (Figure 1-4).



	1998	1999	2000	2001
Total \$'s				
Jan	12,139	11,794	16,338	13
Feb	10,511	26,840	24,009	14
Mar	9,387	13,561	22,115	17
Apr	26,258	17,350	23,774	5
May	22,250	28,583	26,191	43

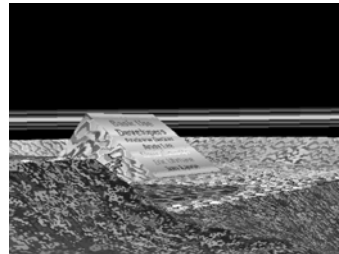


Figure 1-4. In *Excel 97*, if you pressed `Ctrl-G`, typed `L97:X97`, pressed `Enter`, pressed `Tab`, and then clicked on the “Chart Wizard” icon while holding the `Shift` and `Ctrl` keys down, a toy *Flight Simulator* popped up. Expert flying eventually revealed a monolith with the names of the product development team scrolling on it. Such a hidden “feature” is called an “Easter Egg” [1].

Another important skill: a programmer needs to be able to understand *algorithms* and devise his own. An algorithm is a more or less abstract, formal, and general step-by-step recipe that tells how to perform a certain task or solve a certain problem on a computer. We will discuss algorithms in Chapter 4.

The next programmer's skill is using programming languages. People often refer to programmers as "a C++ programmer," "a Java programmer," and so on. They think that knowing a particular programming language is what the programming profession is about. This is not so. First, a programmer is likely to encounter several programming languages over his or her career. Second, as we have seen, a programming job has many different aspects, and mastering a particular language is just one of them.

Still, ultimately a programmer needs to be able to express his or her brilliant structural designs and clever algorithms in a written program. A program written in a high-level language must obey the very precise *syntax rules* of that language and must also follow less precise but as important stylistic conventions established among professionals. We discuss Java syntax and style in Chapter 5.

But even when a program *compiles* with no errors, there is no guarantee that it works as desired. It may show unexpected behavior, produce incorrect output, or just crash. The programmer then has to figure out what is wrong and fix the problem. This takes yet another important skill: solving mysteries. A programmer must analyze the behavior of the program, hypothesize what may be wrong, either in the algorithm or in the code, then verify his hypothesis and make corrections. Of course, there are software tools, called *debuggers*, that help programmers locate *bugs* (errors) in their programs. But it would be too slow and tedious to use a debugger for every little error. Like a good doctor, a good programmer must develop an intuition for diagnosing and locating bugs quickly.

Believe it or not, there was time when programmers wrote their programs on paper and sent them to keypunch operators who would put them on *punch cards* [1, 2]. Now programmers use sophisticated *software development tools*. Being able to use these tools proficiently is another skill in the programmer's skill set.

In addition to general tools, a development environment for a particular programming language includes libraries of pre-programmed and pre-tested reusable software modules. There is no way to remember their technical details, of course. For example, in the 5.0 release, the standard Java library documentation consisted of 10,779 files in 711 folders and took 245 MB of disk space (which is roughly equivalent to 120,000 pages). Luckily the documentation is well organized and available online, so programmers can quickly find the information they need. But like a librarian, a programmer must know where and how to look.

Sometimes programmers forget that there are other people in the world who are not programmers, and that programs are written mostly for those other people, “users.” A programmer must make sure that his or her software is “user-friendly,” that is, easy to use and appears logical and intuitive to a non-technical user. Designing a friendly and effective *user interface* is another very important programming skill. A good user interface designer knows how to lay out screens, use colors, sound, animation, menus, keyboard commands, printouts and reports, and so on. The user interface is the only part of your software that is visible to the outside world, and this is often how your program is judged by non-technical users.

Last but not least, programmers must be familiar with the ethical codes for their profession and strive to uphold the highest professional ethical standards (see Appendix F).

To summarize, these are some of the things that a programmer must be able to do:

- Absorb and use emerging technical information
- Create sound software system architectures
- Understand and devise effective algorithms
- Be proficient with the syntax and style of programming languages
- Diagnose and correct programming errors
- Use software development tools and documentation
- Find and utilize reusable software components
- Design and implement friendly user interfaces
- Uphold the highest standards of professional ethics

Of course, when we said “A programmer must be able to do this; a programmer should know that,” we meant an abstract “programmer” who does not exist in the real world. In reality, a number of people are involved in every software development project, and different people may possess different skills. One person may specialize in designing system architectures, another person may be good at devising complex algorithms, a third person can quickly write and debug code, while someone else may not know much about programming but be an excellent designer of user interfaces. In the software development enterprise, there is a need for different types of work and room for people with different interests and skills.

The discipline that focuses on all aspects of software development is called *computer science*. A computer scientist takes a more theoretical view of one or another aspect of software development, looking for general ways to improve the efficiency of algorithms or for better software development methodologies and new programming

languages. Unlike the computer scientist, a software engineer (or programmer) works on specific software development projects.

1.5 Representation of Information in Computer Memory

Computer memory is a uniform array of bytes that does not privilege any particular type of information. The memory can contain CPU instructions, numbers and text characters, and any other information that can be represented in digital form. Since a suitable analog/digital (A/D) converter can fairly accurately convert any electrical signal to digital form, any information that can be carried over a wire can be represented in computer memory. This includes sounds, images, motion, and so on (but, so far, excludes taste and smell).

CPU instructions are represented in computer memory in a manner specific to each particular brand of CPU. The first byte or two represent the operation code that identifies the instruction and the total number of bytes in that instruction; the following bytes may represent the values or memory addresses of the operands. How memory addresses are represented depends on the CPU architecture, but they are basically numbers that indicate the absolute sequential number of the byte in memory. A CPU may have special *segment* registers and *index* registers that help calculate the actual address in memory for a specified instruction or operand.

The format for numbers is mostly dictated by the CPU, too, because the CPU has instructions for arithmetic operations that expect numbers to be represented in a certain way. Characters (letters, digits, etc.) are represented using one of the several character code systems that have become standard not only for representing text inside computers but also in computer terminals, printers, and other devices. The code assigns each character a number.

Fortunately, high-level programming languages, such as Java, shield computer programmers from the intricacies of how to represent CPU instructions, memory addresses, numbers, and characters.

Representing other types of information is often a matter of a specific application's design. A black-and-white image, for example, may be represented as a sequence of bytes where each bit represents a pixel of the image: 0 for white and 1 for black. The sequence of pixels typically goes from left to right along each horizontal line of the image and then from top to bottom by row. Other memory locations may hold numbers that represent the image dimensions.

1.5.1 Numbers

Integers from 0 to 255 can be represented in one byte using the binary (base-2) system as follows:

Decimal	Binary
0	00000000
1	00000001
2	00000010
3	00000011
4	00000100
5	00000101
6	00000110
7	00000111
8	00001000
9	00001001
10	00001010
...	...
252	11111100
253	11111101
254	11111110
255	11111111

If we use 2 bytes (16 bits), we can represent integers from 0 to $2^{16}-1 = 65535$:

Decimal	Binary
0	00000000 00000000
1	00000000 00000001
2	00000000 00000010
...	...
65534	11111111 11111110
65535	11111111 11111111

In general, k bits can produce 2^k different combinations of 0s and 1s. Therefore, k bits used as binary digits can represent non-negative integers in the range from 0 to 2^k-1 . A 32-bit memory address can identify $2^{32} = 4,294,967,296$ different memory locations. So if we want to be able to address each individual byte, 32-bit addresses cover 4 GB of memory space.



CPUs perform all their arithmetic operations on binary numbers. A CPU may have instructions that perform 8-bit, 16-bit, or 32-bit arithmetic, for instance. Since it is difficult for a human brain to grasp long sequences of 0s and 1s, programmers who have to deal with binary data often use the *hexadecimal* (or simply “*hex*”) representation in their documentation and programs. The hex system is the base-16 system, which uses 16 digits. The first ten digits are the usual 0 through 9, with the eleventh through sixteenth digits represented by the letters A through F. A byte can be split into two four-bit *quads*; each quad represents one hex digit, as follows:

Decimal	Binary	Hex
0	0000	0
1	0001	1
2	0010	2
3	0011	3
4	0100	4
5	0101	5
6	0110	6
7	0111	7
8	1000	8
9	1001	9
10	1010	A
11	1011	B
12	1100	C
13	1101	D
14	1110	E
15	1111	F

Experienced programmers remember the bit patterns for the sixteen hex digits and can easily convert a binary number into hex and back. It is often convenient to use hex representation as an intermediate step for converting numbers from binary to decimal and back. For example:

$$700_{10} = 2 \cdot 16_2 + 11 \cdot 16 + 12 = 2BC_{16} = \underbrace{0010}_{2} \underbrace{1011}_{B} \underbrace{1100}_{C}_2$$

$$\underbrace{1101101}_2 = 6D_{16} = 6 \cdot 16 + 13 = 109_{10}$$

6 D

The following examples show a few numbers represented in the decimal, hex, and 16-bit binary systems:

Decimal	Hex	Binary
0	0000	00000000 00000000
1	0001	00000000 00000001
12	000C	00000000 00001100
32	0020	00000000 00100000
128	0080	00000000 10000000
255	00FF	00000000 11111111
256	0100	00000001 00000000
32767	7FFF	01111111 11111111
32768	8000	10000000 00000000
65535	FFFF	11111111 11111111

For a software developer, knowing the hex system is a matter of cultural literacy. In practice, programmers who use a high-level programming language, like Java, don't have to use it very often and there are calculators and programs that can do conversions.



What about negative numbers? The same bit pattern may represent an unsigned (positive) integer and a negative integer, depending on how a particular instruction interprets it. Suppose we use 32-bit binary numbers, but now we decide that they represent signed integers. Positive integers from 0 to $2^{31}-1$ can be represented as before. These use only the 31 least significant bits. As to negative integers, their representation may be *machine-dependent*, varying from CPU to CPU. Many CPUs, including the Intel family, use a method called *two's-complement arithmetic*. In this method, a negative integer x in the range from -1 to -2^{31} is represented the same way as the unsigned binary number $2^{32} - |x|$ where $|x|$ is the absolute value of x . For example, 21 and -21 will be represented as:

```
00000000 00000000 00000000 00010101 = 2110
11111111 11111111 11111111 11101011 = -2110
```



Real numbers are represented using one of the standard formats expected by the CPU (or a separate floating-point arithmetic unit). Like scientific notation, this representation consists of a fractional part (mantissa) and an exponent part, but here both parts are represented as binary numbers. The IEEE (Institute of Electrical and Electronics Engineers) standard for a 4-byte (32-bit) representation uses 1 bit for the sign, 8 bits for the exponent and 23 bits for the mantissa. 127 is added to the exponent to ensure that negative exponents are still represented by non-negative numbers. This format lets programmers represent numbers in the range from approximately -3.4×10^{38} to 3.4×10^{38} with at least seven digits of precision. Figure 1-5 gives a few examples.

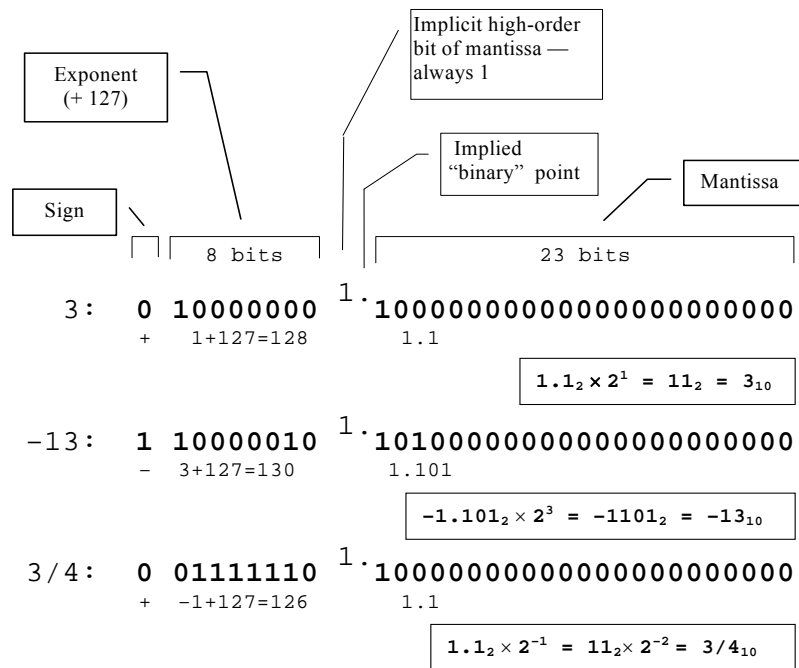


Figure 1-5. IEEE standard representation of 32-bit floating-point numbers

There are dozens of applets on the Internet that illustrate binary representation of numbers and conversions from decimal to binary [1, 2, 3, 4].

1.5.2 Characters

Characters are represented by numeric codes. The two most common character codes are EBCDIC (Extended Binary Coded Decimal Interchange Code) [1], used in IBM mainframes, and ASCII (American Standard Code for Information Interchange, pronounced as'-kee), used in personal computers, printers and other devices. Both of these use one byte per character. In the PC world, the term *ASCII file* refers to a text file (in which characters are represented in ASCII code), as opposed to a *binary file* that may contain numbers, images, or any other digitized information. Normally you won't find EBCDIC-encoded data on a PC unless the file originated on a mainframe.

ASCII code proper defines 128 characters with codes from 0 to 127 and uses only the seven least-significant bits of a byte. The second half of the code, from 128 to 255, is called *extended ASCII* and may vary from machine to machine. Codes from 33 to 127 represent “printable” characters: digits, upper- and lowercase letters, punctuation marks, and so on. 32 (hex 20) is a space.

The first 32 ASCII codes (0-31) are reserved for special control codes. For example, code 13 (hex 0D) is “carriage return” (CR), 10 (hex 0A) is “line feed” (LF), 12 (hex 0C) is “form feed” (FF) and 9 (hex 09) is “horizontal tab” (HT). How control codes are used may depend to some extent on the program or device that processes them. A standard ASCII table, including the more obscure control codes, is presented in Figure 1-6.

Many applications are switching to the platform-independent Unicode standard for encoding characters [1]. For example, the reason you can see the sentence written in Russian — Многие программы используют Юникод [1] — is that the software that displays this document understands the Unicode representation of Cyrillic characters.

Java programs use Unicode internally for representing characters and text strings.

Unicode uses two bytes per character. It includes ASCII codes as a subset (called “C0 Controls and Basic Latin” [1], but generally can encode up to 65,000 characters, enough to encode the alphabets of most world languages and many special characters [1]. Unicode has a provision to extend the character set even further, into millions of different codes.

Hex	0_	1_	2_	3_	4_	5_	6_	7_
_0	0 NUL	16 DEL	32 (SPACE)	48 0	64 @	80 P	96 ,	112 p
_1	1 SOH	17 DC1	33 !	49 1	65 A	81 Q	97 a	113 q
_2	2 STX	18 DC2	34 "	50 2	66 B	82 R	98 b	114 r
_3	3 ETX	19 DC3	35 #	51 3	67 C	83 S	99 c	115 s
_4	4 EOT	20 DC4	36 \$	52 4	68 D	84 T	100 d	116 t
_5	5 ENQ	21 NAK	37 %	53 5	69 E	85 U	101 e	117 u
_6	6 ACK	22 SYN	38 &	54 6	70 F	86 V	102 f	118 v
_7	7 BEL	23 ETB	39 '	55 7	71 G	87 W	103 g	119 w
_8	8 BS	24 CAN	40 (56 8	72 H	88 X	104 h	120 x
_9	9 HT	25 EM	41)	57 9	73 I	89 Y	105 i	121 y
_A	10 LF	26 SUB	42 *	58 :	74 J	90 Z	106 j	122 z
_B	11 VT	27 ESC	43 +	59 ;	75 K	91 [107 k	123 {
_C	12 FF	28 FS	44 ,	60 <	76 L	92 \	108 l	124
_D	13 CR	29 GS	45 -	61 =	77 M	93]	109 m	125 }
_E	14 SO	30 RS	46 .	62 >	78 N	94 ^	110 n	126 ~
_F	15 SI	31 US	47 /	63 ?	79 O	95 _	111 o	127 (NUL)

Figure 1-6. ASCII code used in personal computers and printers

1.6 The Internet

You basically need three things to make computers talk to each other:

1. A wire (or in more recent technology, a radio) link between them;
2. Hardware adapters at each *host* (the term used for a computer connected to a network), switches at the network junctions, and other hardware that controls the network;
3. A common language, called a *protocol*, so that the hosts can understand each other.

Of the three, the protocol is probably the most important. Once a reliable and flexible protocol is designed, the hardware and the connections will somehow follow and fall into place. Actually a protocol standard defines hundreds, even thousands of distinct protocols that function at many different levels, or, as network designers say, layers. (Figure 1-7).

The protocols in the bottom layer deal directly with the hardware — the network technology itself and various devices that are connected to it: modems, switches, high-speed adapters, routers, and so on. These hardware protocols are constantly evolving due to changing and competing new technologies and standards set by manufacturers and professional organizations.

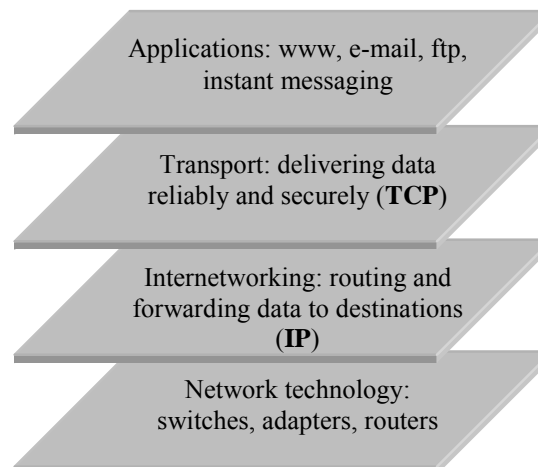


Figure 1-7. TCP/IP protocols in layered network architecture

The next layer deals with routing and forwarding: how to make sure that the information from one host eventually reaches another host. These internetworking protocols must know about the general layout of the network (who is connected to whom), and be efficient and robust for reliable connections. The Internet's internetworking layer is called *IP* (the *Internet Protocol*).

The layer above internetworking, the transport protocol, is responsible for properly handling information from specific applications used on the network and for meeting the requirements of these applications: reliability, security, data compression, and so on. *TCP* (the *Transmission Control Protocol*) is the Internet's transport protocol. The *TCP/IP* combination is what really defines the Internet.

Finally, at the very top layer, there are protocols for network applications such as e-mail (using SMTP — *Simple Mail Transfer Protocol*), the World Wide Web (using HTTP — *HyperText Transfer Protocol*), file transfer (using FTP — *File Transfer Protocol*), instant messaging, remote terminal emulation (*telnet*), and other applications.

If we had to set a birth date for the Internet, it would probably be early September 1969 when the first computer network was tested. The network had only four nodes: University of California in Los Angeles (UCLA), Stanford Research Institute (SRI), University of California in Santa Barbara (UCSB), and the University of Utah in Salt Lake City. Here is how Dr. Leonard Kleinrock [1], a computer science professor at UCLA and one of the Internet pioneers, describes the event. Kleinrock and his group of graduate students hoped to log onto the Stanford computer and try to send it some data. They would start by typing “login” and seeing if the letters appeared on the remote terminal.

“We set up a telephone connection between us and the guys at SRI...We typed the L and we asked on the phone, “Do you see the L?” “Yes, we see the L,” came the response. “We typed the O, and we asked, “Do you see the O.” “Yes, we see the O.” Then we typed the G, and the system crashed! Yet a revolution had begun.” [1, 2]

More precisely, the world's first instance of host-to-host, packet-switched data communications between networked computers had taken place.

The Internet has certainly made a lot of headway since. By 2006, over 204 million people in the US (68%) and over 1 billion people worldwide were online [1]. The best way to explore the Internet and its history is certainly not through a book, but by getting online and browsing [1, 2]. A *browser* is a program that helps its user navigate through the Internet and presents the information that comes from the Internet back to the user. *Internet Explorer* and *Firefox* are the two most popular browsers [1]. Information comes from the Internet in many different formats. The most common one is HTML documents — text files with embedded formatting tags

in them. You can find an HTML tutorial in Appendix C. Other files use standard formats for representing images, sounds, and so on. A browser and its helper modules, called *plug-ins*, know how to handle different formats of data and show (or play) the data to the user.

In a matter of just a few years the Internet has become a vast repository of knowledge and information (and misinformation) of all kinds and from all sources. It would be very difficult to find anything in this ocean without some guidance. *Portals* are popular web sites that arrange a large number of Internet links by category and help users navigate to the subjects they need. *Search engines* are programs based on large Internet hosts that analyze and index the contents of web pages and find web sites relevant to the keywords a user enters.

1.7 Summary

Digital electronics represents information using two states: “on” or “off,” “high” or “low,” “1” or “0.” Digital devices, called gates, implement simple logical operations on signals: AND, OR, NOT. All other logical and arithmetic operations can be implemented using these three simple operations.

The heart of a computer is a CPU (central processing unit) that can perform logical and arithmetic operations. An executable program is a sequence of CPU instructions in machine code. It must be loaded into RAM (random-access memory) before it can run. All instructions and data addresses are encoded in binary sequences of 0s and 1s. The CPU fetches instructions and data from RAM, interprets operation codes, and executes the instructions.

RAM is arranged into bytes; each byte is 8 bits; each bit can hold either 1 or 0. The size of RAM is measured in kilobytes ($1 \text{ KB} = 2^{10} = 1024$ bytes), megabytes ($1 \text{ MB} = 2^{20} = 1,048,576$ bytes), or gigabytes (1 GB is over one billion bytes). The contents of RAM are erased when the power is turned off.

Mass storage devices have a larger memory capacity — many gigabytes — and they can hold the information permanently, but access to them is slower. Data in mass storage is arranged into files by the operating system software. The files are stored in a branching system of directories (represented as folders). The operating system also loads and runs applications, provides a GUI (graphical user interface) to users, and provides system services to programs (such as reading and writing files, supporting input devices, etc.).

All kinds of information are represented in the computer memory as sequences of 0s and 1s. Integers are represented as binary numbers. Real numbers use standard

floating-point binary representation. Characters are represented as their codes in one of the standard coding systems. The most common codes are ASCII and Unicode. The latter includes thousands of characters from virtually all world alphabets.

The Internet is a network of millions of computers connected in many different ways. The Internet is based on the TCP/IP (Transmission Control Protocol / Internet Protocol), which controls information routing on the network and supports various network applications. Higher-level protocols are used in Internet applications such as e-mail, the World Wide Web, file transfer, and remote terminal emulation. A browser is a program on your computer that processes your requests for Internet connections and delivers and displays the received Internet information, web pages in particular. Portals are popular web sites that list many Internet links, arranged by category. Search engines are Internet indexing services that collect and index information from the Internet and help you find web sites relevant to your requests.

Exercises

The exercises for this chapter are in the book (*Java Methods A and AB: Object-Oriented Programming and Data Structures, AP Edition*, ISBN 0-9727055-7-0, Skylight Publishing, 2006 [[1](#)]).