Computer Systems (COVER PAGE TITLE)

Introduction

Computer once meant a person who did computations, but now the term almost universally refers to automated electronic machinery. The first section of this article focuses on modern digital electronic computers and their design, constituent parts, and applications. The second section covers the history of computing. Details on computer architecture, software, and theory would be included.

Abstract

Basic hardware and software components of computers are described and related concepts are defined using examples related to the hospital pharmacy department. Hardware components described include the central processing unit, storage devices, and input and output devices. System software, application software, computer language, and computer characteristics for hospital pharmacy systems are identified. Trends toward computer miniaturization, increased processing capacity at reduced costs, and new software technologies will enable an accelerated pace of new system development and implementation. Applications that are now difficult to justify because of high development cost or low benefit to the hospital will become more easily cost justified.

Computing Basics (Heading One)

The first computers were used primarily for numerical calculations. However, as any information can be numerically encoded, people soon realized that computers are capable of general-purpose information processing. Their capacity to handle large amounts of data has extended the range and accuracy of weather forecasting. Their speed has allowed them to make decisions about routing telephone connections through a network and to control mechanical systems such as automobiles, nuclear reactors, and robotic surgical tools. They are also cheap enough to be embedded in everyday appliances and to make clothes dryers and rice cookers “smart.” Computers have allowed us to pose and answer questions that could not be pursued before. These questions might be about DNA sequences in genes, patterns of activity in a consumer market, or all the uses of a word in texts that have been stored in a database. Increasingly, computers can also learn and adapt as they operate.

Computers also have limitations, some of which are theoretical. For example, there are undecidable propositions whose truth cannot be determined within a given set of rules, such as the logical structure of a computer. Because no universal algorithmic method can exist to identify such propositions, a computer asked to obtain the truth of such a proposition will (unless forcibly interrupted) continue indefinitely—a condition known as the “halting problem.” (See Turing machine.) Other limitations reflect current technology. Human minds are skilled at recognizing spatial patterns—easily distinguishing among human faces, for instance—but this is a difficult task for computers, which must process information sequentially, rather than grasping details overall at a glance. Another problematic area for computers involves natural language interactions. Because so much common knowledge and contextual information is assumed in ordinary human communication, researchers have yet to solve the problem of providing relevant information to general-purpose natural language programs.

Analog computers (Heading One)

Analog computers use continuous physical magnitudes to represent quantitative information. At first they represented quantities with mechanical components (see differential analyzer and integrator), but after World War II voltages were used; by the 1960s digital computers had largely replaced them. Nonetheless, analog computers, and some hybrid digital-analog systems, continued in use through the 1960s in tasks such as aircraft and spaceflight simulation.

One advantage of analog computation is that it may be relatively simple to design and build an analog computer to solve a single problem. Another advantage is that analog computers can frequently represent and solve a problem in “real time”; that is, the computation proceeds at the same rate as the system being modeled by it. Their main disadvantages are that analog representations are limited in precision—typically a few decimal places but fewer in complex mechanisms—and general-purpose devices are expensive and not easily programmed.

Digital computers (Heading One)

In contrast to analog computers, digital computers represent information in discrete form, generally as sequences of 0s and 1s (binary digits, or bits). The modern era of digital computers began in the late 1930s and early 1940s in the United States, Britain, and Germany. The first devices used switches operated by electromagnets (relays). Their programs were stored on punched paper tape or cards, and they had limited internal data storage. For historical developments, see the section Invention of the modern computer.

Mainframe computer (Heading two)

During the 1950s and ’60s, Unisys (maker of the UNIVAC computer), International Business Machines Corporation (IBM), and other companies made large, expensive computers of increasing power. They were used by major corporations and government research laboratories, typically as the sole computer in the organization. In 1959 the IBM 1401 computer rented for $8,000 per month (early IBM machines were almost always leased rather than sold), and in 1964 the largest IBM S/360 computer cost several million dollars.

These computers came to be called mainframes, though the term did not become common until smaller computers were built. Mainframe computers were characterized by having (for their time) large storage capabilities, fast components, and powerful computational abilities. They were highly reliable, and, because they frequently served vital needs in an organization, they were sometimes designed with redundant components that let them survive partial failures. Because they were complex systems, they were operated by a staff of systems programmers, who alone had access to the computer. Other users submitted “batch jobs” to be run one at a time on the mainframe.

Such systems remain important today, though they are no longer the sole, or even primary, central computing resource of an organization, which will typically have hundreds or thousands of personal computers (PCs). Mainframes now provide high-capacity data storage for Internet servers, or, through time-sharing techniques, they allow hundreds or thousands of users to run programs simultaneously. Because of their current roles, these computers are now called servers rather than mainframes.

Supercomputer (Heading two)

The most powerful computers of the day have typically been called supercomputers. They have historically been very expensive and their use limited to high-priority computations for government-sponsored research, such as nuclear simulations and weather modeling. Today many of the computational techniques of early supercomputers are in common use in PCs. On the other hand, the design of costly, special-purpose processors for supercomputers has been supplanted by the use of large arrays of commodity processors (from several dozen to over 8,000) operating in parallel over a high-speed communications network.

Minicomputer (Heading two)

Although minicomputers date to the early 1950s, the term was introduced in the mid-1960s. Relatively small and inexpensive, minicomputers were typically used in a single department of an organization and often dedicated to one task or shared by a small group. Minicomputers generally had limited computational power, but they had excellent compatibility with various laboratory and industrial devices for collecting and inputting data.

One of the most important manufacturers of minicomputers was Digital Equipment Corporation (DEC) with its Programmed Data Processor (PDP). In 1960 DEC’s PDP-1 sold for $120,000. Five years later its PDP-8 cost $18,000 and became the first widely used minicomputer, with more than 50,000 sold. The DEC PDP-11, introduced in 1970, came in a variety of models, small and cheap enough to control a single manufacturing process and large enough for shared use in university computer centres; more than 650,000 were sold. However, the microcomputer overtook this market in the 1980s.

Microcomputer (Heading two)

A microcomputer is a small computer built around a microprocessor integrated circuit, or chip. Whereas the early minicomputers replaced vacuum tubes with discrete transistors, microcomputers (and later minicomputers as well) used microprocessors that integrated thousands or millions of transistors on a single chip. In 1971 the Intel Corporation produced the first microprocessor, the Intel 4004, which was powerful enough to function as a computer although it was produced for use in a Japanese-made calculator. In 1975 the first personal computer, the Altair, used a successor chip, the Intel 8080 microprocessor. Like minicomputers, early microcomputers had relatively limited storage and data-handling capabilities, but these have grown as storage technology has improved alongside processing power.

(IMAGEONE) Image one caption “personal computer and peripherals”

In the 1980s it was common to distinguish between microprocessor-based scientific workstations and personal computers. The former used the most powerful microprocessors available and had high-performance colour graphics capabilities costing thousands of dollars. They were used by scientists for computation and data visualization and by engineers for computer-aided engineering. Today the distinction between workstation and PC has virtually vanished, with PCs having the power and display capability of workstations.

Embedded processors (Heading two)

Another class of computer is the embedded processor. These are small computers that use simple microprocessors to control electrical and mechanical functions. They generally do not have to do elaborate computations or be extremely fast, nor do they have to have great “input-output” capability, and so they can be inexpensive. Embedded processors help to control aircraft and industrial automation, and they are common in automobiles and in both large and small household appliances. One particular type, the digital signal processor (DSP), has become as prevalent as the microprocessor. DSPs are used in wireless telephones, digital telephone and cable modems, and some stereo equipment.

Computer hardware (Heading One)

The physical elements of a computer, its hardware, are generally divided into the central processing unit (CPU), main memory (or random-access memory, RAM), and peripherals. The last class encompasses all sorts of input and output (I/O) devices: keyboard, display monitor, printer, disk drives, network connections, scanners, and more.

The CPU and RAM are integrated circuits (ICs)—small silicon wafers, or chips, that contain thousands or millions of transistors that function as electrical switches. In 1965 Gordon Moore, one of the founders of Intel, stated what has become known as Moore’s law: the number of transistors on a chip doubles about every 18 months. Moore suggested that financial constraints would soon cause his law to break down, but it has been remarkably accurate for far longer than he first envisioned. It now appears that technical constraints may finally invalidate Moore’s law, since sometime between 2010 and 2020 transistors would have to consist of only a few atoms each, at which point the laws of quantum physics imply that they would cease to function reliably.

(IMAGETWO) Image two caption “Moore's law”

Central processing unit (Heading two)

The CPU provides the circuits that implement the computer’s instruction set—its machine language. It is composed of an arithmetic-logic unit (ALU) and control circuits. The ALU carries out basic arithmetic and logic operations, and the control section determines the sequence of operations, including branch instructions that transfer control from one part of a program to another. Although the main memory was once considered part of the CPU, today it is regarded as separate. The boundaries shift, however, and CPU chips now also contain some high-speed cache memory where data and instructions are temporarily stored for fast access.

The ALU has circuits that add, subtract, multiply, and divide two arithmetic values, as well as circuits for logic operations such as AND and OR (where a 1 is interpreted as true and a 0 as false, so that, for instance, 1 AND 0 = 0; see Boolean algebra). The ALU has several to more than a hundred registers that temporarily hold results of its computations for further arithmetic operations or for transfer to main memory.

The circuits in the CPU control section provide branch instructions, which make elementary decisions about what instruction to execute next. For example, a branch instruction might be “If the result of the last ALU operation is negative, jump to location A in the program; otherwise, continue with the following instruction.” Such instructions allow “if-then-else” decisions in a program and execution of a sequence of instructions, such as a “while-loop” that repeatedly does some set of instructions while some condition is met. A related instruction is the subroutine call, which transfers execution to a subprogram and then, after the subprogram finishes, returns to the main program where it left off.

In a stored-program computer, programs and data in memory are indistinguishable. Both are bit patterns—strings of 0s and 1s—that may be interpreted either as data or as program instructions, and both are fetched from memory by the CPU. The CPU has a program counter that holds the memory address (location) of the next instruction to be executed. The basic operation of the CPU is the “fetch-decode-execute” cycle:

(THE FOLLOWING POINTS SHOULD BE IN BULLETS LIST)

Fetch the instruction from the address held in the program counter, and store it in a register.

Decode the instruction. Parts of it specify the operation to be done, and parts specify the data on which it is to operate. These may be in CPU registers or in memory locations. If it is a branch instruction, part of it will contain the memory address of the next instruction to execute once the branch condition is satisfied.

Fetch the operands, if any.

Execute the operation if it is an ALU operation.

Store the result (in a register or in memory), if there is one.

Update the program counter to hold the next instruction location, which is either the next memory location or the address specified by a branch instruction.

(END OF THE BULLET LIST)

At the end of these steps the cycle is ready to repeat, and it continues until a special halt instruction stops execution.

Steps of this cycle and all internal CPU operations are regulated by a clock that oscillates at a high frequency (now typically measured in gigahertz, or billions of cycles per second). Another factor that affects performance is the “word” size—the number of bits that are fetched at once from memory and on which CPU instructions operate. Digital words now consist of 32 or 64 bits, though sizes from 8 to 128 bits are seen.

Processing instructions one at a time, or serially, often creates a bottleneck because many program instructions may be ready and waiting for execution. Since the early 1980s, CPU design has followed a style originally called reduced-instruction-set computing (RISC). This design minimizes the transfer of data between memory and CPU (all ALU operations are done only on data in CPU registers) and calls for simple instructions that can execute very quickly. As the number of transistors on a chip has grown, the RISC design requires a relatively small portion of the CPU chip to be devoted to the basic instruction set. The remainder of the chip can then be used to speed CPU operations by providing circuits that let several instructions execute simultaneously, or in parallel.

There are two major kinds of instruction-level parallelism (ILP) in the CPU, both first used in early supercomputers. One is the pipeline, which allows the fetch-decode-execute cycle to have several instructions under way at once. While one instruction is being executed, another can obtain its operands, a third can be decoded, and a fourth can be fetched from memory. If each of these operations requires the same time, a new instruction can enter the pipeline at each phase and (for example) five instructions can be completed in the time that it would take to complete one without a pipeline. The other sort of ILP is to have multiple execution units in the CPU—duplicate arithmetic circuits, in particular, as well as specialized circuits for graphics instructions or for floating-point calculations (arithmetic operations involving noninteger numbers, such as 3.27). With this “superscalar” design, several instructions can execute at once.

Both forms of ILP face complications. A branch instruction might render preloaded instructions in the pipeline useless if they entered it before the branch jumped to a new part of the program. Also, superscalar execution must determine whether an arithmetic operation depends on the result of another operation, since they cannot be executed simultaneously. CPUs now have additional circuits to predict whether a branch will be taken and to analyze instructional dependencies. These have become highly sophisticated and can frequently rearrange instructions to execute more of them in parallel.

Memory (Heading two)

Main memory (Heading three)

The earliest forms of computer main memory were mercury delay lines, which were tubes of mercury that stored data as ultrasonic waves, and cathode-ray tubes, which stored data as charges on the tubes’ screens. The magnetic drum, invented about 1948, used an iron oxide coating on a rotating drum to store data and programs as magnetic patterns.

In a binary computer any bistable device (something that can be placed in either of two states) can represent the two possible bit values of 0 and 1 and can thus serve as computer memory. Magnetic-core memory, the first relatively cheap RAM device, appeared in 1952. It was composed of tiny, doughnut-shaped ferrite magnets threaded on the intersection points of a two-dimensional wire grid. These wires carried currents to change the direction of each core’s magnetization, while a third wire threaded through the doughnut detected its magnetic orientation.

The first integrated circuit (IC) memory chip appeared in 1971. IC memory stores a bit in a transistor-capacitor combination. The capacitor holds a charge to represent a 1 and no charge for a 0; the transistor switches it between these two states. Because a capacitor charge gradually decays, IC memory is dynamic RAM (DRAM), which must have its stored values refreshed periodically (every 20 milliseconds or so). There is also static RAM (SRAM), which does not have to be refreshed. Although faster than DRAM, SRAM uses more transistors and is thus more costly; it is used primarily for CPU internal registers and cache memory.

In addition to main memory, computers generally have special video memory (VRAM) to hold graphical images, called bitmaps, for the computer display. This memory is often dual-ported—a new image can be stored in it at the same time that its current data is being read and displayed.

It takes time to specify an address in a memory chip, and, since memory is slower than a CPU, there is an advantage to memory that can transfer a series of words rapidly once the first address is specified. One such design is known as synchronous DRAM (SDRAM), which became widely used by 2001.

Nonetheless, data transfer through the “bus”—the set of wires that connect the CPU to memory and peripheral devices—is a bottleneck. For that reason, CPU chips now contain cache memory—a small amount of fast SRAM. The cache holds copies of data from blocks of main memory. A well-designed cache allows up to 85–90 percent of memory references to be done from it in typical programs, giving a several-fold speedup in data access.

The time between two memory reads or writes (cycle time) was about 17 microseconds (millionths of a second) for early core memory and about 1 microsecond for core in the early 1970s. The first DRAM had a cycle time of about half a microsecond, or 500 nanoseconds (billionths of a second), and today it is 20 nanoseconds or less. An equally important measure is the cost per bit of memory. The first DRAM stored 128 bytes (1 byte = 8 bits) and cost about $10, or $80,000 per megabyte (millions of bytes). In 2001 DRAM could be purchased for less than $0.25 per megabyte. This vast decline in cost made possible graphical user interfaces (GUIs), the display fonts that word processors use, and the manipulation and visualization of large masses of data by scientific computers.

Secondary memory (Heading three)

Secondary memory on a computer is storage for data and programs not in use at the moment. In addition to punched cards and paper tape, early computers also used magnetic tape for secondary storage. Tape is cheap, either on large reels or in small cassettes, but has the disadvantage that it must be read or written sequentially from one end to the other.

IBM introduced the first magnetic disk, the RAMAC, in 1955; it held 5 megabytes and rented for $3,200 per month. Magnetic disks are platters coated with iron oxide, like tape and drums. An arm with a tiny wire coil, the read/write (R/W) head, moves radially over the disk, which is divided into concentric tracks composed of small arcs, or sectors, of data. Magnetized regions of the disk generate small currents in the coil as it passes, thereby allowing it to “read” a sector; similarly, a small current in the coil will induce a local magnetic change in the disk, thereby “writing” to a sector. The disk rotates rapidly (up to 15,000 rotations per minute), and so the R/W head can rapidly reach any sector on the disk.

Early disks had large removable platters. In the 1970s IBM introduced sealed disks with fixed platters known as Winchester disks—perhaps because the first ones had two 30-megabyte platters, suggesting the Winchester 30-30 rifle. Not only was the sealed disk protected against dirt, the R/W head could also “fly” on a thin air film, very close to the platter. By putting the head closer to the platter, the region of oxide film that represented a single bit could be much smaller, thus increasing storage capacity. This basic technology is still used.

(IMAGE THREE) Image three caption “computer hard drive”

Refinements have included putting multiple platters—10 or more—in a single disk drive, with a pair of R/W heads for the two surfaces of each platter in order to increase storage and data transfer rates. Even greater gains have resulted from improving control of the radial motion of the disk arm from track to track, resulting in denser distribution of data on the disk. By 2002 such densities had reached over 8,000 tracks per centimetre (20,000 tracks per inch), and a platter the diameter of a coin could hold over a gigabyte of data. In 2002 an 80-gigabyte disk cost about $200—only one ten-millionth of the 1955 cost and representing an annual decline of nearly 30 percent, similar to the decline in the price of main memory.

Optical storage devices—CD-ROM (compact disc, read-only memory) and DVD-ROM (digital videodisc, or versatile disc)—appeared in the mid-1980s and ’90s. They both represent bits as tiny pits in plastic, organized in a long spiral like a phonograph record, written and read with lasers. A CD-ROM can hold 2 gigabytes of data, but the inclusion of error-correcting codes (to correct for dust, small defects, and scratches) reduces the usable data to 650 megabytes. DVDs are denser, have smaller pits, and can hold 17 gigabytes with error correction.