CHAPTER 13
COMPUTER-AIDED DESIGN

Dr. Emory W. Zimmers, Jr., & Technical Staff
Enterprise Systems Center
Lehigh University
Bethlehem, PA

13.1 INTRODUCTION TO CAD

Computer-aided design (CAD) uses the mathematical and graphic-processing power of the computer to assist the engineer in the creation, modification, analysis, and display of designs. Many factors have contributed to CAD technology becoming a necessary tool in the engineering world, such as the computer's speed at processing complex equations and managing technical databases. CAD combines the characteristics of designer and computer that are best applicable to the design process.

The combination of human creativity with computer technology provides the design efficiency that has made CAD such a popular design tool. CAD is often thought of simply as computer-aided

Mechanical Engineers' Handbook, 2nd ed., Edited by Myer Kutz.
ISBN 0-471-13007-9

drafting, and its use as an electronic drawing board is a powerful tool in itself. The functions of a CAD system extend far beyond its ability to represent and manipulate graphics. Geometric modeling, engineering analysis, simulation, and the communication of the design information can also be performed using CAD.

13.1.1 A Historical Perspective of CAD

Graphical representation of data, in many ways, forms the basis of CAD. An early application of computer graphics was used in the SAGE (Semi-Automatic Ground Environment) Air Defense Command and Control System in the 1950s. SAGE converted radar information into computer-generated images on a cathode ray tube (CRT) display. It also used an input device, the light pen, to select information directly from the CRT screen.

Another significant advancement in computer graphics technology occurred in 1963, when Ivan Sutherland, in his doctoral thesis at MIT, described the SKETCHPAD system. The SKETCHPAD system was driven by a Lincoln TX-2 computer. With SKETCHPAD, images could be created and manipulated using the light pen. Graphical manipulations such as translation, rotation, and scaling could all be accomplished on-screen using SKETCHPAD. Computer applications based on Sutherland's approach have become known as interactive computer graphics (ICG). The graphical capabilities of SKETCHPAD showed the potential for computerized drawing in design. The high cost of computer hardware in the 1960s limited the use of ICG systems to large corporations, such as those in the automotive and aerospace industries, which could justify the initial investment. With the rapid development of computer technology, computers became more powerful, using faster processors and greater data storage capabilities. Their physical size and cost decreased, and computers became affordable to smaller companies and personal users. Today it is rare to find an engineering, design, or architectural firm of any size without a working CAD system running on a personal computer or a workstation.

13.1.2 The Design Process

Before any discussion of computer-aided design, it is necessary to understand the design process in general. What is the series of events that leads to the beginning of a design project? How does the engineer go about the process of designing something? How does one arrive at the conclusion that the design has been completed? We address these questions by defining the process (Fig. 13.1) in terms of six distinct stages:

1. Customer input and perception of need
2. Problem definition
3. Synthesis
4. Analysis and optimization
5. Evaluation
6. Final design and specification

A need is usually perceived in one of two ways. Someone must recognize either a problem in an existing design or a customer-driven opportunity in the marketplace for a new product. In either case, a need exists which can be addressed by modifying an existing design or developing an entirely new design. Because the need for change may only be indicated by subtle circumstances, such as noise, marginal performance characteristics, or deviations from quality standards, the design engineer who identifies the need has taken a first step in correcting the problem. That step sets in motion processes that may allow others to see the need more readily and possibly enroll them in the solution process.

Once the decision has been made to take corrective action to the need at hand, the problem must be defined as a particular problem to be solved such that all significant parameters in the problem are defined. These parameters often include cost limits, quality standards, size and weight characteristics, and functional characteristics. Often, specifications may be defined by the capabilities of the manufacturing process. Anything that will influence the engineer in choosing design features must be included in the definition of the problem. Careful planning in this stage can lead to fewer iterations in subsequent design stages.

Once the problem has been fully defined in this way, the designer moves on to the synthesis stage, where knowledge and creativity can be applied to conceptualize an initial design. Teamwork can make the design more successful and effective at this stage. That design is then subjected to various forms of analysis, which may reveal specific problems in the initial design. The designer then takes the analytical results and applies them in an iteration of the synthesis stage. These iterations may continue through several cycles of synthesis and analysis until the design is optimized.

The design is then evaluated according to the parameters set forth in the problem definition. A scale prototype is often fabricated to perform further analysis and to assess operating performance, quality, reliability, and other criteria. If a design flaw is revealed during this stage, the design moves back to the synthesis/analysis stages for reoptimization, and the process moves in this circular manner until the design clears the evaluative stage and is ready for presentation.

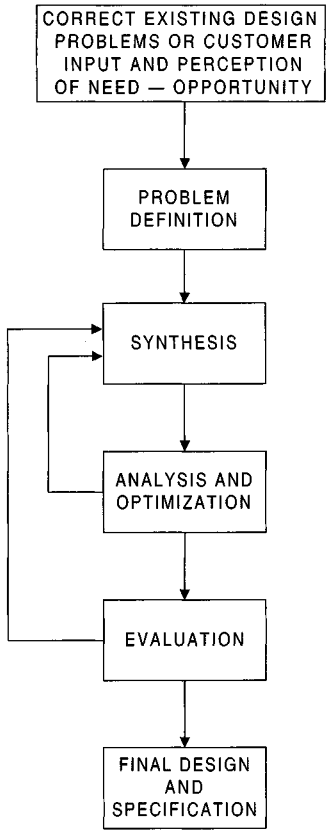

Fig. 13.1 The general design process.

Final design and specification represents the last stage of the design process. Communicating the design to others in such a way that its manufacture and marketing are seen as vital to the organization is essential. When the design has been fully approved, detailed engineering drawings are produced, complete with specifications for components, subassemblies, and the tools and fixtures required to manufacture the product and the associated costs of production. These can then be transferred manually or digitally, using CAD data, to the various departments responsible for manufacture.

In every branch of engineering, prior to the implementation of CAD, design has traditionally been accomplished manually on the drawing board. The resulting drawing, complete with significant details, was then subjected to analysis using complex mathematical formulae and then sent back to the drawing board with suggestions for improving the design. The same iterative procedure was followed and, because of the manual nature of the drawing and the subsequent analysis, the whole procedure was time-consuming and labor-intensive. CAD has allowed the designer to bypass much of the manual drafting and analysis that was previously required, making the design process flow more smoothly and much more efficiently.

It is helpful to understand the general product development process as a step-wise process. However, in today's engineering environment, the steps outlined above have become consolidated into a more streamlined approach called *concurrent engineering*. This approach enables teams to work concurrently by providing common ground for interrelated product development tasks. Product information can be easily communicated among all development processes: design, manufacturing, marketing, management, and supplier networks. Concurrent engineering recognizes that fewer iterations result in less time and money spent in moving from design concept to manufacture and from manufacturing to market. The related processes of Design for Manufacturing (DFM) and Design for Assembly (DFA) have become integral parts of the concurrent engineering approach.

Design for Manufacturing and Design for Assembly methods use cross-disciplinary input from a variety of sources (e.g., design engineers, manufacturing engineers, suppliers, and shop-floor representatives) to facilitate the efficient design of a product that can be manufactured, assembled, and marketed in the shortest possible period of time. Products designed using DFM and DFA are often simpler, cost less, and reach the marketplace in far less time than traditionally designed products. DFM focuses on determining what materials and manufacturing techniques will result in the most efficient use of available resources in order to integrate this information early in the design process. The DFA methodology strives to consolidate the number of parts wherever possible, uses gravity-assisted assembly techniques, and calls for careful review and consensus approval of designs early in the process. By facilitating the free exchange of information, DFM and DFA methods allow engineering companies to avoid the costly rework often associated with repeated iterations of the design process.

13.1.3 Applying Computers to Design

Many of the individual tasks within the overall design process can be performed using a computer. As each of these tasks is made more efficient, the efficiency of the overall process increases as well. The computer is especially well suited to design in four areas, which correspond to the latter four stages of the general design process. Computers function in the design process through geometric modeling capabilities, engineering analysis calculations, automated testing procedures, and automated drafting. Figure 13.2 illustrates the relationship between CAD technology and the final four stages of the design process.

Geometric modeling is one of the keystones of CAD systems. It uses mathematical descriptions of geometric elements to facilitate the representation and manipulation of graphical images on a computer display screen. While the central processing unit (CPU) provides the ability to quickly make the calculations specific to the element, the software provides the instructions necessary for efficient transfer of information between user and the CPU.

Three types of commands are used by the designer in computerized geometric modeling. The first type of command allows the user to input the variables needed by the computer to represent

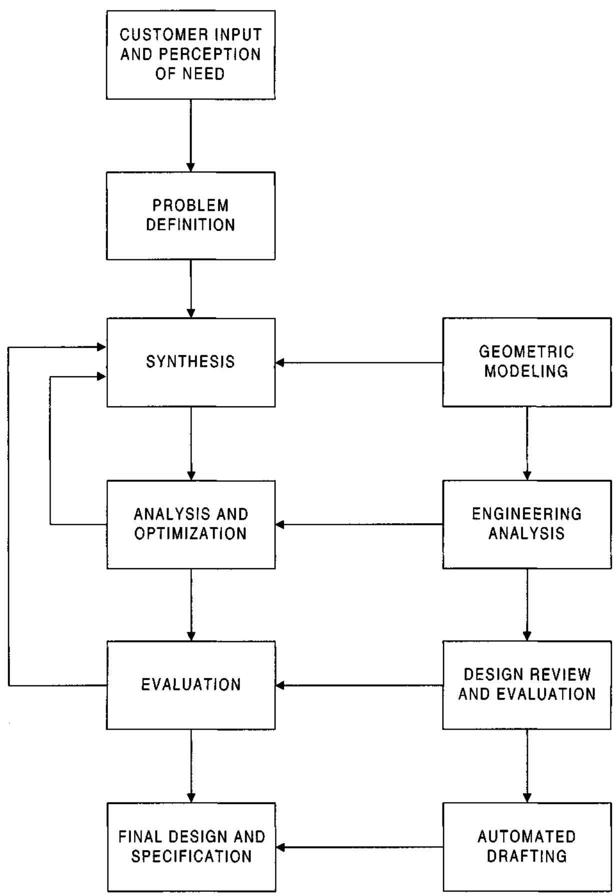

Fig. 13.2 Application of computers to the design process.

basic geometric elements such as points, lines, arcs, circles, splines, and ellipses. The second type of command is used to transform these elements. Commonly performed transformations in CAD include scaling, rotation, and translation. The third type of command allows the various elements previously created by the first two commands to be joined into a desired shape.

During the whole geometric modeling process, mathematical operations are at work that can be easily stored as computerized data and retrieved as needed for review, analysis, and modification. There are different ways of displaying the same data on the CRT screen, depending on the needs or preferences of the designer. One method is to display the design as a two-dimensional representation of a flat object formed by interconnecting lines. Another method displays the design as a three-dimensional representation of objects. In three-dimensional representations, there are four types of modeling approaches:

- Wireframe modeling
- Surface modeling
- Solid modeling
- Hybrid solid modeling

A *wireframe model* is a skeletal description of a three-dimensional object. It consists only of points, lines, and curves that describe the boundaries of the object. There are no surfaces in a wireframe model. Three-dimensional wireframe representations can cause the viewer some confusion because all of the lines defining the object appear on the two-dimensional display screen. This makes it hard for the viewer to tell whether the model is being viewed from above or below, inside or outside.

Surface modeling defines not only the edge of the three-dimensional object, but also its surface. In surface modeling, two different types of surfaces can be generated: faceted surfaces using a polygon mesh and true curve surfaces. NURBS (Non-Uniform Rational B-Spline) is a B-spline curve or surface defined by a series of weighted control points and one or more knot vectors. It can exactly represent a wide range of curves such as arcs and conics. The greater flexibility for controlling continuity is one advantage of NURBS. NURBS can precisely model nearly all kinds of surfaces more robustly than the polynomial-based curves that were used in earlier surface models. The surface modeling is more sophisticated than wireframe modeling. Here, the computer still defines the object in terms of a wireframe but can generate a surface "skin" to cover the frame, thus giving the illusion of a "real" object. However, because the computer has the image stored in its data as a wireframe representation having no mass, physical properties cannot be calculated directly from the image data. Surface models are very advantageous due to point-to-point data collections usually required for Numerical Control (NC) programs in computer-aided manufacturing (CAM) applications. Most surface modeling systems also produce the stereolithographic data required for rapid prototyping systems.

Solid modeling defines the surfaces of an object, with the added attributes of volume and mass. This allows image data to be used in calculating the physical properties of the final product. Solid modeling software uses one of two methods: constructive solid geometry (CSG) or boundary representation (B-rep). The CSG method uses Boolean operations (union, subtraction, intersection) on two sets of objects to define composite models. For example, a cylinder can be subtracted from a cube. B-rep is a representation of a solid model that defines an object in terms of its surface boundaries: faces, edges, and vertices.

Hybrid solid modeling allows the user to represent a part with a mixture of wireframe, surface modeling, and solid geometry. The I-DEAS Master Modeler offers this representation feature.

In CAD software, the hidden-line command can remove the background lines of the object in a model. Certain features have been developed to minimize the ambiguity of wireframe representations. These features include using dashed lines to represent the background of a view, or removing those background lines altogether. The latter method is appropriately referred to as *hidden-line removal*. The hidden-line removal feature makes it easier to visualize the model because the back faces are not displayed. Shading removes hidden lines and assigns flat colors to visible surfaces. Rendering adds and adjusts lights and materials to surfaces to produce realistic effects. Shading and rendering can greatly enhance the realism of the 3D image. Figures 13.3(*a*) and (*b*) show the same object, represented as a pure wireframe and a wireframe with hidden-line removal.

Engineering analysis can be performed using one of two approaches: analytical or experimental. Using the analytical method, the design is subjected to simulated conditions, using any number of analytical formulae. By contrast, the experimental approach to analysis requires that a prototype be constructed and subsequently subjected to various experiments to yield data that might not be available through purely analytical methods.

There are various analytical methods available to the designer using a CAD system. Finite element analysis and static and dynamic analysis are all commonly performed analytical methods available in CAD.

Finite element analysis (FEA) is a computer numerical analysis program (Fig. 13.4) used to solve the complex problems in many engineering and scientific fields, such as structural analysis (stress,

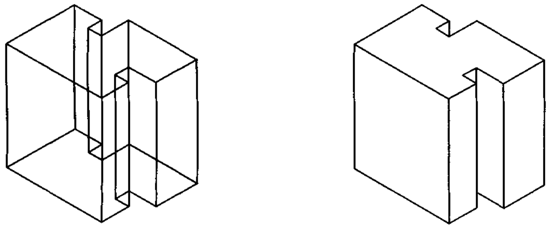

Fig. 13.3 (*a*) Pure wireframe model. (*b*) Wireframe model with hidden-line removal feature.

deflection, vibration), thermal analysis (steady state and transient), and fluid dynamics analysis (laminar and turbulent flow).

The finite element method divides a given physical or mathematical model into smaller and simpler elements, performs analysis on each individual element, using the required mathematics. It then assembles the individual solutions of the elements to reach a global solution for the model. FEA software programs usually consist of three parts: the preprocessor, the solver, and the postprocessor.

The program inputs are prepared in the preprocessor. Model geometry can be defined or imported from CAD software. Meshes are generated on a surface or solid model to form the elements. Element properties and material descriptions can be assigned to the model. Finally, the boundary conditions

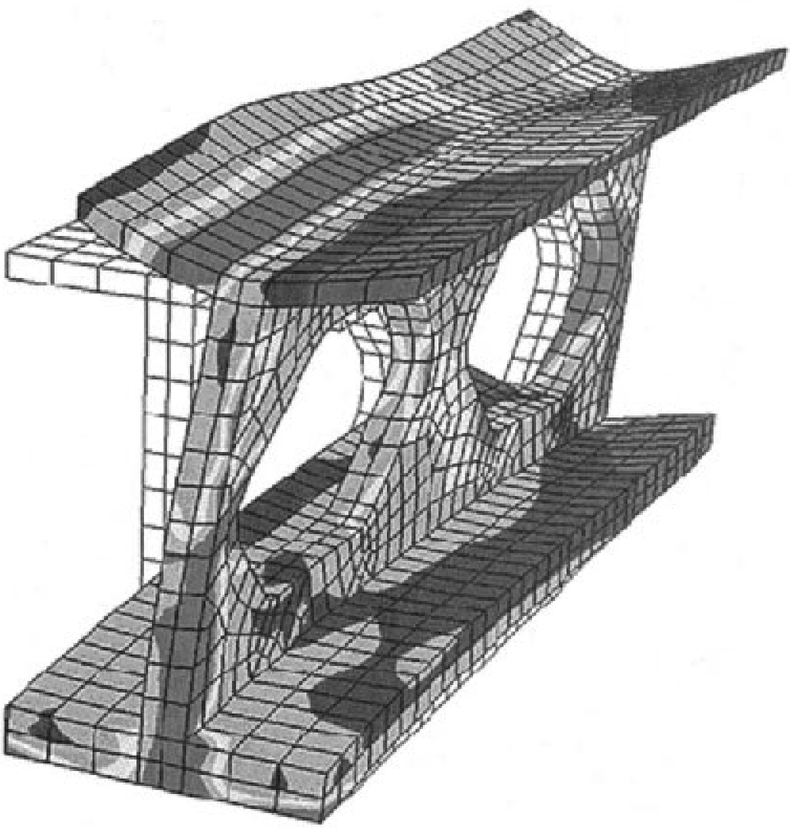

Fig. 13.4 Finite element analysis of random vibration in a beam. Colors or gray scales are often used to show degrees of stress and deflection. The original shape is also outlined without shading for reference (courtesy of Algor, Inc.).

and loads are applied to the elements and their nodes. Certain checks must be completed before the analysis calculation. These include checking for duplication of nodes and elements and verifying the element connectivity of the surface elements so that the surface normals are all in the same direction. In order to optimize disk space and running time, the nodes and elements should usually be renumbered and sequenced. Many analysis options are available in the analysis solver to execute the model. The element stiffness matrices can be formulated and solved to form a global stiffness value for the model solution. The results of the analysis data are then interpreted by the postprocessor in an orderly manner. The postprocessor in most FEA applications offers graphical output and animation displays. Many vendors of CAD software are also developing pre- and post processors that allow the user to visualize their input and output graphically. FEA is a powerful tool in effectively synthesizing a design into an optimized product.

Kinematic analysis and synthesis (Fig. 13.5) studies the motion or position of a set of rigid bodies in a system without reference to the forces causing that motion or the mass of the bodies. It allows engineers to see how the mechanisms they design will function in motion. This luxury enables the designer to avoid faulty designs and also to apply the design to a variety of scenarios without constructing a physical prototype. Synthesis of the data extracted from kinematic analysis in numerous iterations of the process leads to optimization of the design. The increased number of trials that kinematic analysis allows the engineer to perform may have profound results in optimizing the behavior of the resulting mechanism before actual production.

Static analysis determines reaction forces at the joint positions of resting mechanisms when a constant load is applied. As long as zero velocity is assumed, static analysis can be performed on mechanisms at different points of their range of motion. Static analysis allows the designer to determine the reaction forces on whole mechanical systems as well as interconnection forces transmitted to their individual joints. The data extracted from static analysis can be useful in determining compatibility with the various criteria set out in the problem definition. These criteria may include reliability, fatigue, and performance considerations to be analyzed through stress analysis methods.

Dynamic analysis combines motion with forces in a mechanical system to calculate positions, velocities, accelerations, and reaction forces on parts in the system. The analysis is performed stepwise within a given interval of time. Each degree of freedom is associated with a specific coordinate for which initial position and velocity must be supplied. The computer model from which the design

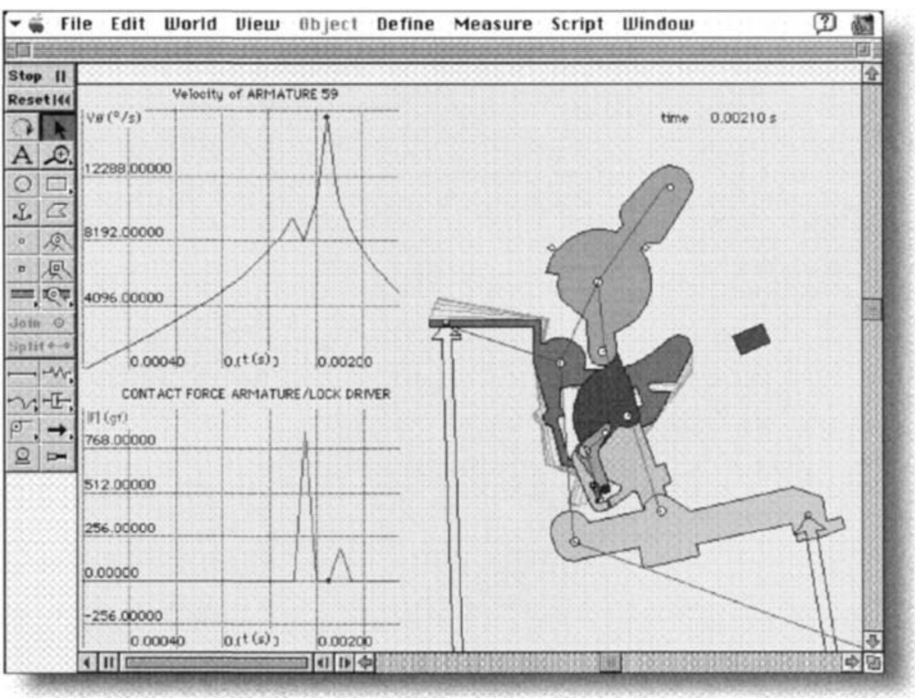

Fig. 13.5 Kinematic analysis of a switch mechanism (image courtesy of Knowledge Solutions, Inc.).

is analyzed is created by defining the system in various ways. Generally, data relating to individual parts, joints, forces, and overall system coordination must be supplied by the user, either directly or through a manipulation of data within the software.

The results of all of these types of analyses are typically available in many forms, depending on the needs of the designer. All of these analytical methods will be discussed in greater detail in Section 13.8.

Experimental analysis involves fabricating a prototype and subjecting it to various experimental methods. Although this usually takes place in the later stages of design, CAD systems enable the designer to make more effective use of experimental data, especially where analytical methods are thought to be unreliable for the given model. CAD also provides a useful platform for incorporating experimental results into the design process when experimental analysis is performed in earlier iterations of the process.

Design review can be easily accomplished using CAD. The accuracy of the design can be checked using automated tolerancing and dimensioning routines to reduce the possibility of error. Layering is a technique which allows the designer to superimpose images upon one another. This can be quite useful during the evaluative stage of the design process by allowing the designer to check the dimensions of a final design visually against the dimensions of stages of the design's proposed manufacture, ensuring that sufficient material is present in preliminary stages for correct manufacture.

Interference checking can also be performed using CAD. This procedure involves making sure that no two parts of a design occupy the same space at the same time.

Automated drafting capabilities in CAD systems facilitate presentation, which is the final stage of the design process. CAD data, stored in computer memory, can be sent to a pen plotter or other hard-copy device (see Section 13.6.2) to produce a detailed drawing quickly and easily. In the early days of CAD, this feature was the primary rationale for investing in a CAD system. Drafting conventions, including but not limited to dimensioning, crosshatching, scaling of the design, and enlarged views of parts or other design areas, can be included automatically in nearly all CAD systems. Detail and assembly drawings, bills of materials (BOM), and cross-sectioned views of design parts are also automated and simplified through CAD. In addition, most systems are capable of presenting as many as six views of the design automatically. Drafting standards defined by a company can be programmed into the system such that all final drafts will comply with the standard.

Documentation of the design is also simplified using CAD. Product Data Management (PDM) has become an important application associated with CAD. PDM allows companies to make CAD data available interdepartmentally on a computer network. This approach holds significant advantages over conventional data management. PDM is not simply a database holding CAD data as a library for interested users. PDM systems offer increased data management efficiency through a client-server relationship among individual computers and a networked server. Benefits of implementing a PDM system include faster retrieval of CAD files through keyword searches and other search features; automated distribution of designs to management, manufacturing engineers, and shop-floor workers for design review; recordkeeping functions that provide a history of design changes; and data security functions limiting access levels to design files (Fig. 13.6). PDM facilitates the exchange of information characteristic of the emerging agile workplace. As companies face increased pressure to provide clients with customized solutions to their individual needs, PDM systems allow an increased level of teamwork among personnel at all levels of product design and manufacturing, cutting the costs often associated with information lag and rework.

Although computer-aided design has made the design process less tedious and more efficient than traditional methods, the fundamental design process in general remains unchanged. It still requires human input and ingenuity to initiate and proceed through the many iterations of the process. Nevertheless, computer-aided design is such a powerful, time-saving design tool that it is now difficult to function in a competitive engineering world without such a system in place. The CAD system will now be examined in terms of its components: the hardware and software of a computer.

13.2 HARDWARE

Just as a draftsman traditionally requires pen and ink to bring creativity to bear on the page, there are certain essential components to any working CAD system. The use of computers for interactive graphics applications can be traced back to the early 1960s, when Ivan Sutherland developed the SKETCHPAD system. The prohibitively high cost of hardware made general use of interactive computer graphics uneconomical until the 1970s. With the development and subsequent popularity of personal computers, interactive graphics applications now are widespread in homes and workplaces.

CAD systems have become available for many hardware configurations. Most CAD systems have been developed for standard computer systems, ranging from mainframes to microcomputers. Others, like turnkey CAD systems, come with all of the hardware and software required to run a particular CAD application, and are supplied by specialized vendors.

13.2.1 Input/Output and Central Processing Unit (CPU)

The above systems all share a dependence on components that allow the actual interaction between computer and users. These electronic components are categorized under two general headings: input

Fig. 13.6 CAD files can be used in conjunction with other applications. The above illustration shows Integraph Corporation's Solid Edge software operating in conjunction with AutoCAD from Autodesk, Inc. and Microsoft Word (image courtesy of Integraph Corporation).

devices and output devices. Input devices transfer information from the designer into the computer's Central Processing Unit (CPU) so that the data, encoded in binary sequencing, may be manipulated and analyzed efficiently. Output devices do exactly the opposite. They transfer binary data from the CPU back to the user in a usable (usually visual) format. Both types of devices are required in a CAD system. Without an input device, no information can be transferred to the CPU for processing, and without an output device, any information in the CPU is of little use to the designer because binary code is lengthy and tedious.

13.3 THE COMPUTER

Although the influence of computer technology is a somewhat recent phenomenon due to the reduced cost of computers over the last two decades, the philosophical basis for the construction and employment of computing systems has a longer history than 20 years.

Charles Babbage, a nineteenth-century mathematician at Cambridge University in England, is often cited as a pioneer in the computing field. Babbage designed an "analytical engine," the capabilities of which would have surprisingly foreshadowed the same basic functions of today's computers had his design not been limited by the manufacturing capabilities of his time. The analytical engine was designed with considerations for input, storage, mathematical calculation, grouping results, and printing results in typeface. Other, less complex mechanical forms of computers include the slide rule and even the abacus.

The vast majority of contemporary computers are digital, although some analog computers do exist. This latter type has been relegated almost to a footnote in contemporary computing due to the overwhelming advances made in digital technology. The difference between digital and analog systems lies in the binary code. Digital computers use a system of switches with two settings, "on" or "off." These settings are typically represented as "0" for "off" and "1" for "on."

Although digital computers vary in size, shape, price, and capabilities, all digital computers have four common features. First, the circuits used can exist in one of two states, either "on" or "off." This characteristic yields the basis for binary logic. Second, all share the ability to store data in binary form. Third, all digital computers can receive external input data, perform various functions relating to that data, and provide the user with the output or result of the performed function. Finally, digital computers can all be operated through the use of instructions organized into sets of separate steps. On a related note, many digital systems possess the ability to perform many different functions at the same time, using a technique known as *parallel processing.*

13.3.1 Computer Evolution

Based on the advances leading to each stage of technological progress, computer systems have commonly been grouped into four generations:

- *First Generation:* Vacuum tube circuitry
- *Second Generation:* Transistors
- *Third Generation:* Small and medium integrated circuits
- *Fourth Generation:* Large-scale integration (LSI) and very large-scale integration (VLSI)

The *first generation* of computers (such as ENIAC in the 1940s) were huge machines both in terms of size and mass. The ENIAC computer at the University of Pennsylvania in Philadelphia was constructed during World War II to calculate projectile trajectories. The circuitry of first-generation computers was composed of vacuum tubes and used very large amounts of electricity (it was said that whenever the ENIAC computer was turned on, the lights all over Philadelphia dimmed). ENIAC weighed 30 tons, occupied 15,000 square feet of floor space, and contained more than 18,000 vacuum tubes. It performed 5000 additions per second and consumed 40 kilowatts of power per hour. Also, due to the vacuum tube circuitry, continuous maintenance was required to change the tubes as they burned out. Input and output functions were performed using punched cards and separate printers. Programming these computers was tedious and slow, usually performed directly in the binary language of the computer.

The *second generation* of computers was developed in the 1950s. These computers used transistors instead of the vacuum tubes of their predecessors, decreasing maintenance requirements as well as electricity consumption. Information was stored using magnetic drums and tapes, and printers were connected on-line to the computer for faster hard-copy output. Unrelated to hardware considerations was the development of programming languages that could be written using more readily understandable commands and then separately converted into the binary data required by the computer.

Third-generation computers were distinguished by the advent of the integrated circuit in the late 1960s, which made computers faster and more compact. Storage, input, and output capabilities also increased dramatically. High-level software languages, such as COBOL, FORTRAN, and BASIC, were developed and gained popularity. These languages were written in a way that the programmer could more readily understand and assembled automatically into a set of instructions for the computer to follow. The most significant development of this period was a downward cost spiral that precipitated the popularity of minicomputers—smaller computers designed for use by one user or a small number of users at a time, as opposed to the larger mainframes of previous generations.

In the *fourth generation* of digital computers, the steady decrease in processing times and cost for computer technology has continued with a corresponding increase in memory and computational capabilities. With large-scale integration (LSI), more than 1000 components can be placed on a single integrated-circuit chip. Very large-scale integration (VLSI) chips contain more than 10,000 components; current VLSI chips have 100,000 or more components on each chip. The semiconductor technology developed in the 1970s condensed whole computers into the size of a single chip, known as a microprocessor. Semiconductors were responsible for the arrival of "personal computers" in the late 1970s and early 1980s.

13.3.2 Categories of Computers

Computers can be divided into categories, depending on their size and capabilities. Traditionally, computers are grouped under the following headings:

- Supercomputers
- Mainframes
- Minicomputers
- Microcomputers

Supercomputers are the world's most powerful computers, often with processing speeds in excess of 20 million computations per second. The performance of the CRAY-2 supercomputer was rated at 100 million floating point operations per second (MFLOPS). Supercomputers are often used to calculate extensive mathematical problems for scientific research purposes. These problems are characterized by the need for high precision and repetitive performance of floating-point arithmetic operations on large arrays of numbers.

Mainframes have large memory capabilities coupled with extremely fast processing speeds. These computers are less powerful systems than supercomputers, but they are used in large CAD systems where a significant amount of highly accurate analysis must occur. Mainframes are highly applicable to analytical methods and are often used in dynamic analysis, stress analysis, heat transfer analysis, and other analytical methods. This type of computer system is used most often in large engineering corporations, such as those in the automotive or aerospace industries, where centralized computing and data storage are essential. Mainframes support multiple users (some over 500) at terminals, giving them access almost instantaneously to the data required to design and share information among the project team. Because of their extensive memory capabilities, mainframes are also used for large database maintenance. Mainframe computers usually require a specialized support staff for maintenance and programming. The typical configuration of a mainframe system is a processor with 32-bit and 64-bit word addressing, 64 megabytes (MB) to 2 gigabytes (GB) of memory, and several gigabytes of storage space.

Minicomputers are somewhat smaller and less powerful than a mainframe, but they nevertheless offer a powerful, less expensive alternative to mainframe systems where a centralized computing environment is desired. They introduced the concept of distributed data processing. A typical minicomputer is available with 16- to 32-bit word addressing, several megabytes of memory, and multiple disk drives amounting to several megabytes to gigabytes of storage space. Turnkey CAD systems were offered as minicomputer systems in the late 1970s and early 1980s. A number of display terminals can be supported by minicomputer systems, and on-line printers for minicomputer systems are capable of delivering between two and three thousand lines of text per minute.

Microcomputers, which include the personal computer and engineering workstations, are desktop-size or smaller computers. These computers have seen the greatest growth in the number of systems being sold and used since the early 1980s. There are various reasons for this trend. Microcomputers are quickly becoming more powerful, with greater memory capabilities. A wide range of microcomputers are available with 8- to 32-bit word addressing, several megabytes of memory, and built-in hard disk, floppy disk, CD-ROM, and tape backup systems.

Many companies operate best using a decentralized approach to computing; however, networks have become increasingly common in microcomputer environments in order to provide some of the advantages of centralized computing when desirable. Powerful servers that support massive client-server networks have largely replaced the huge mainframe computers. Even the power of a contemporary PC exceeds that of a mainframe from the 1960s and 1970s. Furthermore, the computational capability of engineering workstations today exceeds that of most minicomputers. The latest trend is to classify computers as supercomputers, servers, workstations, large PCs, and small PCs.

One common differential between types of computers is the word length. The term *word length* does not refer to words in human language; rather, it signifies the number of places in the base-2 units of the machine language in the various types of computer. Mainframes have traditionally run using 32-bit words, with minicomputers typically having 16-bit capabilities and microcomputers 8-bit capabilities. Word length influences processing speeds and memory-addressing capabilities in computer systems. Longer word length means that more information can be operated on or transferred to a different part of the system in fewer steps, thereby taking less time. The word length also influences memory capabilities by making virtual memory techniques available. While formerly applicable as a general rule for distinguishing the capabilities of the various types, various word lengths are now available in all types of computers. Even some home electronic game systems now employ 32-bit technology in a system about the size of a large textbook.

13.3.3 Central Processing Unit (CPU)

The computer's central processing unit (CPU) is the portion of a computer that retrieves and executes instructions. The CPU is essentially the brain of a CAD system. It consists of an arithmetic and logic unit (ALU), a control unit, and various registers. The CPU is often simply referred to as the *processor.*

The ALU performs arithmetic operations, logic operations, and related operations, according to the program instructions. The control unit controls all CPU operations, including ALU operations, the movement of data within the CPU, and the exchange of data and control signals across external interfaces (e.g., the system bus). Registers are high-speed internal memory-storage units within the CPU. Some registers are user-visible; that is, available to the programmer via the machine instruction set. Other registers are dedicated strictly to the CPU for control purposes. An internal clock synchronizes all CPU components. The clock speed (the number of clock pulses per second) is measured in megahertz (MHz) or millions of clock pulses per second. The clock speed essentially measures how fast an instruction is processed by the CPU.

13.3.4 RISC and CISC Computers

Computers can be divided into two categories, depending on their method of using instructions:

- Reduced Instruction Set Computers (RISC)

- Complex Instruction Set Computers (CISC)

The reasons for designing CISC computers are to simplify compilers and to improve performance. Underlying both of these reasons was the shift to high-level languages (HLLs) in computer programming. Computer architects attempted to design machines that provided better support for HLLs. CISC was expected to yield smaller programs that would execute faster.

RISC technology is very new by comparison. RISC computers use fewer and simpler instructions than conventional CISC computers. Simpler instructions reduce the complexity of the circuits required to implement an instruction, thereby allowing individual instructions to execute quickly. RISC machines generally show a higher level of performance than a comparably complex CISC system, despite the fact that a RISC processor executes more instructions to accomplish a given task than does a CISC processor. The following characteristics are common on all RISC-architecture computers: one instruction per machine cycle, unique register-to-register operations, and an instruction pipeline. RISC architecture often includes more general-purpose registers to maximize the number of operations that take place on the CPU. CISC computers, in contrast, employ more memory referencing. Studies have shown that the compilers on CISC machines tend to favor simpler instructions, such that the conciseness of the complex instruction sets seldom comes into play. The expectation that a CISC computer would produce smaller programs may not be realized because the more complex the instruction set, the more processor time is required to decode and execute each instruction. Longer opcodes required in CISC architecture produce longer instructions. In computationally intensive applications, such as FEA, where calculation times are often measured in hours, RISC processors are generally more efficient performing floating point operations than CISC processors.

Most RISC architecture can be found in workstations that run on the UNIX operating system, such as Sparc MIPS, DEC Alpha, PA-RISC, and PowerPC. The Pentium architecture, however, is an excellent example of CISC design. It represents the result of decades of CISC architecture evolution. Pentium architecture incorporates the sophisticated design principles once found only on mainframes, supercomputers, and servers.

The RISC versus CISC debate continues to drive technology in new directions. There is a growing realization in the industry that RISC and CISC may benefit from each other. Notably, more recent designs, like the PowerPC line of the Apple Macintosh, are no longer "pure" RISC, and the more recent CISC designs, like the Pentium P7, have incorporated some RISC-like features.

Engineering PCs

Computer-aided design projects often range from simple 2D drawings to graphics-intensive engineering applications. Computationally intensive number crunching in 3D surface and solid modeling, photorealistic rendering, and finite element analysis applications demand a great deal from a personal computer. Careful selection of a PC for these applications requires an examination of the capabilities of the CPU, RAM capacity, disk space, operating system, network features, and graphics capabilities. The industry advances quickly, especially in microprocessor capabilities. The following PC configurations list minimum requirements for various CAD applications. It should be noted that because of rapid advances in the industry, this listing may be dated by the time of publication.

For 2D drafting applications, a low-end PC is sufficient. This denotes a PC equipped with a 486 processor running on a 16-bit DOS operating system with 16 MB of RAM, or Microsoft Windows with 32 MB of RAM. (The operating system needs 4 MB, most drafting applications require at least 8 MB of RAM, and Windows holds as much data as possible in RAM.) Eight K of on-board cache memory and 256 K to 512 K of external cache for faster response is recommended. A 500-MB, fast SCSI (Smaller Computer System Interface) hard drive is minimum. SCSI is a type of bus used to support local disk drives and other peripherals. Five hundred MB is considered minimum because CAD and SWAP files require approximately 150 MB, and the operating system itself usually requires about 100 MB. A 16-inch high-resolution (1024 $\times$ 768), 256-color monitor should also be considered a minimum requirement. CD-ROM drives and fast modems with transfer rates of 28,800 baud are essential for non-networked tasks.

For 3D modeling and FEA applications, a Pentium 100 MHz processor running on a 32-bit Windows NT operating system works significantly better. The minimum memory requirement for Windows NT is 32 MB and the operating system requires 160 MB of hard-drive space. The system should have a 16-K internal cache with 256 K to 512 K of external cache for increased performance. The monitor for these applications should be 21 in. with .25 ultrafine dot pitch, high resolution (1600 $\times$ 1280), and 65,536 colors. Peripheral Component Interconnect (PCI) local bus (a high-bandwidth, processor-independent bus) and a minimum 1-GB SCSI-2 hard drive are also required.

Engineering Workstations

The Intel Pentium CPU microprocessor reduced the performance gap between PCs and workstations. A current trend is the merging of PCs and workstations into "personal workstations." Pentium Pro and RISC processors, such as DEC Alpha, Sparc MIPS, PowerPC, PA-RISC, are all powerful personal workstations with high-performance graphics accelerators.

Until recently, operating systems were the main distinction between a low-end workstation and a high-end PC. The UNIX operating system, which supports multitasking and networking, usually ran on a workstation. DOS, Windows, and the Macintosh operating systems, which perform single tasks, usually ran on PCs. That distinction is beginning to disappear because of the birth of Microsoft Windows NT and IBM's OS/2. Both of these operating systems support multitasking and networking, as well. Other operating systems designed for workstations are being modified for use on a PC, such as IBM's AIX. UNIX now runs on laptop computers with the PowerPC microprocessor.

Many CAD and FEA software applications are traditionally UNIX-based applications. Since there are significant differences in price between UNIX and Windows NT, more and more CAD and FEA vendors have released versions of their software for Windows NT. Windows NT is now being offered on workstations with processors such as DEC Alpha and MIPS R4000. Memory capacity for a personal workstation can be up to 256 MB, with up to 2 GB of disk space, such as those from Silicon Graphics, Inc.'s Indy Systems. Thanks to 64-bit technology and a scalable modular platform, high-end workstation performance can now boast supercomputer-like performance at a fraction of the cost of a mainframe or supercomputer. These systems include the HP Series 700, IBM RS/6000, Sun Sparcstation 10, Silicon Graphics Indigo, and certain models of DEC Alpha/AXP. The noted performance gain is a result of using more powerful 64-bit word addressing and up to 350 MHz clock speed. Dual processors available in some engineering workstations with the DEC Alpha system, Intel Pentium chip, or Motorola 68060 chip allow some degree of scalable parallel processing within current engineering workstations.

13.3.5 Parallel Processing

To reach higher levels of productivity in the analysis of complex structures with thousands of components, such as in combustion engines or crash simulations, the application of parallel processing was introduced. Two parallel processing methods used in engineering applications are:

- Massively parallel processing (MPP)
- Scalable parallel processing (SPP)

MPP machines combine a number of processors into one machine and boast large amounts of processing power. While these machines were expected to revolutionize computerized engineering analysis, users could not simply upload existing applications to these half-billion dollar computers. On the contrary, programming MPP machines proved to be difficult even for experienced programmers, and few applications were ever developed for general use.

SPP, however, has shown extraordinary potential for use in general engineering analysis. SPP techniques essentially link PCs or workstations, each with existing memory and disk storage capacities and CPU capabilities, allowing the attributes of each individual machine to be applied to the computations. In addition, there is no lack of application software for SPP users. Applications for thermal, dynamic, stress (both linear and non-linear), and fluid analyses are available for use with SPP computers.

13.4 MEMORY SYSTEMS

Memory systems store program information and other data in a manner that facilitates efficient access. In order to accomplish this task, various technologies and organizational methods are employed. The typical computer system is equipped with a hierarchy of memory subsystems: some (internal to the system) can be directly accessed by the processor and some can be accessed by the processor via input/output devices (external). Internal memory consists of main memory, cache, and registers (the CPU's own local memory). External memory consists of secondary storage devices, such as magnetic disk and tape, and optical memory storage on CD-ROM.

13.4.1 Organizational Methods

Memory systems organize binary data into addressable words where data can be stored or retrieved. Some systems use a method called "interleaved memory," which refers to an ability to access more than one word at a time. Interleaved memory is an advantage in situations where it is likely that the second word will be required by the CPU soon after the first.

The memory circuitry is a separate entity from the processing circuitry of the CPU. Therefore, in order for the two circuit systems to communicate, a memory controller is required. The controller deciphers the requests for memory information or storage access from the CPU and initiates the proper sequence of events. Because a controller can only control a certain amount of memory, multiple controllers are often used within a single system. The implementation of multiple memory controllers allows for interleaved memory transfer. Each controller, with its own memory domain within a system, increases efficient data transfer by allowing sequential data to be stored in different domains such that before one controller is finished transferring one part of the sequence, the next controller is already beginning its operation sequence. In some systems, memory can be shared among different CPUs of the same computer. The controllers used in conjunction with a multiprocessor system in this case would have access ports for each of the individual processors.

13.4.2 Internal Memory and Related Techniques

Registers

Memory within the CPU for data required to perform a specific task, such as the operand or result of a mathematical calculation, is stored in memory devices called *registers*. These registers are accessible by specific commands from the CPU. Other registers in the CPU are inaccessible for memory storage but are included as a part of the working system as a whole. Generally, registers hold the same number of bits, or binary digits, as the word length of the CPU. The number of registers in a CPU is variable, ranging from 1 to 16 or more.

Metal Oxide Semiconductor (MOS) Random Access Memory (RAM)

This type of storage is semiconductor-based memory, with wide-ranging applications in nearly all computer systems. RAM can either be dynamic or static. Dynamic RAM (DRAM) uses circuitry that must be periodically refreshed by rewriting the data in each block of memory. Static RAM (SRAM) does not require the refreshing but is usually more expensive than dynamic RAM. In general, static RAMs are faster than dynamic RAMs. Both types of RAM are volatile; they lose their memory when power is turned off. Thus, RAM can be used only as temporary storage. To compensate, programs can be stored on magnetic disks or tapes or on other forms of solid-state memory, or a battery can be used to supply the necessary power to maintain semiconductor memory when the CPU is not in use.

Often, it is useful to separate memory into two types: Random Access Memory (RAM) and Read Only Memory (ROM). RAM is used in memory blocks from which the CPU must be able to "read" (access stored information) and "write" (store new data). Because information in RAM can be accessed, subsequently modified, and possibly erased, it does not provide the needed security for important programs. For those applications, such as system programs, function tables, and library subroutines, data are best stored in ROM memory. Because access is usually limited to retrieval, data are not easily altered and the integrity of key programs is ensured. Some ROMs can actually be erased and rewritten under certain conditions. These types are useful when a program needs periodic alterations but should be protected from general access and possible accidental erasure.

Cache Memory

Cache memory is designed to hold information relating to frequently used applications and subroutines in active memory. Cache memory, however, is usually a separate piece of hardware between the CPU and main memory. It provides faster data transfer to the CPU than does main memory, but usually at a higher cost as well. This cost is usually well justified, especially if the computer is to be used with repetitive programs, where the cache can dramatically increase the speed with which programs run, and increase the efficiency of the user. Cache sizes of between 1 K and 512 K are usually offered for lower average cost per bit and faster average access time.

Virtual Memory

This technique addresses the problem of very large programs that use extensive address space and operate within a limited memory capacity. Programs use the registers of the CPU to keep the most active applications of the program available quickly. Other, less active applications are stored on magnetic disk space until needed. If needed, the application called for will be directed to occupy a less active register and the data formerly in that register will be saved onto the disk in order to maintain any changes made to the data during execution of the program.

Memory Addressing

Some instructions from the CPU may require an operation involving one or more operands stored in memory. Operands of this type include those for logic, mathematical operations, and so forth. For example, an addressing mode might supply the ability to take operands from various locations in memory and store the result in a separate location. Most CPUs employ a variety of addressing modes, depending on the operation. The variety of addressing modes for different tasks is a benefit in most CPUs. It provides a degree of flexibility in data management that increases efficiency and processing ability. Furthermore, memory addressing is made even more efficient in some systems by the ability to operate on single bits within an 8-digit byte, on the byte itself, and sometimes on words of 16 and 32 bits. A technique known as *extended addressing* can further increase the memory capabilities of the computer. Using this technique, the program running extended addressing functions considers memory as a number of pages. Each page is assigned a relocation constant that is combined with the other addresses on that page to form a longer address than would normally be used. In an 8-bit computer, extended addresses might take the form of 10-bit words, which increases the number of addresses possible within the system.

13.4.3 External Memory

Most computer systems use some sort of magnetic storage system to store data after semiconductor memory has been erased due to a loss in power. Although computer systems of the 1990s have main memories much larger than those commonly used in previous decades, contemporary computer programs have many more capabilities, and hence take up more memory space, than programs of preceding decades. This leaves the main memory and cache memory still requiring external (with respect to the CPU) storage capabilities. Magnetic storage systems are one answer to this problem. These storage technologies provide the added benefit of an ability to be copy data an almost unlimited number of times. Some copies can be stored away from the computer for security, while others may be kept nearby, perhaps on-line to facilitate access to the stored data. The two main magnetic storage configurations are disks and tapes.

13.4.4 Magnetic Disks

Magnetic disks are connected to the CPU through a controller attached to an input/output port. The controller can often control a number of disk drives and is usually programmed with a significant amount of data relating to error detection, data transfer, and other pertinent information.

Disk drives use a drive motor and one or more "heads" to read and write data on the disk. The drive motor turns a spindle on which the disk rests, rotating the disk at a controlled speed. A standard speed allows data written on one disk drive to be read on another drive. The head mechanism holds a read/write head for each recording surface on the disk. The heads are normally held away from the surface of the disk but placed extremely close to the surface or directly on the surface of the disk, depending on the type of drive, during data transfer. The magnetic disk itself is coated with a magnetic oxide, which forms the actual storage medium. The oxide must be organized or "formatted" into closely spaced, concentric circular tracks. The tracks must be positioned accurately and consistently on all disks in order for the head mechanism to position itself accurately over a specific track. Data are recorded and subsequently read as analog variations in the magnetic field of the oxide medium. These data are transferred to the disk drive controller, which converts the analog signal to a digital one for processing by the CPU. Disks are further formatted into blocks, or "sectors," of equal area. On most disks, these blocks are established in the medium by the manufacturer. The seek time is the time needed to position the head at the desired track. Rotation latency time is the time elapsed for an appropriate sector to rotate and line up with the head. The sum of the seek time and the latency time gives the overall access time. This time consideration can be useful during the purchase of a disk drive, since the access time will affect the efficiency of the overall system if frequent storage access is required. Soft-sectoring, a technique in which the disk drive will establish blocks of unequal area on the disk, can also be used. Finding a particular track on a magnetic disk, where the distance between tracks can be as little as 0.01 mm, is no easy task. All hard disk drives use a servo-control mechanism to ensure accurate positioning over a desired track. Many floppy drives, where the distance between adjacent tracks is not as small, use a stepping motor system.

Floppy Disks

The floppy disk was designed as a cheap and simple device to provide quick, reliable access to information stored as a back-up to computer memory. Physically, the floppy disk is a flexible diskette coated with magnetic oxide and contained within a square plastic housing. The housing has openings which allow the spindle of the drive motor to turn the disk and provide space for the head mechanism to make contact with the disk's surface. Floppy disks have continued to shrink in physical size and grow in terms of storage limits since their inception. The earliest floppy diskettes were 8 in. in diameter and capable of storing 256 K of data. These disks also typically used only one side of the diskette. The most commonly used diskettes today are 3.5 in. in diameter, and are capable of storing 1.44 MB of data using both sides of the disk. Access times for floppy disks are 100–500 msec and data transfer rates are generally lower than 300 K/sec. Floppy disk drive systems provide reliable data storage on a medium which can be removed from one computer and used on another. They are among the most common magnetic data storage systems currently in use.

Hard Disks

This term is a general heading for a category of disk drives where the disk remains within the drive. Because the disk and drive are housed within the same sealed unit, hard disks provide some significant advantages over floppy disks. The disk medium remains free of contaminants, resulting in greater reliability and data accuracy. The heads of a hard disk are lightweight and designed to hover aerodynamically and extremely close to the disk surface without touching. This virtually eliminates wear on either the disk or the drive heads. Shorter access times are also seen using hard disk systems because two heads are normally associated with each surface.

Portable hard disks supplying 1 GB of storage space are now available. These disks provide extra security for important files and provide a convenient medium for storing large amounts of data for back-up or other purposes.

13.4.5 Magnetic Tape

Magnetic tape was the first application of magnetic data storage employed with computer systems. This medium is effective, to a large extent, due to existing standards for data format. These standards, like those for magnetic disks, allow data to be used on different types of computers. Various forms of magnetic tape storage systems have been developed and satisfy specific user needs.

Standard industrial tape drives use reels of 12.7-mm-wide tape at lengths of 731 or 365 m, coated with an oxide medium similar to that used in magnetic disks. The tape moves past read/write heads which provide data transfer at 800–6250 bits per linear inch. The tape motion is servo-controlled for a high degree of accuracy, with the tape winding more than 180° around a capstan to provide sufficient physical control. The motors driving the tape reels must also be carefully controlled to ensure proper tape tension. Two mechanisms are typically used to ensure this control. In older tape drives, the tape moves over fixed- and variable-position pulleys. The variable pulleys are attached to a spring-loaded tension arm drawing the tape in a "W" form between the pulleys. The position of the tension arm gives the motor control mechanism the necessary information regarding tape winding and release to ensure the proper tension. Most current tape drive designs have abandoned the tension arm in favor of the following vacuum chamber technique. Between each reel and capstan, a vacuum chamber draws the tape into a loop 1–2 m long. The length of the tape in the chamber is detected using photoelectric sensors and this information is used by the motor controller to govern the movement of the reels. Smaller magnetic tape drives are available for smaller systems such as microcomputers. In these drives, the tape is normally housed in a plastic cartridge that protects the tape medium from contamination. This provides excellent data integrity on a much smaller scale than industrial-type systems. An 8-mm tape is usually used with a workstation system, while PC-based tape drives are usually on a 4-mm format. Because magnetic tape moves in a linear fashion, it is inefficient for applications requiring rapid random access to stored data. It does, however, provide an excellent means for back-up protection of important data.

13.4.6 Optical Data Storage

While magnetic systems are a popular and reliable method for storing large amounts of data, they can become quite cumbersome in terms of physical size as the amount of data increases. There is also the limitation on speed imposed by the need to convert between the digital signal of the CPU and an analog signal. Digital data storage addresses both of these problems. First, the data storage capability of a compact disk (CD), essentially identical to that used in musical recordings, is 680 K, and the technology for downloading the information stored on CDs keeps advancing. The initial transfer speed of CD drives was 150 K/sec. Currently, quad-speed CD drives, capable of data transfer at 600 K/sec, have become standard, and 6X speed drives delivering data at 900 K/sec are already on the market. Second, because the storage medium and the CPU use the same digital data format, there is no need for a controller to convert the signal from an analog signal. Often, compact disk storage in computers is referred to as CD-ROM, because the technology to write information in digital form on the compact disk is still out of reach in terms of price for most computer users. The technology does exist, however, and with time the price will surely make the ability to write data to a CD more generally available. Most CAD software packages are now available on CD-ROM.

13.5 INPUT DEVICES

Commonly used input devices in CAD systems include the alphanumeric keyboard, the mouse, the light pen, and the digitizer. All of these allow information transfer from the device to the CPU. The information being transferred can be alphanumeric or functional (in order to use command paths in the software) or graphic in nature. In either case, the devices allow an interface between the designer's thoughts and the machine that will assist in the design process.

13.5.1 Keyboard

The alphanumeric keyboard is one of the most recognizable computer input devices. Rows of letters and numbers (typically laid out like a typewriter keyboard) are used with other functional keys. These keys are either dedicated to tasks such as control of cursor placement on a display screen or definable by the user, transfer bits of information to the CPU in one of several ways. Key depression can be detected through a simple mechanical switch, a change in magnetic coupling, or a change in capacitance. The alphanumeric keyboard is dedicated to the input of alphanumeric information and special commands via function keys.

Special programmed-function keyboards with 16–32 buttons can also be used in conjunction with a CAD system. These keyboards are often separate from the alphanumeric keyboard, but the keys can similarly be dedicated or definable to specific CAD tasks. Some keyboards will employ cardboard or plastic overlays that show the function of each key. In the case where the keyboard is applicable to several tasks within the general CAD techniques, the overlays show which functions the keys will command different techniques.

13.5.2 Touch Pad

The touch pad is a device that allows command inputs and data manipulation to take place directly on the screen. The touch pad is mounted over the screen of the display terminal, and the user can select areas or on-screen commands by touching a finger to the pad. Various techniques are employed in touch pads to detect the position of the user's finger. Low-resolution pads employ a series of light-emitting diodes (LEDs) and photodiodes in the x- and y-axes of the pad. When the user's finger touches the pad, a beam of light is broken between an LED and a photodiode, which determines a position. Pads of this type generally supply 10–50 resolvable positions in each axis. A high-resolution panel design generates high-frequency shock waves traveling orthogonally through the glass. When the user touches the panel, part of the waves in both directions is deflected back to the source. The coordinates of this input can then be calculated by determining how long after the wave was generated it was reflected back to the source. Panels of this type can supply resolution of up to 500 positions in each direction. A different high-resolution panel design uses two transparent panel layers. One layer is conductive while the other is resistive. The pressure of a finger on the panel causes the voltage to drop in the resistive layer, and the measurement of the drop can be used to calculate the coordinates of the input.

The input of graphical data is somewhat clumsy using a keyboard or touch pad. For this reason, various input devices that are specialized for graphics input have been created and are widely used in CAD.

13.5.3 Mouse

The mouse is used for graphical cursor control. A mouse conveys cursor-placement information in the x–y coordinate plane to the CPU. A spherical roller is housed within the mouse such that the roller touches the plane upon which the mouse is resting. When the mouse is moved along a flat surface, the spherical roller simultaneously contacts two orthogonal potentiometers, each of which is connected to an analog–to–digital converter. The orthogonal potentiometers send x- and y-axis vector information via a connecting wire to the CPU, which performs the necessary vector additions to allow cursor control in any direction on a 2D display screen. Often, a mouse will be equipped with one to three pressure-sensitive buttons that assist in the selection of on-screen command paths. Since the mouse is inexpensive and simple to use, it has become a standard computer input device.

13.5.4 Trackball

This device operates much like a mouse in reverse. The main components of the mouse are also present in the trackball. Like the mouse, the trackball uses a spherical roller that comes into contact with two orthogonally placed potentiometers, sending x- and y-axis vector information to the CPU via a connecting wire. The difference between the mouse and the trackball lies in the placement of its spherical roller. In a trackball, the spherical roller rests on a base and is controlled directly by manual manipulation. As in a mouse, buttons may be present to facilitate the use of on-screen commands.

13.5.5 Light Pen

Another computer input device that can be used in CAD systems is the light pen. Somewhat misnamed, the light pen does not project light; rather, it detects light from a raster-scan cathode ray tube (CRT) screen (see Section 13.6). Light pens used in the SAGE Radar system in the 1950s resembled guns that were pointed at the screen, with input delivered through a trigger-like device. Contemporary light pens are hand-held cylindrical instruments approximately the size of an ink pen. At one end of the cylinder is a lens and a photo-optical sensor. The other end of the cylinder is connected to the computer by a cable. The pen detects the timing of the screen's repeated illuminations (a process so fast that the constant flicker one would expect is absent and the screen appears to maintain a constant 2D image) by detecting the light pulse at the desired location when the screen is illuminated. The pulse is then transferred through the cable to the CPU, which uses the pulse to determine where the light pen is in contact with the screen and uses this information to track its position continuously. The location is determined by correlating the pulse with the graphical display data in the CPU to identify what graphical information was being displayed at the given time and location on the screen. The device can also send a second type of signal to the CPU, indicating the selection of a point on the screen when a button on the pen is depressed. Light pens can thus be used to create lines and shapes that appear instantaneously to the human eye on the computer display. Light pens also select on-screen command paths in a manner similar to the mouse and trackball. Small graphical areas, sometimes called icons, on the screen can be associated with programmed software commands. If the user points the light pen at an icon and depresses the light pen button, the desired command will be executed on-screen, bypassing the need for keyboard inputs.

13.5.6 Digitizer

A digitizer is an input device consisting of a large, flat surface coupled with an electronic tracking device, or cursor. The cursor is tracked by the tablet underneath it and buttons on the cursor act as switches to allow the user to input position data and commands. Digitizing tablets apply different technologies to sense and track cursor position (Fig. 13.7). The three most common techniques use electromagnetic, electrostatic, and magnetostrictive methods to track the cursor. Electromagnetic tablets have a grid of wires underlying the tablet surface. Either the cursor or the tablet generates an alternating current that is detected by a magnetic receiver in the complementary device. The receiver generates and sends a digital signal to the CPU, giving the cursor's position. Despite their use of electromagnetism, these types of digitizers are not compromised by magnetic or conductive materials on their surface. Electrostatic digitizers generate a variable electric field that is detected by the tracking device. The frequency of the field variations and the time at which the field is sensed provide the information necessary to give accurate coordinates. Electrostatic digitizers function accurately in contact with paper, plastic, or any other material with a small dielectric constant. They do lose accuracy, however, when even partially conductive materials are in close proximity to the tablet. Magnetostrictive tablets use an underlying wire grid similar to that used in electromagnetic tablets. These tablets, however, use magnetostrictive wires (i.e., wires which change dimension depending on a magnetic field) in the grid. A magnetic pulse initiated at one end of a wire propagates through the wire as a wave. The cursor senses the wave using a loop of wire and relays a signal to the CPU, which then couples the time of the cursor signal with the time elapsed since the wave originated to give the position. These tablets require periodic remagnetization and recalibration to maintain their functional ability.

Digitizing tablets can usually employ various modes of operation. One mode allows the input of individual points. Other modes allow a continuous stream of points to be tracked into the CPU, either with or without one of the cursor buttons depressed, depending on the needs of the user. A digitizing-rate function, which enables the user to specify the number of points to be tracked in a given period of time, is also often present in CAD systems with digitizers. The rate can be adjusted as necessary to facilitate the accurate input of curves.

Whatever the type, digitizers are highly accurate graphical input devices and strongly suited to drafting original designs and to tracing existing designs from a hard-copy drawing. Resolution can be up to 1000 lines per linear inch. Tablet sizes typically range from 10 × 11 in. to 44 × 60 in.

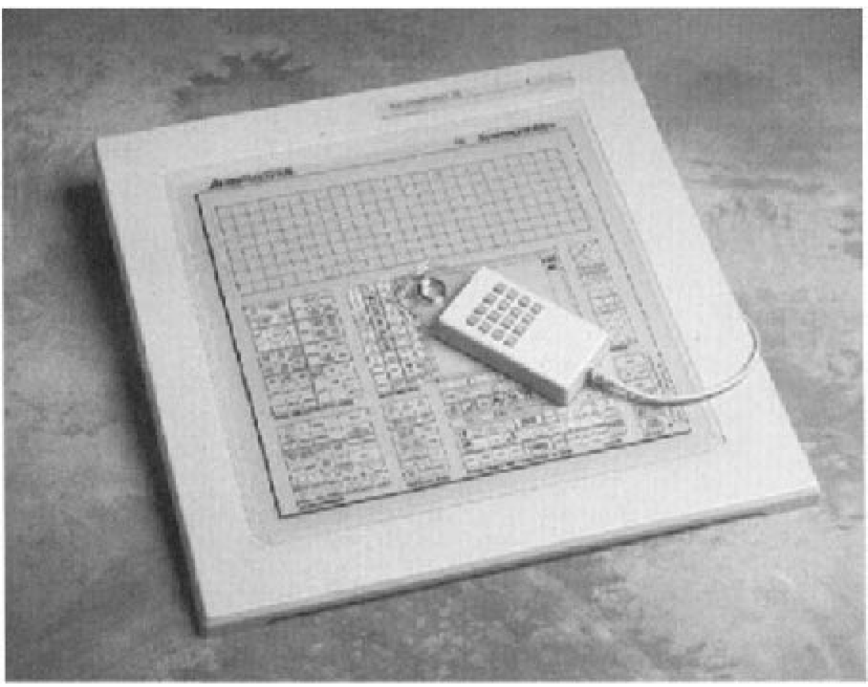

Fig. 13.7 Digitizing Tablet and Cursor (courtesy of Summagraphics, Inc.).

Often, plastic sheets with areas for command functions, such as switching between modes, as discussed above, are laid over the tablet to allow the designer to access software commands directly through the tablet using one of the buttons on the tracking device. Many of these commands deal with the generation of graphical elements, such as lines, circles, and other geometries.

13.5.7 Scanner

Scanners have been used as computer input devices for some time, but their employment in CAD systems has been somewhat limited until recently due to resolution limits. A scanner uses a photo-optical sensor to detect areas of light and dark on the page when illuminated by a bright light source. The information from the sensors is digitized and sent to the CPU through a connecting cable. Special software that supplies optical character recognition (OCR) is capable of recognizing alphanumeric characters on the scanned image so that engineering notes and other text can be maintained with the image once the scanning is complete.

13.6 OUTPUT DEVICES

Just as a CAD system requires input devices to transfer information from the user to the CPU, output devices are also necessary to transfer data in visual form back to the user. Electronic displays provide real-time feedback to the user, enabling visualization and modification of information without hard-copy production. Often, however, a hard-copy is required for presentation or evaluation, and the devices that use the data in the CPU or stored in memory to create the desired copy are a second category of output devices.

13.6.1 Electronic Displays

Contemporary computer graphics displays use a cathode ray tube (CRT) (Fig. 13.8) to generate an image on the display screen. The CRT heats a cathode to project a beam of electrons onto a phosphor-coated glass screen. The electron beam energizes the phosphor coating at the point of contact, causing the phosphor to glow.

CRTs employ two different techniques, stroke-writing and raster-scan, to direct the beam onto the screen (Fig. 13.9). A CRT using the stroke-writing technique directs the beam only along the vectors given by the graphics data in the CPU. In a raster-scan CRT, the electron beam sweeps systematically from left to right and top to bottom at a continuous rate, employing what is known as *rasterization.* The image is created by turning the electron beam on or off at various points along the sweep, depending on whether a light or a dark dot is required at those points to create a recognizable image.

Regardless of the technique used to create an image on a CRT screen, the phosphor glows for a very short period of time after being energized by the electron beam. Three types of graphics terminals make up the vast majority of those used in CAD systems. Each employs a different approach to create a continuous image on the screen.

Vector Refresh Terminals

These early graphics terminals (circa 1960s) use the stroke-writing approach in the CRT. The electron beam continuously refreshes the image at speeds around 40–50 cycles per second to avoid a noticeable flicker. Refresh terminals permit a high degree of movement in the displayed image, as well as high resolution. Selective erasing or editing is possible at any time without erasing and repainting the entire image.

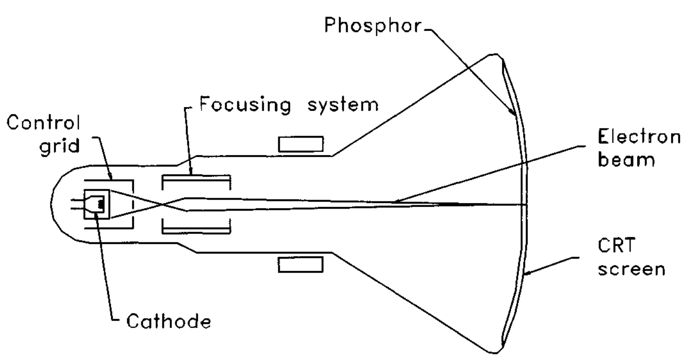

Fig. 13.8 Diagrammatic representation of a cathode ray tube (CRT).

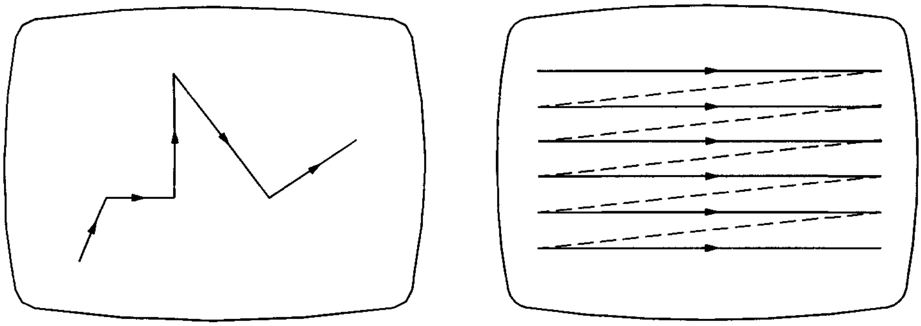

Fig. 13.9 (*a*) Electron beam direction in a stroke-writing CRT. (*b*) Electron beam pattern in a raster-scan CRT.

Direct View Storage Tube (DVST)

DVST was one of the most widely used graphics display devices in the mid- to late 1970s. In a DVST terminal, the CRT uses the stroke-writing technique to generate the desired image on the screen. One or more electron flood-guns continuously supply the necessary energy to maintain the image on the screen. DVST generates a long-lasting, flicker-free image with high resolution and no refreshing. It handles an almost unlimited amount of data. However, display dynamics are limited, since DVST terminals do not permit selective erasure. The image also does not appear as bright as with a vector refresh or raster-scan terminal.

Raster Scan Terminals

As their name suggests, these displays use rasterization to create the image on the CRT screen. Much like a television set (the difference being that a TV input signal is analog while that of a computer terminal is digital), these terminals refresh the image by continuously sweeping over the screen in a constant pattern. This allows real-time changes to be made to the image. Raster-scan is currently a dominant technology in CAD graphics display. Raster-scan terminal features include brightness, accuracy, selective erasure, dynamic motion capabilities, and the potential for unlimited color. The raster-scan terminal can display a large amount of information without flicker, although the resolution is not as high as with a DVST display. Raster-scan technology has moved rapidly since 1982 and, with the introduction of graphic accelerators and additional add-on frame buffers to workstations and PCs, affordable, high-quality graphics have finally reached the end-user.

13.6.2 Hard Copy Devices

Despite the ease with which design files can be managed using computer technology, a hard copy of the work is often required for recordkeeping and presentation. Output devices have been developed to interface with the computer and produce a hard copy of the requisite design or file. The processes used in hard copy devices are analogous to those used in CRT displays. The various types of hard-copy devices are shown in the following outline:

Vector Plotters

1. *Pen Plotter*
   - Drum Plotter
   - Flat-Bed Plotter
2. *Computer Output to Microfilm (COM) Plotter*

Raster Plotters

1. *Electrostatic Plotter*
2. *Inkjet Plotter*
3. *Laser Plotter*

Vector Plotters

Just as storage tube and vector-refresh displays create images through an electron beam directed along vectors defined in the design file, vector plotters create an image using designated vectors. Vector plotters produce very high-resolution hard copies. Two common kinds of vector plotters are the pen plotter and the COM plotter.

Pen Plotter. Pen plotters use mechanical ink pens, directed along design vectors to create images on paper or similar media. Pen plotters are further divided into drum- and flat-bed types.

Drum Plotter. The drum plotter (Fig. 13.10) consists of a cylindrical drum, most often mounted horizontally, and a pen mounted on a slide parallel to the surface of the drum. The plotter matches the motion of the drum and that of the pen along the slide to draw the appropriate vectors and create the desired image. The drum plotter is able to produce hard copies that are limited in length only by the amount of paper on a roll.

Flat-Bed Plotter. Flat-bed plotters act on the same vector-plotting concept, but the paper is attached to a flat surface. The writing tool is again movable along a metal slide. The slide provides one of the two coordinate axes upon which the design will be printed. The parallel tracks along which the slide itself moves provides the other axis. As the pen moves along the slide, the slide moves along the other axis, enabling the two-dimensional vectors to be drawn.

Computer-Output to Microfilm (COM) Plotter. COM plotters constitute a third type of vector plotter. These plotters produce images on film rather than on paper, using light instead of ink. These expensive units facilitate efficient archiving of designs. Designs can be stored using a fraction of the space required for hard-copy filing and enlarged to original size when needed. The cost of producing a plot using a COM plotter can be significantly less than the cost associated with a flat-bed or drum plot, but the quality upon enlargement is generally poorer than that obtained using a pen plotter.

Raster Plotters

Many plotters require rasterization, breaking up the image into a series of dots that will then be reconstituted to recreate the image in much the same way that a raster-scan CRT screen uses these dots to produce a recognizable image on the display terminal. Common raster plotters include the electrostatic plotter, inkjet plotter, and laser plotter. These types of plotters generally produce lower-resolution images than those generated on a pen plotter. However, because the time to produce a rasterized plot is independent of complexity, these have the advantage of being significantly faster and are often used to create preliminary hard copies before a higher-resolution plot is performed.

Electrostatic Plotter. Electrostatic or direct printing forms images for thermosensitive media by application of heat from nibs on thermal printing heads. The paper turns black at the precise points where heat is applied. Electrostatic plotters have few moving parts and thus are reliable and quiet.

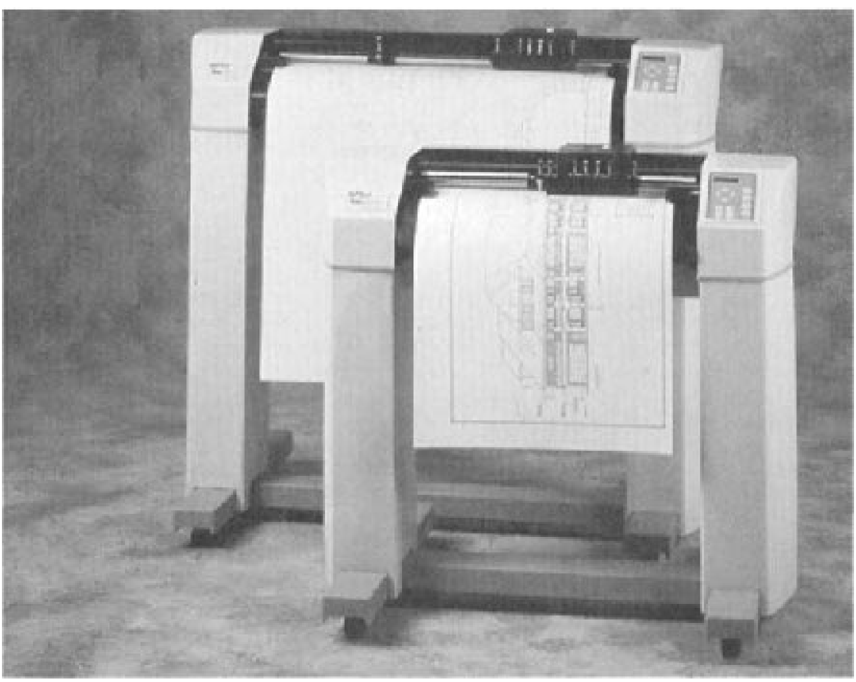

Fig. 13.10 Two drum-type plotters (courtesy of Summagraphics, Inc.).

Resolution can be as high as 400 × 800 dpi, with gray scales ranging from 16–128 values. These are medium- to high-throughput devices, producing complex images in about a minute. On-board computing facilities, such as RISC processors and fast hard disk storage mechanisms, contribute to rapid drawing and processing speeds. Expansion slots accommodate interface cards for LANs or parallel ports.

Inkjet Plotter. Inkjet plotters and printers fire tiny ink droplets at paper or a similar medium from minute nozzles in the printing head. Heat generated by a separate heating element almost instantaneously vaporizes the ink. The resulting bubble generates a pressure wave that ejects an ink droplet from the nozzle. Once the pressure pulse passes, ink vapor condenses and the negative pressure produced as the bubble contracts draws fresh ink into the nozzle. These plotters do not require special paper and can also be used for preliminary drafts. Inkjet plotters are available both as desktop units for 8.5 × 11-in. graphics and in wide format for engineering CAD drawings. Typical full-color resolution is 360 dpi, with black-and-white resolution rising to 700 × 720 dpi. These devices handle both roll-feed and cut sheet media in widths ranging from 8.5–36 in. Also, ink capacity in recently developed plotters has increased, allowing these devices to handle large rolls of paper without depleting any one ink color. Inkjet plotters are very user-friendly, often including sensors for the ink supply and ink flow that warn users of an empty cartridge or of ink stoppage, allowing replacement without losing a print. Other sensors eliminate printing voids and unwanted marks caused by bubbles in the ink lines. Special print modes typically handle high-resolution printing by repeatedly going over image areas to smooth image lines. In addition, inkjet plotters typically contain 6–64 megabytes of image memory and options such as hard drives, an Ethernet interface for networking, and built-in Postscript interpreters for faster processing. Inkjet plotters and printers are increasingly dominating other output technologies, such as pen plotters, in the design laboratory.

Laser Plotter. Laser plotters produce fairly high-quality hard copies in a shorter period of time than pen plotters. A laser housed within the plotter projects rasterized image data in the form of light onto a photostatic drum. As the drum rotates further about its axis, it is dusted with an electrically charged powder known as toner. The toner adheres to the drum wherever the drum has been charged by the laser light. The paper is brought into contact with the drum and the toner is released onto the paper, where it is fixed by a heat source close to the exit point. Laser plotters can quickly produce images in black and white or in color, and resolution is high.

13.7 SOFTWARE

Software is the collection of executable computer programs including operating systems, languages, and application programs. All of the hardware described above can do nothing without software to support it. In its broadest definition, software is a group of stored commands, sometimes known as a program, that provides an interface between the binary code of the CPU and the thought processes of the user. The commands provide the CPU with the information necessary to drive graphical displays and other output devices, to establish links between input devices and the CPU. The commands also define paths that enable other command sequences to operate. Software operates at all levels of computer function. Operating systems are a type of software that provides a platform upon which other programs may run. Likewise, individual programs often provide a platform for the operation of subroutines, which are smaller programs dedicated to the performance of specific tasks within the context of the larger program.

13.7.1 Operating Systems

Operating systems have developed over the past 50 years for two main purposes. First, operating systems attempt to schedule computational activities to ensure good performance of the computing system. Second, they provide a convenient environment for the development and execution of programs. An operating system may function as a single program or as a collection of programs that interact with each other in a variety of ways.

An operating system has four major components: process management, memory management, input/output operations, and file management. The operating system schedules and performs input/output, allocates resources and memory space and provides monitoring and security functions. It governs the execution and operation of various system programs and applications such as compilers, databases, and CAD software.

Operating systems that serve several users simultaneously (e.g., UNIX) are more complicated than those serving only a single user (e.g., MS-DOS, Macintosh Operating System). The two main themes in operating systems for multiple users are multiprogramming and multitasking.

Multiprogramming provides for the interleaved execution of two or more computer programs (jobs) by a single processor. In multiprogramming, while the current job is waiting for the input/output (I/O) to complete, the CPU is simply switched to execute another job. When that job is waiting for I/O to complete, the CPU is switched to another job, and so on. Eventually, the first job completes its I/O functions and is serviced by the CPU again. As long as there is some job to