15

Computer Fundamentals

This chapter focuses mainly on computer hardware fundamentals, with a brief introduction to some of the relevant software-related topics. The chapter begins with a brief description of different types of computer system, from giant supercomputers to tiny digital assistants, which is then followed up by anatomical description of a generalized computer system, with particular reference to microcomputer systems. Other hardware-related topics that are extensively covered include input/output devices and memory devices.

15.1 Anatomy of a Computer

The basic functional blocks of a computer comprise the central processing unit (CPU), memory and input and output ports. These functional blocks are depicted in the block schematic arrangement of Fig. 15.1. As is clear from the figure, these functional blocks are connected to each other by internal buses. The CPU is the brain of the computer. It is basically a microprocessor with associated circuits. Ports are physical interfaces on the computer, through which the computer interacts with the input and output devices. Memories are storage devices used for storing data and instructions. The CPU fetches the data and instructions by sending the address of the memory location on the address bus. The data and the instructions are then transferred to the CPU by the data bus. The CPU then executes the instructions and stores the processed data in the memory or sends them to an output device via the data bus. It may be mentioned here that in most cases the instructions modify the data stored in the memory or obtained from an input device.

15.1.1 Central Processing Unit

As mentioned above, the CPU is the brain of the computer. The fundamental operation of the CPU is to execute a sequence of stored instructions called a program. In other words, it controls the execution

Digital Electronics: Principles, Devices and Applications Anil K. Maini
 ISBN: 978-0-470-03214-5

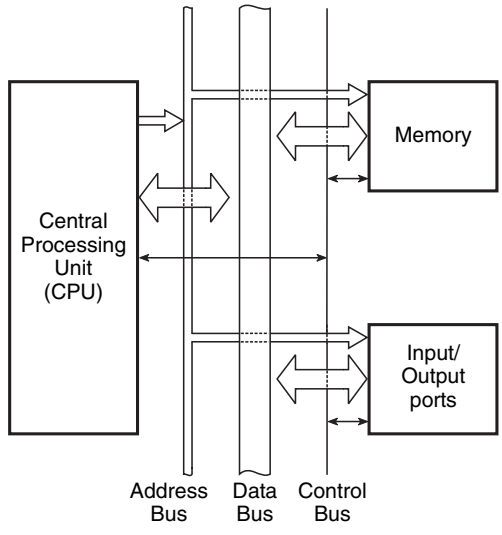


Figure 15.1 Block schematic of a typical computer.

of the computer software programs. It fetches and executes the instructions from the memory in a sequential manner. It may be mentioned here that the CPU can operate on more than one instruction at a time. Early CPUs were custom designed for a particular type of computer. But they have given way to a standardized class of processors that are used for generic applications. Since the advent of microprocessors in the 1970s, single-chip microprocessors have totally replaced all other types of CPU, and today the term 'CPU' refers to a microprocessor.

A microprocessor is a programmable device that accepts binary data from an input device, processes the data according to the instructions stored in the memory and provides results as output. The important functional blocks of a microprocessor are the arithmetic logic unit, the control unit and the register file. Microprocessors were discussed at length in Chapter 13.

15.1.2 Memory

There are several types of memory used in a computer. They can be classified as primary memory and secondary memory. Primary memory is directly connected to the CPU and is accessible to the CPU without the use of input/output channels. Primary memory can be classified into process registers, main memory, cache memory and read only memory (ROM). Process registers are present inside the CPU and store information to carry out the current instruction. Main memory is a random access memory (RAM) that stores the programs that are currently being run and the data related to these programs. It is a volatile memory and is used for temporary storage of data and programs. Cache memory is a special type of internal memory that can be accessed much faster than the main RAM. It is used by the CPU to enhance its performance. ROM is a nonvolatile memory that stores the system programs including the basic input/output system (BIOS), start-up programs and so on.

Secondary or auxiliary memory cannot be accessed by the CPU directly. It is accessed by the CPU through its input/output channels. Secondary memory has a much greater capacity than primary memory, but it is much slower than the primary memory. It is used to store programs and data for future use. Most commonly used secondary memory devices include the hard disk, floppy disks, compact disks (CDs), USB disks and so on. The hard disk is used for storing the high-level operating systems, application software and the user data files. Floppy disks have a limited capacity of 1.44 MB and have been replaced by CDs and USB drives. Floppy disks, CDs and USB drives are also referred to as off-line storage devices as they can be easily removed from the computer. Different types of memory are covered in Section 15.4.

15.1.3 Input/Output Ports

A port is a physical interface on the computer through which the input and output devices are connected to and interact with the computer. Ports are also used as an interface to connect two computers to each other. The ports on the computer can be configured as input and output ports through software. These ports are of two types, namely serial ports and parallel ports. Serial ports send and receive one bit at a time through a single wire pair. Parallel ports send multiple bits at the same time over a set of wires. Serial ports are used to connect devices such as modems, digital cameras, etc., to the computer. Parallel ports are used to connect printers, scanners, CD burners, external hard drives, etc., to the computer. Serial and parallel ports are discussed in detail in Section 15.8.

15.2 A Computer System

Figure 15.2 shows the block diagram of a typical computer system. The diagram basically shows the interconnection of the computer with the commonly used input/output devices. Input devices convert the raw data to be processed into a computer-understandable format. Some of the commonly used input devices include the keyboard, mouse, scanner and so on. Output devices convert the processed data into a format understandable by the user. Commonly used output devices include the monitor, printer, cameras, and so on. Input and output devices are discussed at length in Section 15.7.

15.3 Types of Computer System

Computers can be classified into various types, depending upon the technology used or the size and capacity or the applications for which they are designed.

15.3.1 Classification of Computers on the Basis of Applications

Based on the application or the purpose, computers are often classified as general-purpose computers and special-purpose or dedicated computers. *General-purpose computers* are comparatively more flexible and thus can be used to work on a large variety of problems including business and scientific problems. For instance, banking applications such as financial accounting, pay-roll processing, etc., at the head-office level would require the services of a general-purpose computer. The size and capacity of a general-purpose computer could of course vary, depending upon the quantum of data and the

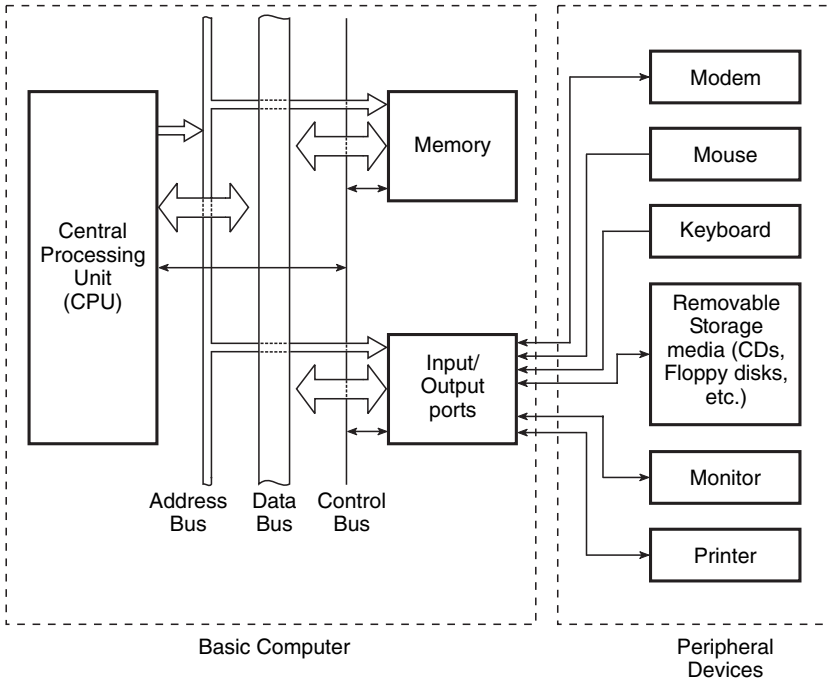


Figure 15.2 Block diagram of a typical computer system.

complexity of data processing to be done. *Special-purpose computers*, on the other hand, are designed for a dedicated application. These computers perform a certain predecided and fixed sequence of operations. Typical applications include the computers used for weather forecasting, aircraft control systems, missile and other weapon guidance systems, etc.

15.3.2 Classification of Computers on the Basis of the Technology Used

Based on the technology used, the computers are classified as analogue computers, digital computers and hybrid computers. In analogue computers, the input data comprise continuously changing electrical or nonelectrical (temperature, pressure, speed, volume, etc.) information. There are numerous examples of analogue computational devices. One such device is the speedometer of an automobile. The input data to this device or machine are the continuously varying rotational speed of its driveshaft. The rotational motion is converted into a linear movement of a needle pointer that indicates the speed in km/h. A tachometer used to measure the rotational speed is another device of the same type. The input data in the case of a digital computer are discrete in nature. They are represented by a binary notation in the form of 0s and 1s. A hybrid computer is a mixture of the two. It attempts to combine the good points of both analogue and digital computers. In a typical hybrid computer, the measuring functions are performed the analogue way while the control and logic functions are digital in nature.

15.3.3 Classification of Computers on the Basis of Size and Capacity

Based on their size and capacity, computers are classified as mainframe computers, minicomputers, microcomputers and supercomputers.

15.3.3.1 Mainframe Computers

A mainframe computer is the largest, fastest and perhaps one of the most expensive computer systems of general use. Before the advent of minicomputers and microcomputers respectively in the third- and fourth-generation periods, all data processing was done on mainframe systems only. Thousands of such machines are still in use in medium- and large-size business houses, universities, hospitals, etc.

These machines have a very large primary storage capability and have a very high processing speed. Because of their size and speed, mainframe systems must be placed on special platforms that allow wiring and cooling systems. These machines are useful not only because they have an enormous storage capacity but also because of their capability to support a large number of terminals. Modern-day mainframe computers are defined by their high-quality internal engineering, reliability, technical support and security features, along with their performance qualities. Their applications include the processing of a huge amount of different kinds of data such as census, industry/consumer statistics, financial transactions processing, etc., in large private and public enterprises, government agencies, etc. Examples of mainframe computers include IBM's zSeries and System z9 servers, Unisys's ClearPath mainframes, the zSeries 800 from Hitachi and IBM, the Nonstop systems from HP and so on.

15.3.3.2 Minicomputers

A minicomputer more or less resembles a mainframe system except that it is comparatively smaller and less expensive. They represent a class of multi-user computers that are used for middle-range computing applications, in between the mainframe systems and the microcomputers. Minicomputers were developed during the third-generation period. PDP-8 and PDP-11 from Digital Equipment Corporation (DEC) are examples of the popular minicomputers developed in the late 1960s. Minicomputers gave way to microcomputers in the mid-1980s and early 1990s.

15.3.3.3 Microcomputers

The microcomputer, the development of which was made possible largely owing to the development of the microprocessor, is a compact, relatively inexpensive and complete computer. The most obvious, though not the only difference between a microcomputer and a mainframe is the physical size. While a mainframe system may fill a room, a microcomputer may be put on a desktop or may even fit into a brief case. Although microcomputers can be distinguished from mainframe and minicomputers on the basis of size, technology used, applications and so on, these dividing lines are hazy and these categories almost overlap with each other owing to brisk advances in technology. Like mainframes and minis, today's microcomputers do data processing, manipulate lists, store, retrieve and sort information. Unlike mainframes and minis, microcomputers do not require any specialized environment for operation and can be effectively made use of by people who do not have any comprehensive formal training in computer techniques. In fact, these machines are designed to be used both at the workplace and at home. The concept of office automation has become feasible only with the advent of microcomputers.

15.3.3.4 Personal Computers

A personal computer, popularly known as a PC, is a stand-alone microcomputer that is used in a varied range of applications, from writing letters to the present-day desktop publishing, from playing video games to enquiring about railway and air schedules, from simple graphics to designing an advertisement, from simple financial accounting to preparing spread sheets and so on.

With the development of microprocessors and related peripherals, the personal computer of today is as powerful as a minicomputer of yesteryears. The processing speed has touched GHz and the hard disk capacity has reached tens of GBs. The contemporary microprocessors for the PC platform offer applications including internet audio and streaming video, image processing, video content creation, speech, computer-aided simulation and design, games, multimedia and multitasking user environments. Depending upon their size and portability, they can be classified as desktops, laptops and palmtops.

Desktops are personal computers for use on a desk in an office or at home. They are currently the most popular type of computer in use. Laptops, also referred to as notebooks, are mobile personal computers that can be carried in a briefcase. They do not always require an external power source and run on rechargeable batteries for 4–5 h. Some of the famous manufacturers of laptops include IBM, Compaq, Acer, Dell, HP and so on.

15.3.3.5 Workstations

Workstations are high-end technical computing desktop microcomputers designed primarily to be used by one person at a time, but they can also be connected remotely to other users if needed. They offer high performance compared with a personal computer, especially with respect to graphics, processing power and multitasking ability. Today, workstations use many technologies common to the personal computers.

15.3.3.6 Supercomputers

Supercomputers are the fastest and most powerful of all computer systems. They are typically 200 times faster than the mainframes. Supercomputers are mainly used for calculation-intensive applications requiring enormous amounts of data to be processed in a very short time. These include weather forecasting, weapons research, breaking secret codes, designing aircraft, molecular modelling, physical simulations and so on. Supercomputers are mainly used in universities, military agencies and scientific research laboratories. Supercomputers are highly parallel systems, i.e. they perform many tasks simultaneously. They generate a lot of heat and need a proper cooling mechanism. Some of the popular supercomputers include Cray-1, Cray X-MP/4, Cray-2, Intel's ASCI Red/9152 and ASCI Red/9632 and IBM's Blue Gene/L.

15.4 Computer Memory

Computer memory refers to components, devices, chips and recording media that are used for temporary, semi-permanent and permanent storage of data. As mentioned in the previous section, there are several types of memory device used in a computer. These include RAM, ROM, cache, flash memory, hard disk, floppy disk, CDs and so on. Memory devices can be broadly classified into two types, namely primary memory and secondary storage. Figure 15.3 shows the various types of memory device present in a typical computer system. It may be mentioned here that, in computer terminology, 'memory' usually refers to RAM and ROM and the term 'storage' refers to hard disks, floppy disks

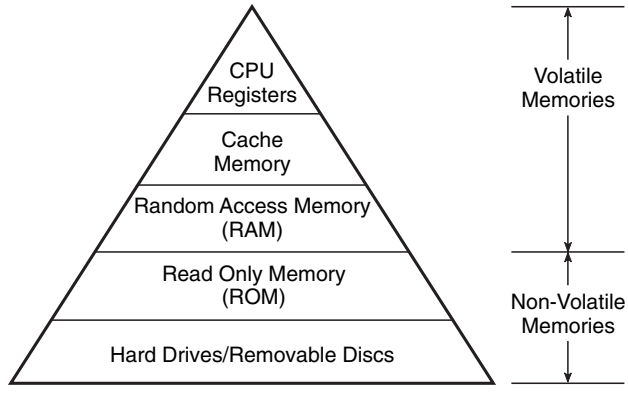


Figure 15.3 Various types of memory present in a typical computer system.

and CDs. Primary memory is described in this section, and secondary storage media are discussed in Section 15.10.

15.4.1 Primary Memory

The primary memory holds the program instructions for the program to be executed, the input data to be processed and the intermediate results of any calculations when processing is being done. Primary memory is also used for storing BIOS and start-up programs.

When a program and data are entered into a computer, the control unit directs them to the primary memory. Each program instruction and each data item is stored in a memory location that has a unique address. These data and instructions are held till new data items and instructions are written over them. Thus, the same data can be accessed repeatedly if so desired and the same instructions can be executed repeatedly if so required. This is what is known as the stored program concept. The primary memory of a computer further comprises process registers, random access memory (RAM), cache memory and read only memory (ROM). Process registers are memory cells built into the CPU that contain the specific data needed by the CPU. Cache memory is basically a type of RAM memory.

15.4.1.1 Random Access Memory

RAM is a read/write memory where the data can be read from or written into any of the memory locations regardless of the order in which they are arranged. Therefore, all the memory locations in a RAM can be accessed at the same speed. RAM is used to store data, program instructions and the results of any intermediate calculations during the execution of a program. Also, the same data can be read any number of times and different data can be written into the same memory location, with every fresh data item overwriting the existing one. It is typically used for short-term data storage as it cannot retain data when the power is turned off.

RAM is available in the form of ICs as well as in the form of plug-in modules. The plug-in modules are small circuit boards containing memory ICs and having input and output lines connected to an edge connector. They are available as single in-line memory modules (SIMMs) and dual in-line memory modules (DIMMs). More than one memory IC (or chip) can be used to build the RAM for

larger systems. The capacity or size of a RAM is measured in bytes. RAM chips are available in the memory capacities ranging from 2 kB to as much as 32 MB. 1 kB of memory equals $2^{10} = 1024$ bytes and 1 MB of memory equals 2^{20} bytes. The terms 'kilo' (k) and 'mega' (M) have been used, as 2^{10} and 2^{20} are approximately equal to 1000 and 1 000 000 respectively. As an illustration, a microcomputer with a 64 kB of RAM has $64 \times 2^{10} = 2^6 \times 2^{10} = 2^{16} = 65\ 536$ bytes of memory. The two categories of RAM are static RAM (SRAM) and dynamic RAM (DRAM). RAM is discussed in detail in Section 15.5.

15.4.1.2 Read Only Memory

In the case of ROM, instructions can be written into the memory only once at the manufacturer's premises. These instructions can, however, be read from a ROM as many times as desired. Once it is written, a ROM cannot be written into again. The contents of a ROM can thus be accessed by a CPU but cannot be changed by it. The instructions stored on a ROM vary with the type of application for which it is made. The ROM for a general-purpose microcomputer, for instance, would contain system programs such as those designed to handle operating system instructions.

In the case of some special types of ROM, it is possible for users to have their own instructions stored on the ROM as per their requirements. Such ROM chips are called PROMs (Programmable Read Only Memory). PROM contents, once programmed, cannot be changed. But then there are some special types of PROMs whose contents can be erased and then reprogrammed. These are known as EPROMs (Erasable Programmable Read Only Memory). ROM memories are discussed in detail in Section 15.6.

15.5 Random Access Memory

In this section we will discuss at length the types of RAM and their basic construction, properties, applications and so on.

RAM has three basic building blocks, namely an array of memory cells arranged in rows and columns with each memory cell capable of storing either a '0' or a '1', an address decoder and a read/write control logic. Depending upon the nature of the memory cell used, there are two types of RAM, namely static RAM (SRAM) and dynamic RAM (DRAM). In SRAM, the memory cell is essentially a latch and can store data indefinitely as long as the DC power is supplied. DRAM on the other hand, has a memory cell that stores data in the form of charge on a capacitor. Therefore, DRAM cannot retain data for long and hence needs to be refreshed periodically. SRAM has a higher speed of operation than DRAM but has a smaller storage capacity.

15.5.1 Static RAM

As mentioned before, the basic element of SRAM is a latch memory cell. Figure 15.4 shows a basic SRAM memory cell. The memory cell is selected by setting the 'select' line active. The data bit is written into the cell by placing it on the 'data in' line and is read from the 'data out' line.

SRAMs can be broadly classified as asynchronous SRAM and synchronous SRAM. Asynchronous SRAMs are those whose operations are not synchronized with the system clock, i.e. they operate independently of the clock frequency. 'Data in' and 'data out' in these RAMs are controlled by address transition. Synchronous SRAMs are those whose timings are initiated by clock edges. 'Address', 'data in', 'data out' and all other control signals are synchronized with the clock signal. Synchronous

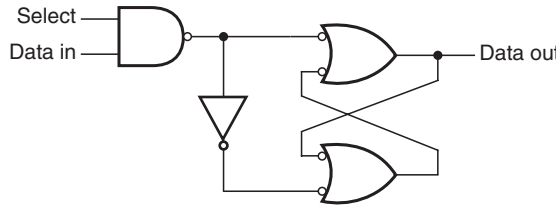


Figure 15.4 Basic SRAM memory cell.

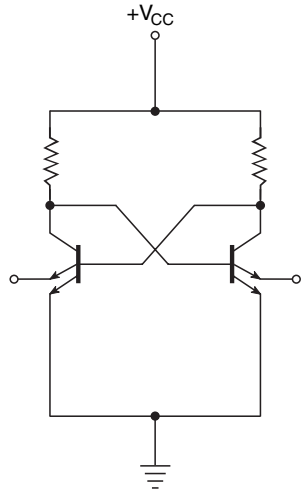
SRAMs normally have an address burst feature, which allows the memory to read and write at more than one location using a single address. Both synchronous and asynchronous SRAMs are available in bipolar, MOS and BiCMOS technologies. While bipolar SRAM offers a relatively higher speed of operation, MOS technology offers a higher capacity and reduced power consumption. Figures 15.5(a) and (b) show the basic bipolar memory cell and the MOS (NMOS more specifically) memory cell respectively.

15.5.1.1 Asynchronous SRAM

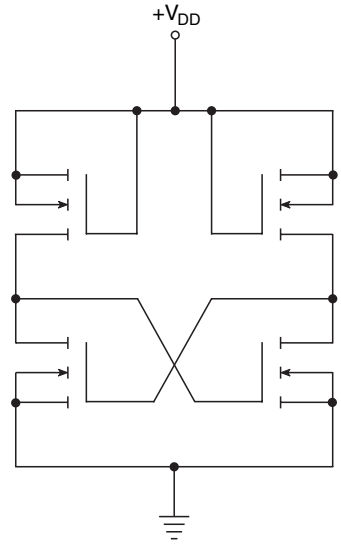
Figure 15.6 shows the typical architecture of a 64×8 asynchronous SRAM. It is capable of storing 64 words of eight bits each. The main blocks include a 6-to-64 line address decoder, I/O buffers, 64 memory cells and control logic for read/write operations. The memory cells in a row are represented as a register. Each register is an eight-bit register and can be read from as well as written into. As can be seen from the figure, all the cells inside the same register share the same decoder output line, also referred to as 'row line'. The control functions are provided by $R/\overline{W}$ (read/write) and $\overline{CS}$ (chip select) inputs. $R/\overline{W}$ and $\overline{CS}$ inputs are also referred to as $\overline{WE}$ (write enable) and $\overline{CE}$ (chip enable) inputs respectively. The 'data input' and 'data output' lines are usually combined by using common input/output lines in order to conserve the number of pins on the IC package.

The memory is selected by making $\overline{CS} = 0$. During the 'read' operation the status of the $R/\overline{W}$ and $\overline{CS}$ pins is '1' and '0' respectively, while during the 'write' operation it is '0' and '0' respectively. During the 'read' operation the input buffers are disabled and the contents of the selected register appear at the output. During the 'write' operation the input buffers are enabled and the output buffers are disabled. The contents of the input buffers are loaded into the selected register, the previous data of which are overwritten by the new data. The output buffers, being tristate, are in the high-impedance state during the write operation. $\overline{CS} = 1$ deselects the chip, and both the input and the output data buffers get disabled and go to the high-impedance state. The contents of the memory in this case remain unaffected. 'Chip select' inputs are particularly important when more than one RAM memory chip is combined to get a larger memory capacity.

In the case of larger SRAM memories, there are two address decoders, one for rows and one for columns. They are referred to as row decoders and column decoders respectively. Some of the address lines are fed to the row decoder and the rest of the address lines are fed to the column decoder. Figure 15.7 shows the architecture of a typical 16K×8 asynchronous SRAM. The memory cells are arranged in eight arrays of 128 rows and 128 columns each. Memories with a single address decoder are referred



(a)



(b)

Figure 15.5 (a) Basic bipolar memory cell and (b) a basic MOS memory cell.

to as two-dimensional memories, and those with two decoders are referred to as three-dimensional memories.

Figures 15.8(a) and (b) show the timing diagrams during 'read' and 'write' operations respectively. The diagrams are self-explanatory. Read and write cycle time intervals of a few nanoseconds to a few tens of nanoseconds are common in the case of asynchronous SRAMs.

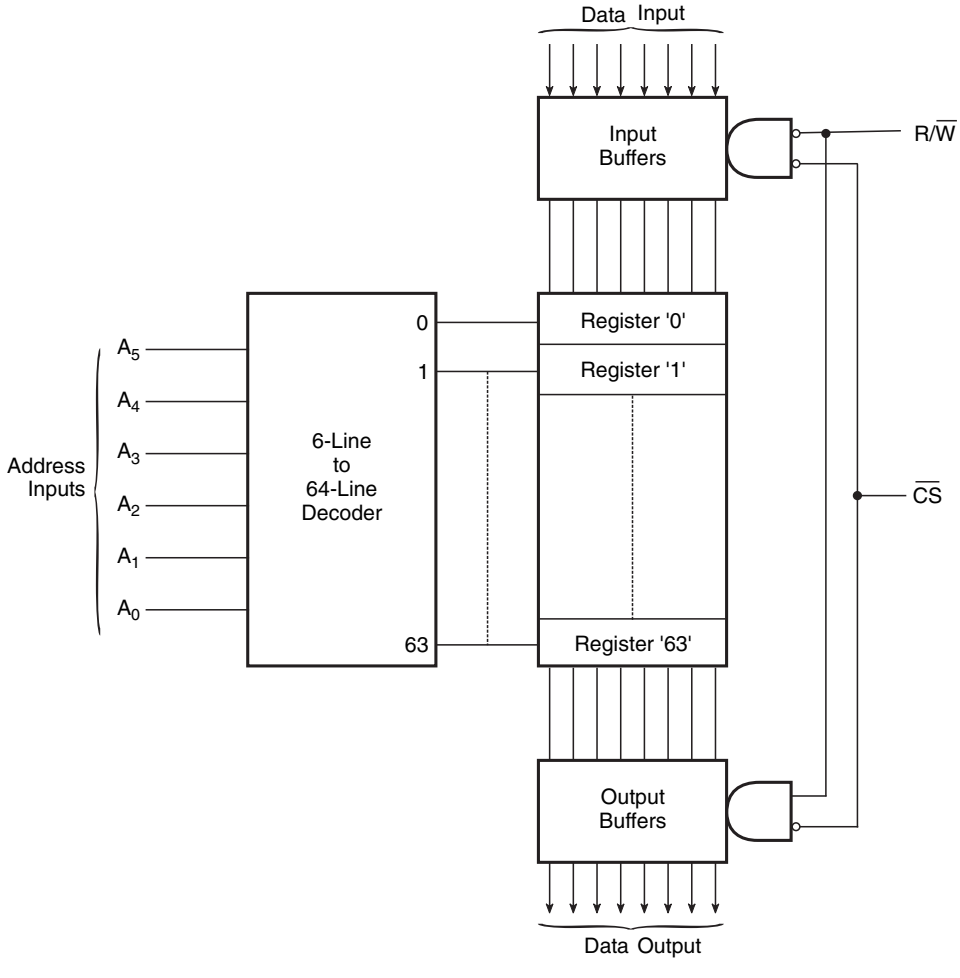


Figure 15.6 Typical architecture of a 64×8 asynchronous SRAM.

The different timing intervals shown in the diagram are defined as follows:

- Complete read cycle time t_{RC}. This is defined as the time interval for which a valid address code is applied to the address lines during the 'read' operation.
- RAM access time t_{ACC}. This is defined as the time lapse between the application of a new address input and the appearance of valid output data.
- Chip enable access time t_{CO}. This is defined as the time taken by the RAM output to go from the Hi-Z state to a valid data level once $\overline{CS}$ is activated.
- Chip disable access time t_{OD}. This is defined as the time taken by the RAM to return to the Hi-Z state after $\overline{CS}$ is deactivated.

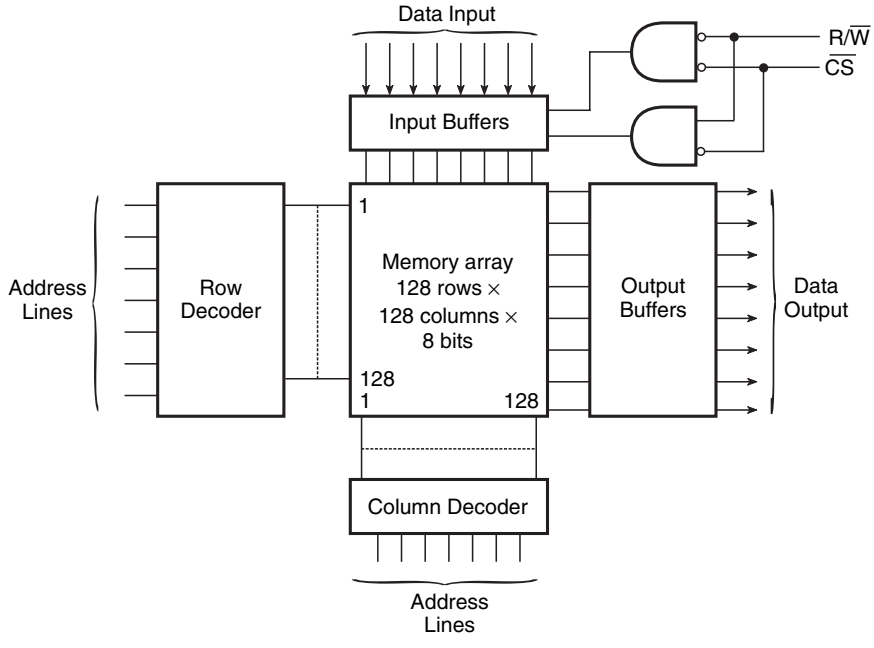
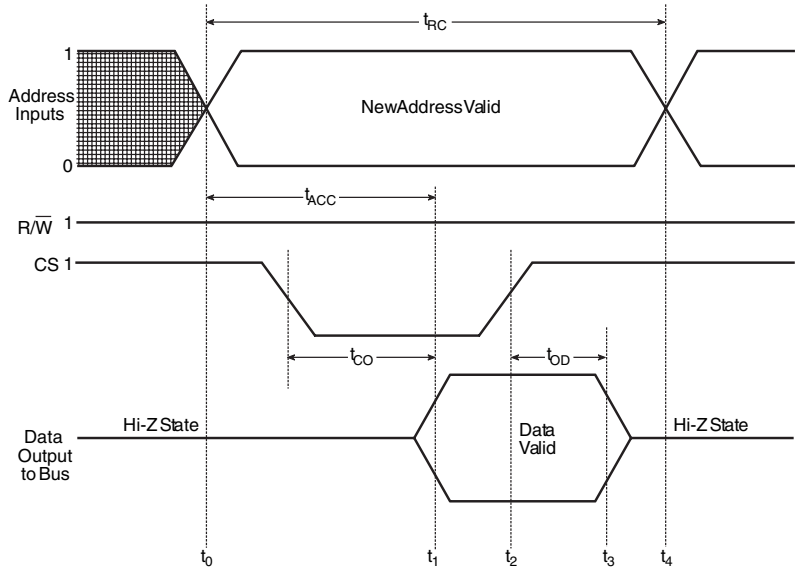


Figure 15.7 Typical architecture of a 16K×8 asynchronous SRAM.

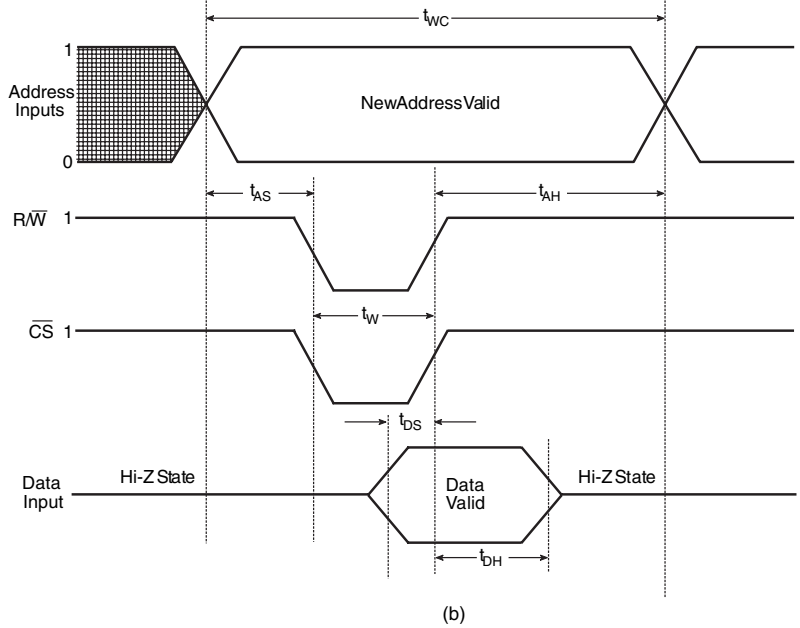
- Complete write cycle time t_{WC}. This is defined as the time interval for which a valid address code is applied to the address lines during the 'write' operation.
- Write pulse width t_W. This is the time for which $R/\overline{W}$ is held LOW during the 'write' operation.
- Address set-up time t_{AS}.This is the time interval between the appearance of a new address and $R/\overline{W}$ going LOW.
- Data set-up time t_{DS}. This is defined as the time interval for which the $R/\overline{W}$ must remain LOW after valid data are applied to the data inputs.
- Data hold time t_{DH}. This is defined as the time interval for which valid input data must remain on the data lines after the $R/\overline{W}$ input goes HIGH.
- Address hold time interval t_{AH}. This is defined as the time interval for which the valid address must remain on the address lines after the $R/\overline{W}$ input goes HIGH.

15.5.1.2 Synchronous SRAM

Synchronous SRAM, as mentioned before, is synchronized with the system clock. In the case of a computer system it operates at the same clock frequency as the microprocessor. This synchronization of microprocessor and memory ensures faster execution speeds. The basic difference between the architecture of synchronous and asynchronous SRAMs is that the synchronous SRAM makes use of clocked registers to synchronize 'address', $R/\overline{W}$, $\overline{CS}$ and 'data in' lines to the system clock. Figure 15.9 shows the basic architecture of a 32K × 8 synchronous SRAM with a burst feature. As we can see from the figure, the memory array block, the address decoder block and $R/\overline{W}$ and $\overline{CS}$ are the same



(a)



(b)

Figure 15.8 (a) Timing diagram during a READ operation and (b) the timing diagram during a WRITE operation.

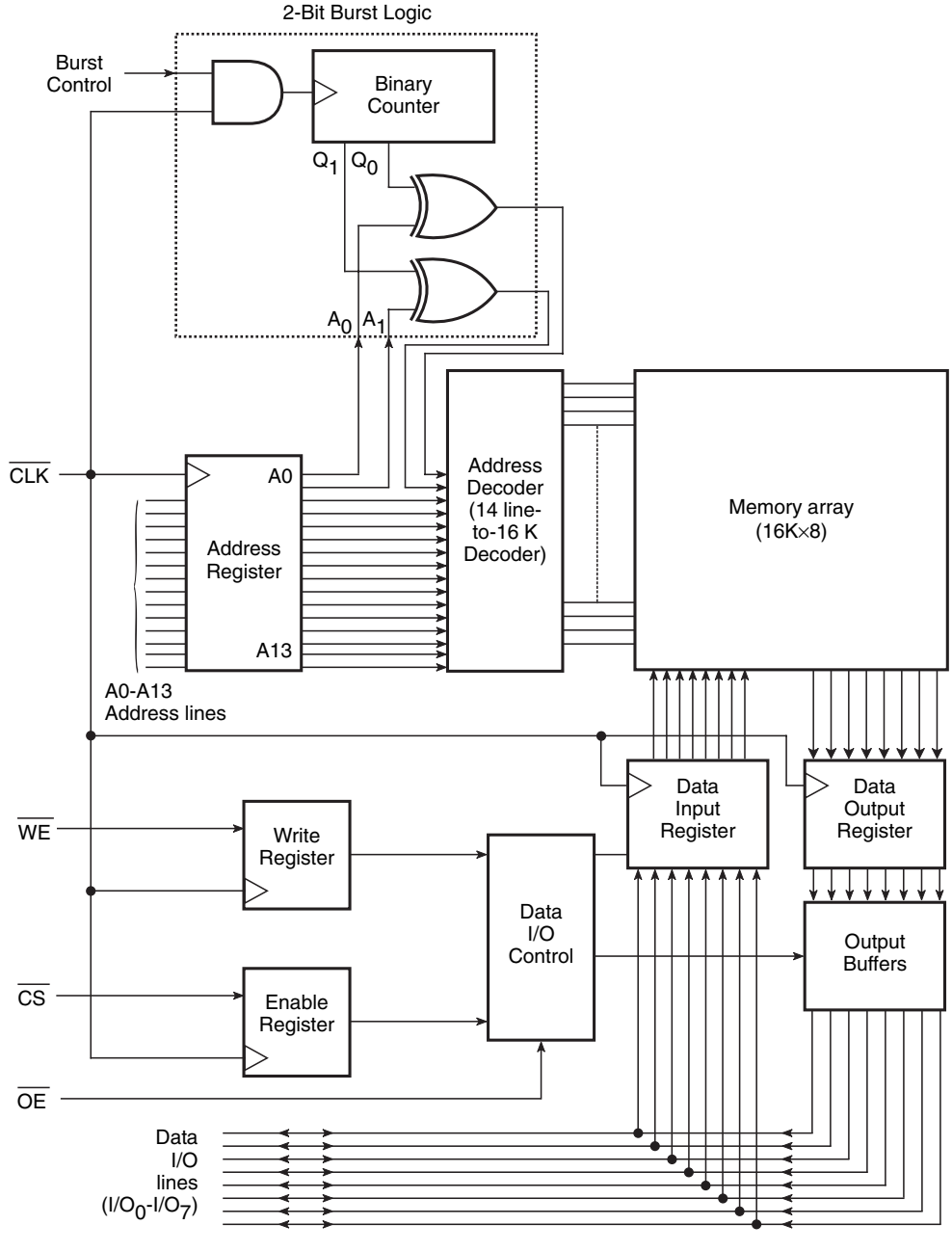


Figure 15.9 Architecture of a 16K ×8 synchronous SRAM.

as in the case of an asynchronous SRAM. As mentioned before, most synchronous SRAMs have an address burst feature. In this case, when an external address is latched to the address register, a certain number of lowest address bits are applied to the burst logic. Burst logic comprises a binary counter and EXCLUSIVE-OR gates. The output of the burst logic, which basically produces a sequence of internal addresses, is fed to the address bus decoder. In the case of a two-bit burst logic, the internal address sequence generated is given by A_1A_0, $A_1\overline{A_0}$, $\overline{A_1}A_0$, $\overline{A_1}\ \overline{A_0}$, where A_0 and A_1 are the address bits applied to the burst logic. The burst logic shown in Fig. 15.9 is also a two-bit logic.

15.5.2 Dynamic RAM

The memory cell in the case of a DRAM comprises a capacitor and a MOSFET. The cell holds a value of '1' when the capacitor is charged and '0' when it is discharged. The main advantage of this type of memory is its higher density, or more bits per package, compared with SRAM. This is because the memory cell is very simple compared with that of SRAM. Also, the cost per bit is less in the case of a DRAM. The disadvantage of this type of memory is the leakage of charge stored on the capacitors of various memory cells when they are storing a '1'. To prevent this from happening, each memory cell in a DRAM needs to be periodically read, its charge (or voltage) compared with a reference value and then the charge restored to the capacitor. This process is known as 'memory refresh' and is done approximately every 5–10 ms.

Figure 15.10 shows the basic memory cell of a DRAM and its principle of operation. The MOSFET acts like a switch. When in the 'write' mode ($R/\overline{W} = 0$), the input buffers are enabled while the output buffers are disabled. When '1' is to be stored in the memory, the 'data in' line must be in the HIGH state and the corresponding 'row line' should also be in the HIGH state so that the MOSFET is switched ON. This connects the MOSFET to the 'data in' line, and it charges the capacitor to a positive voltage level. When '0' needs to be stored, the 'data in' line is LOW and the capacitor also acquires the same level. When the 'row line' is taken to the LOW state, the MOSFET is switched OFF and is disconnected from the bit line. This traps the charge on the capacitor. In 'read' mode ($R/\overline{W} = 1$), the output buffers are enabled while the input buffers are disabled. When the 'row line' is taken to HIGH logic, the MOSFET is switched ON and connects the capacitor to the 'data out' line through the output

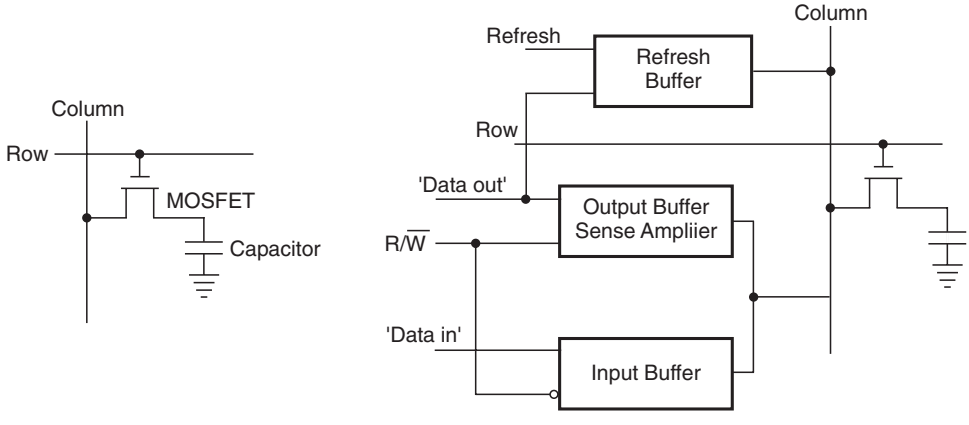


Figure 15.10 Basic memory cell of a DRAM.

buffer. The refresh operation is performed by setting $R/\overline{W} = 1$ and by enabling the refresh buffer. There are two basic modes of refreshing the memory, namely the burst refresh and distributed refresh modes. In burst refresh mode, all rows in the memory array are refreshed consecutively during the refresh burst cycle. In distributed refresh mode, each row is refreshed at intervals interspaced between 'read' and 'write' operations.

15.5.2.1 DRAM Architecture

The architecture of DRAM memory is somewhat different from that of SRAM memory. Row and column address lines are usually multiplexed in a DRAM. This is done to reduce the number of pins on the package. Row address select (RAS) and column address select (CAS) inputs are used to indicate whether a row or a column is to be addressed. Address multiplexing is particularly attractive for higher-capacity DRAMs. A 4 MB RAM, for instance, would require 22 address inputs ($2^{22} = 4M$).

Figure 15.11 shows the architecture of a 16K × 1 DRAM. The heart of a DRAM is an array of single-bit memory cells. Each cell has a unique position as regards row and column. Other important blocks include address decoders (row decoder and column decoder) and refresh control and address latches (row address latch and column address latch). As can be seen from the figure, seven address lines are time multiplexed at the beginning of the memory cycle by the RAS and CAS lines. Firstly, the seven-bit address (A_0–A_6) is latched into the row address latch, and then the seven-bit address is latched into the column address latch (A_7–A_{13}). They are then decoded to select the particular memory location. Larger word sizes can be achieved by combining more than one chip. This is discussed in the next section. Figures 15.12(a) and (b) respectively show the timing diagrams for read and write operations. A DRAM is relatively slower than a SRAM. The typical access time is in the range 100–250 ns.

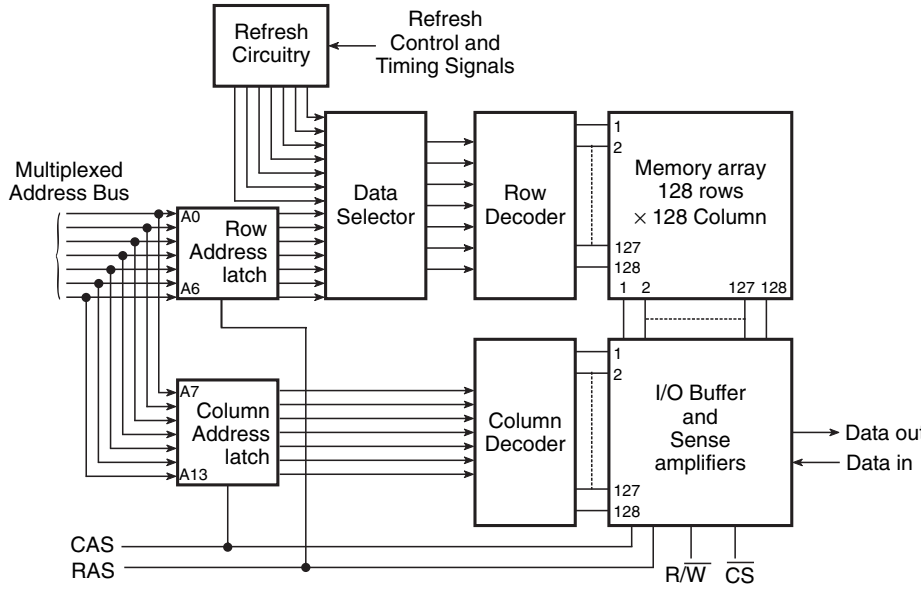


Figure 15.11 Architecture of a 16K × 1 DRAM.

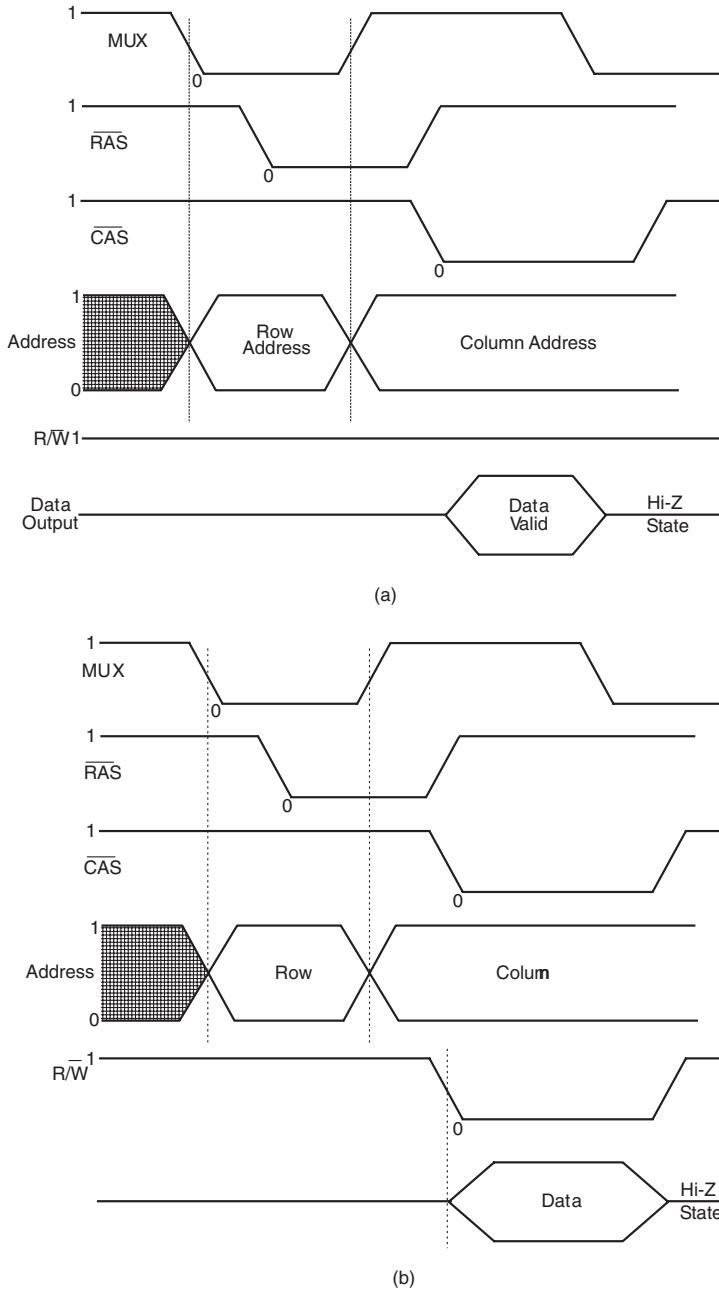


Figure 15.12 (a) Timing diagrams for a READ operation and (b) timing diagrams for a WRITE operation.

15.5.2.2 Types of DRAM

DRAM memories can be further classified as fast page mode (FPM) DRAM, extended data output (EDO) DRAM, burst extended data output (BEDO) DRAM and synchronous (S) DRAM. In FPM DRAM, the row address is specified only once for access to several successive column addresses. Hence, the read and write times are reduced. EDO DRAM is similar to FPM DRAM, with the additional feature that a new access cycle can be started while keeping the data output of the previous cycle active. BEDO DRAM is an EDO DRAM with address burst capability. All the types of DRAM discussed hitherto are asynchronous DRAMs, and their operation is not synchronized with the system clock. SDRAM, as the name suggests, is a synchronous DRAM whose operation is synchronized with the system clock.

15.5.3 RAM Applications

One of the major applications of RAM is its use in cache memories. It is also used as main memory to store temporary data and instructions in a computer.

15.5.3.1 Cache Memory

Advances in microprocessor technology and also the software have greatly enhanced the application potential of present-day computers. These enhanced performance features and increased speed can be optimally utilized to the maximum only if the computer has the required capacity of main (or internal) memory. The computer's main memory, as we know, stores program instructions and data that the CPU needs during normal operation. In order to get the maximum out of the system, this would normally require all of the system's main memory to have a speed comparable with that of the CPU. It is not economical for all the main memory to be high speed. This is where the cache memory comes in.

Cache memory is a block of high-speed memory located between the main memory and the CPU. The cache memory block is the one that communicates directly with the CPU at high speed. It stores the most recently used instructions or data. When the processor needs data, it checks in the high-speed cache to see if the data are there. If they are there, called a 'cache hit', the CPU accesses the data from the cache. If they are not there, called a 'cache miss', then the CPU retrieves them from the relatively slower main memory. Cache memory mostly uses SRAM chips, but it can also use DRAM.

There are two levels of cache memory. The first is the level 1 cache (L1 or primary or internal cache). It is physically a part of the microprocessor chip. The second is the level 2 cache (L2 or secondary or external cache). It is in the form of memory chips mounted external to the microprocessor. It is larger than the L1 cache. The L1 and L2 cache memories range from 2 to 64 kB and from 256 kB to 2 MB in size respectively. Some systems have higher-level caches (L3, L4, etc.), but L1 and L2 are the most common. Figure 15.13 shows the use of L1 and L2 cache memories in a computer system.

15.6 Read Only Memory

ROM is a nonvolatile memory that is used for permanent or semi-permanent storage of data. The contents of ROM are retained even after the power is turned off. In this section we will be discussing at length the ROM architecture, types of ROM and typical applications.

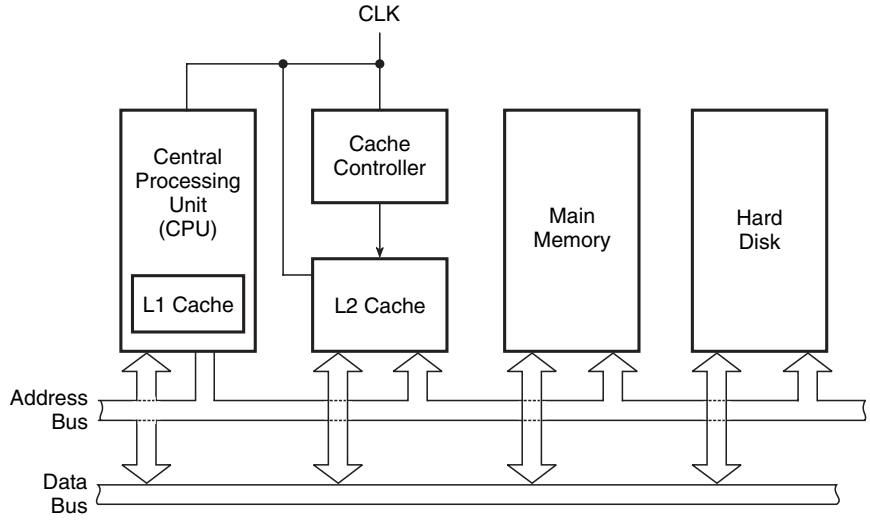


Figure 15.13 Cache memory in a computer system.

15.6.1 ROM Architecture

The internal structure or architecture of a ROM comprises three basic parts, namely the array of memory cells, the address decoder and the output buffers. The address decoder comprises a single decoder in the case of small memories. In the case of large memories it comprises two decoders referred to as row and column decoders. The operation of a ROM can be best explained with the help of the simplified representation of a 32×8 ROM, as shown in Fig. 15.14.

The array of memory cells stores the data to be programmed into the ROM. The number of memory cells in a row equals the word size, and the number of memory cells in a column equals the number of such words to be stored. In the memory shown in Fig. 15.14, the word size is eight bits and the number of words is 32. The data outputs of each of the memory cells in the array are connected to an internal data bus that runs through the entire circuit. The address decoder, a 1-of-32 decoder in this case, sets the corresponding 'row line' HIGH when a binary address is applied at its input lines. A five-bit address code ($A_4A_3A_2A_1A_0$) is needed to address 32 memory cells. As an illustration, an address code of 10011 will identify the nineteenth row.The output is read from the column lines. The data placed on the internal data bus by the memory cells are fed to the output buffers. $\overline{CS}$ is an active LOW input used to select the memory device. In the case of larger memories, the address decoder comprises row as well as column decoders. Let us consider a 2K-bit ROM device with 256×8 organization. The memory is arranged in the format of a 32×64 matrix instead of a 256×8 matrix. Five of the address lines are connected to the row decoder, and the remaining three lines are connected to the column decoder. The row decoder is a 1-of-32 decoder, and it selects one of the 32 rows. The column decoder comprises eight 1-of-8 decoders. It selects eight of the total 64 columns. Thus, an eight-bit word appears on the data output when the address is applied and $\overline{CS} = 0$.

Figure 15.15 shows the typical timing diagram of a ROM read operation. It shows that there is a time delay that occurs between the application of an address input and the availability of corresponding data at the output. It is this time delay that determines the ROM operating speed. This time delay is

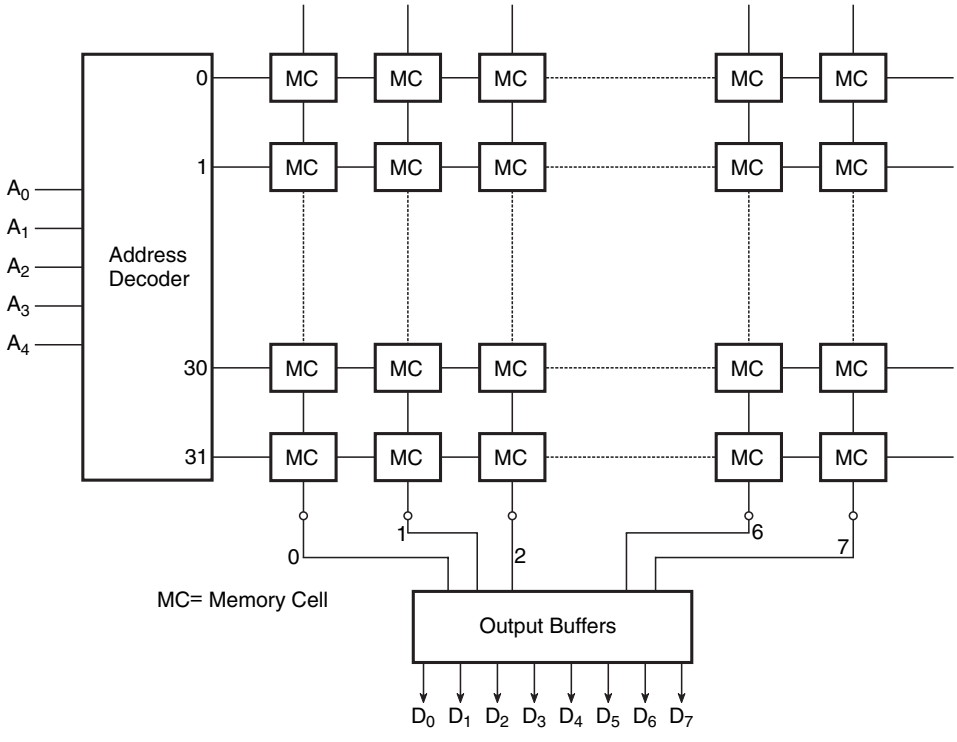


Figure 15.14 Architecture of 32 × 8 ROM.

known as the access time, t_{ACC}. Another useful timing parameter is the output enable time, t_{OE}, which is the time delay between application of input and appearance of valid data output.

Typical bipolar ROMs have access times of 30–90 ns. In the case of NMOS devices, the access times range from 35 to 500 ns. The output enable time, t_{OE}, in the case of bipolar ROMs is in the range 10–20 ns. For MOS-based ROMs, t_{OE} is in the range 25–100 ns.

15.6.2 Types of ROM

Depending upon the methodology of programming, erasing and reprogramming information into ROMs, they are classified as mask-programmed ROMs, programmable ROMs (PROMs) and erasable programmable ROMs (EPROMs) [ultraviolet-erasable programmable ROMs (UV EPROMs) and electrically erasable programmable ROMs (EEPROMs)].

15.6.2.1 Mask-programmed ROM

In the case of a mask-programmed ROM, the ROM is programmed at the manufacturer's site according to the specifications of the customer. A photographic negative, called a mask, is used to store the required data on the ROM chip. A different mask would be needed for storing each different set

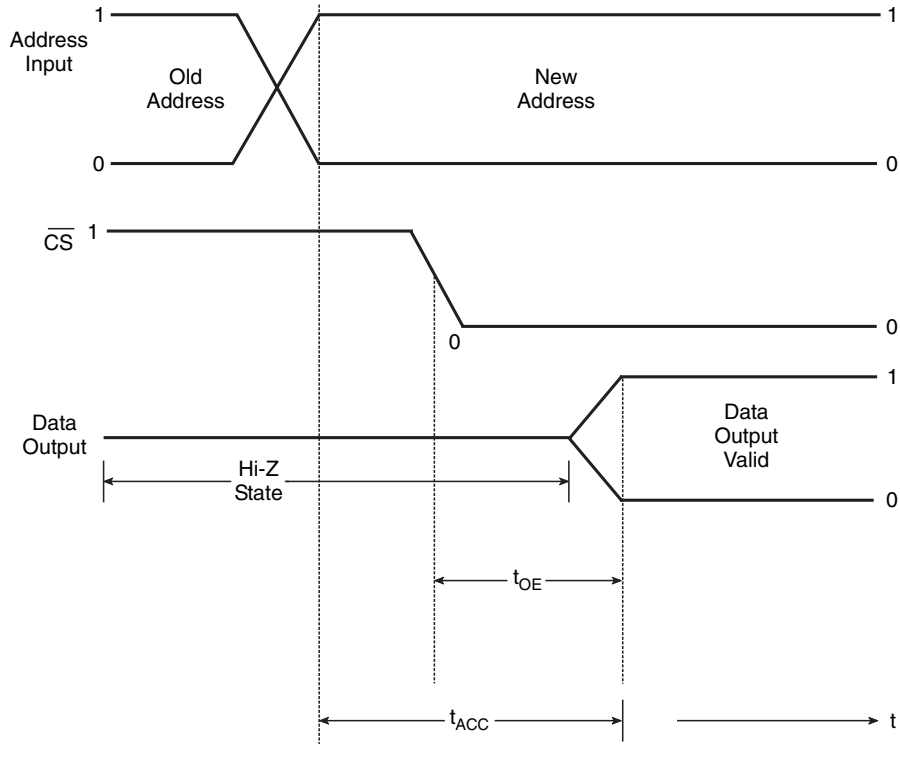


Figure 15.15 Typical timing diagram of a ROM READ operation.

of information. As preparation of a mask is an expensive proposition, mask-programmed ROM is economical only when manufactured in large quantities. The limitation of such a ROM is that, once programmed, it cannot be reprogrammed.

The basic storage element is an NPN bipolar transistor, connected in common-collector configuration, or a MOSFET in common drain configuration. Figures 15.16(a) and (b) show a MOSFET-based basic cell connection when storing a '1' and '0' respectively. As is clear from the figure, the connection of the 'row line' to the gate of the MOSFET stores '1' at the location when the 'row line' is set to level '1'. A floating-gate connection is used to store '0'. Figures 15.16(c) and (d) show the basic bipolar memory cell connection when storing a '1' and '0' respectively.

Figure 15.17 shows the internal structure of a 4×4 bipolar mask-programmed ROM. The data programmed into the ROM are given in the adjoining truth table. The transistors with an open base store a '0', whereas those with their bases connected to the corresponding decoder output store a '1'. As an illustration, transistors Q_{30}, Q_{20}, Q_{10} and Q_{00} in row 0 store '1', '0', '1' and '0' respectively. The stored information in a given row is available at the output when the corresponding decoder is enabled, and that 'row line' is set to level '1'. The output of the memory cells appears at the column lines. For example, when the address input is '11', row 3 is enabled and the data item at the output is 0110.

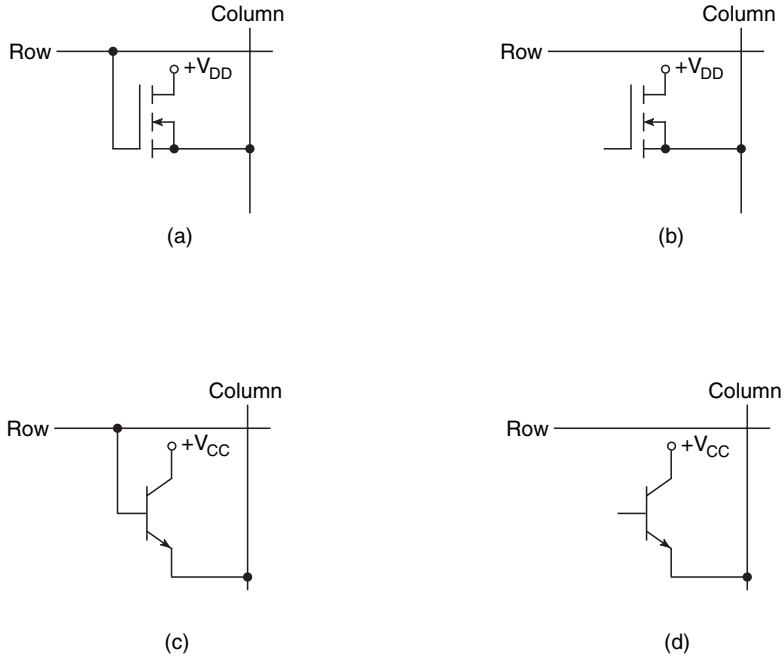


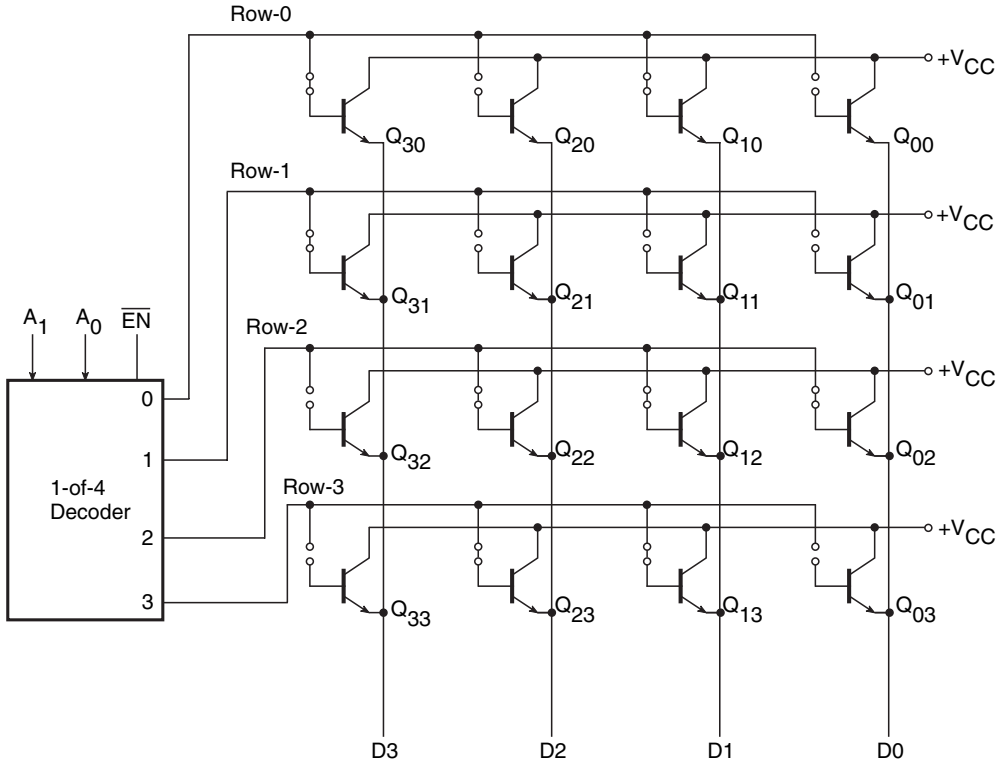
Figure 15.16 Basic cell connection of a mask-programmed ROM.

In the ROM architecture shown in Fig. 15.17, the number of memory cells in a row represents the word size. The four memory cells in a row here constitute a four-bit register. There are four such registers in this ROM. In a 16×8 ROM of this type there will be 16 rows of such transistor cells, with each row having eight memory cells. The decoder in that case would be a 1-of-16 decoder.

15.6.2.2 Programmable ROM

In the case of PROMs, instead of being done at the manufacturer's premises during the manufacturing process, the programming is done by the customer with the help of a special gadget called a PROM programmer. Since the data, once programmed, cannot be erased and reprogrammed, these devices are also referred to as one-time programmable ROMs.

The basic memory cell of a PROM is similar to that of a mask-programmed ROM. Figures 15.18(a) and (b) show a MOSFET-based memory cell and bipolar memory cell respectively. In the case of a PROM, each of the connections that were left either intact or open in the case of a mask-programmed ROM are made with a thin fusible link, as shown in Fig. 15.18. The different interconnect technologies used in programmable logic devices are comprehensively covered in Chapter 9. Basic fuse technologies used in PROMs are metal links, silicon links and PN junctions. These fusible links can be selectively blown off to store desired data. A sufficient current is injected through the fusible link to burn it open to store '0'. The programming operation, as said earlier, is done with a PROM programmer. The PROM chip is plugged into the socket meant for the purpose. The programmer circuitry selects each address of the PROM one by one, burns in the required data and then verifies the correctness of the



Truth Table

Address		Data			
A_1	A_0	D_3	D_2	D_1	D_0
0	0	1	0	1	0
0	1	1	0	0	0
1	0	0	1	1	1
1	1	0	1	1	0

Figure 15.17 Internal structure of a 4×4 bipolar mask-programmed ROM.

data before proceeding to the next address. The data are fed to the programmer from a keyboard or a disk drive or from a computer.

PROM chips are available in various word sizes and capacities. 27LS19, 27S21, 28L22, 27S15, 24S41, 27S35, 24S81, 27S45, 27S43 and 27S49 are respectively 32×8, 256×4, 256×8, 512×8, $1K \times 4$, $1K \times 8$, $2K \times 4$, $2K \times 8$, $4K \times 8$ and $8K \times 8$ PROMS. The typical access time in the case of these devices is in the range 50–70 ns. MOS PROMs are available with much greater capacities than bipolar PROMs. Also, the power dissipation is much lower in MOS PROMs than it is in the case of bipolar PROMs with similar capacities.

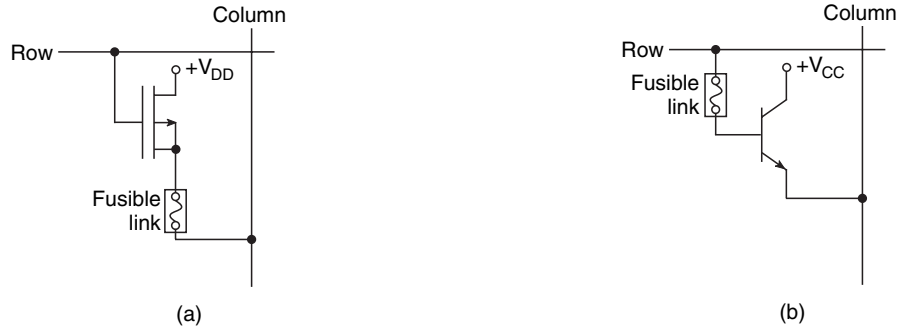


Figure 15.18 Basic memory cell of a PROM.

15.6.2.3 Erasable PROM

EPROM can be erased and reprogrammed as many times as desired. Once programmed, it is nonvolatile, i.e. it holds the stored data indefinitely. There are two types of EPROM, namely the ultraviolet-erasable PROM (UV EPROM) and electrically erasable PROM (EEPROM).

The memory cell in a UV EPROM is a MOS transistor with a floating gate. In the normal condition, the MOS transistor is OFF. It can be turned ON by applying a programming pulse (in the range 10–25 V) that injects electrons into the floating-gate region. These electrons remain trapped in the gate region even after removal of the programming pulse. This keeps the transistor ON once it is programmed to be in that state even after the removal of power. The stored information can, however, be erased by exposing the chip to ultraviolet radiation through a transparent window on the top of the chip meant for the purpose. The photocurrent thus produced removes the stored charge in the floating-gate region and brings the transistor back to the OFF state. The erasing operation takes around 15–20 min, and the process erases information on all cells of the chip. It is not possible to carry out any selective erasure of memory cells. Intel's 2732 is 4K × 8 UV EPROM hardware implemented with NMOS devices. Type numbers 2764, 27128, 27256 and 27512 have capacities of 8K × 8, 16K × 8, 32K × 8 and 64K × 8 respectively. The access time is in the range 150–250 ns. UV EPROMs suffer from disadvantages such as the need to remove the chip from the circuit if it is to be reprogrammed, the nonfeasibility of carrying out selective erasure and the reprogramming process taking several tens of minutes. These are overcome in the EEPROMs and flash memories discussed in the following paragraphs.

The memory cell of an EEPROM is also a floating-gate MOS structure with the slight modification that there is a thin oxide layer above the drain of the MOS memory cell. Application of a high-voltage programming pulse between gate and drain induces charge in the floating-gate region which can be erased by reversing the polarity of the pulse. Since the charge transport mechanism requires very low current, erasing and programming operations can be carried out without removing the chip from the circuit. EEPROMs have another advantage – it is possible to erase and rewrite data in the individual bytes in the memory array. The EEPROMs, however, have lower density (bit capacity per square mm of silicon) and higher cost compared with UV EPROMs.

15.6.2.4 Flash Memory

Flash memories are high-density nonvolatile read/write memories with high density. Flash memory combines the low cost and high density features of an UV EPROM and the in-circuit electrical

erasability feature of EEPROM without compromising the high-speed access of both. Structurally, the memory cell of a flash memory is like that of an EPROM. The basic memory cell of a flash memory is shown in Fig. 15.19. It is a stacked-gate MOSFET with a control gate and floating gate in addition to drain and source. The floating gate stores charge when sufficient voltage is applied to the control gate. A '0' is stored when there is more charge, and a '1' when there is less charge. The amount of charge stored on the floating gate determines whether or not the MOSFET is turned ON.

It is called a flash memory because of its rapid erase and write times. Most flash memory devices use a 'bulk erase' operation in which all the memory cells on the chip are erased simultaneously. Some flash memory devices offer a 'sector erase' mode in which specific sectors of the memory device can be erased at a time. This mode comes in handy when only a portion of the memory needs to be updated.

Figure 15.20 shows the basic array of a 4×4 flash memory. As in the case of earlier memories, there is an address decoder that selects the row. During the read operation, for a cell containing a '1' there is current through the bit line which produces a voltage drop across the active load. This is compared with the reference voltage, and the output bit is '1'. If the memory cell has a '0', there is very little current in the bit line. Memory sticks are flash memories. They are available in 4, 8, 16, 32, 64 and 128 MB sizes.

To sum up, while PROMs are least complex and low cost, they cannot be erased and reprogrammed. UV EPROMs are a little more complex and costly, but then they can be erased and reprogrammed by being taken out of the circuit. Flash memories are in-circuit electrically erasable either sectorwise or in bulk mode. The most complex and most expensive are the EEPROMs, but then they offer byte-by-byte electrical erasability in circuit.

15.6.3 Applications of ROMs

The majority of ROM applications originate from the need for nonvolatile storage of data or program codes. Some of the common application areas include firmware, bootstrap memory, look-up tables, function generators and auxiliary memory.

The most common application of ROM chips is in the storage of data and program codes that must be made available to microprocessor-based systems such as microcomputers on power-up. This component of the software is referred to as firmware as it comes embedded in the hardware with the machine. Even consumer products such as CD players, microwave ovens, washing machines, etc., have embedded microcontrollers that have a microprocessor to control and monitor the operation according to the information stored on the ROM.

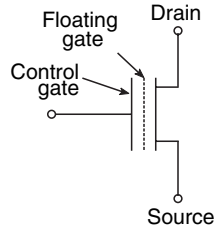


Figure 15.19 Basic cell of flash memory.

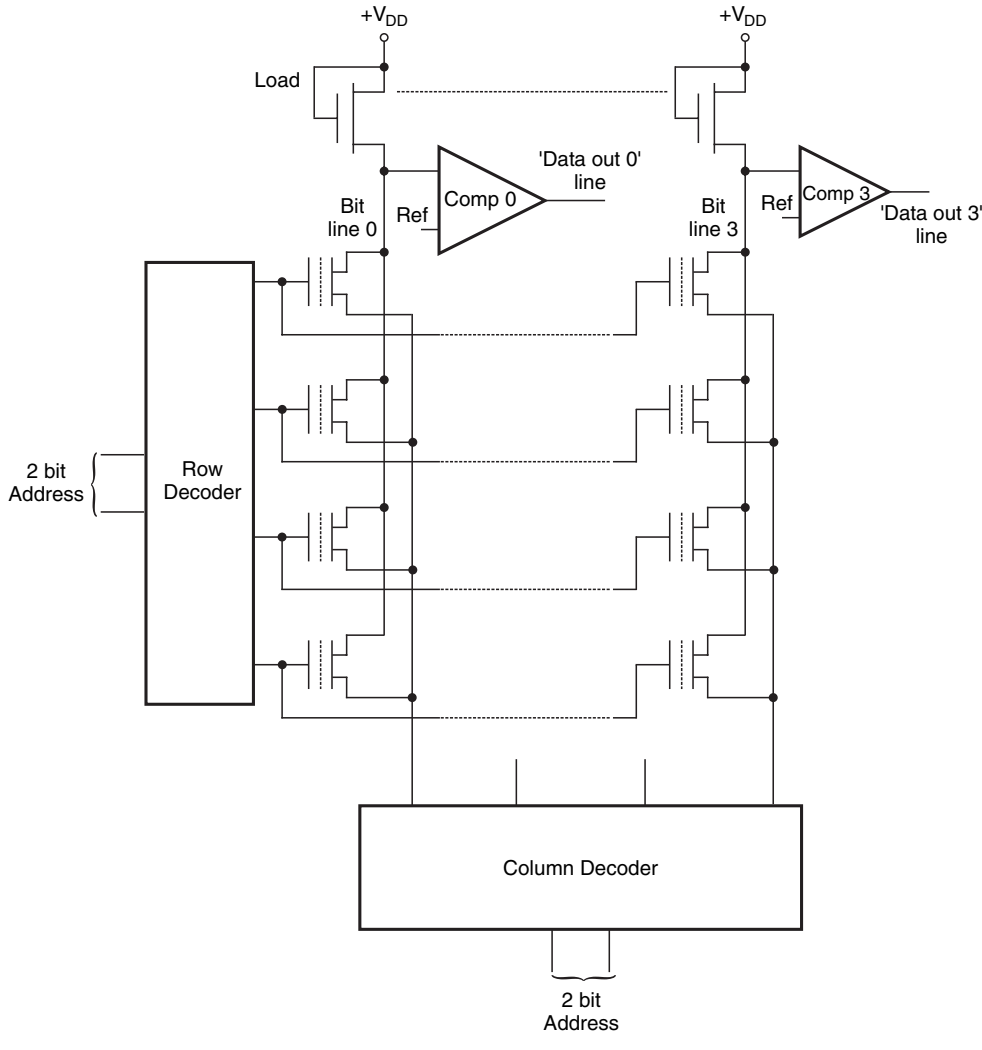


Figure 15.20 Basic array of 4×4 flash memory.

ROMs are also used to store the 'bootstrap program' in computers. It is a relatively small program containing instructions that will cause the CPU to initialize the system hardware after it is powered on. The bootstrap program then loads the operating system programs stored in the secondary memory into its main internal memory. The computer then begins to execute the operating system program. This start-up operation is also called the 'booting operation'.

ROMs are frequently used as 'look-up tables'. There are two sets of data, one constituting the address and the other corresponding to the data stored in various memory locations of the ROM. Corresponding to each address input, there is a unique data ouput. One typical application is that of

code conversion. As an illustration, a ROM can be used to build a binary-to-BCD converter where each memory location stores the BCD equivalent of the corresponding address code expressed in binary.

A ROM can be an important building block in a waveform generator. In a typical waveform generation set-up, ROM is used as a look-up table, with each of its memory locations storing a unique digital code corresponding to a different amplitude of the waveform to be generated. The address inputs of the ROM are fed from the output of a counter. The data outputs of ROM feed a D/A converter whose output constitutes the desired analogue waveform. This concept is also utilized in speech synthesizers, where the digital equivalent of speech waveform values are stored in the ROM.

Today, ROMs have become a viable alternative to the use of magnetic disks for auxiliary storage, more so for lower-capacity requirements. The low power consumption of flash memories, for instance, makes them particularly attractive for notebook computers.

Example 15.1

A certain ROM is capable of storing 16 kB of data. If the internal architecture of the ROM uses a square matrix of registers, determine (a) the number of registers in each row, (b) the number of registers in each column, (c) the total number of address inputs, (d) the type of row decoder and (e) the type of column decoder.

Solution

(a) The ROM capacity = 16K = 16 × 1024 = 16 384 bytes. Therefore, the total number of registers = 16 384. Since the registers are arranged in a square matrix, the number of rows equals the number of columns. The number of registers in each row = 128.
(b) The number of registers in each column = 128.
(c) The total number of memory locations = 16 384 = 2^{14}. Therefore, the total number of address inputs = 14.
(d) 1-of-7 decoder.
(e) 1-of-7 decoder.

Example 15.2

Determine the minimum size of a ROM required to convert a four-bit straight binary code into a Gray code equivalent. Also, write data to be programmed in various memory locations of the ROM.

Solution

- Table 15.1 shows the four-bit straight binary numbers and their Gray code equivalents.
- It is clear from the table that the MSB of the straight binary number is the same as the MSB of the Gray code equivalent.
- This can therefore be passed on as such to the output.
- In that case, each memory location of the ROM needs to store only three-bit data as the fourth bit is available as such from the input.
- The required size of the ROM is therefore 16 × 3.
- The three-bit data to be programmed into 16 different memory locations of the ROM corresponding to address inputs of 0000 to 1111 in the same order would be 000, 001, 011, 010, 110, 111, 101, 100, 100, 101, 111, 110, 010, 011, 001 and 000.
- Figure 15.21 shows this in ROM representation.

Table 15.1 Example 15.2.

Binary code				Gray code			
A_3	A_2	A_1	A_0	D_3	D_2	D_1	D_0
0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	1
0	0	1	0	0	0	1	1
0	0	1	1	0	0	1	0
0	1	0	0	0	1	1	0
0	1	0	1	0	1	1	1
0	1	1	0	0	1	0	1
0	1	1	1	0	1	0	0
1	0	0	0	1	1	0	0
1	0	0	1	1	1	0	1
1	0	1	0	1	1	1	1
1	0	1	1	1	1	1	0
1	1	0	0	1	0	1	0
1	1	0	1	1	0	1	1
1	1	1	0	1	0	0	1
1	1	1	1	1	0	0	0

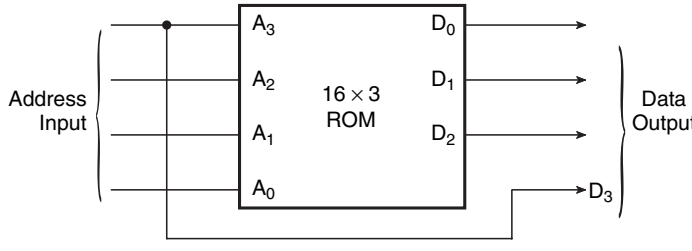


Figure 15.21 Solution to problem 15.2.

15.7 Expanding Memory Capacity

When a given application requires a RAM or ROM with a capacity that is larger than what is available on a single chip, more than one such chip can be used to achieve the objective. The required enhancement in capacity could be either in terms of increasing the word size or increasing the number of memory locations. How this can be achieved is illustrated in the following paragraphs with the help of examples.

15.7.1 Word Size Expansion

Let us take up the task of expanding the word size of an available 16×4 RAM chip from four bits to eight bits. Figure 15.22 shows a diagram where two such RAM chips have been used to achieve the

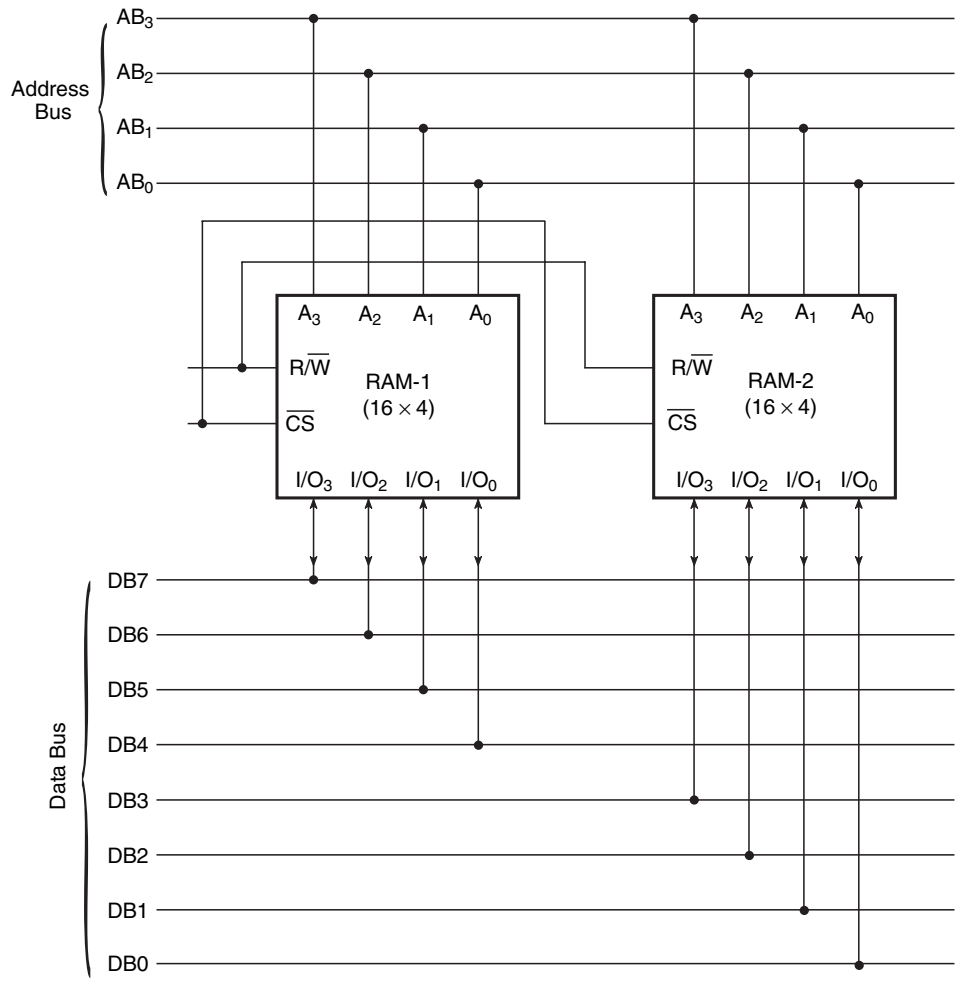


Figure 15.22 Word size expansion.

desired effect. The arrangement is straightforward. Both chips are selected or deselected together. Also, the input that determines whether it is a 'read' or 'write' operation is common to both chips. That is, both chips are selected for 'read' or 'write' operation together. The address inputs to the two chips are also common. The memory locations corresponding to various address inputs store four higher-order bits in the case of RAM-1 and four lower-order bits in the case of RAM-2. In essence, each of the RAM chips stores half of the word. Since the address inputs are common, the same location in each chip is accessed at the same time.

15.7.2 Memory Location Expansion

Figure 15.23 shows how more than one memory chip can be used to expand the number of memory locations. Let us consider the use of two 16×8 chips to get a 32×8 chip. A 32×8 chip would need five address input lines. Four of the five address inputs, other than the MSB address bit, are common to both 16×8 chips. The MSB bit feeds the input of one chip directly and the input of the other chip after inversion. The inputs to the two chips are common.

Now, for first half of the memory locations corresponding to address inputs 00000 to 01111 (a total of 16 locations), the MSB bit of the address is '0', with the result that RAM-1 is selected

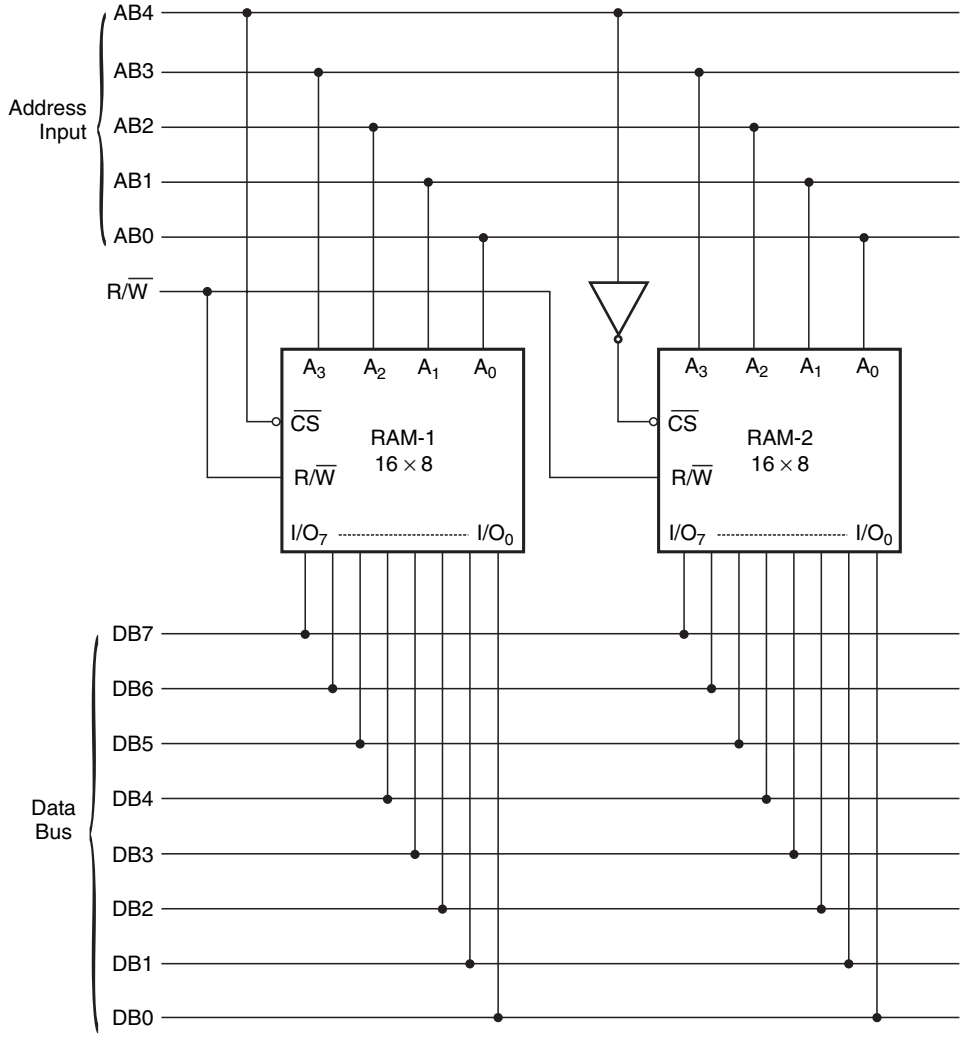


Figure 15.23 Memory location expansion.

and RAM-2 is deselected. For the remaining address inputs of 10000 to 11111 (again, a total of 16 locations), RAM-1 is deselected while RAM-2 is selected. Thus, the overall arrangement offers a total of 32 locations, 16 provided by RAM-1 and 16 provided by RAM-2. The overall capacity is thus 32×8.

Example 15.3

Two 16 MB RAMs are used to build a RAM capacity of 32 MB. Show the configuration and also state the address inputs for which the two RAMs will be active. The two RAMs have common I/O pins, a WRITE ENABLE input that is active LOW and a CHIP SELECT input that is active HIGH.

Solution

Figure 15.24 shows the arrangement. Since the overall RAM capacity is 32 MB, it will have 25 address inputs (AB0 to AB24) as $32\text{M} = 2^{25}$. For address inputs $(0000000)_{\text{hex}}$ to $(0\text{FFFFFF})_{\text{hex}}$, which account for 16M ($=2^{24}$) memory locations, RAM-1 is enabled and 16 M locations of RAM-1 are available. RAM-2 is deselected for these address inputs. For address inputs $(1000000)_{\text{hex}}$ to $(1\text{FFFFFF})_{\text{hex}}$, the total number of addresses in this group again being equal to 16M, RAM-2 is selected and RAM-1 is deselected. 16M locations of RAM-2 are available. Thus, out of 32 MB, 16 MB is stored in RAM-1 and 16 MB is stored in RAM-2.

Example 15.4

What is available is a 1K$\times$ 8 chip of the type shown in Fig. 15.25. This chip, as shown in the diagram, gets activated only when select input $\overline{CS1}$ is LOW and select input CS2 is HIGH. Show how two such ROMs can be connected to get $2K \times 8$ ROM without using any additional logic.

Solution

- Figure 15.26 shows the arrangement.
- The address bit AB_{10} is low for the first 1024 address inputs (from 00000000000 to 01111111111) and ROM-1 is selected.
- For the remaining 1024 address inputs (from 10000000000 to 11111111111), the AB_{10} bit is HIGH, thus enabling ROM-2.

Example 15.5

Figure 15.27 shows an arrangement of four memory chips, each 16 $\times$4 RAM with an active LOW chip select input. Determine the total capacity and the word size. Which RAMs will put data on the data bus when the address input is 00001101. Also, determine the address input range for which RAM-1 and RAM-2 will be active.

Solution

- For address inputs $(00000000)_2$ to $(00001111)_2$, RAM-1 and RAM-2 are selected.
- RAM-1 stores four higher bits and RAM-2 stores four lower bits of data words corresponding to the 16 address inputs mentioned above.

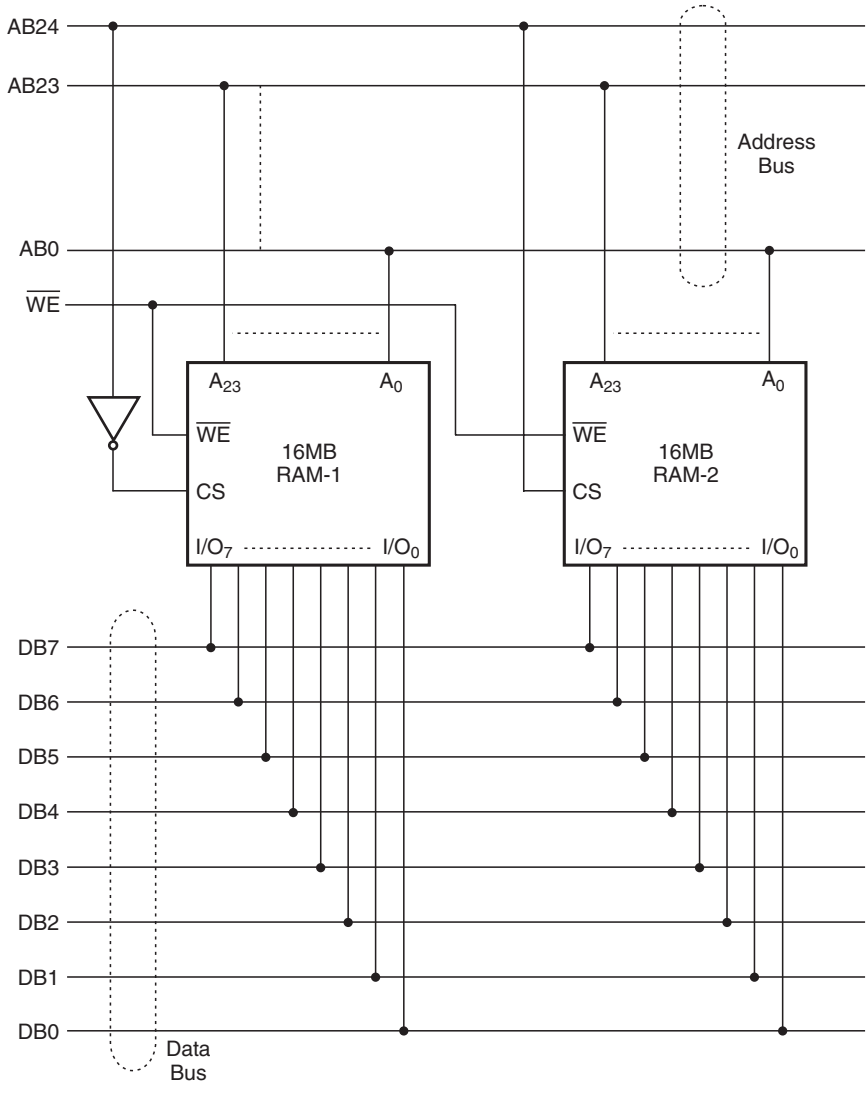


Figure 15.24 Solution to example 15.3.

- This gives us a capacity of 16×8.
- Now, for address inputs $(00010000)_2$ to $(00011111)_2$, RAM-3 and RAM-4 are selected.
- Similarly, RAM-3 and RAM-4 respectively store four upper bits and four lower bits of data words corresponding to these address inputs.
- This again gives a capacity of 16×8.
- Thus, the overall capacity is 32×8.
- The word size is 8.

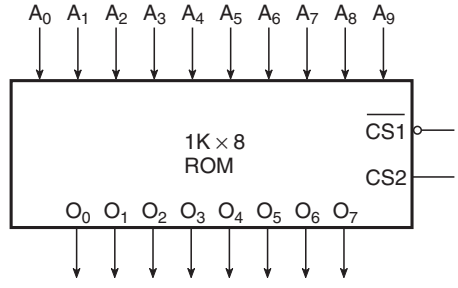


Figure 15.25 Example 15.4.

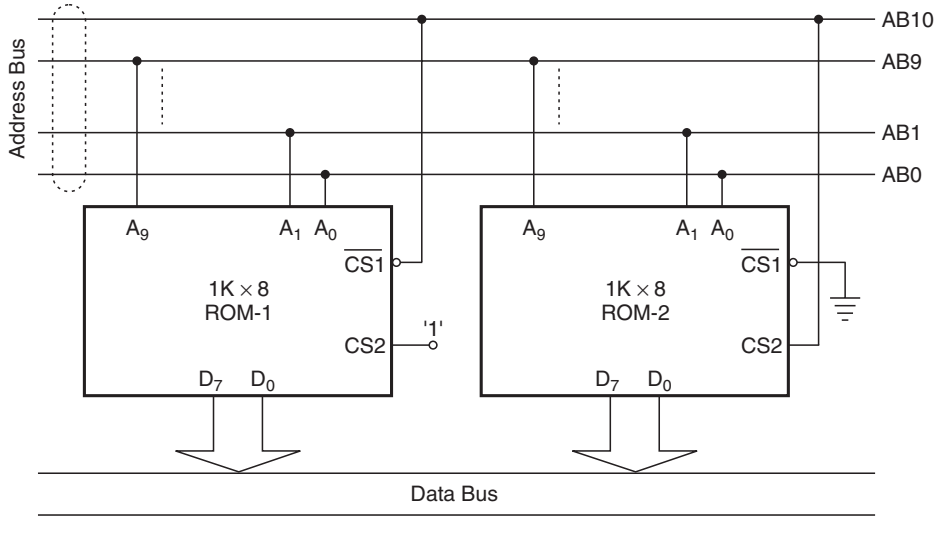


Figure 15.26 Solution to example 15.4.

- For an address input 00001101, RAM-1 and RAM-2 will be selected.
- The address input range for which RAM-1 and RAM-2 are active is $(00000000)_2$ to $(00001111)_2$.

15.8 Input and Output Ports

Input and output ports were briefly introduced in the earlier part of the chapter in Section 15.1.3. As outlined earlier, these are categorized as serial and parallel ports. The commonly used serial and parallel ports are described in the following paragraphs.

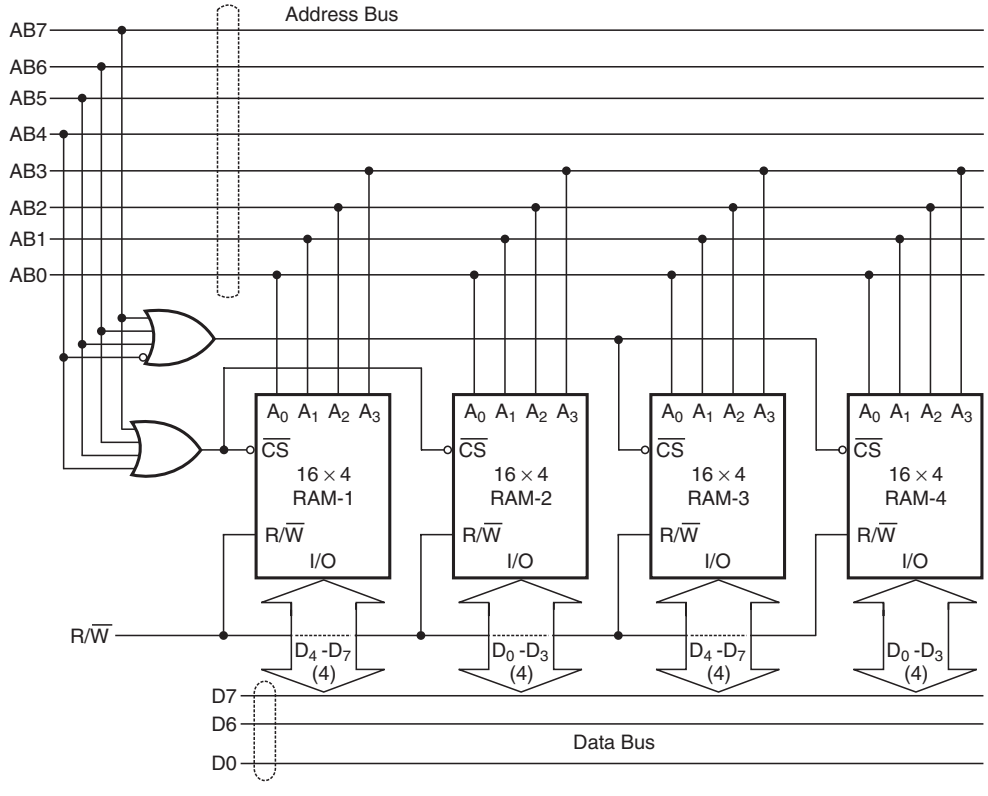


Figure 15.27 Example 15.5.

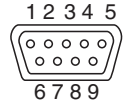
15.8.1 Serial Ports

A serial port is a physical communication interface through which the information transfer takes place one bit at a time. Serial ports are used to connect mouse, keyboard and modems to the computer. Some of the commonly used serial standards include the RS-232C port, PS/2, FireWire and USB.

15.8.1.1 RS-232C Port

RS-232 is one of the oldest and most well-known standards for serial interfaces approved by the Electronic Industries Association (EIA). It was developed to interface data terminal equipment (DTE) with data communication equipment (DCE). RS-232C, a variant of the RS-232 standard, is the most relevant for the computer world. RS-232C is mostly used to connect modem and other communication devices to the computer. In this case the computer is referred to as the DTE and the attached device as the DCE.

The RS-232C standard specifies 25 communication lines between the DTE and the DCE. Hence, the standard RS-232C connector is a 25-pin connector (DB-25). For personal computer applications,



1. DCD (Data Carrier Detect)
2. RD (Recieve Data)
3. TD (Transmit Data)
4. DTR (Data Terminal Ready)
5. GND (Ground)
6. DSR (Data Set Ready)
7. RTS (Request To Send)
8. CTS (Clear To Send)
9. RI (Ring Indicator)

Figure 15.28 DE-9 connector.

not all the 25 pins are required. Hence, most personal computers have a nine-pin connector (DE-9). Figure 15.28 shows the DE-9 connector along with its pin assignments.

The maximum specified cable length for the RS-232C interface is 50 ft for a data transmission rate of 20 kbaud. As the cable length increases, the transmission rate decreases. The RS-422 and RS-423 standards have higher transmission speeds than RS-232C. They also support larger cable lengths. However, RS-232C remains the most commonly used serial port.

15.8.1.2 FireWire

FireWire is the name of the interface specified by the IEEE standard 1394. This high-speed serial bus standard is used for interfacing graphics and video peripherals such as digital cameras and camcoders to the computer. FireWire can be used to connect up to 63 devices in a cyclic topology. It supports both plug-and-play and hot swapping. It is available in two versions, namely FireWire 400 and FireWire 800. FireWire 400 hardware is available in six-pin and four-pin connectors and can support data rates of 100, 200 and 400 Mbits/s. The four-pin connector is used mostly in consumer electronic goods and the six-pin connector is used in computers.

FireWire 800 is based on the IEEE 1394b standard and supports a data rate of 786.432 Mbits/s. It has a nine-wire connection.

15.8.1.3 Universal Serial Bus (USB)

The USB port was introduced in the year 1997 and is used to connect printers, mouse, scanners, digital cameras and external storage devices to the computer. Different versions of the USB standard include 0.9, 1.0, 1.1 and 2.0, with USB 2.0 being the latest. Another variant of the USB standard is the radio spectrum based USB implementation, known as Wireless USB.

A USB port can be used to connect 127 devices. It supports two data rates of 1.5 Mbits/s (low speed) and 12 Mbits/s (full speed). Most of the USB 2.0 devices also support data rates of 480 Mbits/s (Hi speed). USB is a four-wire connection and is available in two standard types referred to as type A and type B. Miniature versions of the USB connector are also available, namely Mini-A and Mini-B. Figure 15.29 shows different types of USB connector, along with their pin details.

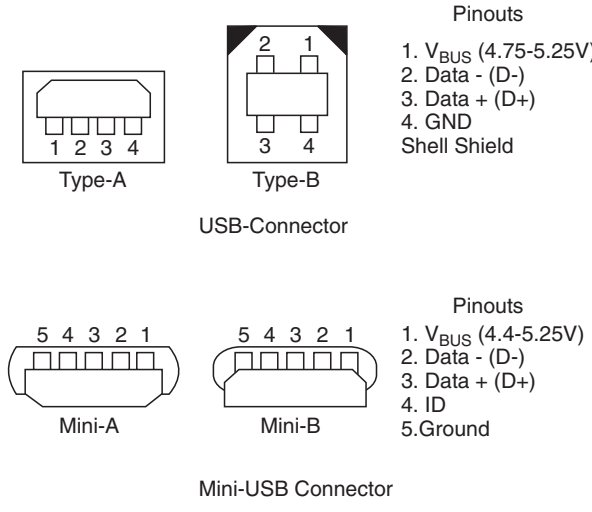


Figure 15.29 USB connector.

15.8.1.4 PS/2 Connector

PS/2 connectors are used for connecting the keyboard and mouse to a personal computer. The PS/2 mouse and PS/2 keyboard connectors are similar to each other, except for the fact that the PS/2 keyboard connector has an open-collector output. PS/2 mouse and keyboard connectors have replaced the DE-9 and five-pin DIN connectors respectively. Figure 15.30 shows the PS/2 connector with the pin details.

15.8.2 Parallel Ports

Parallel ports send multiple bits at the same time over a set of wires. They are used to connect printers, scanners, CD burners, external hard drives, etc., to the computer. Commonly used standard parallel ports include IEEE-488, the small computer system interface (SCSI) and IEEE 1284.

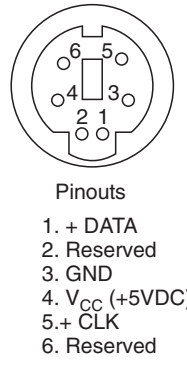


Figure 15.30 PS/2 connector.

15.8.2.1 IEEE-488

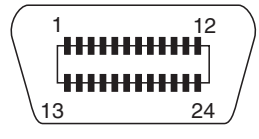
IEEE-488 is a short-range parallel bus standard widely used in test and measurement applications. It is also referred to as a general-purpose interface bus (GPIB). The IEEE 488 standard specifies a 24-wire connection for transferring eight data bits simultaneously. Other connections include eight control signals and eight ground lines. The maximum data rate is 1 MB/s in the original standard and about 8 MB/s with the modified standard (HS-488). Figure 15.31 shows the pin connections and pin details.

15.8.2.2 Small Computer System Interface (SCSI)

SCSI is a widely used standard for interfacing personal computers and peripherals. SCSI is a standard given by the American National Standards Institute (ANSI). There are several variations of this standard, and one variant may not be compatible with another. Some of the SCSI versions include SCSI-1, SCSI-2, Wide SCSI, Fast SCSI, Fast Wide SCSI, Ultra SCSI, SCSI-3, Ultra SCSI-2 and Wide Ultra SCSI-2. Description of all these interfaces is beyond the scope of this book.

15.8.2.3 IEEE-1284

IEEE 1284 is a standard that defines bidirectional parallel communications between computers and other devices. It supports a maximum data rate of 4 MB/s. It supports three types of connector: DB-25 (type A) for the host connection, Centronics 36-pin (type B) for the printer or device connection and Mini Centronics 36-pin (type C), a smaller alternative for the device connection. IEEE 1284-I



1EEE 488 Connector

Pinouts

1. Data 1/O1
2. Data 1/O2
3. Data 1/O3
4. Data 1/O4
5. EOI (End or Identity)
6. DAV (Data Valid)
7. NRFD (Not Ready for Data)
8. NDAC (Not Data Accepted)
9. IFC (Interface Clear)
10. SRQ (Service Request)
11. ATN (Attention)
12. Sheild Ground
13. Data 1/O5
14. Data 1/O6
15. Data 1/O7
16. Data 1/O8
17. REN (Remote Enable)
18. P/O Twisted Pair with 6
19. P/O Twisted Pair with 7
20. P/O Twisted Pair with 8
21. P/O Twisted Pair with 9
22. P/O Twisted Pair with 10
23. P/O Twisted Pair with 11
24. Signal Ground

Figure 15.31 Pin connections and details of the IEEE-488 connector.

devices use IEEE 1284-A and IEEE 1284-B connectors, while IEEE 1284-II devices use IEEE-1284-C connectors. The type C connector is not very popular.

15.8.3 Internal Buses

Input/output ports are used to connect the computer to external devices. Input and output standards described in the previous sections are referred to as external bus standards. In addition to these external buses, computers also have internal buses that carry address, data and control signals between the CPU, cache memory, SRAM, DRAM, disk drives, expansion slots and other internal devices. Internal buses are of three types, namely the local bus, the PCI bus and the ISA bus.

15.8.3.1 Local Bus

This bus connects the microprocessor to the cache memory, main memory, coprocessor and PCI bus controller. It includes the data bus, the address bus and the control bus. It is also referred to as the primary bus. This bus has high throughput rates, which is not possible with buses using expansion slots.

15.8.3.2 PCI Bus

The peripheral control interconnect (PCI) bus is used for interfacing the microprocessor with external devices such as hard disks, sound cards, etc., via expansion slots. It has a VESA local bus as the standard expansion bus. Variants of the PCI bus include PCI 2.2, PCI 2.3, PCI 3.0, PCI-X, PCI-X 2.0, Mini PCI, Cardbus, Compact PCI and PC/104-Plus. The PCI bus will be superseded by the PCI Express bus. PCI originally had 32 bits and operated at 33 MHz. Various variants have different bits and data transfer rates.

15.8.3.3 ISA Bus

The industry-standard architecture (ISA) bus is a computer standard bus for IBM-compatible computers. It is available in eight-bit and 16-bit versions. The VESA local bus was designed to solve the bandwidth problem of the ISA bus. It worked alongside the ISA bus where it acted as a high-speed conduit for memory-mapped I/O and DMA, while the ISA bus handled interrupts and port-mapped I/O. Both these buses have been replaced by the PCI bus.

15.9 Input/Output Devices

Input/output devices are human–machine interface devices connected to the computer. Input devices are used for entering data into the computer. They convert the raw data to be processed into a computer-understandable format. Output devices convert the processed data back into a user-understandable format. This section briefly describes the commonly used input/output devices.

15.9.1 Input Devices

As mentioned before, input devices convert the raw data to be processed into a computer-understandable format. Input devices can be broadly classified into various types, depending upon the type of input data they handle. Commonly used input devices include keyboard devices, pointing devices, image and video input devices and audio input devices.

15.9.1.1 Keyboard Devices

Keyboards are designed for the input of text and characters and also to control the operation of a computer. Keyboards have an arrangement of keys where each press of a key corresponds to some action. Keyboards are available in different types and sizes. Keyboard and pointing devices are also referred to as data entry input devices.

15.9.1.2 Pointing Devices

These include the computer mouse, trackball, joystick, touch screen, light pen and so on. The mouse is a handheld device whose motion is translated into the motion of a pointer on the display. It is one of the most popular input devices used with microcomputers. A joystick consists of a handheld stick that pivots about one end and transmits its angle information to the computer. Touch screens are input devices that sense the touch event and send processing signals to the computer. Touch screens are available in various types including resistive, surface wave, capacitive, infrared, strain gauge, optical imaging and so on. Light pens are devices that transmit their coordinates to the machine when placed against the CRT screen of the machine. Hence, they allow the user to point to displayed objects on the screen or to draw on the screen, similarly to a touch screen but with greater position accuracy.

15.9.1.3 Image and Video Input Devices

These devices, as the name suggests, take some image or video as the input and convert it into a format understandable by the computer. These include magnetic ink character recognition (MICR), optical mark recognition (OMR), optical character recognition (OCR), scanners, digital cameras and so on. MICR devices are used to detect the printed characters with magnetically charged ink and convert them into digital data. They are widely used in the banking industry for the processing of cheques. An OMR device senses the presence or absence of a mark but not the shape of the character. It is a very popular input device for surveys, census compilations and other similar applications. OCR devices are used for translating images of text or handwritten data into a machine-editable text or for translating pictures or characters into a standard encoding scheme (ASCII or Unicode).

A scanner is a device that analyses an image such as a photograph, printed text, etc., of an object and converts it to a digital image. OCR, OMR and image scanners are also referred to as data automation input devices. A digital camera is an electronic device used to capture and store photographs electronically instead of using photographic film.

15.9.2 Output Devices

Output devices convert the processed data back into a user-understandable format. Like an input device, an output device, too, acts as a human–machine interface. Printers, plotters and displays are

the commonly used output devices. Computer output microfilm (COM) is another form of computer output where huge amounts of data can be outputted and stored in a very small size.

15.9.2.1 Printers

A printer is a device that produces a hard copy of the documents stored in electronic form, usually on a physical print medium such as paper. Printers can be broadly classified as 'impact printers' and 'nonimpact printers'. An 'impact printer' is one where the characters are formed by physically striking the type-device against an inked ribbon. Dot-matrix printers, daisy wheel printers, ball printers and drum and chain printers belong to this category. The dot-matrix printer is the most popular in this category. The 'dot matrix' is the basis of the printing mechanism in dot-matrix printers. The dot matrix is formed by arranging a number of small rods in a specified number of rows and columns. The number of rows and the number of columns in the dot matrix may vary from printer to printer. In order to print a character, the corresponding configuration of rods are stricken. The larger the number of dots in the dot matrix, the better is the printer quality.

Impact printers have been largely replaced by nonimpact printers. In this case, there is no physical contact with the paper. The characters are formed by using heat (in thermal printers), laser beam (in laser printers), ink spray (in inkjet printers), photography (in xerographic printers) and so on. Thermal printers are low-cost serial printers that use a number of small heating elements to construct each character from a dot-matrix print head. They use a special kind of heat-sensitive paper that turns black when heated. An inkjet printer sprays small droplets of ink rapidly from tiny nozzles onto the surface of the paper to form characters. A laser printer consists of a toner and a light-sensitive drum and works in a similar manner to a photocopier machine, except that, instead of working photographically from a printed document, the laser printer uses a laser beam to create the image.

15.9.2.2 Plotters

A plotter is a printer-like device used for producing hard-copy outputs of maps, charts, drawings and other forms of graphics. It is a vector graphics printing device that operates by moving a pen over the surface of the paper. Different types of plotter include pen plotters, electrostatic plotters and dot-matrix plotters. There are two types of pen plotter, namely the flat-bed plotter and the drum plotter. In the case of flat-bed plotters the pens move and the paper is stationary, whereas in the case of drum plotters the pens are stable and the paper is moved on a drum.

The electrostatic plotter works like a nonimpact-type electrostatic printer. It electrostatically charges the surface of a special kind of paper at the desired points and then passes the paper through a toner containing ink particles of opposite charge. The ink adheres to the paper surface only at charged points. The dot-matrix plotter works on the same principle as the impact-type dot-matrix printer.

15.9.2.3 Displays

Displays are devices used to display images on the screen in accordance with the signals generated by the computer. Displays are of various types including cathode ray tube (CRT) displays, liquid crystal displays (LCDs), plasma displays and organic light-emitting diode (OLED) displays. The CRT is a vacuum tube employing a focused beam of electrons from the cathode to hit the luminescent screen. The LCD is a display device made up of a number of colour or monochrome pixels arrayed in front of a light source or reflector. Each pixel comprises a liquid crystal molecule.

The plasma display is a flat-panel display where visible light is created by a phosphorus screen excited by discharged inert gases. The OLED is a special type of LED in which the emissive layer comprises a thin film of organic compounds.

15.9.2.4 Computer Terminals

Computer terminal in general refers to the entire range of devices that are connected to a computer and can be used to enter data into the computer system and receive the processed data as output. A computer terminal is used both as an input and as an output device. Typically, it consists of a keyboard and a CRT. Based on the capabilities and performance features, terminals are classified as dumb, smart and intelligent terminals. Depending upon the type of data the terminals are capable of displaying, they are classified as alphanumeric and graphic terminals. Detailed description of the various types of terminal is beyond the scope of this book.

15.10 Secondary Storage or Auxiliary Storage

Secondary storage devices are used for the mass nonvolatile storage of data and programs. It is often not practical to build a very large-sized primary memory to meet all the storage requirements of the system as it will increase the size and cost. That is where secondary storage is useful. Usually, it is located physically outside the machine. Although it is not an essential component in theoretical terms, the secondary storage is almost indispensable if one wants to exploit the full potential of a computer. Secondary storage devices are also referred to as auxiliary storage devices.

Owing to its semiconductor nature, the primary storage can be accessed much faster than any of the storage media used for secondary storage. The secondary storage on the other hand is economical as far as cost per unit data stored is concerned and has an unlimited storage capacity. It is also safe from getting tampered with by any unauthorized persons. Commonly used secondary storage devices include magnetic, magneto-optical and optical storage devices. Another emerging secondary storage device is the USB flash drive.

15.10.1 Magnetic Storage Devices

Magnetic storage devices include magnetic hard disks, floppy disks and magnetic tapes.

15.10.1.1 Magnetic Hard Disks

Hard disks are nonvolatile random access secondary data storage devices, i.e. the desired data item can be accessed directly without actually going through or referring to other data items. They store the data on the magnetic surface of hard disk platters. Platters are made of aluminium alloy or a mixture of glass and ceramic covered with a magnetic coating. Figure 15.32 shows the internal structure of a typical hard disk. As can be seen from the figure, there are a few (two or more) platters stacked on top of each other on a common shaft. The shaft rotates these platters at speeds of several thousand rpm. Each platter is organized into tracks and sectors (Fig. 15.33), both having a physical address used by the operating system to look for the stored data. Tracks are concentric circles used to store data. Each track is further subdivided into sectors so that the total number of sectors per side of the magnetic disk is the product of the number of tracks per side and the number of sectors per track. And if it is a

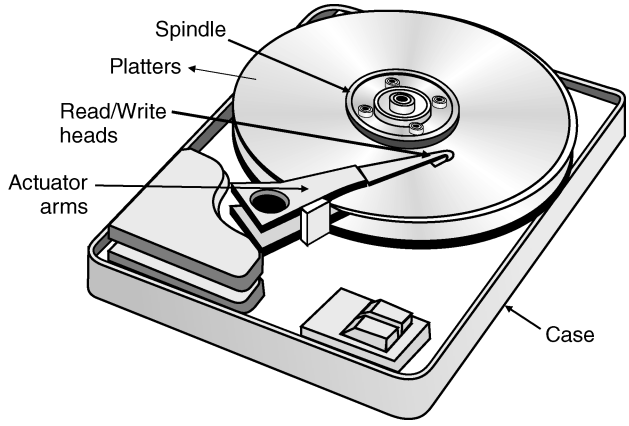


Figure 15.32 Internal structure of a typical hard disk.

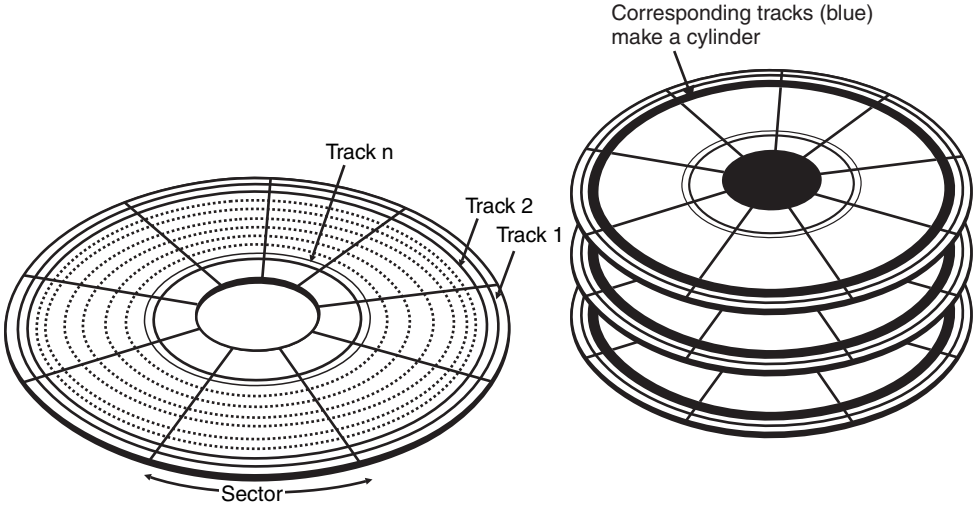


Figure 15.33 Tracks and sectors in a hard disk.

double-sided disk, the total number of sectors gets further multiplied by 2. From known values of the total number of sectors and the number of bytes stored per sector, the storage capacity of the disk in bytes can then be computed.

There is a read/write head on one or both sides of the disk, depending upon whether it is a single-sided or a double-sided disk. The head does not physically touch the disk surface; it floats over the surface and is close enough to detect the magnetized data. The direction or polarization of the magnetic domains on the disk surface is controlled by the direction of the magnetic field produced by the write head according to the direction of the current pulse in the winding. This magnetizes a small spot on

the disk surface in the direction of the magnetic field. A magnetized spot of one polarity represents a binary '1', and that of the other polarity represents a binary '0'.

One of the most important parameters defining the performance of the hard disk is the size of the disk. Disks are available in various sizes ranging from 20 GB to as large as 80 GB. Other parameters defining the hard disk performance include seek time and latency time. *Seek time* is defined as the average time required by the read/write head to move to the desired track. *Latency time* is defined as the time taken by the desired sector to spin under the head once the head is positioned over the desired track.

15.10.1.2 Floppy Disks

Floppy disks are removable disks made of flexible polyester material with magnetic coating on both sides. Important parts of a floppy disk are shown in Fig. 15.34. Floppy disks are also organized in the form of tracks and sectors similar to a hard disk. A floppy disk drive unit is required to read data from or write data into a floppy disk. A read/write head that forms a part of the drive unit does this job. During a read or write operation, the disk rotates to the appropriate position and the head makes a physical contact with the disk to do the desired operation.

Earlier floppy disks were available in 5.25 inch size with a storage capability of 360 kB. They were known as double-sided double-density (DSDD) floppy disks. They have been superseded by 3.5 inch floppy disks having a storage capability of 1.44 MB. Floppy disks are fast being replaced by CD disks and USB drives.

15.10.1.3 Magnetic Tapes

Magnetic tapes are sequential access secondary storage devices used for storing backup data from mass storage devices. In sequential access storage devices, in order to access a particular data item, one has to pass through all the data items stored prior to it. The magnetic tapes are run on machines called tape

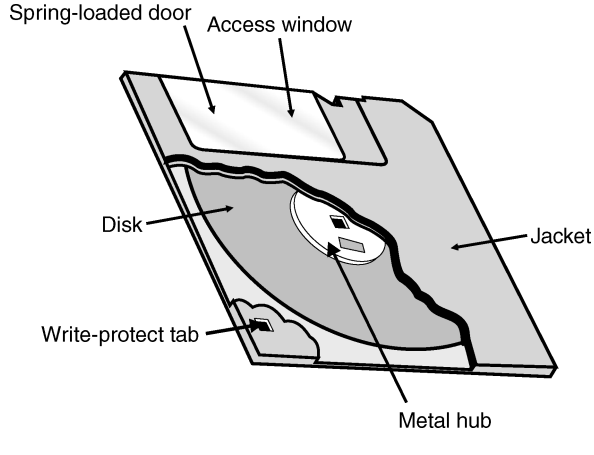


Figure 15.34 Important parts of a floppy disk.

drive units. The data on the tape are represented by tiny magnetized spots, with the presence of a spot representing a binary digit '1' and its absence representing a binary digit '0'. A simple and commonly used method of representing data on tapes is by using EBCDIC code. Magnetic tape is available in the form of reels, cassettes and cartridges. Reels are the most popular type.

15.10.2 Magneto-Optical Storage Devices

Magneto-optical storage devices use a combination of magnetic and optical technologies for data storage. The magnetic coating used in the case of these devices requires heat to alter the magnetic polarization, making them extremely stable at ambient temperatures. For the data write operation, a laser beam having sufficient power is focused onto a tiny spot on the disk. This raises the temperature of the spot. Then the magnetic field generated by the write head changes the polarization of the magnetic particles of that spot, depending upon whether a '1' or a '0' needs to be stored.

For the read operation, a laser beam with less power is used. It makes use of the 'Kerr effect', where the polarity of the reflected beam is altered depending upon the polarization of the magnetic particles of the spot.

15.10.3 Optical Storage Devices

One of the most significant developments in the field of storage media has been that of optical storage devices. Having arrived on the scene in the form of CD-Audio (Compact Disk-Audio) in the early 1980s, since then optical disks have undergone tremendous technological development. These are available in various forms, namely CD-ROM (Compact Disk Read Only Memory), WORM disks (Write Once Read Many), CD-R (Compact Disk Read), CD-RW (Compact Disk Read/Write) and DVD-ROM (Digital Versatile Disk Read Only Memory).

An optical disk differs from a conventional hard disk (solid magnetic disk) in the method by which information is stored and retrieved. While hard disks use a magnetic head to read and write data, in the case of an optical disk this is done with a laser beam. The high storage density of optical disks primarily results from the ability of the coherent laser beam to be focused onto a very tiny spot. The main advantages of optical disks include their vast storage capacity, immunity to illegal copying and their easy removability. Also, they do not transfer viruses from one user to the next.

15.10.3.1 CD-ROM

A CD-ROM is a disk comprising three coatings, namely polycarbonate plastic on the bottom, a thin aluminium sheet for reflectivity and a top coating of lacquer for protection. It can store up to 660 MB of data. It is formatted into a single spiral track having sequential sectors. CD-ROMs are prerecorded at the factory and store data in the form of pits and lands.

These are classified by the access time and data transfer rate. The performance of CD-ROM disks is enhanced by spinning them faster to achieve a higher transfer rate and faster access time. These are rated as 2X, 4X, 6X, 16X, 24X and so on. A 16X CD-ROM drive will be 16 times faster than the original drives. The spinning rate of the drive is the number of revolutions per minute. Its seek time is the time the drive takes to locate a track where desired data are stored. The time for which the drive has to wait for data to rotate under it is the latency. The sum of seek time and latency is the access time.

The read operation (Fig. 15.35) is performed by using a low-power laser beam. The laser beam is focused onto pits and lands. Laser light reflected from a pit is 180°out of phase with the light reflected from land. This light is detected by a photodiode followed by a processing circuitry. As the disk rotates, a series of pits and lands are sensed and the data stored in them is read.

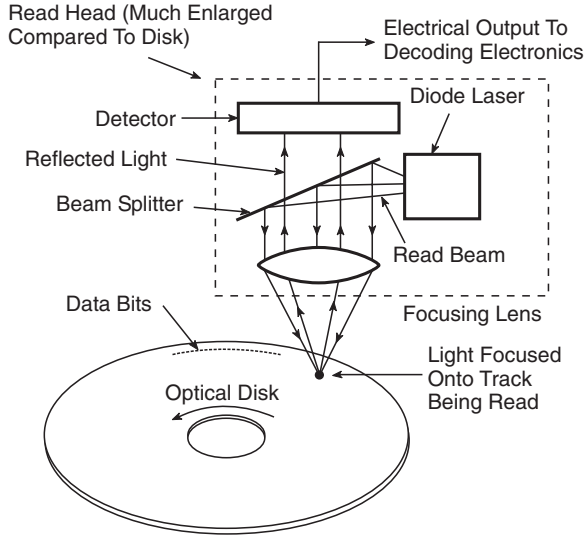


Figure 15.35 Use of a laser beam for CD READ operation.

15.10.3.2 WORM

This is a type of optical storage device where the data can be written once by the user, cannot be erased later but can be read many times. A low-power laser beam is used to burn microscopic pits on the disk surface. Burned surfaces represent a binary '1' and unburned areas represent a binary '0'.

15.10.3.3 CD-R

This is a type of WORM that allows multiple write sessions to different areas of the disk. In this case a laser is used to write data in the form of microscopic pits in an organic dye layer. The presence and absence of a bump indicate a '1' and '0' respectively.

15.10.3.4 CD-RW

In this case, data can be recorded, erased, rewritten and read many times. Recording of data is done by changing the state of the material from a well-structured crystalline state to a less ordered amorphous state.

15.10.3.5 DVD-ROM

Originally the term DVD was an abbreviation of Digital Video Disk, but today it is used for referring to Digital Versatile Disks. It has a much higher storage density than a CD-ROM. This is because the pit size is smaller in the case of DVD-ROMs.

CD-ROMS are single-side storage devices, whereas DVD-ROMs are available in single-sided as well as double-sided formats. As against the 660 MB storage capacity of a CD-ROM, a single-sided DVD of the same size offers 4.7 GB in a single layer. A double-layer or double-sided DVD would offer 9.4 GB of storage capacity, and a double-sided, double-layer DVD would have up to 17 GB, which is about 30 times the storage capacity available on a CD-ROM. DVD-R and DVD-RAM are the counterparts of CD-R and CD-RW.

15.10.4 USB Flash Drive

USB flash drives are plug-and-play flash-memory data storage devices integrated with the USB interface. They are lightweight, rewritable, erasable devices with storage capacities ranging from 8 MB to 64 GB.

Review Questions

1. With the help of a block schematic, describe the role of various elements in a computer system.
2. Explain the difference between:

   (a) a *sequential access memory* and a *random access memory*;
   (b) a *memory write* operation and a *memory read* operation;
   (c) EEPROM and UVEPROM;
   (d) synchronous SRAM and asynchronous SRAM.

3. Explain in brief the concept of cache memory.
4. With the help of a diagram, describe the functioning of different parts of a typical SRAM.
5. Compare the performance features of an SRAM and a DRAM. What is DRAM refreshing? Which type of RAM would you expect in battery-operated equipment?
6. Why do we need to have secondary storage devices when the computer already has a primary storage? Distinguish between magnetic tape and magnetic disk as a secondary storage device.
7. Briefly describe the following:

   (a) a serial port and a parallel port;
   (b) the internal bus system of a computer;
   (c) auxiliary storage devices.

8. What are the commonly used input and output ports in a computer system? Briefly describe the applications of each one of them.

Problems

1. A certain memory is specified as $16K \times 8$. Determine (a) the number of bits in each word, (b) the number of words being stored and (c) the number of memory cells.

   (a) 8; (b) 16 384; (c) 131 072

2. A certain memory is specified as $32K \times 8$. Determine (a) the number of address input lines, (b) the number of data input lines, (c) the number of data output lines and (d) the type of decoder.

(a) 15; (b) 8; (c) 8; (d) 1-of-15 decoder

3. It is desired to construct a $64K \times 16$ RAM from an available RAM chip specified as $16K \times 8$. Determine the number of RAM chips required for the same.

8

4. The following data refer to a hard disk: number of tracks per side = 600; number of sides = 2; number of bytes per sector = 512; storage capacity in bytes = 21 504 000. Determine the number of sectors per track for this hard disk.

35

Further Reading

1. Tocci, R. J. and Ambrosio, F. J. (2002) *Microprocessors and Microcomputers: Hardware and Software*, Prentice-Hall, NJ, USA.
2. Rafiquzzaman, M. (1990) *Microprocessors and Microcomputer-based System Design*, CRC Press, FL, USA.
3. Keeth, B. and Baker, J. (2000) *DRAM Circuit Design: A Tutorial* (IEEE Press Series on Microelectronic Systems), John Wiley & Sons–IEEE Press, New York, USA.
4. Prince, B. (1999) *High Performance Memories: New Architecture DRAMs and SRAMs – Evolution and Function*, John Wiley & Sons, Ltd, Chichester, UK.
5. Axelson, J. (1997) *Parallel Port Complete: Programming, Interfacing and Using the PC's Parallel Port*, Lakeview Research, Madison, WI, USA.
6. Axelson, J. (1998) *Serial Port Complete*, Lakeview Research, Madison, WI, USA.