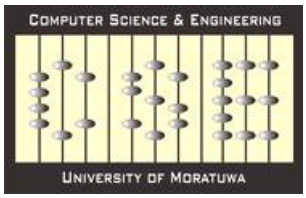


CS 101 – Computer Systems

Department of Computer Science and Engineering
Faculty of Engineering
University of Moratuwa
Sri Lanka

By: Dilum Bandara
Proof Read By: Mr. Samantha Senaratna

Table of Content

Chapter 6 – Operating Systems and Application Software

Chapter 7 – Introduction To Networking

List of Figure

List of Tables

1 – Introduction

Today computers are everything and everywhere. Some of you may have used a computer; some may have seen a computer or at least have seen it from a picture or television. Rest of the chapter provides a brief introduction about computers, its usage and some background information.

1.1 What Is a Computer?

The answer to the question "What a computer is?" depends on what you want to do with it and how you look at it. The way a small kid looks at a computer is different from a student or school leaver. A professional (doctor, engineer or scientist) may use it in a completely different manner.

"Computer is an electronic device for analyzing and storing data, making calculations, etc."[1]. The word computer comes from the word compute which means *calculate.* Today we use computers for almost everything however in early days computers were used only for mathematical calculations. In simple words computer is a device (or machine) which can perform a given task. Most important features of a computer are:

- It is a machine
- Able to execute (understand) a finite set of 'instructions'.
- Able to process 'data' according to those 'instructions'.
- Able to execute a 'sequence of instructions' that are 'stored' within the machine in a specified order.
- Able to deviate from the 'sequence' based on 'result' of a previous operation.

The computer is a combination of mechanical, electrical and electronic components. Computers cannot think as humans or animals. Computers are programmed in such a way where the users will feel like it is thinking. The ability of a computer to think (or act as thinking) depends on the smartness of the programmer therefore what a computer can do it limited only by the imagination of the programmer. When we give her[2] (computer) a problem she can analyze all possible solutions, constrains, wrong answers and come up with a reasonable solution millions times faster than an average human. Computers can overtake humans in its high speed and accuracy of a decision not with the ability to think.

Central Processing Unit (CPU) is the brain of a computer. The CPU does all the calculations and logical decision-making. However like our brain the CPU cannot remember past events. Therefore we need special form of memory. Most of the components used in computers can be matched with parts of the human body.

1.2 Features

As mentioned earlier it was initially used only for complex calculations however with the advancement of technology it has evolved immensely. It has already influenced all forms of industrial and business needs.

A student can use the computer to find information for his/her studies from the Internet or educational CDs (edutainment material), read computer based books (e-Books), follow computer based study packs (e-Learning), play computer games, listen to music, watch movies (multimedia), he/she can also create his/her won music, movies and animations, draw pictures, design web pages, chat with friends (messaging), pen pals, etc.

An employee can use computers to do his day to day work such as preparing reports and letters, printing bills, managing stock, accounting, in communication (e-mails, online messaging, file transferring) and preparing multimedia presentations. Doctors use computers to keep track of their

[1] Oxford English Dictionary

[2] The gender of a computer is considered as female, hence it is referred to as "she" rather than "he" or "it". Since most computer scientists were male, they decided that a computer should be female.

patients' history, study about diseases and build computerised models of the human body for various simulations. Engineers and Scientists use it to perform complex calculations, simulations, technical drawings, build computerized models, weather forecasting, analyzing pictures from outer space, sending rockets to outer space, etc. Musicians and artists use computers to produce high quality music, movies, 3D animations, pictures, photographs, cartoons, etc.

Basically computers can do almost every task that the user wants it to perform, faster than any human being. If something cannot be done with today's technology, with the advancement of technology it would certainly be possible within next few years. The field of computing is evolving so rapidly that no one will be able to learn everything in his/her lifetime.

1.3 The History

Compared with other engineering disciplines Computer Engineering is still very immature. However over the last 60 years it has evolved rapidly and the rate of technology change is unmatched by any other engineering profession.

Many researches and scientists have contributed to the development of the computer. It is a complex piece of machinery made up of many parts, each of which can be considered as a separate invention.

The history of computing is closely related with mathematics. The story begins as a calculating machine. Early man used his/her fingers to represent numbers. Then numbers were represented by stones arranged in heaps of ten. This led to the development of the *abacus* which was invented by the Chinese 3000 years ago.

By the beginning of the 17th century the Arabic system of numeration for both calculating and recording was widespread. Later in the 17th century Oughtred developed the first *slide rule*. During this period Pascal produced the first mechanical calculating machine. This machine could only perform addition and subtraction. Later it was enhanced by Leibnitz in 1673 to perform multiplication and division.

In 19th century Charles Babbage put forward detailed proposals for an automatic calculating machine. In 1822, he built the first *differential engine* to calculate values of polynomial functions. From 1842 to 1848 he worked on a design of a general purpose digital calculating machine which he called an *analytical engine*. He adopted the idea of *punch cards* as an input mechanism to this analytical engine. Lady Lovelace, a friend of Babbage produced many programs using his idea to perform advanced mathematical calculations. She is considered to be the first programmer.

ENIAC (Electronic Numeral Integrator and Calculator), built in 1946 was the first completely electronic calculator. This machine had 18,000 valves. The EDSAC (Electronic Delay Storage Automatic Calculator) and EDVAC (Electronic Discrete Variable Automatic Calculator) were the first automatic *general purpose* electronic computers. These were implemented based on the Von Neumann's architecture. Later the first commercial computer named EDSAC was introduced In 1953 IBM entered the field by introducing the IBM 701 machine.

Since 1949, there has been considerable progress in the development of hardware and software. The most significant advances in hardware technology have been the invention of the transistor, magnetic tape and Integrated Circuits (ICs). With these advances, the speed and reliability of computers has increased significantly. Combination of these inventions, low power consumption and reduced machine size has led to the current era of computers.

1.4 Different Types of Computers

Before the introduction of the *Personal Computers* (PCs), people used (and sometimes still do) *mainframes* and *miniframes*. In those machines several users had to collectively use the same set of centralised resources such as; processing power, memory, and storage through different terminals, in a multi user environment. A personal computer is, typically, a single user machine, in which all the resources necessary to accomplish the tasks required by a user, are located. To be precise, however, today's PCs are capable of serving multiple users as well.

The first personal computer produced by IBM was called the PC (which stands for Personal Computer). Increasingly the term PC came to mean IBM or IBM-compatible personal computers, to the excluding other types of personal computers, such as Macintosh.

In recent years, the term PC has become more and more difficult to define or classify. In general, however, it applies to any personal computer based on an Intel or an Intel-compatible microprocessor (referred as the x86 architecture). For nearly every other component, including the operating system, there are several options, all of which fall under the rubric of PC.

Today there are at least three basic types of personal computers; Desktops, Laptops, Personal Digital Assistants (PDAs).

1.5 Components of a Computer?

Components of a computer can be subdivided as *software* and *hardware.* Hardware is the physical electrical, electronic and mechanical components of a computer. In chapter 5 hardware components are discussed in detail.

Software resides on top of hardware and it controls the hardware. Software is what users see, feel and work with; but it can never be touched (intangible). When the user types a letter, draws a picture or listen to music he/she make use of software. Software refers to any *program* that tells (instructs) the computer (hardware) what to do.

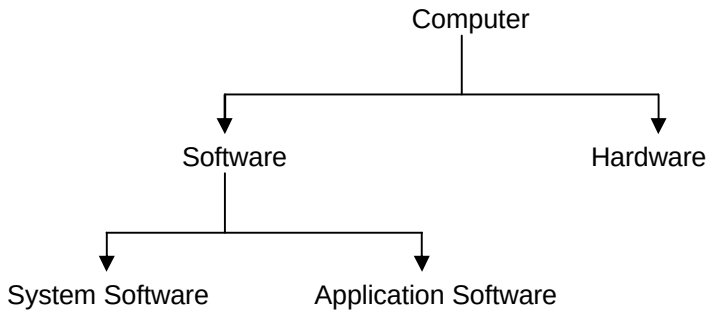


Figure 1.1 – Components of a computer

A program is a well ordered set of *instructions* that instructs the computer on what to do. A program can instruct the computer saying do this first, then do something else, after that do this, jump here, jump there, etc. A program needs to follow a proper sequence in order to get the work done. Programs are written using special languages (some are very similar to English) called *Programming Languages*. Software falls into two sub categories namely System software and Application software (figure 1.1). In chapter 8 and 9 these are described in detailed.

1.5.1 System Software

System software takes control of the computer when it starts and then plays the central role in controlling everything happens after that. It manages the system, maintain and control all the resources. Operating Systems belongs to the category of system software.

1.5.2 Application Software

Application software is designed to perform specific personal, business or scientific tasks. Programs that allow users to type letters, draw pictures, play a song or movie are called application software. Application software includes programs that support word processing, desktop publishing, multimedia, database management, electronic commerce (e-Commerce), communication, personal and organisational information management, etc.

1.6 Information Technology

The word IT stands for Information Technology. It is a new trend in communication technology which allows sharing of information all around the world making it a Global Village. It is all about information; collecting data, summarising data, extracting information from data, storing, making it

available to every one over the Internet, transmission of information, making decisions based on available information, etc. Today it is not just producing information it is also about communicating it with the rest of the world. So the modern term is ICT (Information Communication Technology).

It is not wrong to say that IT came to for-front because of the evolution of computers. IT and computers are interlinked so that there would not be IT without computers. Some people use the word IT to indicate both computer applications and information management. This is why most students say they are following an IT course or degree not simply computer course. Both words are interchangeable although they have slight differences in meanings.

1.7 The Users Role

As a computer user you have some responsibilities, that no one would force you to fallow but it is good to know your responsibilities and fallow them.

Technology could either be used for good or bad. Computers could be the most expensive single piece of equipment in a house (other than a vehicle). So you must be careful in using it. But you should not be afraid to use a computer, because it is much more robust than most of the other equipments that you encounter in our day to day life. It is essential that you give away your fear towards it.

If you understand your responsibilities and ethics you could be an excellent user who makes use of the technology in an optimum way. Some of the ethics could be slightly change depending on the profession or type of a user. However you should not use it for any unethical activity or support such activities.

A code of ethic provides direction for IT professionals and users so they can apply computer technology responsibly. Thousands of private and public sector organizations maintain data on individuals, educational, tax, medical, weapons, credit card information, etc. so we must preserver the privacy of such information. It is not ethical to glimpse such information unless you are authorized to do so. It is also your responsibility not to keep (or store) such private information in a way that everyone can easily access. Respect to the privacy of others, you should never keep your eyes on the keyboard when some one is typing a password and no one should type a password when someone lucking at him/her. Honour and fallow policies and rules fallowed by your school, institute, university or office.

It is not good to use or keep illegal copies of programs, songs or movies[3]. If cost of software is your concern you can always use the Free and Open Source Tools (FOSS) where you do not really need to pay even a single rupee.

With your knowledge you should contribute to the society and human well-being and not to build a virus that would destroy every piece of software and hardware.

1.8 Future of Computers

It is bit hard to predict where computing would be in another 5 or 10 years. Because computer technologies (both hardware and software) are evolving at a rapid rate that we sometimes cannot plan or imagine.

In several years from now everything would be fully computerised and world will not live without it. Whether you run your own business or work for someone else basic computer literacy would be a must and as a member of the engineering profession you need more than some basic understanding of it.

You need to learn more if you need to go extra mile in the Information Super Highway.

IT/ICT would offer you the opportunity to improve the quality of your life. It is your challenge to combine your skills and direct the technology for the betterment of the society.

[3] Currently some of the copyright related laws have being introduced in the country. Therefore uses of illegal copies are prohibited.

2 – Number Systems

Numbers are represented using different number systems. Some of the common number systems are *decimal*, *binary*, *octal* and *hexadecimal*. Decimal is the number system used by humans in their day to day life while binary is the number systems used by computers. Other 2 systems were earlier used in computers but not much today.

It is not surprising that our (i.e. humans') number system is decimal which is based on units of ten, when nature provides us with 10 fingers. On the other hand computers have only electronic or electromechanical switches. These switches have only 2 states, either ON or OFF as a result computers are based on units of 2 which is the base of binary number system.

2.1 Decimal Number System

The characteristic which distinguish one number system from another is called the *base* (or *radix*) of a number system. The base or radix, of a number system is defined as the number of different digits that can occur in each position in the number system. In simple terms, this is the number of symbols in a system. The base of the decimal number system is 10 and the symbols are 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Table 2.1 summarises different number systems and their symbols.

Table 2.1- Number bases

Number System	Base	Symbols
Binary	2	0, 1
Octal	8	0, 1, 2, 3, 4, 5, 6, 7
Decimal	10	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
Hexadecimal	16	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F

The number 4567 in base 10 means:

$$\begin{aligned} 4567 &= \text{four thousand five hundred and sixty seven} \\ &= 4000 + 500 + 60 + 7 \\ &= (4 \times 1000) + (5 \times 100) + (6 \times 10) + (7 \times 1) \\ &= (4 \times 10^3) + (5 \times 10^2) + (6 \times 10^1) + (7 \times 10^0) \end{aligned}$$

In number systems numbers are represented by means of positional notations. As each digit moves one place left, it is multiplied by the base (in this case by 10) and when it moves right, it is divided by the base (in this case 10). The number 512.49 in base 10 means:

$$\begin{aligned} 512.49 &= 500 + 10 + 2 + 0.40 + 0.09 \\ &= (5 \times 100) + (1 \times 10) + (2 \times 1) + (4 \times 0.1) + (9 \times 0.001) \\ &= (5 \times 10^2) + (1 \times 10^1) + (2 \times 10^0) + (4 \times 10^{-1}) + (9 \times 10^{-2}) \end{aligned}$$

A number **N** is base **b** is written as:

$$N = a_n a_{n-1} a_{n-2} \ldots a_1 a_0 . a_{-1} a_{-2} a_{-3} \ldots a_{-m} \qquad \text{--- (2.1)}$$

and is defined as:

$$N = a_n b^n a_{n-1} b^{n-1} a_{n-2} b^{n-2} \ldots a_1 b^1 a_0 b^0 . a_{-1} b^{-1} a_{-2} b^{-2} a_{-3} b^{-3} \ldots a_{-m} b^{-m} \qquad \text{--- (2.2)}$$

The **a** in the above equation is called the *digit* and **b** is the *base*. The positional notation (**.**) employs the radix point to separate the *integer* and *fractional* part of the number. These are called decimal points and in binary they are called binary points.

2.2 Binary Number System

The base of the binary number system is 2 and numbers are represented using symbols 0 and 1. The number 1101.11 in base 2 means:

$$1101.01 = (1 \times 2^3) + (1 \times 2^2) + (0 \times 2^1) + (1 \times 2^0) + (0 \times 2^{-1}) + (1 \times 2^{-2})$$

Based on equations (2.1) and (2.2) a = [0, 1] and b = 2.

2.3 Octal Number System

The base of the octal number system is 8 and the symbols are 0, 1, 2, 3, 4, 5, 6 and 7. The number 456.41 in base 8 means:

$475.01 = (4 \times 8^2) + (7 \times 8^1) + (5 \times 8^0) + (0 \times 8^{-1}) + (1 \times 8^{-2})$

Based on equations (2.1) and (2.2) a = [0, 1, 2, 3, 4, 5, 6, 7] and b = 8.

2.4 Hexadecimal Number System

The base of the hexadecimal number system is 16 and numbers are represented using symbols 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F. Since there are no digits beyond digit 9 first 6 characters of the English language is used to represent rest of the digits. The number 1FA.4C in base 16 means:

$$1FA.4C = (1 \times 16^2) + (F \times 16^1) + (A \times 16^0) + (4 \times 16^{-1}) + (C \times 16^{-2})$$
$$= (1 \times 16^2) + (15 \times 16^1) + (10 \times 16^0) + (4 \times 16^{-1}) + (12 \times 16^{-2})$$

In this case a = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F] and b = 16.

2.5 Number Base Conversion

It is necessary to convert from one number system to another. Suppose you want add 2 numbers using the Windows Calculator, the 2 numbers are input to the calculator in decimal number system. However *microprocessors* (*Central Processing Unit - CPU*) can only understand binary numbers. So the calculator has to convert from decimal to binary before performing the addition. After the addition the final answer will also be in binary. However the answer should be converted to a human readable format therefore the program needs to convert the answer back to decimal number system.

2.5.1 Decimal to Binary Conversion – Integers

To convert a decimal number to binary, divide the number successively by 2. When it cannot be further divide record the reminder in the reverse order.

Example 2.1: Convert 8_{10} to binary.

```
8/2   =     4     r = 0
4/2   =     2     r = 0
2/2   =     1     r = 0
1/2   =     0     r = 1
```

When the reminder is read from bottom to up it becomes 1000 in binary. So $8_{10} = 1000_2$

Example 2.2: Represent 123_{10} in binary.

```
123/2 =     61    r = 1
61/2  =     30    r = 1
30/2  =     15    r = 0
15/2  =     7     r = 1
7/2   =     3     r = 1
3/2   =     1     r = 1
1/2   =     0     r = 1
```

Then $123_{10} = 1111011_2$

2.5.2 Decimal to Binary Conversion – Fractions

When numbers include fractions the integer portion of the number is converted in the same manner as above. Then the fraction is multiplied by 2 and the integer part of the result is noted. The integer (which will be either 1 or 0) is then striped out from the answer and fraction is multiplied again. This process continues until the process ends or sufficient degree of precision has been reached. Consider the following example.

Example 2.3: Represent 0.5_{10} in binary.

```
0.5×2       =     1.0
0.0×2 (this ends the process)
```

Therefore $0.5_{10} = 0.1_2$

Example 2.4: Represent 0.25_{10} in binary.

```
0.25×2      =     0.50
0.50×2      =     1.00
0.00×2 (this ends the process)
```

So $0.25_{10} = 0.11_2$ (Read the integer portion from top to bottom)

Example 2.5: Represent 0.1_{10} in binary.

```
0.1×2       =     0.2
0.2×2       =     0.4
0.4×2       =     0.8
0.8×2       =     1.6
0.6×2       =     1.2
0.2×2       =     0.4
0.4×2       =     0.8
0.8×2       =     1.6
0.6×2       =     1.2
0.2×2       =     0.4
0.4×2       =     0.8
0.8×2       =     1.6
```

this process never ends therefore $0.1_{10} = 0.000110011001_2$

The number 0.1 in decimal cannot be accurately represented in binary even you try infinite number of times. This can happen for any number base conversion. Also note that bit sequence 1101 is getting repeated, such as number is represented as; $0.000\overline{1101}$.

2.5.3 Binary to Decimal Conversion - Integers

Binary to decimal conversion can be performed by adding together the necessary powers of two.

Example 2.6: Represent 101001_2 in decimal.

$$101001 = (1 \times 2^5) + (0 \times 2^4) + (1 \times 2^3) + (0 \times 2^2) + (0 \times 2^1) + (1 \times 2^0)$$
$$= (1 \times 32) + (0 \times 16) + (1 \times 8) + (0 \times 4) + (0 \times 2) + (1 \times 1)$$
$$= 32 + 8 + 1$$
$$= 41 \rightarrow 101001_2 = 41_{10}$$

To perform the same number based conversion a more methodical *algorithm* can also be used First take the left most non-zero bit, double it and add it to the bit on its right. Now take the result, double it and add it to the next bit on the right. Continue in this way until the *least significant bit* (LSB) has been added in.

Example 2.7: Represent 111001_2 in decimal.

```
1     1     1     0     0     1
1
 ⟶   2
       3
        ⟶  6
             7
              ⟶ 14
                  14
                   ⟶ 28
                       28
                        ⟶ 56
                          57
```

$111001_2 = 57_{10}$

2.5.4 Binary to Decimal Conversion – Fractions

The algorithm for converting binary fractions to their decimal equivalent is based on the fact that a bit in one column is worth half the value of a bit in the column on its left. Staring at the right most non-zero bit, take that bit and halve it. Now add the result to the next bit on its left. Halve this result and add it to the next bit on the left. Continue this process until the binary period symbol is reached.

Example 2.8: Represent 0.111001_2 in decimal.

```
0.    1     1     1     0     0     1
                                    1
                                1/2 ←
                                1/2
                          1/4 ←
                          1/4
                    1/8 ←
                    9/8
              9/16 ←
              25/16
        25/32 ←
        57/32
57/64 ←
```

$0.111001_2 = 57/64_{10} = 0.890625_{10}$

2.5.5 Decimal to Octal Conversion

The process of converting decimal number to octal is similar to decimal to binary conversion, instead of dividing by 2 divide the decimal number by 8.

Example 2.9: Represent 123_{10} in octal.

```
123/8 =     15    r = 3
15/8  =     1     r = 7
1/8   =     0     r = 1
```

$123_{10} = 173_8$

2.5.6 Octal to Decimal Conversion

This is similar to binary to decimal conversion in here power of 8 is considered instead of power of 2.

Example 2.10: Represent 432_8 in decimal.

$$\begin{aligned} 432_8 &= (4 \times 8^2) + (3 \times 8^1) + (2 \times 8^0) \\ &= (4 \times 64) + (3 \times 8) + (2 \times 1) \\ &= 256 + 24 + 2 \\ &= 282 \rightarrow 432_8 = 282_{10} \end{aligned}$$

Other method introduce in binary to decimal conversion can also be applied here.

2.5.7 Decimal to Hexadecimal Conversion

The conversion is similar to decimal to binary conversion, instead of dividing by 2 the decimal number is divided by 16.

Example 2.11: Represent 1234_{10} in hexadecimal.

```
1234/16     =     77    r = 2
77/16       =     4     r = 13 = D
4/16        =     0     r = 4
```

$1234_{10} = 4D2_{16}$

2.5.8 Hexadecimal to Decimal Conversion

Example 2.12: Represent $6A8_{16}$ in decimal.

$$\begin{aligned} 6A8_{16} &= (6 \times 16^2) + (\mathbf{A} \times 16^1) + (8 \times 16^0) \\ &= (6 \times 256) + (\mathbf{A} \times 16) + (8 \times 1) \\ &= 1536 + (\mathbf{10} \times 16) + 8 \\ &= 1536 + 160 + 8 \\ &= 1704 \rightarrow 6A8_{16} = 1704_{10} \end{aligned}$$

2.5.9 Binary to Octal Conversion

Binary to octal conversion can be performed in two steps; first converting from binary to decimal and then converting the decimal value to octal. However this approach can be tedious. Instead a shortcut can be used and it makes use of the fact that $8 = 2^3$. This implies that a number represented with 3 binary digits is equal to a number represented by a single octal digit.

Therefore to convert a given binary number to octal, group the bits into 3 digit block starting from the least significant bit and move towards the left. Replace each block of 3 digits with the corresponding octal digit.

Example 2.13: convert 11010011_2 into octal

```
 11 | 010 | 011
011 | 010 | 011    - add extra bit since each block should be 3 digits
 3     2     3
```

$11010011_2 = 323_8$

2.5.10 Octal to Binary Conversion

It is also possible to convert an octal number into a binary number. In here also we can make use of the fact that $8 = 2^3$. This implies that a number represented with an octal digit is equal to a number represented by a three binary digits. Therefore to convert an octal to binary, form the digits into groups of 1 octal number and replace each group with the corresponding 3 digit binary number.

Example 2.14: convert 276_8 into binary

```
 2  |   7 |   6
010 | 111 | 110    - add extra bit since each block should be 3 digits
```

$276_8 = 010111110_2$

2.5.11 Binary to Hexadecimal Conversion

The same shortcut used in binary to octal conversion can also be applied here. Instead of forming groups of 3 digits form 4 digit groups. This is based on the fact that $16 = 2^4$.

Example 2.15: convert 100111010011 into hexadecimal

```
1001 | 1101 | 0011
   9     13      3
   9      D      3
```

$100111010011_2 = 9D3_{16}$

2.5.12 Hexadecimal to Binary Conversion

It is also possible to convert a hexadecimal number into a binary number. In here also we can make use of the fact that $16 = 2^4$. Therefore to convert a hexadecimal number to a binary number, form the digits into groups of 1 hexadecimal number and replace each group with the corresponding 4 digit binary number.

Example 2.16: convert 276_8 into binary

```
  2  |    A |    F
0010 | 1010 | 1111    - add an extra 0 digit since each block should be 3 digits
```

$2AF_{16} = 001010101111_2$

2.5.13 Octal to Hexadecimal Conversion

Octal to hexadecimal conversion can be performed IN 2 steps; first convert from octal number to binary and then convert the binary number to hexadecimal. In order to convert octal to binary represent each octal digit with appropriate binary number. Then form groups of 4 bits starting from the LSB.

Example 2.17: convert 1234_8 into hexadecimal

```
   1     2     3     4
 001   010   011   100 →      001010011100
    0010 | 1001 | 1100
       2 |    9 |    C
```

$1234_8 = 29C_{16}$

Similarly hexadecimal numbers can be converted to octal numbers using the binary representation of the number.

2.6 Arithmetic Operations on Binary Numbers

Arithmetic operations on binary numbers are performed in the same way as decimal numbers. In order to perform any arithmetic operation all operands should be in the same number base. If not numbers should be converted to a common number base before performing any mathematical operation. Since decimal arithmetic is obvious the focus is only on the binary number system.

2.6.1 Binary Addition

Adding two binary numbers is simple, instead of digits from 0 to 9 we have only 0 and 1 to deal with.

Example 2.18: Add 01010_2 and 10001_2

```
  01010
  10001 +
  11011
```

2.6.2 Binary Subtraction

Binary subtraction is similar to decimal subtraction.

Example 2.19: Subtract 11110_2 from 11000111_2

```
  11000111
     11110 –
  10101001
```

2.6.3 Binary Multiplication

Binary multiplication is similar to decimal multiplication. Multiply the multiplicand by one bit at a time starting from the LSB. This is a series of left shifting and addition operations. Then left shift the next stages by one bit and continue with the 2nd LSB of the multiplier. Continue this process until you multiply by all the bits in the multiplier.

Example 2.20: Multiply 1110_2 and 1011_2

```
    1110
    1011 ×
    1110
   1110
  0000
 1110____
10011010__
```

2.6.4 Binary Division

Binary division is similar to the decimal division.

Example 2.21: Divide 1110_2 by 10_2

```
       111
   ____
10 | 1110
     10
     ---
      11
      10
      --
       10
       10
       --
        0
```

The same answer can be obtained by right shifting and subtracting.

3 – Data Representation

Computers store and process *data* (the processed data is called *information*). Physical devices used to store and process data in computers are two-state devices. A switch is a two-state device which can either be ON or OFF and its state can be represented by the two symbols 0 and 1. Such a state which is represented by one of those two symbols is known as a *Bit* (an abbreviation for Binary digiT). Bit is the smallest unit of data representation. Data is stored in *memory*, *disks* and CPU *registers* as strings of bits. When 8 bits are combined together it is called a *byte* and 1024 (2^{10}) bytes are called a *Kilobyte*. Table 3.1 summarises various units used to measure data.

Table 3.1 – Metrics used to measure quantities of data

Abbreviation	Symbol		Bytes	Power of 2
Bit	Bit		-	-
Byte	Byte	8 bits	1	2^0
Kilobyte	KB	1024 Bytes	1024	2^{10}
Megabyte	MB	1024 Kilobytes	1,048,576	2^{20}
Gigabyte	GB	1024 Megabytes	1,073,741,824	2^{30}
Terabyte	TB	1024 Gigabyte	1,099,511,627,776	2^{40}
Kilobit	Kb	1000 bits	125	-
Megabit	Mb	1000 Kilobits	125,000	-

Data can be classified into two categories namely *qualitative* data and *quantitative* data (figure 3.1). Qualitative data represents quality or characteristics. Data such as student's name, national identity card number, address and telephone numbers are examples for qualitative data. Quantitative data can be quantified and they are proportional to a given value. Data such as number of students in a class, students' grade for CS101 and students GPA, interest rate of a bank are examples for quantitative data.

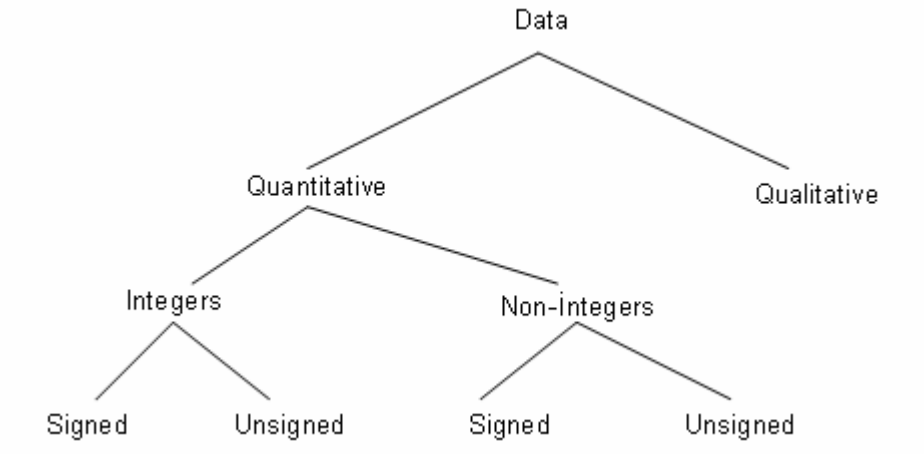


Figure 3.1 – Classification of quantitative data

3.1 Quantitative Data

As stated earlier quantitative data represents some sort of a quantity. Therefore these data types are used in mathematical calculations such as addition, subtraction, multiplication and division. Quantitative data can be further categorised as *integers* and *non-integers* (also referred as *floating point* numbers). Integers denote whole numbers while non-integers denote fractions. Both integers and non-integers can be further categorised as *unsigned* and *signed*. Unsigned numbers denote only the positive values while signed numbers denote both positive and negative numbers.

3.1.1 Unsigned Integers

Unsigned integers denote whole numbers which are only positive. With a single byte (8 bits) we can denote 2^8 (256) numbers. Where zero (00000000_2) is the minimum and 255 (11111111_2) is the maximum.

Data in a microprocessor is stored in *registers*. Register is an array of transistors which can hold particular number of bits. A register which is 8-bit wide can store any number from 0 to 255. Similarly a 16-bit wide register can represent 2^{16} (65536) possible values, where 0 is the minimum and 65535 is the maximum. The general equation for n-bit register is:

Possible number of values	2^n		
Minimum	0	(when all bits are zero)	---(3.1)
Maximum	2^n-1	(when all bits are one)	

Definition: Most Significant Bit and Least Significant Bit

Given a string of bits, the right most bit is called the Least Significant Bit (LSB) because it provides the least contribution to the unsigned value of the number indicated by the bit string. The left most bit is called the Most Significant Bit (MSB) because it provides the highest contribution to the unsigned value of the number. As an example consider the bit string $1001011_2 (75_{10})$.

Here the right most bit (LSB) contributes 1 out of 75 while the left most bit (MSB) contributes 64 out of 75.

Consider the following binary addition.

Example 3.1: Add 111110_2 and 11000111_2

```
  111110
11000111 +
100000101
```

In this case although two 6-bit and 8-bit numbers are added result is a 9-bit number. If the microprocessor has only 8-bit registers it will not be able to hold the final answer. It holds only the last 8 bits and as a result the computerised answer will be 00000101_2. In this case the MSB is called the *carry bit* and such a situation is called an *overflow.* Unless it is specifically stated always assume you are given only 8-bit registers.

3.1.2 Signed Integers

Signed integers denote whole numbers which are both positive and negative. There are three well know approaches namely; Sign & Magnitude representation (S&M), Bias notation (also know as Excess notation) and Complement method.

Definition: Numbering Bits

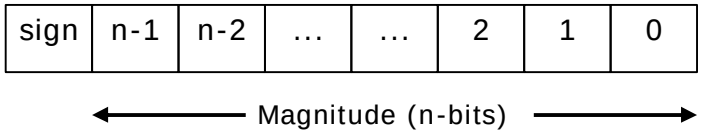
Individual bits on a bit string can be labelled as 0^{th} bit, 1^{st} bit , 2^{nd} bit….. n^{th} bit. Bits are numbered starting from right most bit where the LSB is considered to be bit 0.

Bit Position	n	n-1	….	4	3	2	1
Bit Number	n-1	n-2	….	3	2	1	0

.

Sign and Magnitude Representation

The Sign & Magnitude representation uses the MSB to represent the sign (i.e. whether the number is positive or negative) of the number. Rest of the bits are used to represent the magnitude of the number.



In an n+1 bit register, n^{th} bit is used to represent the sign while rest of the bits are used to represent the magnitude. Following convention is used to in represent the sign. If sign is:

- 0 – the integer is positive or zero
- 1 – the integer is negative or zero

Example 3.2: Represent -29_{10} in S&M method. Assume the register size as 8-bits.

First convert the given decimal number to the corresponding binary number.

$29_{10} = 11101_2$

Since we are given an 8-bit register 7th bit (we start) is used to denote the sign. Since the given number is negative sign bit should be 1. The remaining 7 bits need to be filled with the magnitude (absolute value). Then the answer is:

10011101

Since we are using the MSB to represent sign, the number range gets reduced. Now we have only n-1 bits represent the magnitude (if the register is 8-bit).

Example 3.3: Find the minimum, maximum and the number range that can be stored in an 8-bit register using S&M method.

Since the register is 8-bit wide we can have 7 bits to represent the magnitude (equation 3.1).

$$\text{Minimum number} = -2^n-1 = -2^7-1 = -127$$
$$\text{Maximum number} = +2^n-1 = +2^7-1 = +127$$
Therefore the number range is from -127 to +127.

Within the above range we have the following bit sequences:

0000 0000 = +0
1000 0000 = -0

Now there is 2 definitions for zero which we call the *positive zero* and *negative zero*. This conflicting definition of zero is an issue in computerized calculations. Therefore this method is not that much popular in computerized mathematics.

Concept of the S&M is simple compared to the other two approaches but it causes some problems in calculations and in designing circuits. The Complement method is the most commonly used sign number representation in practice.

3.2 Qualitative Data

As mentioned above qualitative data represent some sort of a quality or characteristics rather than some numerical value. Early computers were mainly processing quantitative data, but today most computers process lot more qualitative data than quantitative data. Qualitative data mostly involves alphanumeric codes. These codes includes letters such as; A to Z, digits such as; 0 to 9, symbols such as; &, %, <, >, @, #, $ and control characters such as; <CR>, <LF>, <BEL>, <ESC>, <DEL>[4].

There are several well know alphanumeric code standards such as; ASCII, EBCDIC and Unicode.

3.2.1 ASCII

American Standard Code for Information Interchange (ACSII) is the most widely used alphanumeric code set. It is used to represent uppercase and lower case Latin letters, numbers and punctuations. There are 128 standard ASCII codes which are represented using 7-bits (also called as 7-bit ASCII). In actual data storage and transmission 8-bits are used and the MSB is called the *parity*[5] bit. Selected set of ASCII codes are given in table 3.2.

3.2.2 EBCDIC

Extended Binary Coded Decimal Interchanged Code (EBCDIC) is used primarily by large IBM computers (super computer and main frames) and compatible equipments. It is also referred as 8-bit ASCII. EBCDIC is not so popular compared to ASCII.

Although ASCII is heavily used it does only allow more than 128 characters. However most languages have more than 128 characters (in Sinhala it is more than 300 characters and some of the Japans and Chinese languages include more than 3000 characters). Therefore there was a need for new coding system which can supports large number of characters as well as multiple languages. This leads to the introduction of Unicode.

[4] CR – Carriage Return, LF – Line Feed, BEL – Bell, ESC – Escape, DEL - Delete
[5] The parity bit is used to check errors in data transmission.

Table 3.2 – Selected set of ASCII codes

ASCII code	Hexa	Symbol
0	0	NUL
1	1	SOH
2	2	STX
3	3	ETX
4	4	EOT
5	5	ENQ
6	6	ACK
7	7	BEL
8	8	BS
9	9	TAB
10	A	LF
11	B	VT
12	C	FF
13	D	CR
14	E	SO
15	F	SI

ASCII code	Hexa	Symbol
48	30	0
49	31	1
50	32	2
51	33	3
52	34	4
53	35	5
54	36	6
55	37	7
56	38	8
57	39	9
58	3A	:
59	3B	;
60	3C	<
61	3D	=
62	3E	>
63	3F	?

ASCII code	Hexa	Symbol
65	**41**	**A**
66	42	B
67	43	C
68	44	D
69	45	E
70	46	F
…	…	…
…	…	…
87	57	W
88	58	X
89	59	Y
90	**5A**	**Z**
….	…..	…
97	**61**	**a**
98	62	b
99	63	c

3.2.3 Unicode

Unicode[6] was designed to overcome the limitation of characters in both ASCII and EBCDIC. It is a 16-bit character representation therefore it can represent 65536 (2^{16}) characters. Because of this large range it can assign unique character codes to characters in a wide range of languages. Unicode provides a unique number for every character no matter what the; platform, program or language.

A singe Unicode character set can also represent multiple languages. It even includes Sinhala and Tamil. The range 0x0D80 (in hexadecimal) to 0x0DFF is assigned for Sinhala (see figure 3.2).

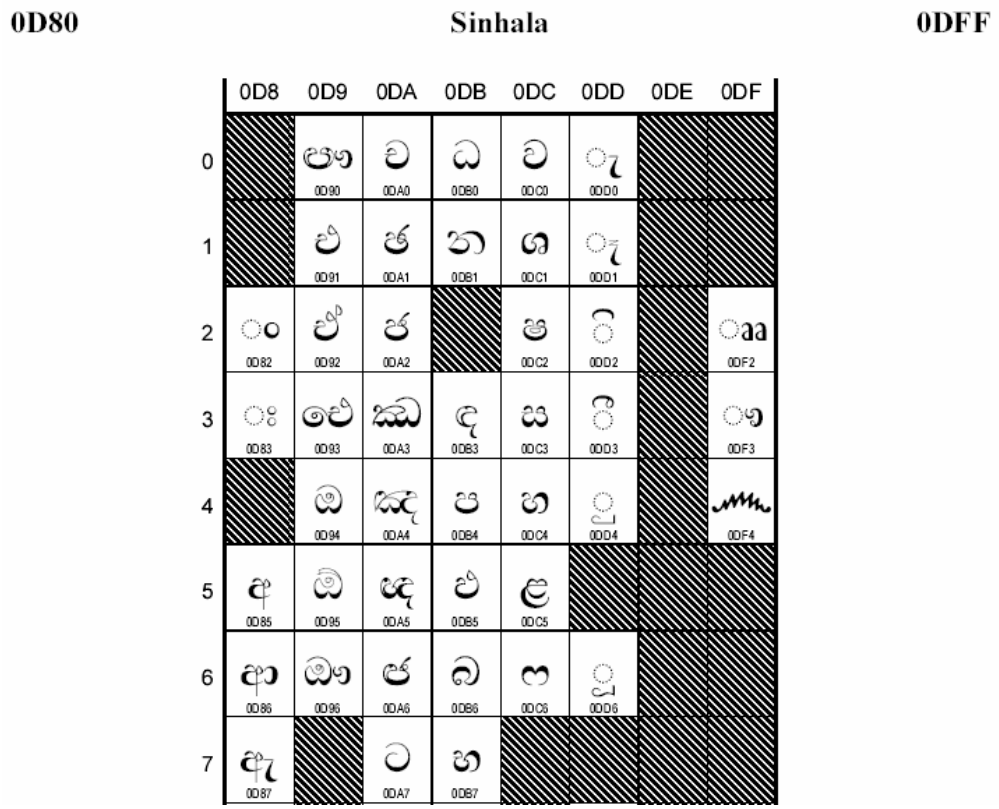


Figure 3.2 – Representation of Sinhala in Unicode

[6] The official Unicode website - www.unicode.org

4 – Logic Gates

4.1 Boolean Algebra

George Boole (1815-1864) a Logician, developed an algebra know an as the Boolean algebra, to examine a given set of propositions (statements) with a view to checking their logical consistency and simplifying them by removing redundant statements or clauses. He used symbols to represent simple propositions and compound propositions were expressed in terms of these symbols and connections.

Consider the statement "*Kamal is a clever student and he passes exams well*". In this statement proposition "Kamal is clever" can be denoted by symbol 'A' and "he passes exams well" can be denoted by symbol 'B'. Then this statement can be represented as; 'A and B'. The symbols 'A' and 'B' are connected by the connective AND.

When a variable is used in algebraic formula, it is generally assumed that the variable may take any numerical value. For an example; in the formula $y = 2x + 3$, it is assumed that x and y may range through the entire filed of real numbers. A variable in Boolean algebra can only have one of 2 possible values 0 or 1 where 0 denotes the 'FALSE' condition while 1 denotes 'TRUE' condition.

In above example; if the statement "Kamal is a clever student" is TRUE, the statement "he passes exams well" is also TRUE. But if the statement is FALSE (if Kamal is not clever) the second statement is also FALSE. This can be represented in the following table which is called the *truth table.*

Kamal is a clever student	He passes exams well
0 (FALSE)	0 (FALSE)
1 (TRUE)	1 (TRUE)

Consider another statement "Saman will go to the party if both Kamal and Mala are going". Following truth table represents the given statement.

K	M	S
0	0	0
0	1	0
1	0	0
1	1	1

K – Kamal going to the party
M – Mala going to the party
S – Saman going to the party

Above condition is satisfy (when S = 1) only if both the inputs are 1 (TRUE). This is a *logical AND* operation. In Boolean algebra AND operation is indicated by the period (**.**) symbol. This is also called the *Boolean multiplication.*

If the statement is changed as "Saman will go to the party if either Kamal or Mala is going". This can be represented in the following truth table.

K	M	S
0	0	0
0	1	1
1	0	1
1	1	1

K – Kamal going to the party
M – Mala going to the party
S – Saman going to the party

If either Kamal or Mala or both of them goes to the party, Saman will also attend the party. This is a *logical OR* operation. In Boolean algebra OR operation is indicated by the plus (+) symbol. This is also called *Boolean addition*.

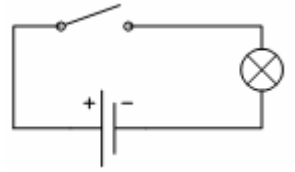
Following is a summary of Boolean multiplication and addition operations.

Boolean addition
0+0 = 0
0+1 = 1
1+0 = 1
1+1 = 1

Boolean multiplication
0.0 = 0
0.1 = 0
1.0 = 0
1.1 = 1

4.2 Logic Gates

Computers are developed using bi-stable devices such as transistors which can either be in ON or OFF states. A simple switch is an example for such a device (figure 4.1).



Switch	Bulb
0	0
1	1

Figure 4.1 - A switching circuit and its Truth table

If the switch is ON the lamp will glow. If the switch is OFF there want be any light. This is a simplest form of a logic circuit. In binary representation logic '0' can be used to indicate the OFF state while '1' can be used to indicate the ON state. Using these notations the operation of a bulb can be represented in a truth table as follows:

Computers are built using whole lot of circuits which include thousands or even millions of *Logic Gates*. Logic gates manipulate the TRUE and FALSE states (0's and 1's) by allowing or disallowing electrons to flow through them.

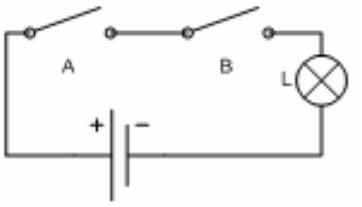
4.2.1 Fundamental Gates

There are 3 fundamental logic gates namely:

1) AND are all inputs are true?
2) OR is at least one input is true?
3) NOT flip the truth value

AND Gate

Let us consider another switching circuit and its truth table (figure 4.2).

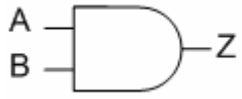


A	B	L
0	0	0
0	1	0
1	0	0
1	1	1

Figure 4.2 - A switching circuit for AND operation and its Truth table

The bulb will glow (TRUE) only if both A and B switches are closed (TRUE). This is the purpose of an AND gate. It should produce a TRUE output only if both inputs are TRUE. This is equivalent to the Boolean multiplication. There are AND gates with more than 2 inputs and to produce a TRUE output all inputs much be TRUE.

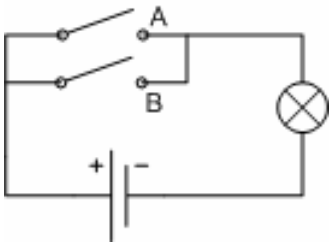
Symbol



Truth Table (Z = A.B)

A	B	Z
0	0	0
0	1	0
1	0	0
1	1	1

OR Gate

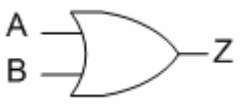


A	B	L
0	0	0
0	1	1
1	0	1
1	1	1

Figure 4.3 - A switching circuit for OR operation and its Truth table

Consider another switching circuit given in figure 4.3 with two parallel switches. The bulb will glow (TRUE) if either switch A, switch B or both switches are is closed (TRUE). This is the purpose of an OR gate. It should produce a TRUE output if either or both inputs are TRUE. This is equivalent to Boolean addition.

Symbol

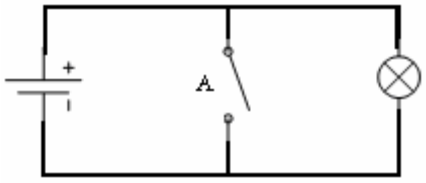


Truth Table (Z = A+B)

A	B	Z
0	0	0
0	1	1
1	0	1
1	1	1

NOT Gate

In the circuit given in figure 4.4 the bulb will glow (TRUE) when the switch is on open state (FALSE). When the switch is closed (TRUE) the path through the switch has a lower resistance than the path through the bulb. Therefore current conducts only through the switch not through the bulb. As a result the bulb will not produce any light (FALSE). See the corresponding truth table. This is equivalent to Boolean complement or the negation.

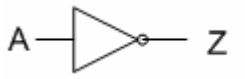


A	Z
0	1
1	0

Figure 4.4 - A switching circuit for NOT operation and its Truth table

The NOT gates inverts its input. If the input is TRUE the output is FALSE. If the Input is FALSE the output is TRUE. A NOT gate has only a single input and output.

Symbol



Truth Table (Z = $\overline{A}$)

A	Z
0	1
1	0

Based on these 3 fundamental gates several other gates can be derived.

4.2.2 Gate Networks

The AND, OR and NOT gates can be interconnected together to form another set of gates and logic networks. These are also called *combinational networks*. Based on these fundamental gates NAND, NOR, XOR and XNOR gates are formed.

NAND Gate

NAND gate is a combination of an AND gate and a NOT gate (figure 4.5).

A	B	X (A.B)	Z ($\overline{X}$)
0	0	0	1
0	1	0	1
1	0	0	1
1	1	1	0

Figure 4.5 – Formation of a NAND gate and its Truth table

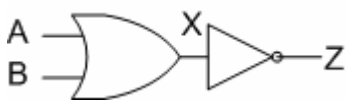
Symbol

Truth Table (Z = $\overline{A.B}$)

A	B	Z
0	0	1
0	1	1
1	0	1
1	1	0

NOR Gate

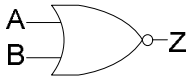
NOR gate is a combination of an OR gate and a NOT gate (figure 4.6). Following tables illustrates how the truth table of a NOR gate is formed.

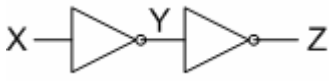


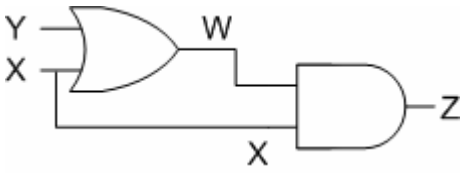
A	B	X (A+B)	Z ($\overline{X}$)
0	0	0	1
0	1	1	0
1	0	1	0
1	1	1	0

Figure 4.6 – Formation of a NOR gate and its Truth table

Symbol **Truth Table (Z = $\overline{A.B}$)**







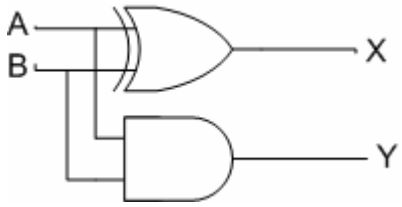
X	Y	W =(X+Y)	Z = (X.W) = X.(X+Y)
0	0	0	0
0	1	1	0
1	0	1	1
1	1	1	1

From above truth table we can see that both Z and X is equal.

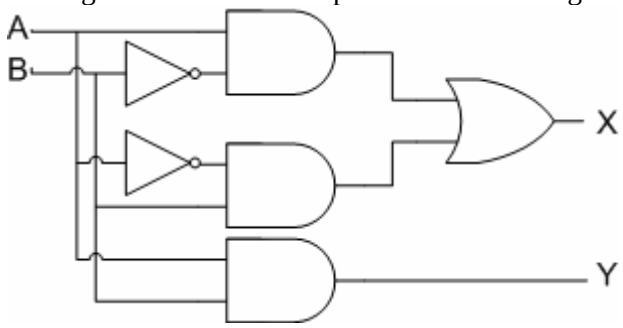
Example 4.3: Consider the following set of binary addition operations performed on a single bit. Deign a gate network to perform this operation.

A	B	C = A+B	Sum (X)	Carry (Y)
0	0	0	0	0
0	1	1	1	0
1	0	1	1	0
1	1	10	0	1

If you are given only a single bit register the last operation will generate an overflow. It will set the carry bit. Based on above results we can design a gate network to perform single bit addition operation. Where $X = A.\overline{B} + \overline{A}.B$ = A⊕B and Y = A.B.



Example 4.4: Redraw the same gate network in example 4.3 without using a XOR gate.



The logic networks inside a computer are much more complex than these. Regardless of their complexity they make use the same fundamental logic gates such as AND, OR and NOT.

Exercise 4.1 – Draw a logic network that implements the equation $Z = \overline{(A \bullet B)}$.

Exercise 4.2 – Draw a logic network that implements the equation $Z = \overline{(A + B)}$.

Exercise 4.3 – Draw a logic network that implements the equation $Z = \overline{A} + \overline{B}$.

Exercise 4.4 – Draw a logic network that implements the equation $Z = \overline{A} \bullet \overline{B}$.

Exercise 4.5 - Prove that gate network in exercise 4.1 is equivalent to gate network in exercise 4.3.

Exercise 4.6 - Prove that gate network in exercise 4.2 is equivalent to gate network in exercise 4.4.

5 – Introduction to Computer Hardware

Hardware is the combination of all physical components that makes up a microcomputer, monitor, printer, etc. A modern computer consists of various hardware components. Some of these components are essential for the proper functionality of a computer while some are optional. These hardware components are governed by various electrical, electronic and mechanical engineering principles. All these principles combined with engineering excellence produces state of the art hardware components.

This chapter introduces the major hardware components in modern personal computers, their usage, design concepts, operations, capacities, speeds, etc. Rather than following the standard approach of introducing the microprocessor first, we will start with introducing the motherboard and memory before introducing the microprocessor.

5.1 Computer

The computer is a machine, which is able to execute a finite set of instructions. She[8] can process data based on those instructions. It is not easy to provide a clear-cut definition of a computer. What is possible is to identify a set of features associated with computers. If a device satisfies these features, it could be classified as a computer.

Significant features of a computer are:

1. It is a machine.
2. Able to execute (understand) a finite set of '*instructions*'.
3. Able to process '*data*' according to those 'instructions'.
4. Able to execute a 'sequence of instructions' that are 'stored' within the machine in a specified order.
5. Able to deviate from the 'sequence' based on a 'result' of a previous operation.

From the above list it is clear that computers are generally associated with '*processing data*' and '*executing instructions*'. However, there are several machines capable of data processing and executing instructions. For example, a calculator is capable of both processing data and executing instructions.

Then the question arises as to what the difference between a calculator and a computer. The answer is points 4 and 5 identified as above. Normally a calculator is not capable of executing a "sequence of instructions" that are "stored" within the machine in a specified order or deviating from the "sequence" based on a "result" of a previous operation. Hence, a calculator cannot be considered as a computer. Everything in a calculator is hard-coded (i.e. developed as physical electronic circuits).

However, there exists a category of calculators known as "programmable calculators" which satisfy both the above requirements. Therefore logically a distinction cannot be made between such calculators and computers. Hence, such programmable calculators can be termed computers.

This means that the term "computer" has a wider meaning. Computers are not just the machines that are placed on desks, which are generally known as 'Personal Computers' or 'Desktops' and machines people carry with them known as 'Laptops'. Personal Desktop/Digital Assistants (PDA's) or 'Pocket PCs' and embedded controllers used in equipment ranging from aircrafts to washing machines are categorized under computers as well.

5.2 Personal Computer

It is difficult to identify the first ever Personal Computer (PC). Altair 8800, produced by MITS in 1975, using the Intel 8080 microprocessor is considered one of the first commercially successful

[8] The gender of a computer is considered as female, hence it is referred to as "she" rather than "he" or "it". Since most computer scientists were male, they decided that a computer should be female.

'personal computers' (cost about $ 375). It contained 256 bytes of memory and had to be programmed using a switch panel. In 1975, Bill Gates and Paul Allen founded Microsoft and developed BASIC 2.0 language, which was used to program the Altair 8800. It was the first *high-level* language available for a personal computer. Within the next decade, many more personal computers were developed and introduced to the market.

5.3 Exponential Growth of Computer Hardware Technology

Since the invention of the Integrated Circuit (IC) and the Microprocessor, the growth of computer hardware technology has been exponential. In 1965 Gordon Moore, co-founder of Intel, observed that the number of transistors per square inch on integrated circuits had doubled every year. Now the rate has slowed down to doubling in every 18 months. This is known as the Moore's Law (figure 5.1).

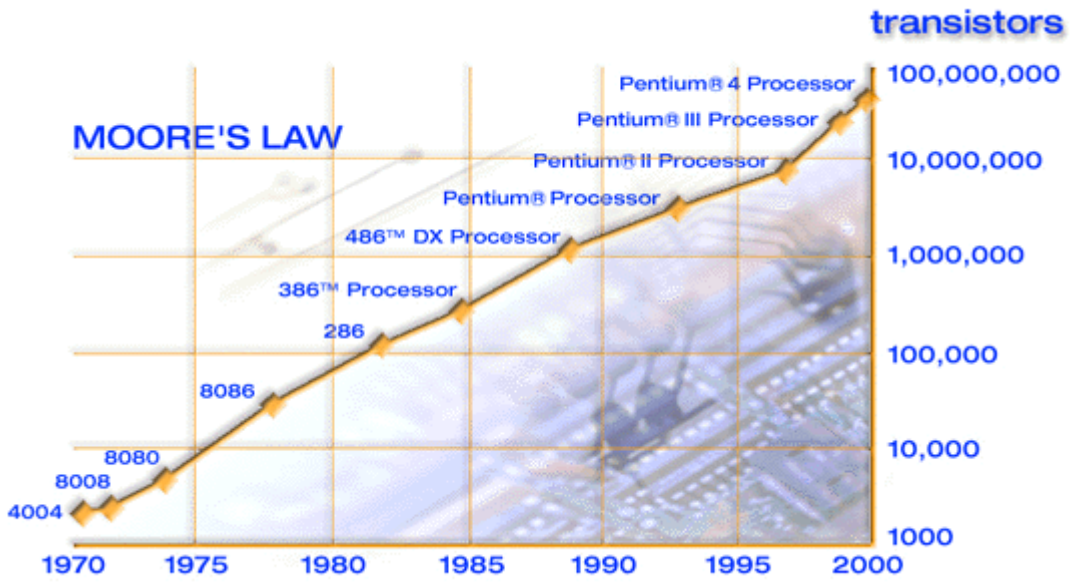


Figure 5.1 – Illustration of Moore's Law

5.4 Major Components of a Personal Computer System

A personal computer system consists of several different parts. Some of these parts can be considered as subsystems within the personal computer system. The following are the common subsystems and parts within a PC (figure 5.2 and 5.3).

- Central Processing Unit (CPU)
- Motherboard
- Storage subsystem
  - Magnetic disk drives – Hard disks, floppy drives
  - Optical disk drives – CD/DVD drives
  - Tape drives
  - Punch card machines etc.
- Input output subsystem
  - Video (Graphics) card
  - VDU – Video Display Unit
  - Keyboard
  - Mouse
  - Printer
  - Scanner
- Memory subsystem
  - RAM – Random Access Memory
  - ROM – Read Only Memory

- Multimedia components
  - Sound card
  - Speakers
  - Game cards
  - Joysticks
- Case/Chassis
- Power supply

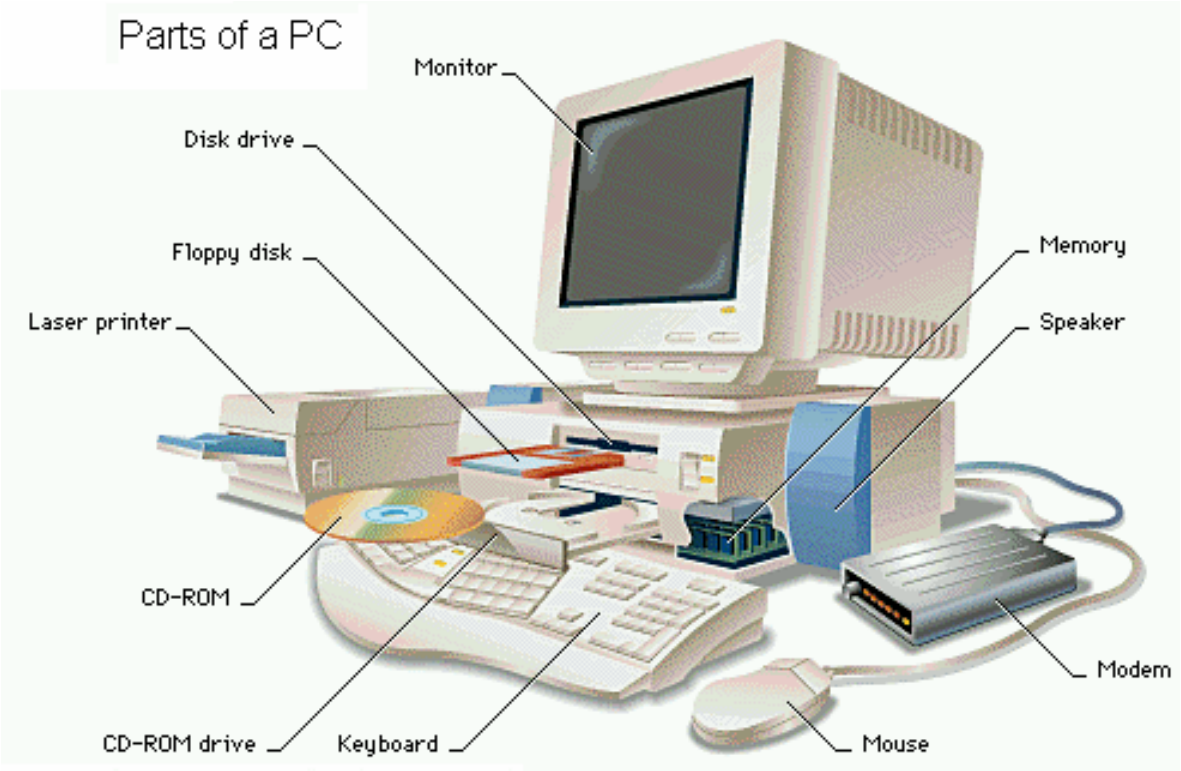


Figure 5.2 – External view of a PC

5.5 The Traditional View of a Computer System

The architecture of modern computers is somewhat different to traditional ones. However, traditional computers are easier to understand compared to the modern ones. Therefore first we will look at concepts of a traditional computer.

The traditional view of a computer system can be classified as; *single user computer systems* and *multi-user computer systems*.

5.5.1 The Traditional View of a Single User Computer System

These are said to be single user computer systems because only a single user can use them at a time. Figure 5.4 illustrates different subsystems of a single user computer system.

As shown in figure 5.4 the Central Processing Unit (CPU) communicates with all the other subsystems and manages the overall functionality of the computer system. The CPU is like the processing elements of a human brain. The Video Display Unit (VDU) is responsible for producing the visual output. The VDU, keyboard and printer belongs to the Input Output (IO) subsystem. These are analogues to the sensors (e.g. nose) and actuators (e.g. hands) of a human.

The memory subsystem of the computer is composed of the Random Access Memory (RAM) and the Read Only Memory (ROM). Like the human brain, computers also need a temporary memory system (provided by RAM) and a permanent memory system (provided by ROM).

The RAM keeps all the current data and instructions in memory while ROM keeps all the data and instructions that need to be there for a long-time. RAM looses all the stored data and instructions when

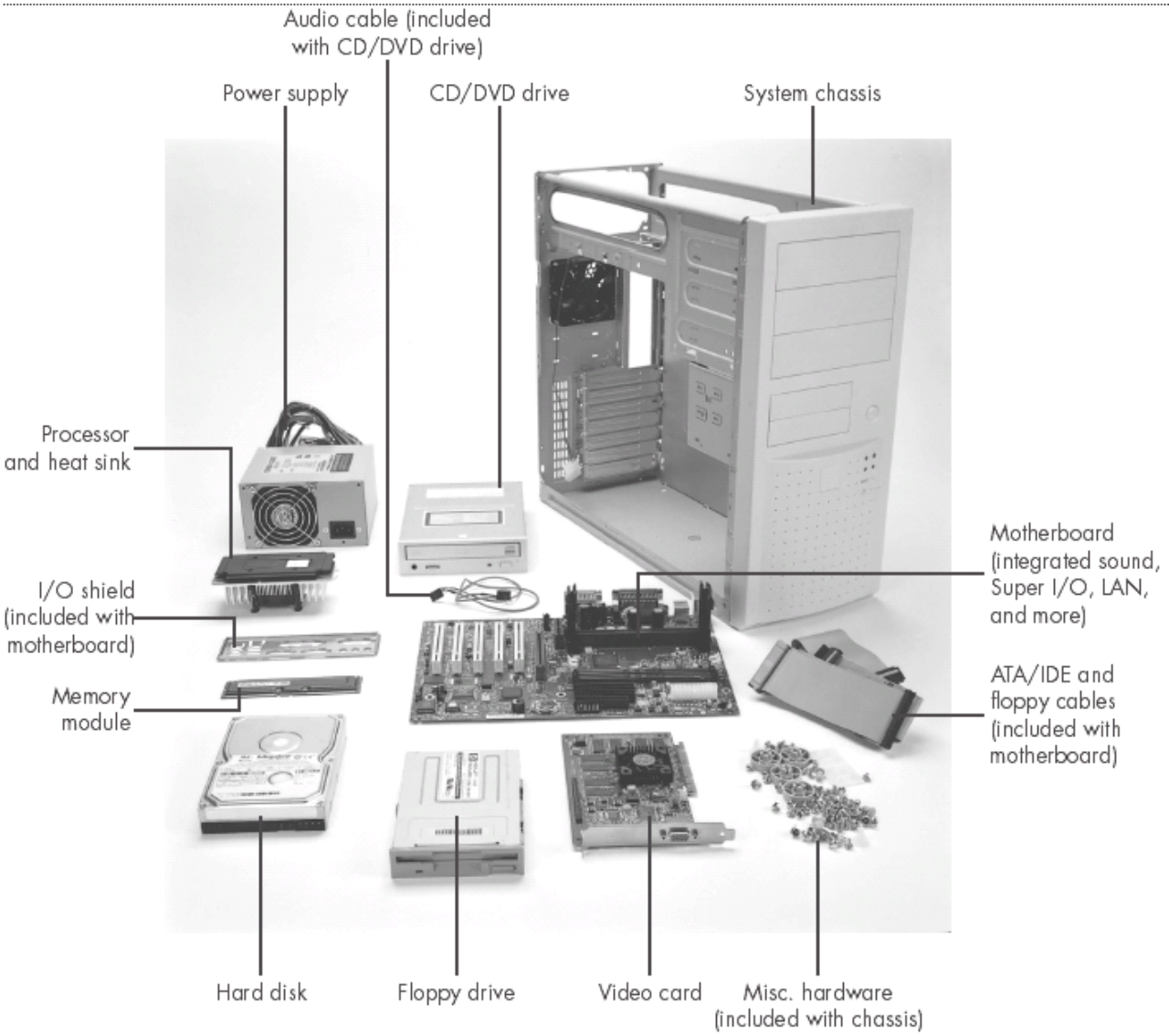


Figure 5.3 – Parts of a PC

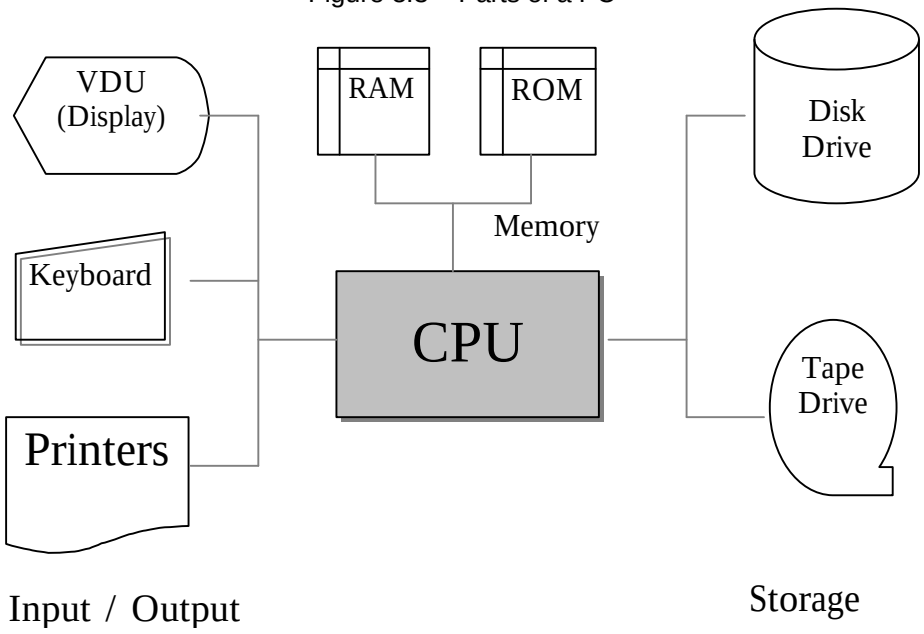


Figure 5.4 – Traditional view of a single user computer system

the computer is switched off. The capacity of both RAM and ROM is limited but computers demand for more data and instruction storage. As a result, secondary storage was introduced.

Secondary storage mainly includes disk drives (hard disks and floppy disks) and tape drives. Hard disks are a high capacity, fast and permanent storage medium. The concepts behind the hard disk will be discussed later. Tape drives (also called DAT drives) are used store data on magnetic tapes. Magnetic tapes were introduced before hard disks and it was used to store both data and instructions. Magnetic tape is a high capacity and cost effective data storage solution. However the data reading/writing speed of a magnetic tape is slower therefore nowadays tapes are used only to back up large amounts of data.

The CPU is the central unit and it communicates with all the other systems and subsystems. It is also responsible for controlling and managing rest of the system. This is a very simple approach to understand and implement but it does not scale well since everything depends on the CPU.

5.5.2 The Traditional View of a Multi User Computer System

In this approach, multiple users use a single physical computer simultaneously. Each user has a separate VDU and a keyboard to interact with the rest of the system (figure 5.5). Rest of the system is same as the single user system. All the users share the CPU, memory, disk drives, printers, etc.

These types of computers are used in places like Banks where multiple 'Dumb Terminals'[9] are connected to a single computer. This approach allows users to easily share resources among them. This is useful in a Bank were several banking officers trying to access the same customer accounts database. If one officer makes changes in any account, it should be immediately visible to the other banking officers.

5.5.3 A Modern Computer System

With the introduction of personal computers and miniaturization, most of the subsystems were included in a box, which is called the 'Machine' (see figure 5.6). Most people refer to this box as the CPU, which is an incorrect term. The CPU is just a single piece of component inside the box.

Modern computer systems contain many additional parts that are not identified with the traditional view of a computer system. These include; Video sub-system, Multimedia devices such as Sound cards, speakers, microphone, CD-ROM/DVD-ROM, networking adapters such as Network Interface Cards (NICs) and Modems, and communication ports such as serial and parallel ports.

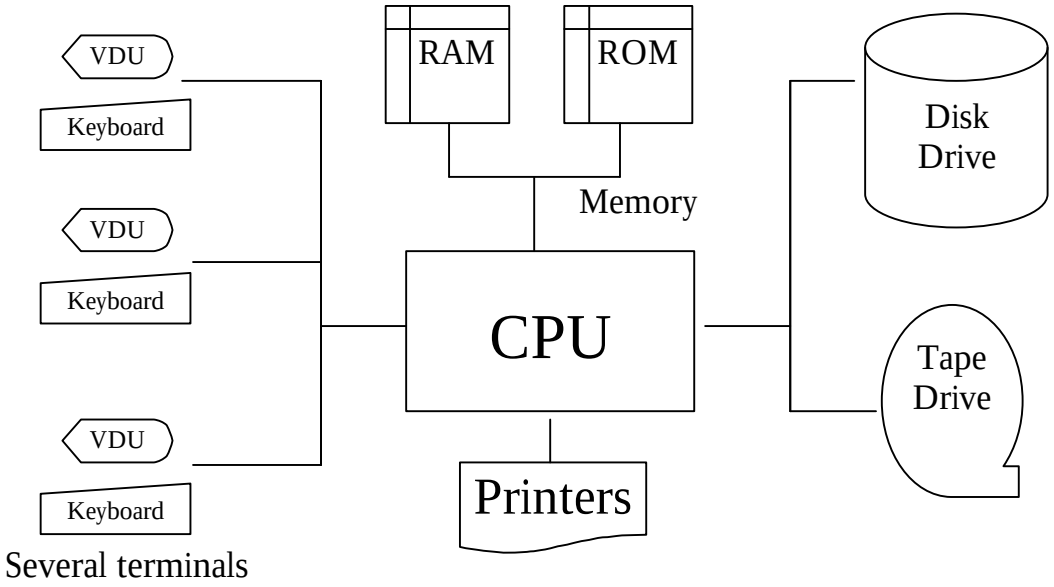


Figure 5.5 – Traditional view of a multi user computer system

[9] These are said to be 'Dumb Terminals' because these terminals just display anything asked by the CPU of the central computer and accept any input from a keyboard and pass that to the CPU for processing. They do not perform any data processing.

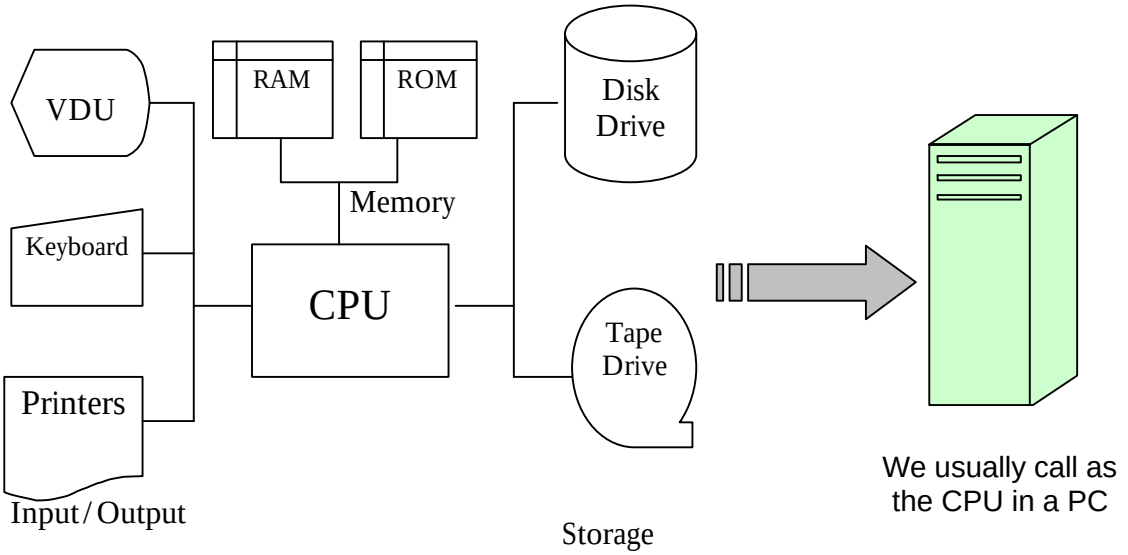


Figure 5.6 – From Traditional View to a modern Computer System

5.6 The Motherboard

The motherboard is the most important component in a PC and it is also called the 'main board'. The motherboard is a large circuit board where the processor, memory and other electronic components are attached. Several other ICs called controllers are attached to the motherboard as well. These controllers are responsible for:

- controlling the hard disks and floppy diskette drives
- communicating with keyboard, mouse, printers, modems, etc.
- supporting operations carried out by the micro-processor (Direct Memory Access controller)

The motherboard provides the path through which the processor communicates with memory, disks, expansion cards, keyboard and other components. Figure 5.7 illustrates the layout of a typical motherboard.

5.7 Memory

As human beings have memory computers also have some sort of memory. Part of this memory is permanent (non-volatile) and other portion is volatile. The volatile memory looses its content whenever power is switched off.

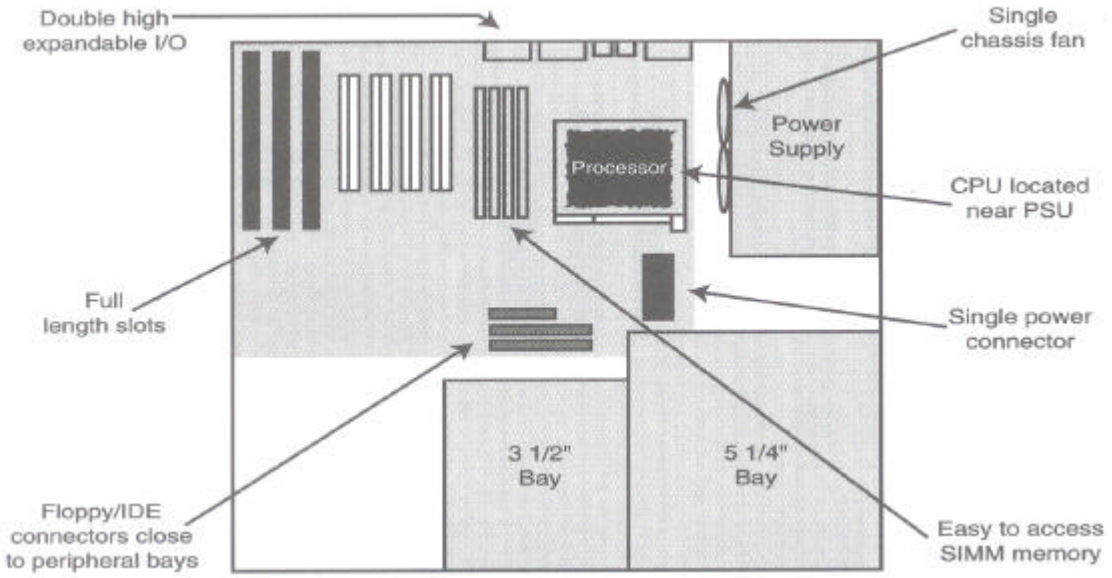


Figure 5.7 – Layout of a typical PC motherboard

Memory is a temporary storage area for programs and data being operated by those programs. Whenever the microprocessor wants to do some processing it gets both data and instructions from the memory and executes them. After the execution, the results are sent back to the memory. Therefore the memory is considered as the workspace (or the play ground) for the microprocessor.

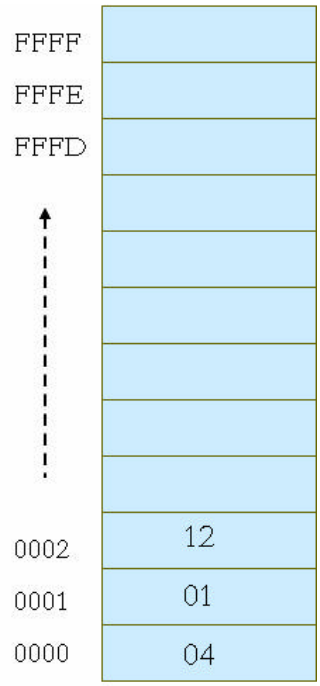


Figure 5.8 – Array of memory locations

Memory consists of an array of consecutive memory locations. Each such location stores a single piece of data (usually a byte). Memory locations are identified by a unique memory address. The first memory location is labelled as memory address zero (0x0000) and rest of the memory locations are labelled one after another. The addresses are normally represented as hexadecimal numbers. In figure 5.8 there are 65536 (2^{16}) memory locations where the first address is 0x0000 and the last address is 0xFFFF. Memory address 0x0000 holds the value 0x04 while memory addresses 0x0001 and 0x0002 hold integers 0x01 and 0x12.

The microprocessor uses memory to store Instructions (programs) and Data (characters and digits). At a time, the microprocessor either reads or writes only one memory location. In the long run it may need to access each and every memory location within the memory array. However, it is impractical to connect each and every memory location directly to the microprocessor using a separate set of electrical connections (i.e. wires). In figure 5.8 there are 65536 memory locations, therefore to connect each memory location 65537 wires are required (assuming a common ground wire). To overcome this issue all the memory locations are connected using shared electrical connections, called the "*Memory Bus*".

5.7.1 Memory Bus

Following is a hypothetical example (figure 5.9) of a bus which runs on a two way road and which accommodates only one passenger. Assume there is a passenger '**A**' in 'house ①', who wants to go to 'house ③'. '**A**' can go to the road and get the bus to 'house ③'. When he arrives at 'house ③' he can get of the bus. During this time bus is dedicated only to passenger '**A**' and no one else can travel in the bus. Suppose passenger '**B**' also wants to go the 'house ③'. Then she has to wait until the bus is free. When it is free, she can use the bus.

Figure 5.9 – Illustration of a Bus

Now suppose that both '**A**' and '**B**' wants to go to 'house ③' at the same time. In a conventional public transport system '**A**' will always get the bus first since he is at the starting point of the road. However, the bus in a computer is slightly different. There is no specific starting point. Whoever comes first to the road (even just fraction of a second before other passenger) will always get the bus

and the other one has to check from time to time whether the bus is free. While the second passenger is waiting if a third passenger comes and tries access the road (just after bus gets free) he/she will get a chance, where the waiting passenger may still need to wait. If the waiting one is unlucky, he/she has to wait forever. This is not a fair deal. Therefore the microprocessor (or bus controller) decides who is going to access the bus at what time.

The bus inside the computer is designed based on the above algorithm. The road is bidirectional and at a time only one memory location can access the bus. A bus is a set of electrical connections (parallel set of wires or set of copper strips on the motherboard) that connects the memory and the microprocessor. There are 3 types of busses; namely *Address Bus*, *Control Bus* and *Data Bus*. All of the 3 buses are enclosed within the memory bus (figure 5.10).

- Address Bus – is used by the microprocessor to indicate the memory location (indicated by the memory address) that it is interested in. Address bus goes from CPU to memory.
- Control Bus – is used to send control information such as read request (RD) or write request (WR) to the memory. Control bus goes from CPU to memory.
- Data Bus – used for actual data transmission. This is a bidirectional bus.

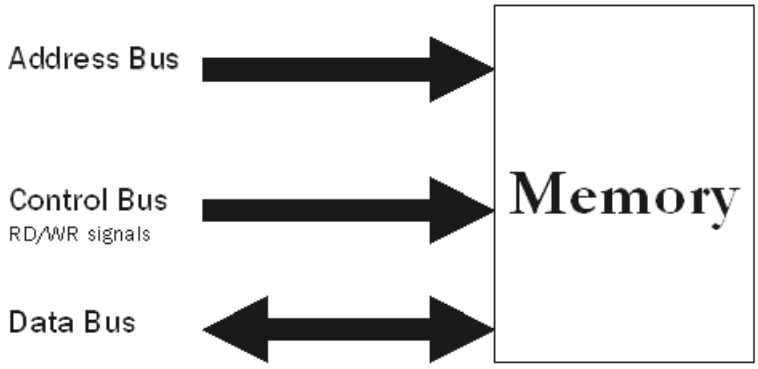


Figure 5.10 – Communicating with the memory

The operations of writing data to the memory and reading from the memory make use of all three buses.

Writing data to the memory

1. First the address of the memory location is placed on the Address Bus
2. Then data is placed on the Data Bus
3. Finally the CPU activates the Write Signal (WR) in the control bus

Reading data from the memory

1. First the address of the memory location is placed on the Address Bus
2. Then the CPU activates the Read Signal (RD) on the control bus
3. Finally data is fetched (read) from the Data Bus

5.7.2 Types of Memory

Several different types of memory are used for various purposes within a computer or microprocessor controlled system. These differences are characterised based on their manufacturing process, volatility, programming methodology and cost.

Read Only Memory – ROM

As its name implies this memories is read only. They are written (actually developed as pre-wired electronic circuits, referred as *hard coded*) when they are being fabricated as ICs. The ICs that are available in the open market with various melodies (i.e. piece of music) belong to this category. In computers, these memories are used to store initial start-up programs of the computer[10]. Since these memories are hard coded, it is not economical to produce these in small quantities.

[10] Referred as the POST – Power On Self Test. A series of tests run by the computer at power-on to check whether the attached hardware components are working correctly.

Programmable Read Only Memory – PROM

These memories are same as ROMs, but their contents can be written once after the individual ICs are manufactured. The programming process requires special equipment. If a smaller number of ROMs are required for a particular application, programming several PROMs is a more cost effective.

UV Erasable PROM – UVEPROM

UVEPROM is similar to PROM but contents can be written several times. Before writing new content existing program should be erased by exposing the IC to Ultra Violet (UV) light. For this purpose, there is an opening at the top of the IC and this is normally coved by a sticker, which does not allow UV light to penetrate (figure 5.11). During the content erasing process, this protective cover should be removed before exposing the IC to the UV light. If this protective cover is damaged the content in the IC may get corrupted after a while. Both erasing and programming processes require special equipment.

Figure 5.11 – UVEPROM

Electrical Erasable PROM – EEPROM

Same as the UVEPROM, except that the content (the program) of the IC can be erased by applying a special high voltage to some of the pins. The programming process requires the EEPROM to be removed from the application circuit and plugged in to a special device before erasing the content.

In modern televisions, we need this sort of memories to keep track of the current channel, volume level, colour configurations, etc. However, it is not practical to remove the EEPROM from the TV and connect to a special programming circuit. The flash ROM was introduced as a solution.

Flash ROM

Flash ROM is a special type of EEPROM that can be erased or programmed while in the application circuit. Once programmed the contents remains unchanged even after a power failure. Flash ROMs are commonly used in modern PCs, various networking devices such as routers and firewalls and memory pens (also referred as memory sticks or USB pens).

Read Write Memory – RWM

These are traditionally known as the RAM (Random Access Memory). Contents in these memories are erased when power is disconnected. There are two major types of RAMs; the *Static RAM* (SRAM) and *Dynamic RAM* (DRAM). Static RAM is developed using transistors while dynamic RAM is developed using capacitors. Therefore they reflect the properties of transistors and capacitors (see table 5.1).

Table 5.1 – Transistors vs. Capacitors

Transistors	Capacitors
Developed using Silicon and other semiconductors	Developed using Silicon and other semiconductors
High speed switching	Slower performance
Will retain state forever (if power is supplied)	Automatically discharges after sometime, need refreshing
More reliable	Less reliable
Low density (no of transistors per cm^2)	High density (no of capacitors per cm^2)
High power consumption	Low power consumption
High cost (per bit)	Low cost (per bit)

Static Ram – SRAM

As said earlier static RAMs are built using transistors. Each transistor represents a single unit of data, which is a bit. Arrays of transistors are combined to produce SRAM. Since these are developed using transistors they inherit all the properties of transistors in table 5.1.

Dynamic RAM – DRAM

DRAMs are developed using capacitors. Each capacitor represents a single unit of data and they process all the characteristic of capacitors listed in table 5.1. Capacitors are cheap but they must be re-charged from time to time (certain DRAMs are needed to be refreshed at 15μs intervals). Due to its lower cost, bulk of the PC memory is made out of DRAM.

SRAMs are more reliable but expensive than DRAMs. They also consume more electrical power compared to DRAMs. These are mostly used as cache memories.

5.7.3 Memory Modules

Memory modules combine set of memory ICs together and present it to the motherboard as a single memory block. Various memory modules such as SIM (Single Inline Memory Module), DIMM (Dual Inline Memory Module) and DDR-DIMM (Double Data Rate-DIMM) are available. Different memory modules have different number of connections, speeds and access times. Figure 5.12 illustrates a DIMM (Dual Inline Memory Module)[11].

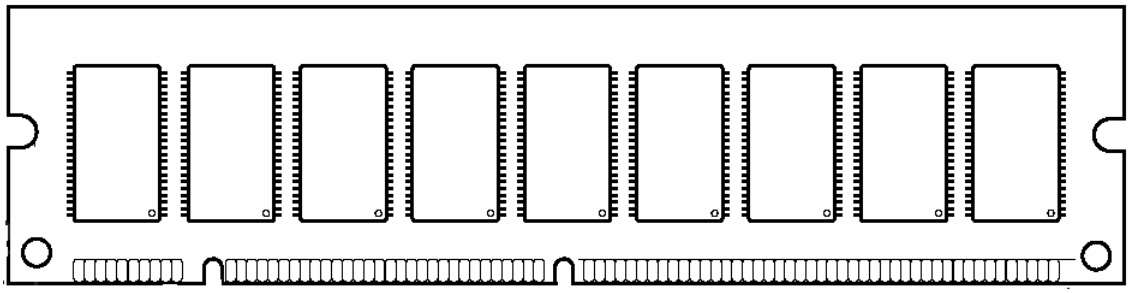


Figure 5.12 – A Dual Inline Memory Module (DIMM)

5.7.4 Memory Characteristics

In order to compare different types of memories, their characteristics need to be considered. A single type of memory does not process all the preferred characteristics. For example, Static RAM has a very high access speed but it is also high in cost. Therefore, a mix of different types of memories has to be used to achieve an optimum result. Listed below are some of the important characteristics of memory:

- Access speed – time taken for the CPU to read from or write to memory
- Cycle time – time taken to complete one memory access operation
- Packing Density - memory capacity per unit area
- Power consumption
- Cost - cost per unit of memory capacity

A memory controller is used to control the communication between different types of memory and the microprocessor (see figure 5.13).

5.7.5 Memory Hierarchy

Modern CPU's are much faster than the speed of memory. The memory has to be organised in such a way that its slower speed does not reduce the performance of the overall system. Some memory types such as Static RAM are faster but expensive in contrast, Dynamic RAM is cheaper but slow. The ultimate objective of having a memory hierarchy is to have a memory system with a sufficient capacity and which is as cheap as the cheapest memory type and as fast as the fastest memory type. The main idea is to use a limited capacity of fast but expensive memory types and a larger capacity of slow but cheap memory types. Special methods are used to store the frequently used items in the faster devices and others in slower devices.

[11] DIMM is a 64-bit wide 168-pin memory module used in new PCs.

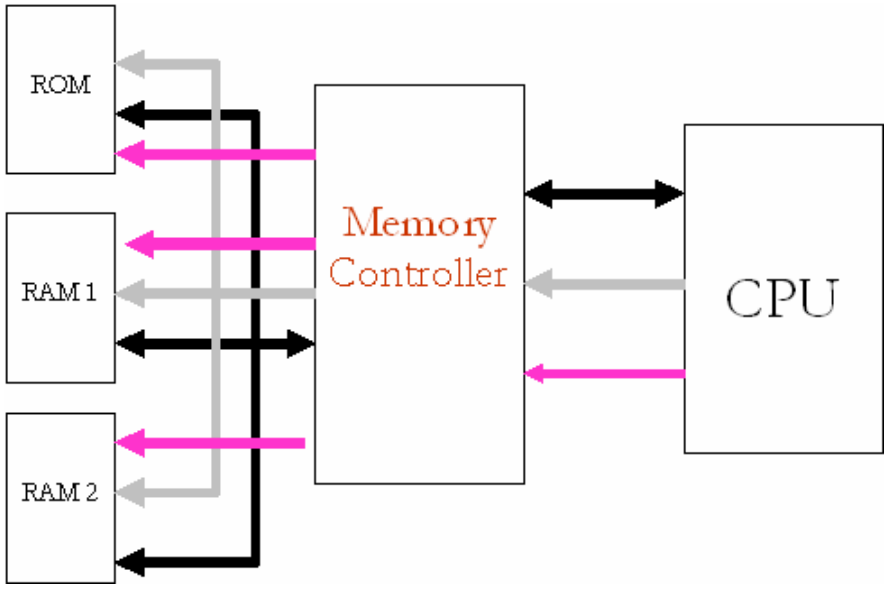


Figure 5.13 – Connecting memory and the microprocessor

Traditional Memory Hierarchy

The traditional memory hierarchy is shown in figure 5.14. The memory access speed gap between secondary storage and registers are filled up by the main memory. Going up the memory hierarchy memory access speed and cost per Megabyte increase. However, the memory capacity (size) increases from top to bottom. In this hierarchy, area of each level is proportional to the memory capacity. Capacities of registers are given in bits (8-bit, 16-bit, 32-bit, 64-bit and 128-bit registers) while capacity of main memory is given in Megabytes (8MB, 16MB, 32MB, 64MB, 128MB, 256MB, etc.). Secondary storage capacities are given in Gigabytes (GB) or some times in Terabytes (TB).

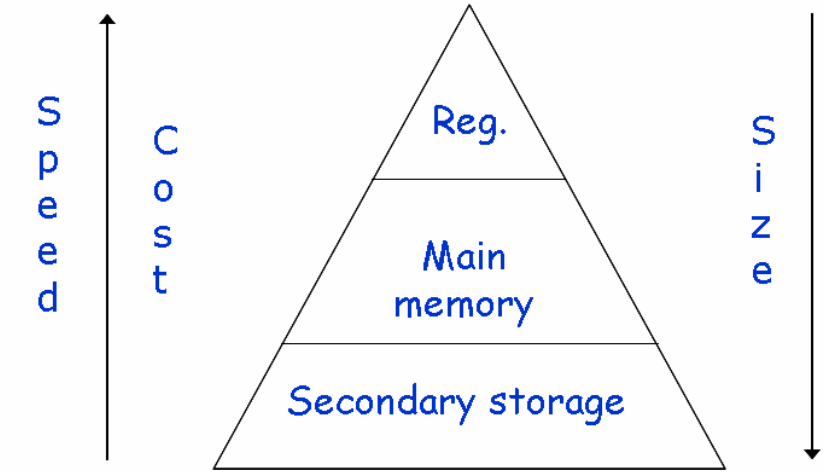


Figure 5.14 – Traditional memory hierarchy

In early days of computing, the speed gap between the memory and the microprocessors was not a big issue. Currently speed of microprocessor increases much faster (speed doubles in every 18 months) compared to the speed of memory access. Therefore the speed gap is widening day-by-day. Although microprocessors are becoming faster and faster they have to depend on memory since memory is the workplace for the microprocessor. This will slow down the fast microprocessors resulting in overall performance degradation. To fill up this widening gap another level called the *Cache memory* is added to the memory hierarchy.

Modern Memory Hierarchy

Modern memory hierarchy includes another level called the 'Cache memory' in-between registers and main memory (figure 5.15). It is a small amount of memory (capacities are given in Kilobytes; common capacities are 128KB, 256KB and 512KB) which is faster than the main memory but slower than the registers. Cache memory fills up the speed gap between main memory and registers.

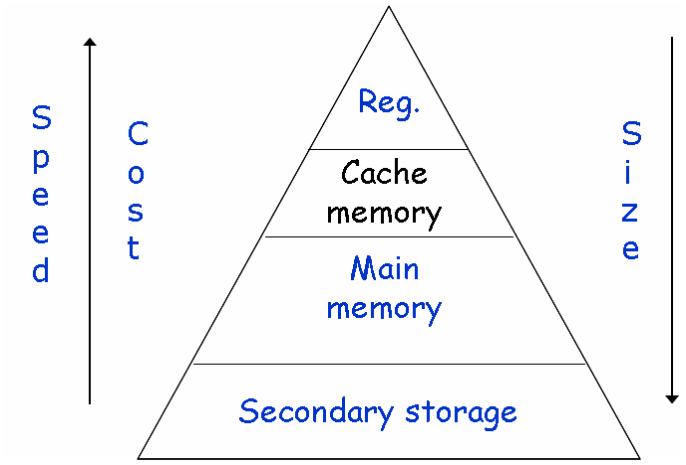


Figure 5.15 – Modern memory hierarchy

Cache is used by the microprocessor to store frequently used instructions and data. Since the cache memory capacity is limited, it cannot hold all the data and instructions that a program needs. Therefore it needs to swap data and instructions between other levels in the memory hierarchy. Consider a case where you want to execute the Notepad in Microsoft Windows.

Step 1: The Notepad.exe file is stored in the hard disk (i.e. secondary storage)

Step 2: When Notepad is needed to be executed the microprocessor loads it into the memory.

Step 3: Before executing a specific instruction, the microprocessor needs that instruction and associated data to be in the cache memory. Therefore block of adjacent instructions (not just a single instruction) and related data from the main memory will be send into the cache memory.

Step 4: Then the microprocessor get a specific instruction into the registers from the cache memory and execute the instruction.

Step 5: After execution, the results will be send back into the cache memory.

Step 6: When new instructions (which are not in the cache memory) are needed the microprocessor loads them from the memory into the cache.

Step 7: Repeating "Step 6" several times will fill up the cache memory. This is an issue when new data or instructions are needed to be loaded into cache from the main memory. To accommodate more space in cache memory some of the infrequently used data and instructions are sent back to the main memory (this is called *swapping*).

Step 8: When both step 6 and 7 repeat after sometime the main memory may also fill up. Then some of the data and instructions are copied back to the secondary storage (actually set of memory blocks are sent) and this process is called the *paging*.

If any of these paged up memory locations are need again they are loaded back into main memory by paging some other set of memory blocks in to the hard disk.

To further enhance the performance, cache memory is split into 2 levels of cache called; *Level 1* cache (L1 cache) and *Level 2* cache (L2 cache). L1 cache is located in-between L2 cache and registers where as L2 cache is located in-between L1 cache and the main memory. L1 cache is very fast and usually built into the microprocessor. L2 cache is slower than the L1 cache but faster than the main memory. L2 cache is usually included in the motherboard. However in the latest microprocessors such as Pentium IVs L2 cache is also part of the microprocessor chip (also referred as the core).

5.8 Central Processing Unit

CPU is also called the Microprocessor. It is the brain or engine of the computer. The CPU performs arithmetic and logic operations and it controls the other components of the computer as well. CPU is the most expensive single piece of component in a PC. Over the last 35 years, microprocessors have

Table 5.2 – Generations of microprocessors

Generation	Microprocessor(s)
1st generation	Intel 8080/8086/8088
2nd generation	80286
3rd generation	80386 (DX/SX)
4th generation	80486 (SX/DX/DX2/DX4)
5th generation	Pentium, AMD K5
6th generation	Pentium Pro, AMD K6
7th generation	Pentium IV/IV HT/IV D, AMD Athlon/Duran/Opteron

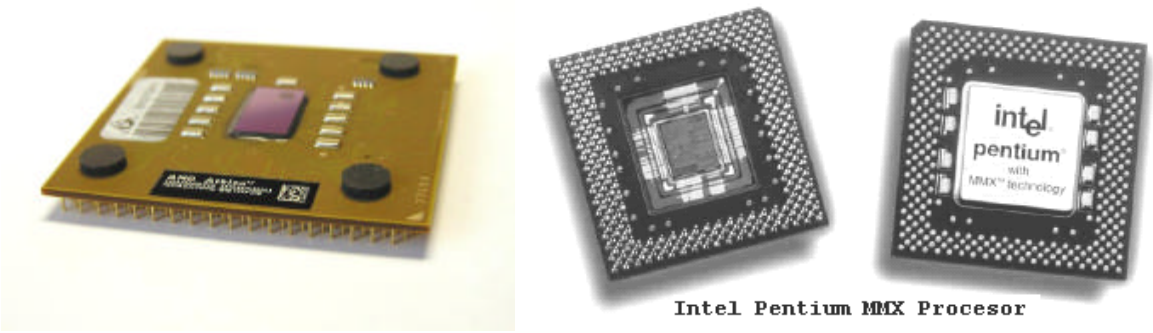


Figure 5.16 – External view of microprocessors

gone through tremendous enhancements and evolved by several generations. The first microprocessor was introduced in the 1970s and currently we are in the 7th generation (refer table 5.2).

5.8.1 Heating and Cooling

With the advancement of microprocessors there is a tremendous increase in speed. High speed is being achieved by having more and more transistors (figure 5.1) inside the CPU core. When the number of transistors is increased, the heat dissipated by the microprocessor also increases. This excessive heat can burn out the microprocessor. To prevent this, the microprocessor needs to be cooled by some external force. In early days with little bit of heat being dissipated heat sink were used as the cooling mechanism. However with the excessive heat generated by faster microprocessors jus having a heat sink was not enough. Therefore cooling fans are being used with the introduction of Pentiums. The size of heat sink and the cooling fan is getting bigger and bigger with each new microprocessor.

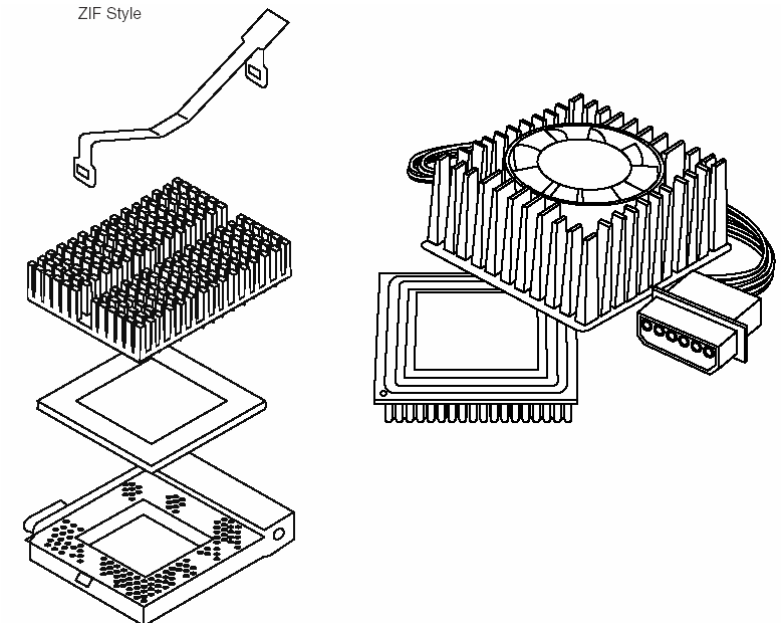


Figure 5.17 – Passive cooling (Left) and Active cooling (Right)

Cooling can be divided into 2 categories namely; *passive cooling* and *active cooling*. Passive cooling uses only a heat sink for heat dissipation. Passive cooling was used in x486 microprocessors and some of the early versions of Pentiums. In modern computers, passive cooling is used to cool various controllers (also called the CPU support chips) such as the Chipset and Video Controllers. Active cooling uses a cooling fan and a heat sink. Heat sink acquires the heat from the microprocessor while cooling fans cools the heat sink by forced airflow. Modern computers use active cooling to cool the microprocessor. In some of the high end computers you may also find water based cooling systems like a car. Proper care should be given to the cooling fan since it will make sure CPU temperature does not rise beyond the nominal temperature (normally this range is around 60-70^{o}C). If the fan fails and you continue to run your computer for one to two minutes it could burn out the microprocessor.

5.8.2 Components of a CPU

As shown in figure 5.18 three major components can be found inside the microprocessor; the Arithmetic and Logic Unit (ALU), the Control unit and a set of Registers.

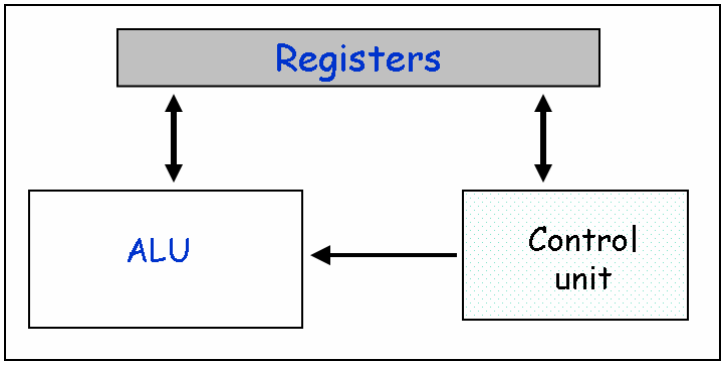


Figure 5.18 – Components of a CPU

Arithmetic and Logic Unit – ALU

This is the data processing unit of the CPU. The arithmetic unit is capable of performing arithmetic operations while the logic unit performs logical operations.

Control Unit

This unit controls the operation of the CPU and rest of the machine.

Registers

Registers are type of memory located inside the CPU that can hold unite of data. This data is useful for both data processing and control functionalities. Registers hold data before for processing and after processing the results are send back to registers. Note the bidirectional arrows that connect the registers with the ALU and control unit in figure 5.18. Since registers are located inside the CPU they can be accessed faster by the ALU and control unit. Several types of CPU registers are used:

- Program Counter (PC)
- Instruction Register (IR)
- Accumulator (A)
- Flag register (F)
- General Purpose Registers (GPR)

Program Counter – PC

The Program Counter (PC) is used to keep track of memory address of the *next instruction to be executed*. It will always point to the next instruction to be executed. When an instruction is fetched[12], always the instruction pointed by the PC is loaded into the CPU. Once the instruction is fetched, the program counter is automatically updated so that it will points to the next instruction (it will always be incremented by 1). The PC can also be incremented or decremented by the programmer.

[12] Fetching is the process of loading an instruction into the CPU. This stage is called the fetch cycle.

Instruction Register – IR

Once an instruction is fetched into the CPU it is stored in the IR for execution. The IR is located closely to the control unit, which decodes the instruction. The control unit decode (understand) the instruction and ask the respective circuit to handle the instruction accordingly. Keeping IR closer to the control unit make this process must faster[13].

Accumulator – A

Results of arithmetic and logical operations always go to the accumulator. Accumulator is connected directly to the output of the ALU. Accumulator is normally indicated by symbol 'A'.

Flag Register – F

Flag Register stores the status of the last operation carried out by the ALU. After an arithmetic or logic operation there can be various states such as; overflow, division by zero, final result is a zero, positive or negative result, results of comparisons, etc. In certain parts of a program before preceding to the next instruction this state information is checked to determine the next operation. Such state information is indicated by setting up various bits in the flag register.

General Purpose Registers – GPRs

These registers are said to be general purpose because they can be used for various tasks. They are used to store intermediate results of the ALU operations. A limited number of GPRs are available inside a CPU and programmers have to use those registers in an effective manner to run even a very complex program. They are normally labelled as register 'B', 'C', 'D', etc (number of GPRs may vary depending on the CPU).

Identify different components and their locations insider the CPU using the diagram given in figure 5.19. Three (3) buses (data, control and address) run all around the CPU. Both control and address buses go out of the CPU while data bus is bidirectional. Also, note that IR is located near the control unit. Flag registers are set by the main ALU but they are accessible only to the control unit.

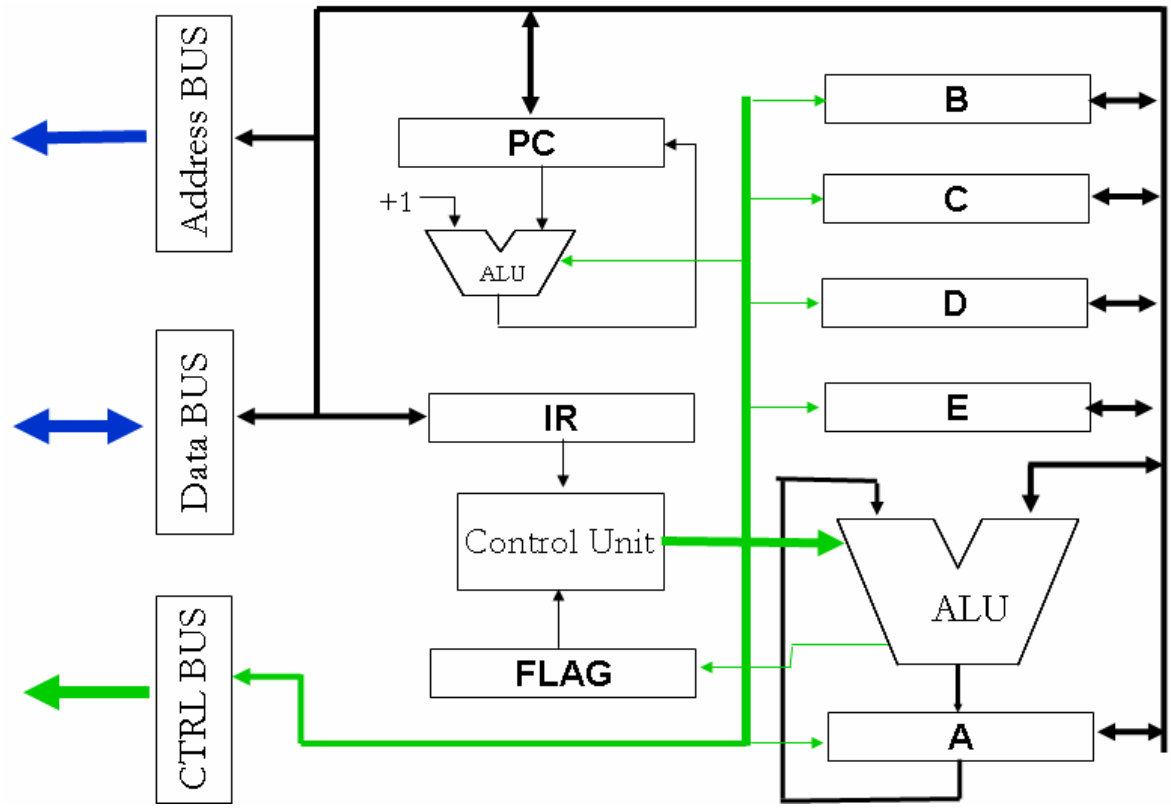


Figure 5.19 – Internal Structure of the CPU

[13] Inside a CPU having more than 100 million transistors in less than a square inch several micro meters (μm) is a long distance.

Notice the secondary ALU after the PC. This ALU increment the PC after fetching a new instruction. Accumulator ('A') is just after the main ALU and all the results of the ALU operations are automatically send to the accumulator. The main ALU accepts a maximum of 2 inputs (*operands*). However one input should come through the accumulator.

5.8.3 Execution of a Program

Consider the following program segment, which is written in Assembly language[14].

```
100: Load A,10        ; A ← 10
101: Load B,15        ; B ← 15
102: Add A,B          ; A ← A+B
103: STORE A,[20]     ; [20] ←A
```

The above program stores two values (0x10 and 0x15) in registers A and B and add them together. The final answer is stored in memory address 0x20. Let us go through the code line by line. An Assembly language statement can be divided into several sections.

<Identifier>	Operation	<Operand(s)>	<;Comment>
100:	Load	A, 10	; A ← 10

Identifier – is an optional element, which allows pointing to a line in the code or memory location.
Operation – indicate the specific assembly instruction (i.e. the action to be performed).
Operand – provides data for the operation to act upon (for certain instructions this is optional).
Comment – can be used by the programmer to keep notes within the program (optional).

Consider the first line of code in the above program. `100: Load A,10 ; A ← 10`

In here the `100:` says store the first instruction at memory address 0x100. `Load A,10` says load the value 0x10 to the accumulator. In this case `Load` is the operation while `A` and `0x10` are the operands. The comment `; A ← 10` indicates move value 0x10 to register A. The direction of data flow is indicated by the direction of the arrow.

Figure 5.20 indicates how these instructions are stored in side the memory. Normally, memory is divided into two portions called the *data* memory and *program* (or instruction) memory. Since the program has not executed the memory location pointed by memory address 0x20 is zero.

Let us observe how various registers are changed when we run the above program. Each instruction in the above given program is executed in a separate instruction cycle. To follow the execution of an instruction cycle, the values of the registers have to be observed in each step of the instruction cycle.

Figure 5.21 indicates status of various registers just before starting the first fetch cycle. Since the program has not executed yet, the PC will point to the next memory address where the instruction to be executed. Then in the next stage the CPU gets the first instruction from memory address 0x100 and stores it in IR register (this is the end of first fetch cycle).

Figure 5.22 indicates status of various registers just after the first fetch cycle. Now the Assembly instruction `Load A,10` is loaded into the instruction register. Meanwhile PC automatically gets incremented and starts pointing to the memory address of the next instruction to be executed (0x101).

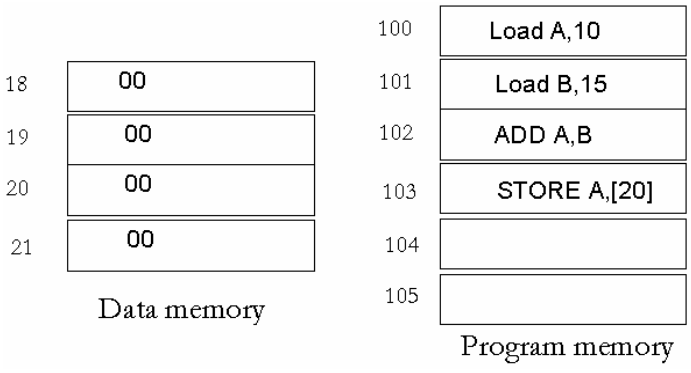


Figure 5.20 – Content of memory when the program is loaded

[14] Assembly language is the lowest level of programming that a programmer can do.

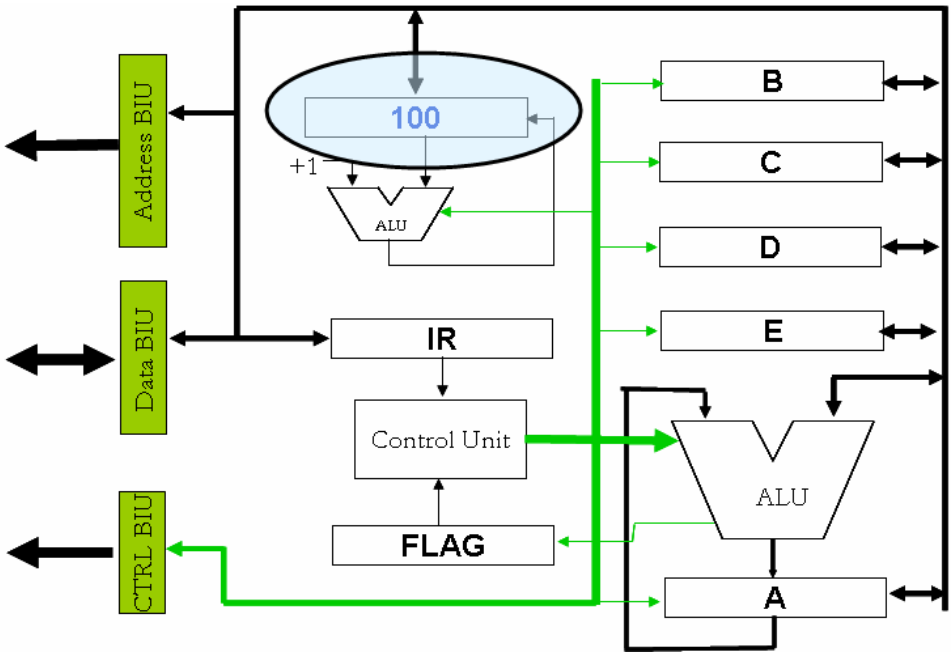


Figure 5.21 – Before execution of the 1st fetch cycle

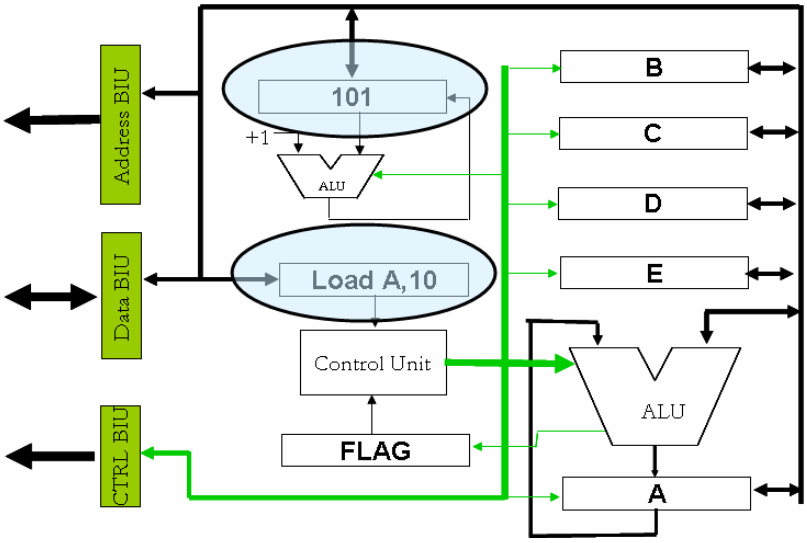


Figure 5.22 – After the 1st fetch cycle

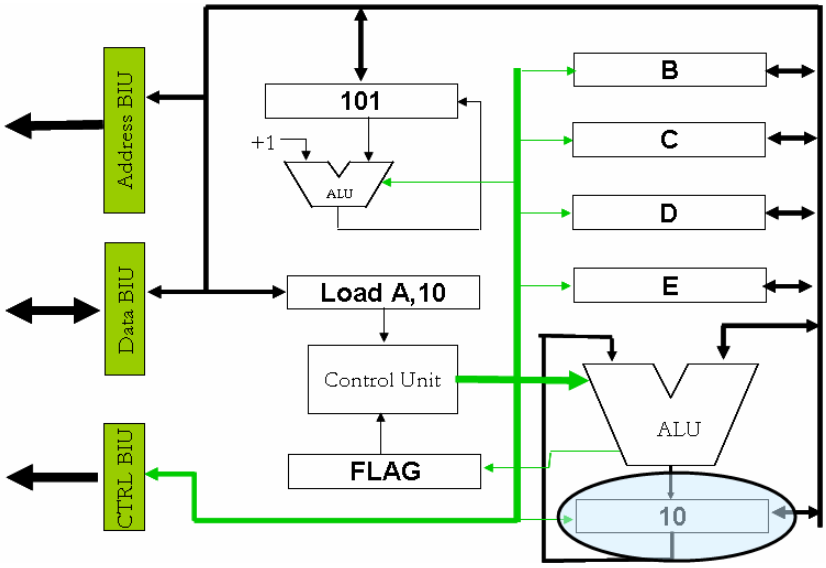


Figure 5.23 – After the 1st instruction cycle

Then the control unit decodes the instruction and understand it as a Load operation. Then after the first instruction cycle (also referred as the *execution cycle*) it will set the accumulator (register A) as 0x10 (figure 5.23). Now the CPU has fully executed the first Assembly language instruction (`Load A,10`).

Next the second assembly instruction (`Load B,15`) which is stored in memory address pointed by PC (0x101) is loaded into the CPU. Figure 5.24 indicates the status of registers after the second fetch cycle. At the end of the second fetch cycle the PC is automatically incremented and starts pointing to the next memory address (0x102). Now IR will hold the second Assembly instruction. During the second instruction cycle the control unit identify this as another Load instruction and at the end of the cycle integer 0x15 is stored in general purpose register B (figure 5.25).

Up to now, two instructions have being executed. Then after the third fetch cycle the next instruction pointed by PC (which is in memory address 0x102) will be loaded into the IR (see figure 5.26). At end of this stage, the PC will point to the next memory location, which is 0x103.

Then at the third instruction cycle, the control unit understands this is an addition operation so it calls the addition circuit to add the two operands, which are stored in registers A and B. After the execution of the third instruction cycle the result of the add operation (0x10+0x15=0x25) will be saved in the accumulator (A) (see figure 5.27).

In the fourth fetch cycle instruction pointed by the PC will be loaded into the IR. Then at the end of the fourth instruction cycle the value stored in the accumulator will be saved in the memory location pointed by memory address 0x20.

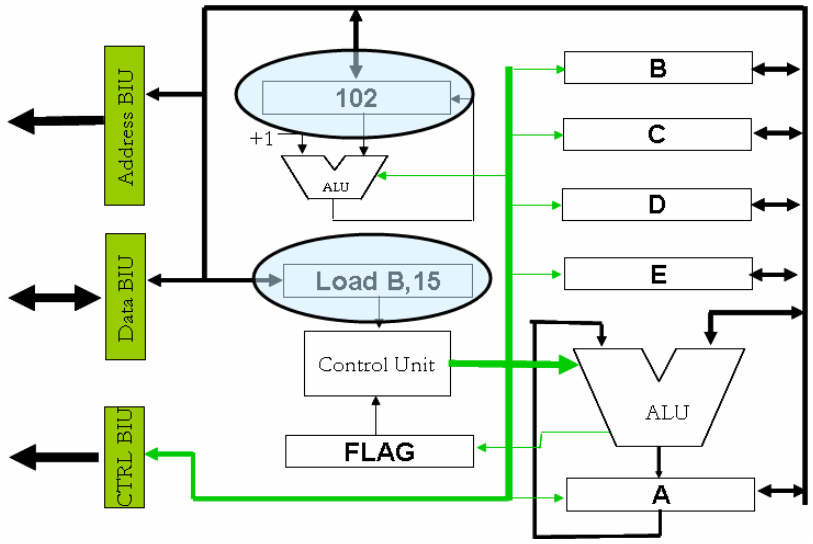


Figure 5.24 – After the 2nd fetch cycle

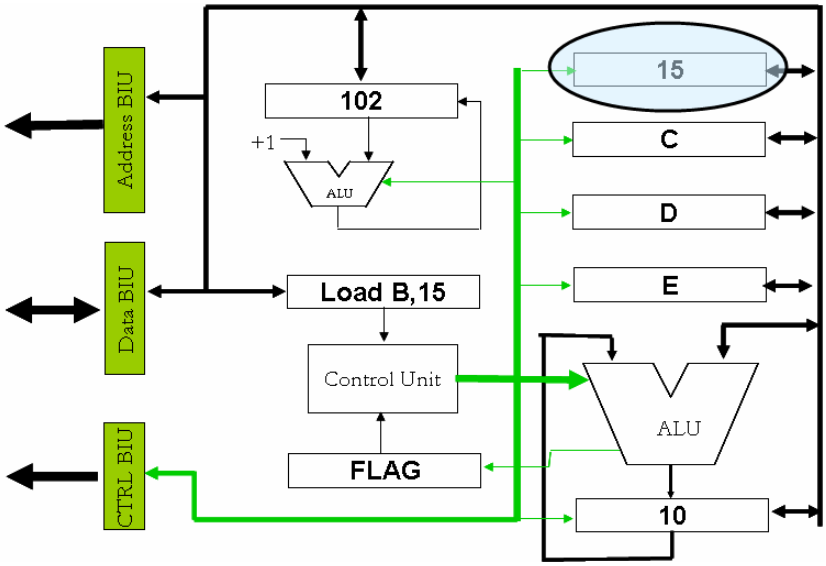


Figure 5.25 – After the 2nd instruction cycle

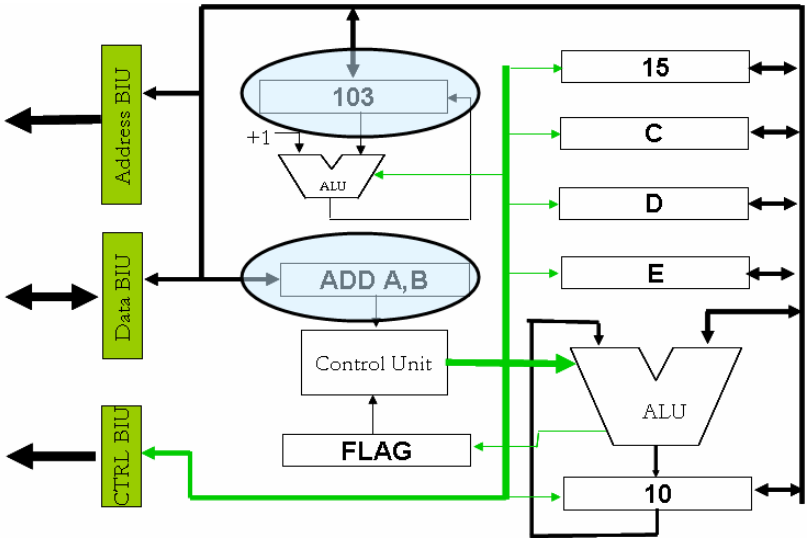


Figure 5.26 – After the 3rd fetch cycle

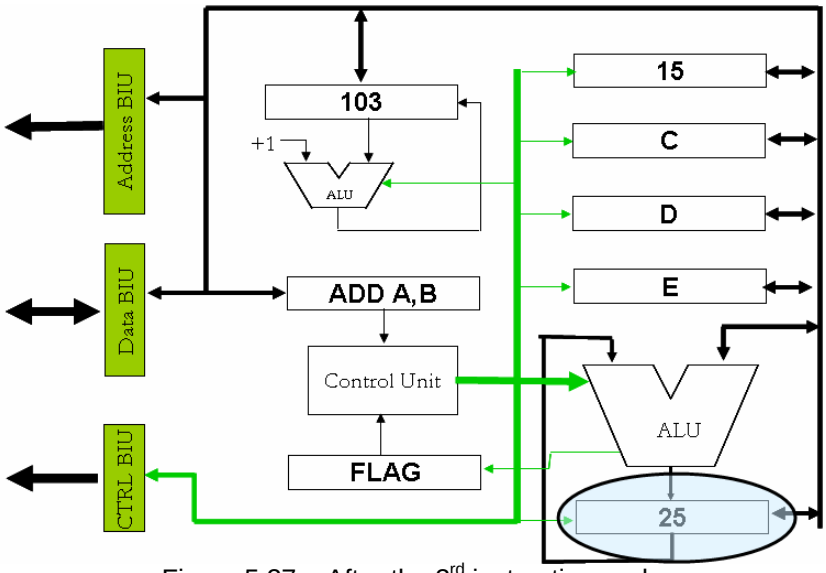


Figure 5.27 – After the 3rd instruction cycle

5.8.4 Enhancing The CPU Performance

With the advancement of electronics microprocessors became smaller and faster. Faster CPUs were developed by having faster transistors. When the industry demands for faster and faster CPUs manufacturers has to use faster and faster transistors. However, over the last several years speed of transistors has been saturated. Therefore having faster transistors is no longer the solution to achieve faster CPUs. Then manufactures came up with several optimisations to improve the performance of CPUs. Some of the approaches are:

- Instruction pre-fetching
- Instruction Pipelining
- Hyper Threading (HT)
- Dual Core

Instruction Pre-fetching

In section 5.8.3 when we studied how a program is executed, we realised there are 2 cycles called the; fetch cycle and the instruction cycle (also referred as execution cycle). When an instruction is executed, it first goes through the fetch cycle then through the execution cycle. Then only next instruction is fetched into the instruction register. During each cycle, either the ALU or the fetching circuit (mainly the IR and PC registers) is not utilized.

In figure 5.28 instructions are first executed in the conventional way (upper half of the figure) and later they are executed using pre-fetching (lower half of the figure). In conventional execution, fetch

cycle of the 2nd instruction needs to wait until the execution cycle of the 1st instruction is fully complete. Similarly, 3rd instruction needs to wait until the 2nd instruction is fully complete.

In the second approach when one instruction is in the execution stage, the next instruction is fetched into the CPU. The fetch cycle of the 2nd instruction starts while the 1st instruction is in the execution cycle. However, the execution cycle of the 2nd instruction will not start until the execution of the 1st instruction is complete. Similarly, 3rd instruction starts its fetch cycle while 2nd instruction is in its execution cycle. Also note that the 3rd instruction will not start its execution cycle until the 2nd instruction if fully executed.

At the end of the execution total time took to execute 3 instructions using the instruction pre-fetching is lesser than the conventional approach. Therefore by using instruction pre-fetching the CPU can do more work within a given time period (i.e. this approach improves overall CPU performance).

Instruction Pipelining

The term pipelining is used because this process is derived from an industrial assembly line where output of one-step is fed to the next step as an input. In an assembly line of a car production company, multiple cars are assembled at the same time. This is achieved by dividing the car assembling process into multiple sub-stages and carrying out each sub stage at the same time. Someone observing at the end of the assembly line will realise that multiple cars are being manufactured within one hour, whereas to assemble a single car, it may take several hours. If the production line can be divided into multiple stages, more cars can be in the production line. This approach increases the number of cars assembled per day, but do not reduces the time spent on a single car.

The same concept is used inside the CPU. Pipelining divides the instruction cycle into a series of sub-operations and a separate segment of the CPU is dedicated to one sub-operation. Since there are many stages, multiple instructions can be in the instruction cycle at the same time. Each of these instructions should be in a different stage or a sub-operation. To achieve pipelining more electronic circuits are needed inside the CPU core.

This is an extension of the idea of instruction pre-fetching. This approach will increase the throughput of the microprocessor[15]. Pipelining will not speedup a single instruction, it will speed up a set of instructions.

Suppose there are 2 CPUs A and B. A has 5 stage (i.e. number of sub-operations) pipeline while B has a 9 stage pipeline. Since A has 5 stages 5 instructions can be in the instruction cycle at the same time. In CPU B, 9 instructions can be in the instruction cycle. Therefore CPU B executes more instructions compared to CPU A in a given time interval (i.e. CPU B has a higher throughput so it has better performance than A).

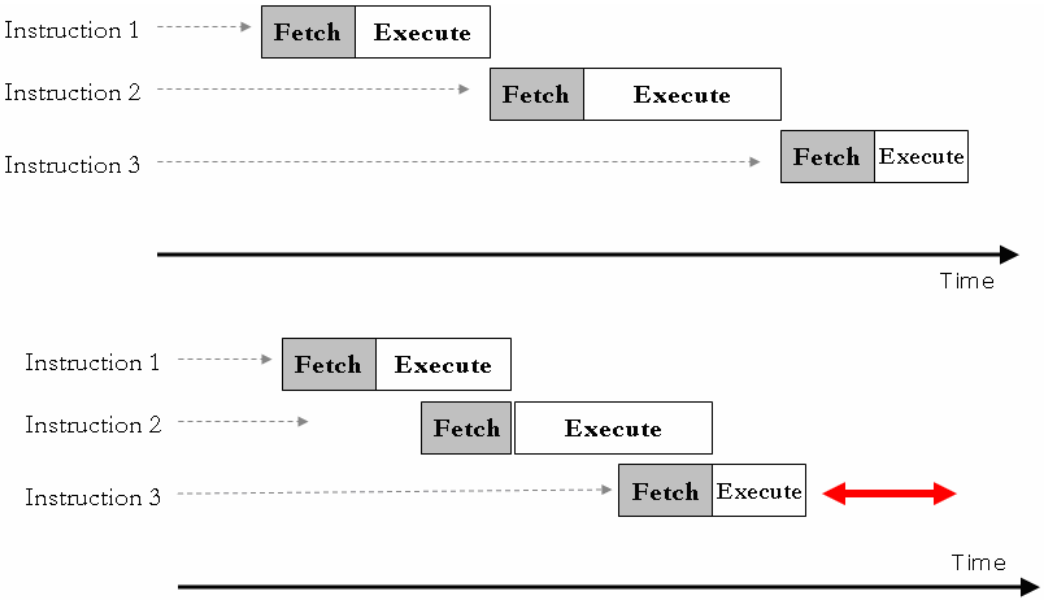


Figure 5.28 – Instruction pre-fetching

[15] Throughput is the number of outputs per unit time

Hyper Threading

Hyper Threading (HT) is another approach that improves the performance of a CPU. HT was first introduced by Intel with their Pentium IV 2.8GHz processors. It allows two different resources of the CPU to be used at the same time. For an example when one thread[16] (instruction) makes use of the integer unit of the ALU another thread (instruction) can makes use of the floating point unit. However the two threads cannot use the same resource at the same time; therefore HT does not always allow two threads (instructions) to execute at the same time (i.e. if two threads require the same CPU resource then they have to execute one after the other). When hyper threading is possible the operating systems will feel that it is running on top of a two CPU computer.

Hyper Threading is achieved by having a mix of shared, replicated and partitioned chip resources, such as registers, arithmetic units and cache memory. However in order to make use of HT the computer should satisfy following four conditions:

a) The CPU should supports HT technology
b) HT technology enabled chipset
c) HT technology enabled BIOS
d) HT technology enabled/optimized operating system

Dual Core

Dual core is the latest approach towards enhanced performance. The idea of dual core was introduced with the IBM Power4 microprocessor however AMD is the one who actually brought it to the consumer market. As its name implies it combines two independent microprocessors and their respective caches onto a single Silicon chip (figure 5.29). Since it is two different microprocessors not just set of replicated components performance gained by dual core is higher than HT.

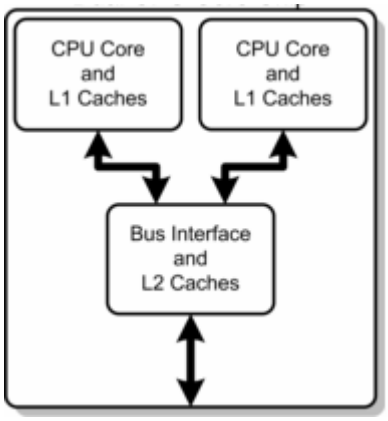


Figure 5.29 – Components of a Dual Core chip

5.8.5 Improving Overall System Performance

Section 5.8.4 introduced some of the approaches that can be used to improve the performance of the CPU. What really matters is how we can improve the performance of the overall computer rather than just the CPU. Having a faster CPU will help the overall performance, but it is not the only factor that will decide the performance of the entire machine. Consider 2 computers; 'A' and 'B' with 1GHz and 2GHz CPUs. Other than the CPU, both 'A' and 'B' have the same hardware configuration. Computer 'B' is faster but it does not mean that it is twice as fast as computer 'A'.

The overall performance of a computer can be improved by:

- using a high performance CPU
- having an effective memory hierarchy
- using buses with different speeds
- using CPU support chips

[16] Thread is an operating system construct that actually executes inside the CPU. In simple terms it can be considered as the execution part of a program.

Performance is mostly degraded due to slowness in memory. Therefore CPU performance can be highly improved by having an effective memory hierarchy. The cache memory plays a major role in improving performance of the modern CPU.

Some CPU's such as Intel Pentiums use different speeds for its internal operations and for communication with memory. It uses a high-speed bus for its internal communication while a slower bus is used to communicate with main memory. This is analogous to having a conventional road and a super highway. Vehicles that can go faster can use the highway while slower ones have to use the conventional road. In such an arrangement, a faster vehicle is not slowed because of a slower one. In modern motherboards Front Side Bus (FSB) is used to communicate with the slower memory. Usually the speed of FSB ranges from 133MHz to 833MHz.

5.8.6 CPU Support Chips

Von Neumann's Architecture

Von Neumann, a consultant who worked in the ENIAC project, published a paper in 1945, which described all the parts of a stored-program computer:

- A memory, containing both data and instructions
- A calculating unit, capable of performing both arithmetic and logical operations on the data
- A control unit, which could interpret an instruction retrieved from the memory and select alternative courses of action based on the results of previous operations

The computer structure resulting from these criteria is popularly known as a Von Neumann Machine[17]. The Von-Neumann Architecture is a CPU centric system where:

- Each operation is carried out only by the CPU
- Every movement of data must be made via the CPU
- Memory is the only "Direct Access" storage device for the CPU
- Only a single operation is carried out by the CPU at any time

Von Neumann's architecture is a simple and implementable proposal. However, for some operations such as; disk access, giving attention to peripheral devices and for periodic or time critical operations this is not the best mechanism.

If the CPU needs to access a file, it should be first loaded into the memory. Therefore the CPU has to instruct the hard disk that it needs a file. Then the file is divided in blocks and each block is sent from the hard disk into the CPU. Then the CPU has to store it in the memory. Then only it can access the file from the memory. In this case, all the communication is done through the CPU so it has to stop all other work and help the communication. The CPU time is precious thus, it should not be wasted unnecessarily. Instead of this approach, if another controller can get the file from the hard disk and put it in memory on behalf of the CPU it will save valuable CPU time. Then the CPU can access the file from memory. CPU support chips are used to carryout such tasks on behalf of the CPU.

CPU support chips are used to improve the overall performance of a system. In this approach, the CPU is like the manager of an organization while CPU support chips are like supervisors under him. Manager will give orders to supervisors and they will ask the workers (in a computer workers are components such as; hard disk, floppy disk, keyboard, communication ports, etc.) to carryout the actual work.

If this approach is compared with the conventional Von Neumann's architecture, it is like having an organization with a single employee, where that person is the manager as well as the only worker. If a file is required, the manager has to go through all the files and find out the correct one. This approach wastes manager's time that could have been used for something more useful. In an organization with a well-established management hierarchy, the manager can request a clerk to find the file on behalf of him/her self. When the file is available, the manager can do his/her work and when it is finished, the clerk can put the file back again. These approaches allow the manager to concentrate on matters that are more important. Some of the CPU support chips are:

[17] Virtually all the computers produced since were Von Neumann machines.

- Direct Memory Access (DMA) controllers
- Interrupt Controllers
- Real-Time Clock (RTC)
- Other devices
    - Disk controllers
    - Communication controllers
    - Display controllers

Direct Memory Access (DMA) Controller

The DMA controller provides a way of bypassing the CPU when transferring data between memory and Input/Output (IO) devices. This is in contrast to Von Neumann's architecture where everything has to go through the CPU (figure 5.30-A). The DMA controller resides between the CPU and memory. When the CPU needs something to be in memory, it informs the DMA controller. Then the DMA controller accesses the particular resource on behalf of the CPU and loads it into the memory (figure 5.30-B).

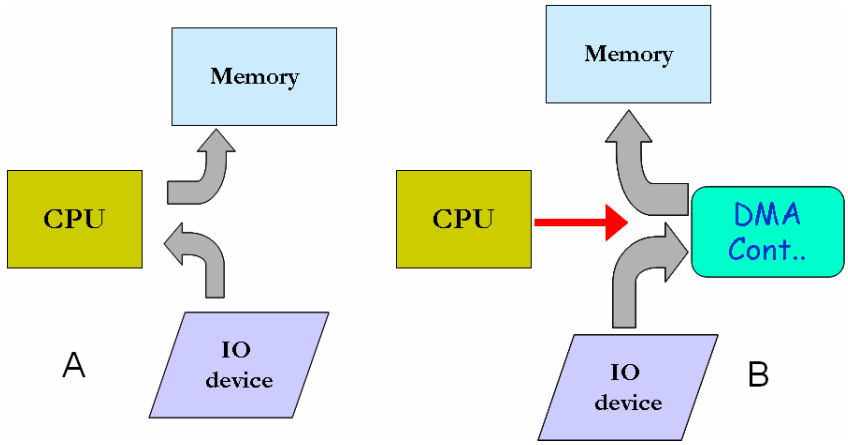


Figure 5.30 – Communication with and without DMA controller

Disk Controllers

Some of the common disk controllers are; Floppy Disk Controller (FDC) and ATA Controller (ATAC) for hard disks. When reading/writing to/from a disk the CPU will create a special memory area (called a *buffer*) containing the sector address and the data to be written or read. Then the CPU informs the FDC (or ATAC) about the location of the buffer. The disk controller then transfers the content of the buffer directly from memory to the disk sector.

Real Time Clocks - RTC

The RTC is used to keep track of time of the day. It is usually backedup by an extra power source (generally by a lithium battery). Additionally it is used to store some of the configuration information such as CMOS[18] setup memory.

5.9 Display Controllers

Display controllers are used to generate images and text that you see on the displaying device on behalf of the CPU. Video controllers are used display the image that you see on monitor. Display controllers are either available as a separate expansion card or integrated into the motherboard (figure 5.31). Display commands are given by the CPU and they are carried out by the display controller. Display controller generates the actual image in its memory; called the *Refresh Buffer*, with 1's and 0's (figure 5.32). Then it is passed through to the video controller to generate the actual image.

There are several video standards:

- VGA – Video Graphics Array. Graphics are supported at a minimum resolution of 320x240 pixels in 256 colurs and also for 640x480 in 16 colours.

[18] CMOS (Complementary Metal-Oxide Semiconductor) memory is used to store system configuration data.

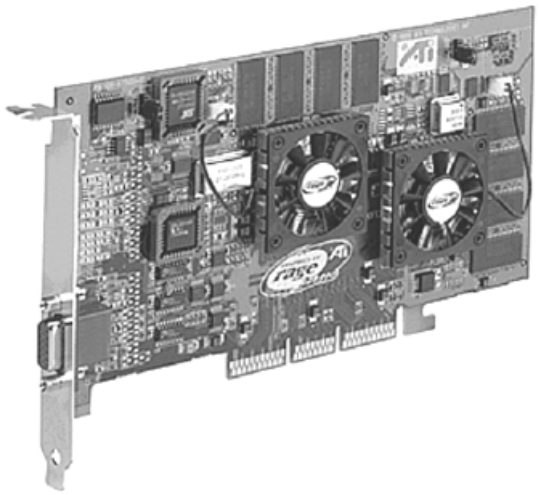


Figure 5.31 – A video card

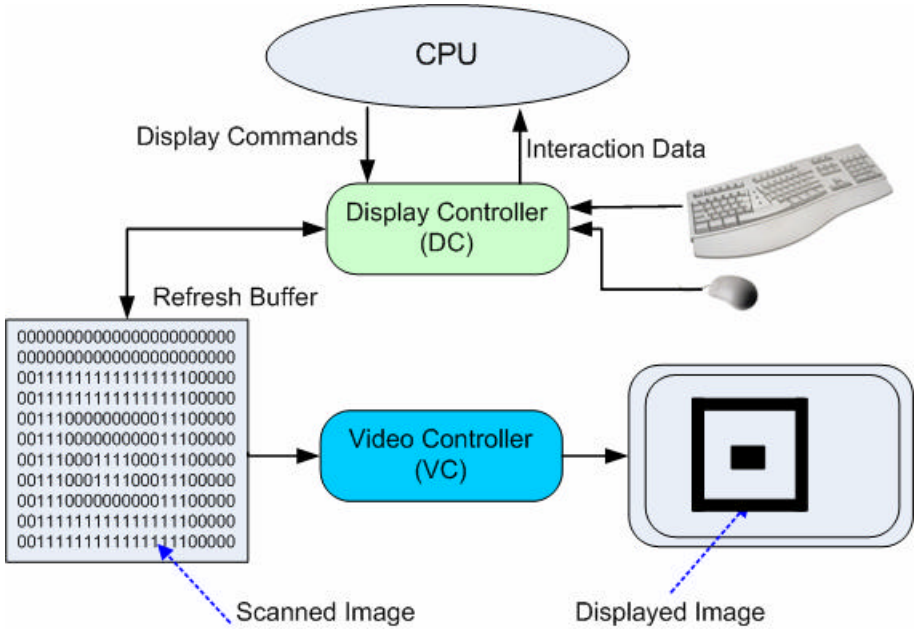


Figure 5.32 – Display and Video controllers

- SVGA – Super VGA. Supports resolution up to 800x600.
- VESA SVGA - Video Electronic Standards Association SVGA. Was developed to standardise SVGA. Also includes a video standard for connecting high-speed adaptors directly to the processor bus.

Video cards are classified based on their video processor and video memory. High-speed video processors can render (i.e. colour) 2D and 3D images much faster (suitable for 3D animations and gaming). High capacity video memory produces better quality images without distortion or flicker.

5.10 Video Display Unit

The video display unit is also called the monitor or the display. It is the device used for viewing images generated by the video controller. Displaying devices are characterised by the displaying mechanism. There are three major mechanisms namely; Cathode Ray Tube (CRT), Thin Film Transistor (TFT) and Liquid Crystal Display (LCD)

5.10.1 Cathode Ray Tube

The CRT uses the concept of a cathode tube. It uses the raster-scan technology, in which a beam of electrons is sent from back of the tube to the screen (figure 5.33). The screen is a “phosper” coated surface and it emits light when the electrons hit its surface.

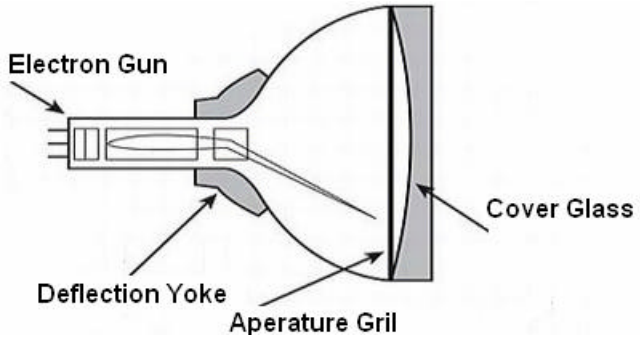


Figure 5.33 – Cross section of a CRT monitor

CRTs are relatively cost effective compared to other displaying devices. They can represent colours that are more accurate and has a larger viewing angle. However, they consume considerable amount of electricity and occupy lot of desk space.

Because of the raster scan technology and the generated light having a very small lifetime, image produced by CRTs needs to be refreshed periodically (indicated by the refresh rate). This is called the *raster scan* technology. A CRT in a television generates 25 frames/second while a CRT used in a computer generates around 60-70 frames/second. You may have seen computer monitors through the television. Have you ever noticed that the computer display goes up and down when you look at them through the TV? This happens since the two frame rates do not coincide with each other. The human eye scans an object at a rate of 10Hz (i.e. 10 frames/second). Due to the difference in frame rates, human eye also faces the same problem (although we do not realise it). Therefore in the long run CRTs are not really good for our eyes.

5.10.2 Thin Film Transistor

TFTs use set of transistors which emit light instead of a cathode ray tube. Each pixel[19] is represented by a transistor. In a colour TFT each pixel is represented by 3 transistors (one each for three fundamental colours; Red, Green and Blue) which is placed on top of another. A matrix of transistors forms a TFT screen.

TFTs consume less power and utilise less desk space. However, they cost a lot compared to CRTs and does not support high quality graphics (hopefully it will improve within next few years). Viewing angle of a TFT is also an issue. These transistors will retain their light, if power is continuously supplied. Therefore they do not need any refreshing. Because of this reasons TFTs are much more suitable for the human eye.

TFTs are mainly used in portable computers, PDAs, mobile phones and some video games.

5.10.3 Liquid Crystal Display

LCD displays make use of liquefied crystal. They do not produce light but they control the reflection of light by changing the direction of liquefied crystals by applying a small electric field (figure 5.34).

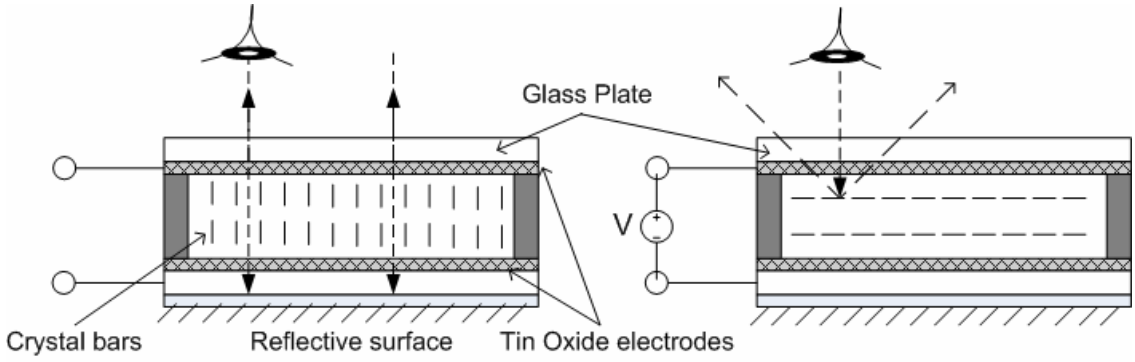


Figure 5.34 – The principle of liquid crystal

[19] A term meaning picture element. Any of the tiny elements that form a picture on a video display screen is called a pixel. In other words it can be considered as the fundamental element of a picture.

LCDs are cost effective and consume very little power. Most LCD screens are 2 colour but multicolour LCD screens are also available. LCD displays are mostly used in *Consumer Electronic* (CE) products such as; digital watches, calculators, mobile phones, etc.

Displays are categorized based on their screen size (i.e. the length of the diagonal), which is normally given in inches. They are also categorised based on the colour depth and resolution. Resolution is a measure of the number of horizontal and vertical pixels. It determines the amount of information that appears on the screen. Normally, resolution is given as a multiplication of horizontal and vertical pixels. A modern computer display should support at least a resolution of 1024x768. This resolution is supported only by SVGA or other modern video standards such as VESA-SVGA. The VGA standard only supports a maximum resolution of 640x480 pixels.

5.11 Secondary Storage

Secondary storage devices are needed in addition to the volatile memory as a permanent and high capacity storage solution. Cost per a Megabyte of secondary storage is lower than the main memory. Different secondary storage devices such as; hard disks, floppy disks, CD-ROMs, DVD-ROMs and ZIP disks are used in personal computer systems.

These storage solutions can be divided into two broader classes namely; *Magnetic Storage* and *Optical storage*. A combination of magnetic and optical technology called *floptical* technology is also available.

5.11.1 Hard Disk Drive

Hard disks are one form of magnetic storage. Hard disk is the main secondary storage device used in a computer. Its operation is identical to a conventional radio cassette tape. However, hard disk uses a disk coated with magnetic medium rather than a plastic tape. Hard disks contain a rigid disk shaped *platter*, which is constructed using glass or Aluminium. A magnetic medium is applied on this platter and read/write heads are used to read and write data to and from each platter. Heads are connected to the head arm and it is controlled by the head actuator (figure 5.35).

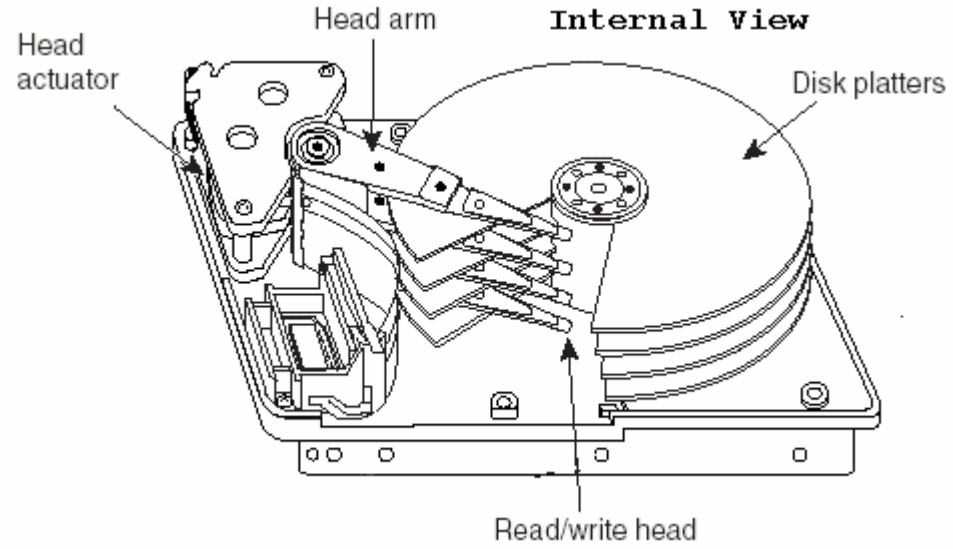


Figure 5.35 – Internal view of a hard disk

Compared to most other secondary storage devices hard disk can store large capacity of data and currently have capacities in tens of Gigabytes. It also has a much higher data transfer speed. Higher capacity is achieved by having multiple platters in the same hard disk and storing data on both sides of the platter.

Hard disks are categorized based on their capacity, controller and platter rotation speed. Common capacities include 40, 60, 80 and 120GB. There are several hard disk controllers such as IDE – Integrated Device Electronics, SCSI – Small Computer System Interface and Serial ATA – Serial AT Attachment Interface. Common rotation speeds are 3600, 5400 and 7200 RPM. Platters are kept in a dust free environment inside the hard disk casing. When the disk is spinning the read/write heads

move very close to the surfaces of the platters at a considerable speed. Therefore even a tiny dust particle trying to pass this gap could damage the head, platter or both.

A *track* is a single ring of data on one side of a platter (figure 5.36). A disk track is too large to manage data effectively as a single storage unit. Some disk tracks can store more than 100 KB of data which is very ineffective when managing small files. Therefore tracks are divided into several fixed size divisions called *sectors*. These sectors represent arc shaped pieces of the track (figure 5.36). In a typical hard disk there are more than 900 sectors in a single track and typical size of a sector is 512 bytes. The set of tracks on a disk that are on each side of all the platters in a stack and are at the same distance from the centre of the disk is called a *cylinder* (figure 5.36).

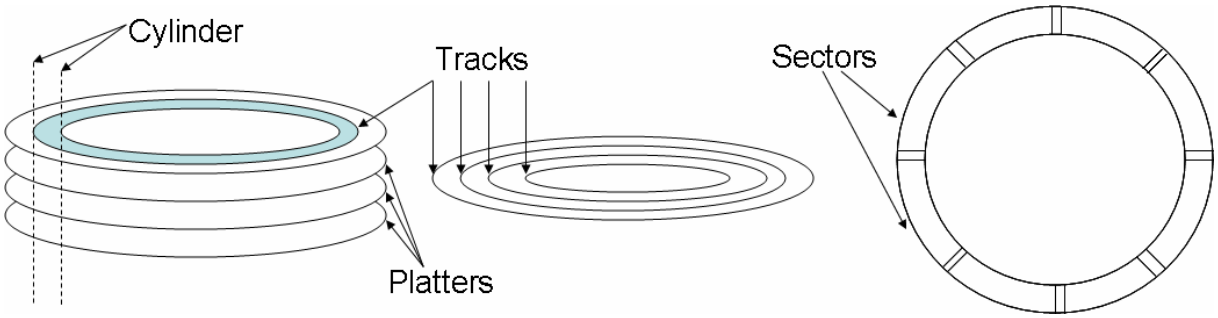


Figure 5.36 – Tracks, sectors and cylinders

Most commodity hard disks have only a single platter therefore both sides of the platter are used to store data in order to gain higher capacity. Other hard disks use a stack of platters. In such cases top and bottom most surfaces of the platters are not used for data storage. For an example consider a hard disk with 4 platters (figure 5.37). It uses only 6 surfaces to store data out of the 8 available surfaces therefore such a hard disk needs only 6 read/write heads.

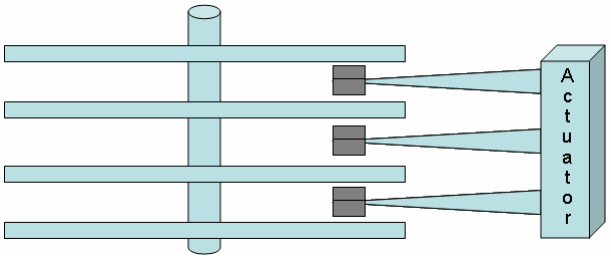


Figure 5.37 – A four platter hard disk drive

5.11.2 Floppy Disk Drive

Floppy disk is a removable disk which has a flexible magnetic medium that is enclosed in a semi rigid or rigid plastic case. The principle behind the floppy disk is identical to the hard disk. In later 1960's IBM developed the first floppy disk drive. Early floppies were having a diameter of 8" and its capacity was limited to 300KB. Later 5¼" *minifloppy* was introduced which was having capacity of either 720KB or 1.2MB. However currently we use the 3½" High Density (HD) floppy disk which comes in a rigid plastic case (not really as flexible as earlier floppies) and having a capacity of 1.44MB.

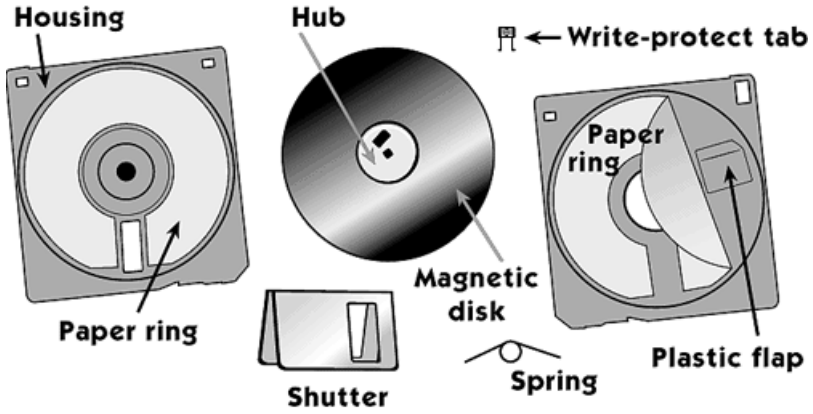


Figure 5.38 – Parts of a floppy disk

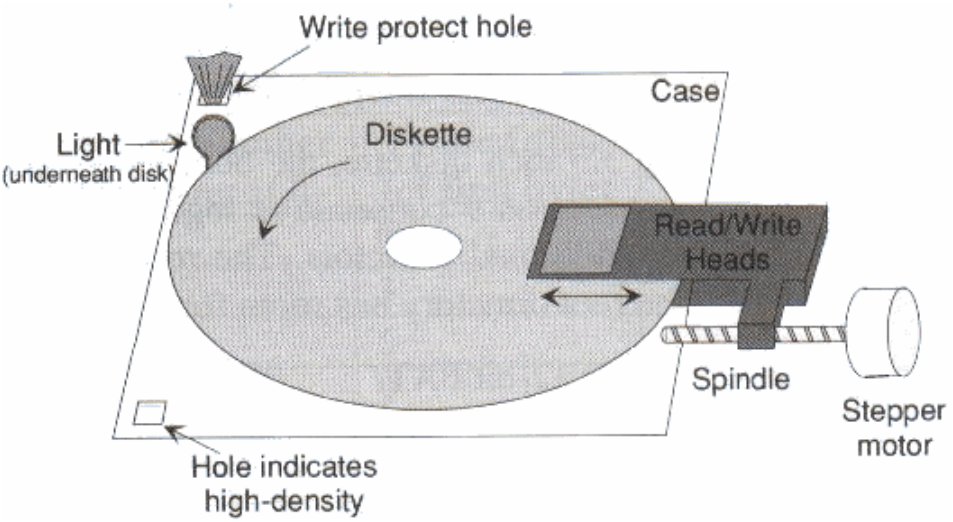


Figure 5.39 – Functionality of a floppy disk

Figure 5.38 shows the parts of a 3½" floppy disk. The magnetic coated, semi rigid, plastic disk is kept in a plastic housing and its two surfaces are covered by two paper rings (for protection). The disk is mounted on a *hub* and a rectangular shaped *cut-out* is used by the disk driver to firmly grab the disk while rotating. The read/write head access the disk through a small opening called the *flap*. The *spring* loaded *shutter* is used to cover up the flap so that prevents any damages to the disk by dust particles. The shutter will open up only when the disk is inside the disk driver and when it is ejected the spring will automatically close the shutter. A *write protect tab* is used to prevent the disk been overwritten and it either opens or closes the *write protect hole* (figure 5.39). If the write protect hole is closed the disk can not be overwritten (then light cannot penetrate through the hole) and if the hole is open (light can penetrate through the hole) it can be overwritten. The head actuator mechanism is slightly different to the mechanism in a hard disk. In a hard disk the head arm moves laterally from centre of the disk towards to the edge while in a floppy disk the movement is horizontal (figure 5.39). The read/write head is mounted on a *spindle* and the spindle is controlled by a stepper motor. The *high density hole* is only available in High Density floppy disks which have a higher data density.

5.11.3 Optical Storage

Optical storage devices make use of light instead of magnetism. There are different forms of optical storage such as; CD-ROM (Compact Disk – Read Only Memory), CD-R (CD – Recordable), CD-RW (CD-Rewritable), DVD (Digital Versatile/Video Disk), DVD-R, DVD-RW, etc. Most of these storage mechanisms use tiny visible light beams or laser.

A CD is made of polycarbonate wafer, 120mm in diameter and 1.2mm thick, with a 15mm hole in the centre. This wafer base is stamped or moulded with a single physical track in a spiral configuration starting from the inside of the disk and spiralling outwards. If you examine the spiral track under a microscope, you would see that along the track are raised bumps, called *pits*, and flat areas between the pits called *lands* (figure 5.40).

The light beam used to read the disk would pass through the clear plastic, so that the stamped surface is coated with a reflective layer of aluminium to make it reflective. Then the aluminium is coated with a thin protective layer of acrylic lacquer and finally a label or printing is added. Data recorded on the

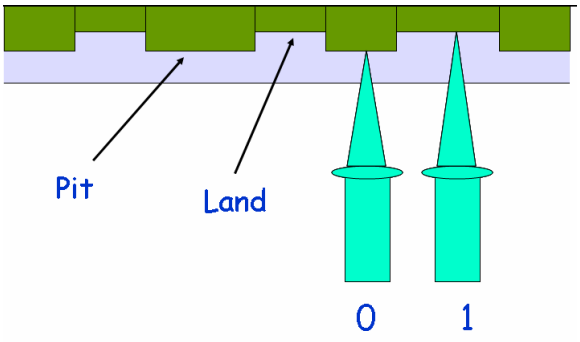


Figure 5.40 – Geometry of a compact disk

CD is read based on the reflection of light from pits and lands. However, if the aluminium layer is damaged the light passes through the CD and will not reflect back resulting data loss. Therefore special care must be given to protect both sides of the CD.

The pit's height above a land is specially critical as it relates to the wavelength (λ) of the light beam used to read the disk. The pit height is exactly ¼λ above the land. Therefore a light beam striking the land travels ½λ (¼λ + ¼ λ) further than a light beam striking the top of the pit. This means a light beam reflected from a pit is ½λ out-of-phase with the rest of the light being reflected from the disk. The out-of-phase waves cancel each other out. Therefore a light beam hitting a pit will not return back to the light sensor and it will appear as dark spot. A dark light spot represents logical '0' where a visible light spot represents logical '1'.

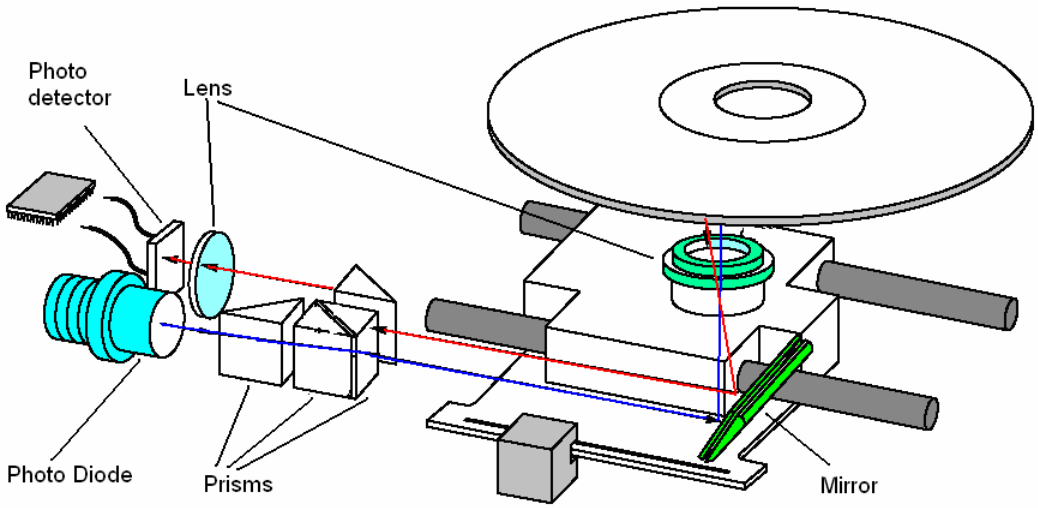


Figure 5.41 – Components of a CD-ROM drive

Figure 5.41 shows various components of a CD-ROM drive. The light beam is generated by a light source (photo diode) and it is directed through several prisms. Then the reflecting mirror rotates the light beam by 90°. The servomotor, positions the beam onto the correct track on the CD by moving the reflective mirror. Then the reflected light from the surface of the CD is sent out through a focusing lens to the same mirror. Mirror rotates the beam by 90° and it is send through a beam splitter (set of prisms). The beam splitter directs the returning light towards another focusing lens. Then the directed light is detected by a photo detector (phototransistor) and it will invert the light into set of electrical impulses. These electrical impulses will indicate whether each bit is a '1' or a '0'.

5.12 Input Devices

Input devices are used to provide inputs to the computer. The keyboard and the mouse are the two most common used input devices. There are several other input devices such as track ball, touch pad, digitiser, joystick, barcode readers, optical character recognition, magnetic ink character recognition, microphone, etc. These devices can be broadly classified as keyboard entry, pointing devices, document readers and data capture devices.

5.12.1 Keyboard Entry

Keyboard is the mostly heavily used input device. All form of computers (except PDA) has a physical keyboard for entering text and commands. A computer keyboard is similar to a typewriter keyboard but has few more special purpose keys like; CTRL, ALT, Windows and function keys. By using a combination of keys the keyboard can produce all the characters, digits and symbols in character codes such as ASCII. A typical Microsoft Windows compatible keyboard has 107 keys which include characters, numbers, punctuation marks, symbols, numeric keypad and set of special purpose keys.

All physical keyboards use a bank of push buttons whose individual states can be detected by the keyboard controller. In devices such as PDA uses will find a virtual keyboard which is displayed on a touch sensitive screen. When the user wants to press a key he/she has to touch the displayed character on the screen and it will be detected by *firmware* (explained in chapter 9).

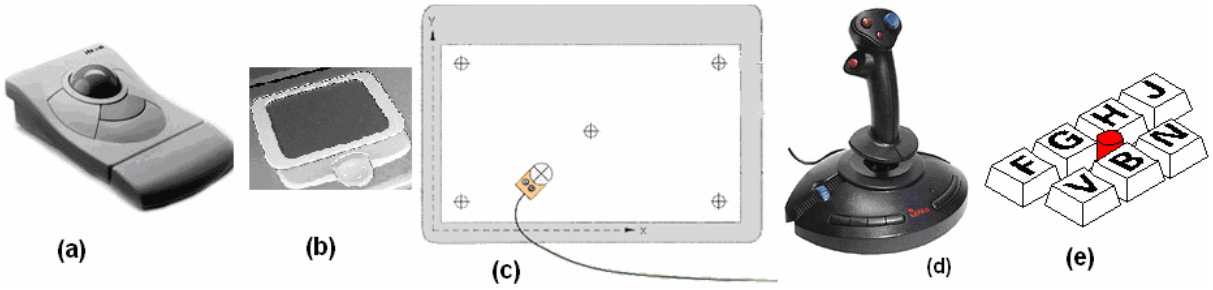
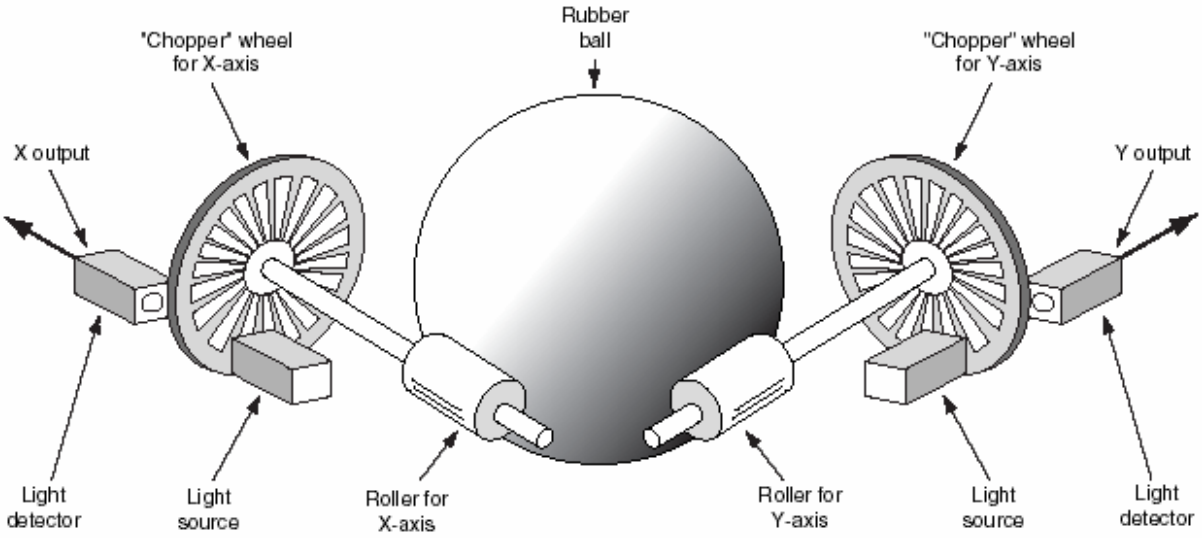


Figure 5.42 – Pointing devices. (a)–Trackball, (b)–Touch Pad, (c)–Digitizer, (d)–Joy stick, (e)–Stick.

5.12.2 Pointing Devices

When users' wants to draw pictures or technical drawings, play games a device is needed that points to various locations on the screen. Such devices are called pointing devices and they are used to indicate positions (points) in a Graphical User Interface (GUI). The most common point and draw device is the mouse (or mice) while other devices like the Trackball, Touchpad, Digitiser and Joystick is used for various purposes.

The mouse is a versatile device and it can be used for any pointing or drawing requirement. It can be used to move the courser, select an item, double click on an Icon, drag and drop a file, drawing images, etc. In terms of functionality a trackball is a stationary upside down mouse (figure 5.42-a). Trackballs were mostly used in portable computers where they had an integrated mouse and a keyboard. Today those are mostly used for special applications like controlling movement of devices and locating places in digital maps. Touchpad is a small touch sensitive pad (not a touch sensitive screen) which detects the user's finger movements (figure 5.42-b). The user has to keep his/her finger on the pad and needs to move it towards the direction that he/she wants her cursor to move. User can also press the touch pad to simulate the action of a mouse click. Digitiser (figure 5.42-c) is used to manipulate special drawings like plans and maps. The joystick (figure 5.42-d) operates like a helicopter joystick that can be moved towards any direction and when released, it returns to a central position. Those are mostly used for game playing and controlling robots. Stick (figure 5.42-e) is another tiny, flexible, pointing device that is mostly found in IBM portable computers (located in between keys in the keyboard) and mobile phones. Its functionality is same as the joystick however it uses the tension of the stick to determine the direction of movement.



5.43 – Mechanism of an roller ball mouse

The mice can be further classified based on the movement detection mechanism. Friction or roller ball mouse uses a heavy rubber ball to determine the direction of movement while the optical mouse uses a light beam and a tiny CCD (Charged Coupled Device) camera to detect the movement. The principle behind the friction mouse is very simple (figure 5.43). Two rollers are attached to the rubber ball one for the X-axis movement and other for the Y-axis movement. A chopper wheel with shutters is attached to the end of each roller. An Infra Red (IR) light source and a detector are kept in either side

of the chopper wheels to detect the movement of the wheel. When the mouse moves towards the X-axis the ball rotates towards that direction. Then because of the friction the X-axis roller and the chopper wheel attached to the ball rotates. This rotating chopper wheel chops the light beam resulting a blinking action. This blinking is detected by the IR receiver and it translates it as a movement along the X-axis. When the mouse moves towards Y-axis same thing happens where the Y-axis roller and chopper wheel becomes active. If the mouse moves towards both directions at the same time (some where in between X and Y axis) both rollers get activated.

5.12.3 Document Readers

A number of mechanisms have been developed to read data directly from a document. The mark readers are designed to understand a specific mark(s) on a document. These are used specially in MCQ based exams where the students have to put a small black circle using a pencil around the respective answer. Then the scanned image of the answer sheet is imaged processed to detect such marks. Then detected answers are compared with the model answer sheet.

Certain devices are developed to detect characters or digits written in a specific format or style. These characters are created by variety of printers and one such application is the number plate of a vehicle. This mechanism is called the Optical Character Recognition (OCR). A further extension of this approach is used even to detect hand written documents.

In Magnetic Ink Character Recognition (MICR) special magnetic ink is used when printing or typing documents. The content of such documents can be extracted by deterring the variation in magnetic field. A cheque is one such example where the cheque number, bank code and the amount are encoded using special form of digits and printed with magnetic ink.

5.12.4 Data Capture Devices

These devices are used in special purpose applications for direct data extraction. Barcode is one such example where data is encoded in a set of printed bars with different widths (variation of width may not be easily detectable by human eye) (figure 5.44-a). These bars can be identified by an optical barcode reader (figure 5.44-b). Barcodes are used almost everywhere including grocery items, membership cards, books in a library, spare part of all forms of vehicles, computer accessories, etc.

5.44 – Barcode (a) and barcode readers (b)

Data can also be directly recoded on small magnetic strips (like in a floppy disk) and can be read by magnetic readers. This is a very convenient and cost effective form of data storage therefore used in ATM or credits cards, driving licenses, identify cards, etc.

5.13 Output Devices

These are the devices that visualise the output of a computer. The video display unit is the major output device which was explained in section 5.10. Other output devices includes the printers, multimedia projectors and speakers. Speakers produce audio output while the VDUs, projectors and printers produce the visual output.

5.13.1 Printers

The printer produces a permanent hardcopy output from a computer by converting digital data into marks on a paper. Printers are built using mechanical, electromechanical and electronic components which are precisely placed and timely controlled. Printers can be broadly classified based on the printing mechanism as *impact* and *non impact* printers. In impact printers the printing head strikes onto the paper through an ink ribbon forming some mark on the paper. In non impact printers such direct contact is not available. Dot matrix, Daisy wheel, tape and cylinder printers belongs to the

category of impact printers while ink jet, bubble jet, thermal and laser printer belongs to the non impact category. Each of these printers has different speed, cost, quality and permanency characteristics.

Dot Matrix Printer

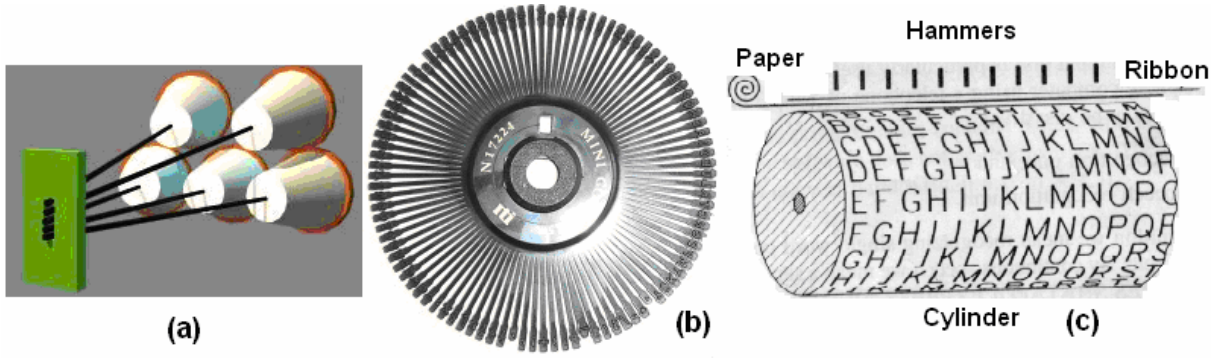
Dot matrix printer is an impact printer that print characters composed of matrix of dots. These dots are formed by number of needles pressing an ink ribbon onto the paper (figure 5.45-a). These printers have relatively smaller number of moving parts and easier to maintain. Until the introduction of other printers such as inkjet and bubble jet dot matrix printers were used for all forms of printing needs. First generation dot matrix printers had only 7 needles (7 dots per column) therefore quality of their output was rather poor. Later 24 needle printer head was introduced and it was able to produce near letter quality printouts. However, even with the 24 needle head quality of images is not so satisfactory. Dot matrix printers are moderate in terms of cost and in terms of cost per page it is the lowest. However these printers are relatively slow and noisy. The printing speed depends on the expected quality of the characters. Because of these concerns today these printers are used only for special purposes like printing multiple copies of forms and bills and cutting tracings.

Daisy Wheel Printer

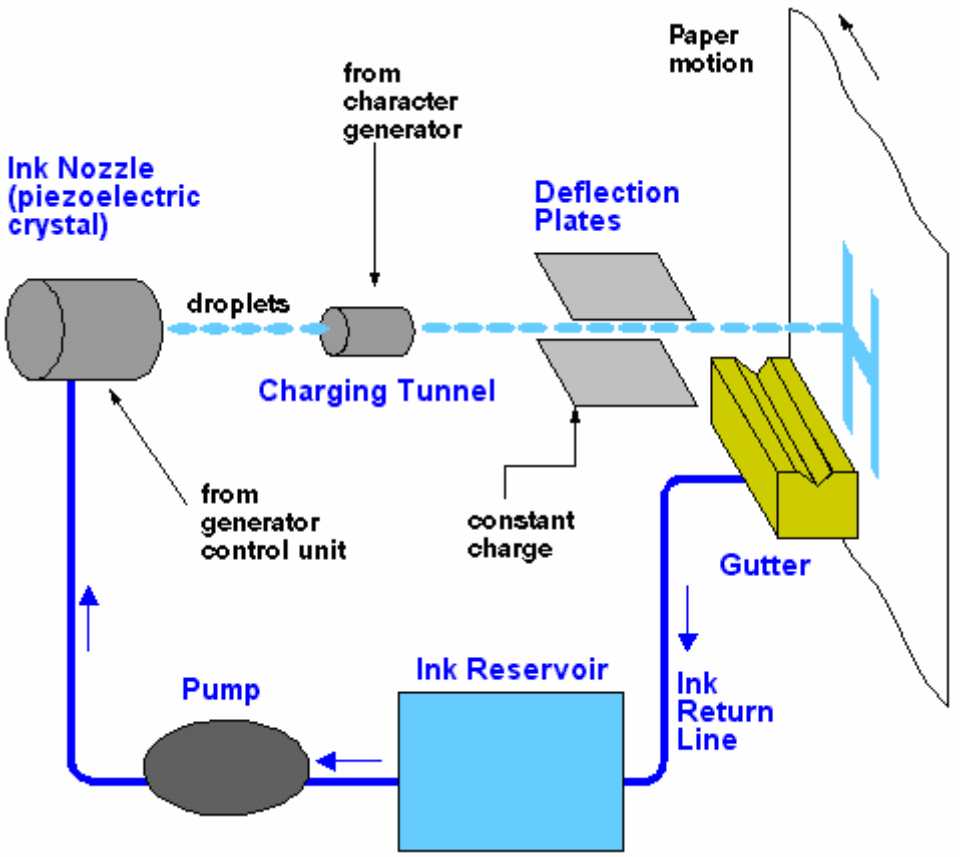
The Daisy wheel printer has a disk with large number of spokes attached to its periphery (figure 5.45-b). Each end of the spoke (tip) is embossed with a character, digit or a symbol. This wheel is built using a light weight metal or plastic therefore has a lower inertia. As a result it can be rotated much faster. The wheel rotates in a vertical plane and it is mounted in front of an ink ribbon. On the back side of the spoke a solenoid driven hammer is mounted. When a character needs to be printed by rotating the wheel the desired embossed character is moved to the print position. Then the solenoid is energised and the hammer presses the tip of the spoke against the paper through the ribbon. This action forms a mark on paper. Daisy wheel printers are slow and noisy. However the print quality is much better but it cannot printer any characters or symbols which are not embossed in the Daisy wheel. Therefore these printers cannot be used to print graphics or images.

Cylinder Printer

These printers make use of a rotating cylinder and in terms of functionality it is like a printer with multiple Daisy wheels (figure 5.45-c). The surface of the cylinder is embossed with various characters, digits and symbols. These logical wheels are position at different locations within the cylinder so that they can printer multiple characters at the same time. In these printers it is not the cylinder that hits the paper (when the hammer is activated) it is the ribbon that hits against the cylinder through the paper (figure 5.45-c). Multiple hammers are used to activate set of logical wheels. Since it can print multiple characters at the same time these are the fastest printers in the world. But cost of these printers is much higher. The noise is the main concern therefore these printers are generally mounted inside soundproof cabinet. These printers cannot print any graphics or images. In certain printers instead of a rotating cylinder a rotating ribbon is used. The mechanism behind those printers is also same. The ribbon is embossed with character, digits and symbols.



5.45 Impact printer mechanisms. (a) – Dot matrix printer head, (b) – Daisy wheel, (c) – Cylinder printer mechanism



5.46 – Mechanism of an ink jet printer

Ink Jet Printer

In ink jet printers, a tiny jet of ink emitted from a nozzle is used for printing (figure 5.46). When needed, ink which is kept in an ink reservoir (referred as the ink cartridge) is transferred to the ink nozzle by a suction pump. The nozzle is a piezoelectric crystal that vibrates forming small drops of ink (referred as the droplets) from the ink stream. Each droplet that leaves the nozzle is given an electric charge by the charging tunnel. Then these droplets can be electrostatically deflected by the deflection plates just like the beam of electrons in a CRT. There are four deflection plates; two for the horizontal movement (not shown in figure) and another two for the vertical movement. By controlling the deflections characters can be printed on the paper. Droplets that do not hit the paper are collected by the gutter and then sent back to the ink reservoir for reuse.

These printers can form better quality printouts (both text and graphics) at a reasonable speed. In terms of cost these printers are moderate but cost per paper is bit on the high side.

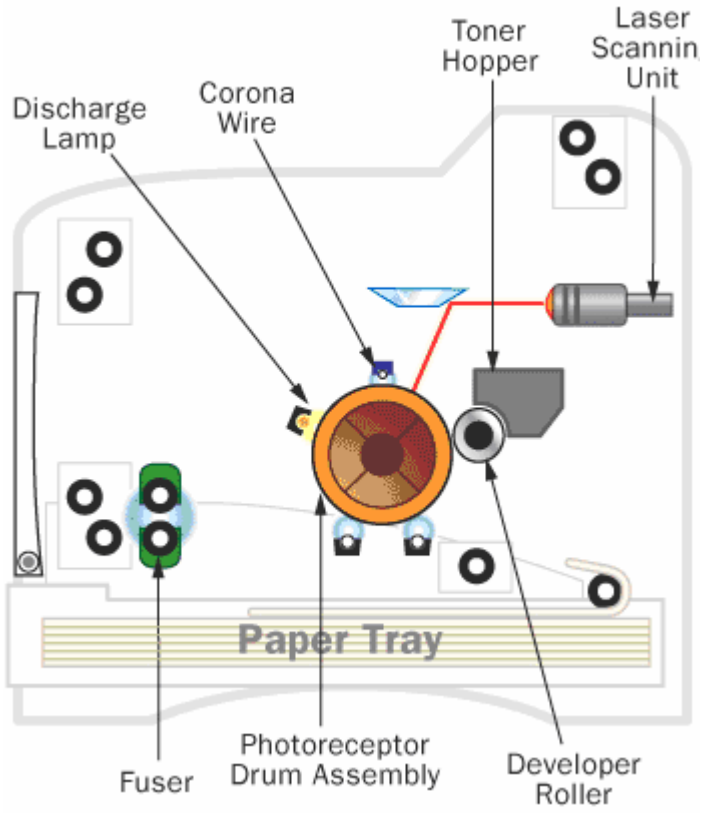
Laser Printer

Laser printers are the fastest document printing machines and can also produce high quality printouts. Cost of a printer is much higher but on the other hand cost per paper is low. Most laser printers are single colour (back and white) however colour laser printers are also immerging.

The anatomy of a laser printer is almost identical to a photocopier which is modified to accept inputs form a computer. Figure 5.47 illustrates the components of a laser printer. The heart of the printer is the photoreceptor drum assembly (normally referred as the drum). This drum assembly is made out of highly photoconductive material that is discharged by light photons. Corona wire is used to charge the drum. Laser scanning unit and the rotating hexagonal mirror is used to form the image on the drum. Toner hopper holds a black powder called the toner (toner is the powder form of ink that is used for printing in this case). The ionised drum pulls out the toner from the toner hopper through the rotating developer roller.

When something is to be printed the laser scanning unit send a laser beam to the drum forming an electrostatic image on the drum. The rotating mirror is used to control and guide the laser beam that hits the drum. The ionised image on the drum is called the positive image. The drum rotates clock wise

and when it moves closer to the developer toner is attached to the ionised areas of the drum. When it further rotates toner is transferred to the paper which is having even more electrostatic charge. The surface of the paper is pre heated to fix toner on to it (not shown in figure 5.47). Any remaining toner on the drum is scraped off by a very fine blade mounted before the discharge lamp (not shown in figure 5.47). Finally the discharge lamp is used to recharge the photons on the drum removing any positive image on the drum. When the drum comes back to the position of the corona wire process starts repeating forming another positive image and printing that on paper.



5.47 – Components of a laser printer

6 – Operating Systems and Application Software

In early days all the programs were written from the beginning to run on a specific computer. If the same program is to be used in a slightly different machine (in terms of hardware) it was required to be rewritten from the beginning. Each of these programs had to control the CPU scheduling, memory management, storage management, Input/Output, etc. rather than its actual business or scientific functionality. With the advancement of technology, programs are becoming more and more complex and it was not possible to develop software from the beginning to a specific hardware platform.

A layer of software called Operating System (OS) was introduced to overcome this issue. The OS acts as a virtual machine on top of the hardware. Writing programs for the virtual machine was much easier than writing programs for pure hardware. The OS provides certain set of services to programs which are not tightly coupled to specific hardware components. The OS manages all the devices and provides user programs with a simple interface to deal with the hardware.

The OS is a special type of software application which is referred as *system software*. The rest of the handout illustrates; concepts behind OSs, different type of OSs and some of the well known commercial OSs. In the final section, the handout introduces different types of application software.

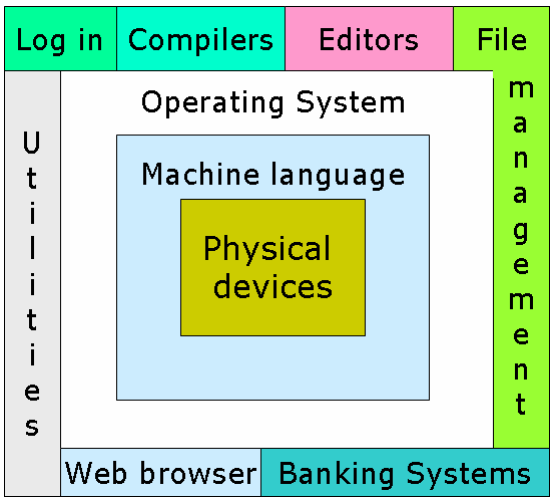


Figure 6.1 – OS as a virtual machine

6.1 An Operating System

An operating system provides two major services; it acts as an *extended machine* and a *resource manager*.

6.1.1 Operating System as an Extended Machine

It hides all the messy details about the underlying hardware and how to deal with them. It presents users with a virtual machine that is easier to use (figure 6.1). It offers uniform way of doing something. Although the approach in saving a file in a floppy disk, hard disk or magnetic tape is physically different the OS users do not need to be aware of those issues. For them saying “save this file in that place” is enough. All the matters relating to; how to save, where to save and at what speed to write is handled by the OS. And again when a user wants to open a file he/she can just say “open the file at that place” regardless of knowing how to read from a floppy, hard disk or a CD-ROM. It is up to the OS the do all the hard work and opens the file.

6.1.2 Operating System as a Resource Manager

The OS is responsible for managing resources in a computer. It is its responsibility to manage them in the most effective manner while ensuring user satisfaction. OS will decide which program runs at which time, how much of memory to be allocated for the program, where to save a file so that the disk space is optimally utilised, how to deal with concurrent users, how to enforce security, etc.

The readers should clearly understand programmes such as; the Schell (e.g. DOS prompt), editors (e.g. Notepad, WordPad, vi, eMac), compilers, file management systems (e.g. Windows Explorer) and multimedia applications which are bundled with the OS installation are not part of the actual OS. They are just set of *utility programs* running on top of the OS (figure 6.1).

The core or the nucleus of an OS is the *kernel.* Kernel loads before any of the user programs and remains in the memory. It is responsible for managing CPU, memory, disk, processes[20], etc.

6.2 History of Operating Systems

First generation (1945-1955) operating systems cannot be considered as real operating systems. Those days' programmes were written by changing the wire in a *plugboards*[21]. When a different program is to be executed the plugboard needs to be rewired.

The second generation (1955-1965) OSs were mainly batch systems. Figure 6.2 illustrates steps in a batch system. Users prepared jobs and submitted them to the operator. Normally jobs are usually submitted in the form of punch cards (figure 6.2 step *a*). It takes lot of time to read a punch card therefore jobs are transferred from punch cards into a magnetic tape (figure 6.2 step *b*). To further reduce the overhead multiple jobs with similar needs are batched together. Then the tape is inserted into the main computer (figure 6.2 step *c*) and it does the processing. When the outputs of the jobs are ready (figure 6.2 step *e*) it is sent back to the user mostly in a form of a hardcopy (figure 6.2 step *f*).

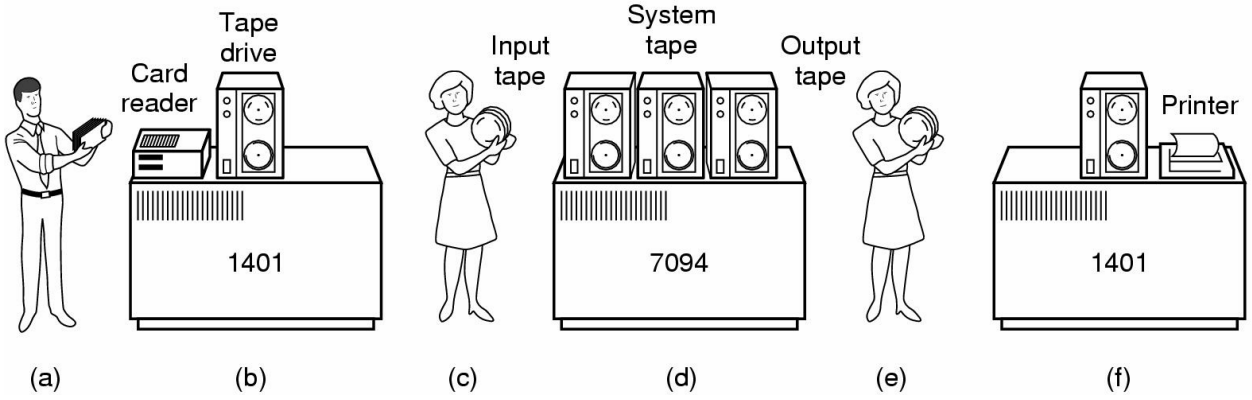


Figure 6.2 – Batch systems

Some of the jobs take a long time, possibly a day or even more. Most of these jobs are highly I/O intensive (needs to read/write lot of data from/to the magnetic tape) hence the CPU is mostly idle. A typical computer has large number of resources but a single job rarely uses all of these resources. While a computer being several million Dollars (at those days) keeping CPU idle and lower resource utilization was a major issue. Therefore to make sure CPU has enough work and to enhance the resource utilization *Multiprogramming* was introduced, which led to the third generation (1965-1980) of OSs.

In multiprogramming the OS access number of jobs from a magnetic disk rather than from a magnetic tape. When the computer is running the OS keeps multiple jobs in the memory. The CPU executes one of the jobs in the memory and it continues executing the selected one, until it gets blocked due to some I/O activity. During this time, to maximize the CPU utilization the OS switches to one of the remaining jobs in the memory and executes the newly selected one. When the current job gets blocked due to some I/O activity the CPU switches to yet another job in the memory. Eventually the 1st job will finish its I/O activity and gets the CPU back. This switching will continue until all the jobs are finish. Due to this approach CPU is never idle and resource utilization is improved.

[20] Process is an instance of a running program. A process is more than the program code, it also includes the state of the current activities represented by CPU registers, variables, memory addresses, function arguments, etc.
[21] Plugboard is a large circuit board where different vacuumed tubes are connected to form a circuit. The tip of the wire includes a plug, therefore instead of soldering those wires they were directly plugged into the Plugboard.

Later a variation of multiprogramming called *time-sharing systems* (also referred as *Multitasking*) was introduced. In multiprogramming there is no user interaction. The user submits his/her job to the operator and then operator has the control; even the operator has a very little control while the batch is being executed. In time sharing systems, the user provides instructions directly from the keyboard to the computer and some of the results will be shown through the terminal at the same time.

These interactive I/O runs at "human speed" but for a computer it is pretty slow. A user would be really happy if he/she can type about 7 characters per second, but a computer can execute thousands of instructions within that time. Therefore the CPU can execute another program by the time a user types the next character. During this small period of time the CPU executes multiple jobs by switching among them at a rapid speed. Each user is given a particular timeslot to access the CPU. When the timeslot get expired the CPU will switch to another user's job. After sometime the first user will get the CPU for another timeslot.

Consider the example given in figure 6.3 where three users having three unequal jobs; A, B and C[22]. Suppose each user is given a timeslot worth 100ms CPU time. The CPU executes user A's job for 100ms then it will switch to user B. Then B's job is executed for another 100ms and the CPU switches to job C. Suppose that C's job requires some I/O activity during its timeslot (after 60ms). Then the CPU stops executing job C and again switches back to job A. This switching among jobs will continue either a job expire its timeslot or get blocked due to some I/O activity.

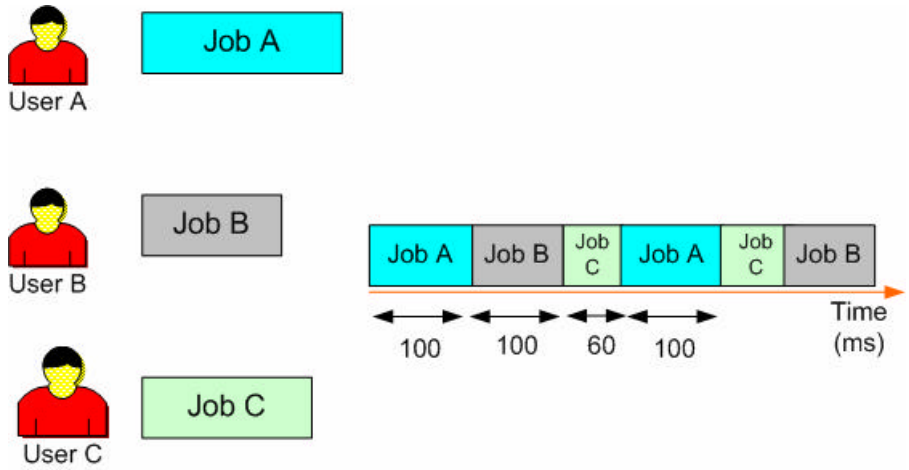


Figure 6.3 – A time-sharing system with 3 users

Time-sharing system allows users to interact with their programs while it is being executed. Timesharing improves resources utilization and CPU is never idle. None of the individual users will realise that there are other users in the system and everyone feels that they own the computer.

Currently we are in the fourth generation (1980 - present) of OS which is referred as Personal Computer OSs. Sometimes these OSs are also referred as *Desktop Systems*. These were introduced in late 1970's. The jobs handled by these systems are neither CPU intensive nor I/O intensive and they make use of commodity hardware components. These systems are more interested on maximising user convenience and responsiveness. Anyway they also incorporate many of the features available in larger OSs. Some of the features such as security and reliability which were not initially incorporated were added later.

6.3 Types of Operating Systems

Operating Systems can be classified into several categorise such as:

1 Mainframe operating systems
2 Server operating systems
3 Multiprocessor operating systems
4 Personal computer operating systems
5 Real time operating systems

[22] Length of the job indicates the required CPU time.

6 Embedded operating systems
7 Smart card operating systems

6.3.1 Mainframe Operating Systems

Mainframe OSs are used with very large computers called *Mainframes*. They process many jobs at once requiring large amount of I/O activities. These OSs offer; batch processing, time-sharing and transaction processing. OS/360 and OS/390 are some of the examples for mainframe OSs developed by IBM which is used in IBM mainframes and mid-range servers.

6.3.2 Server Operating Systems

Servers are scale down version of mainframes and can accommodate multiple users at once. Servers are much cheaper compared to mainframes therefore most organizations prefer servers than mainframes. Servers include; print services, file services, web servers, mail servers, etc. Multiple clients connect to these servers through a network. The operating systems used by these servers are called server operating systems. Windows 2000/2003 Server/Advanced Server, Enterprise Linux Advanced Server and Sun Solaris are some of the well known server operating systems.

6.3.3 Multiprocessor Operating Systems

In some computers multiple CPUs are connected to the same motherboard and these are referred as multiprocessors computers. These computers can execute multiple jobs at the same time where each CPU executes a different job. Operating systems used in these computers are called multiprocessor OSs and those are variations of Server OSs with special features for communication and connectivity among multiple jobs and CPUs. Windows 2000/2003 Server/"Advanced Server", "Enterprise Linux Advanced Server" and Sun Solaris are some of the well known server operating systems which support multiprocessors.

6.3.4 Personal Computer Operating Systems

These are OSs used in PCs. They are mainly interactive systems and currently these are also called *Multimedia operating systems.* They offer; multimedia features, word processing, desktop publishing, gaming, Internet access, etc. Windows 98, Millennium (Me)/XP, Red Hat Linux 7/8/9, Fedora 2/3/4/5, Ubuntu, Apple Macintosh are some of the well known OSs.

6.3.5 Real Time Operating Systems

In these systems, time is the key parameter. These OSs need to respond to various events with strict time constraints (i.e. they must meet deadlines). Operating systems used in these types of systems are called real time OSs. They are used in industrial process control, automobile, digital audio and video systems.

Real time systems can be divided into two categories called; *hard real time* and *soft real time* systems. Hard real time systems must meet deadlines under whatever circumstances where as in soft-real time systems occasional miss is acceptable. VxWorks and QNX are some of the well known real time OSs.

6.3.6 Embedded Operating Systems

These operating systems are used in devices such as palmtop computers and embedded systems. They are used in PDAs, microwave ovens, mobile telephones, networking devices, etc. These operating systems have to deal with lower processing power, small storage capacity, memory and power restrictions. PalmOS, Symbian, Windows CE, Windows Mobile are some of the examples.

6.3.7 Smart Card Operating Systems

Smart cards are used for various purposes such as; payphone cards, ATM cards and crypto cards. These are credit or ATM size cards with an inbuilt microprocessors and small amount of memory (figure 6.4). Smart card OSs runs on these tiny microprocessors. They have very limited processing power, space and memory (both volatile and non-volatile). Java cards with inbuilt Java Virtual Machine (JVM) is one such example.

Figure 6.4 – A Smart card

6.4 Functions of an Operating System

Functionality of an OS slightly varies based on the type of the OS. Following are common list of tasks carried out by a typical OS:

- **Memory management** – OS is responsible for allocating memory for programs. The objective is to utilize limited memory in the most effective manner. It is also responsible for sending data and instruction from memory to the hard disk (i.e. swapping) when memory is filled up. The OS is also responsible to send the data back to the memory from secondary storage when required for processing.
- **Spooling print jobs** – all the print jobs will be collected onto the hard disk and later they are scheduled by the OS for printing.
- **Configuring devices** – OS allows easy access to devices, their installation and configuration.
- **Monitoring system performance** – Most modern OSs can display (either graphically or numerically) how system resources such as CPU time and memory is utilized. This information can be used to identify whether resources are under utilized or over utilized.
- **Administering security** – OS support multiple concurrent users while making sure each user has its own privacy. It also enforces authentication[23] and authorization[24].
- **Managing storage media and files** – The OS will make sure whatever the secondary storage devices are optimally used. It will do all the hard work of saving and retrieving files to and from disks, organizing them, putting the date stamp, setting various file attributes such as read-only, hidden, etc

6.5 Popular Operating Systems

There are several classes of operating systems and many variations; mainly DOS, UNIX and Mainframe OSs from IBM. DOS (Disk Operating System) later became Windows while UNIX led to different variants such as; Linux, Sun Solaris, Free BSD, AIX, etc. IBM operating systems were mainly targeted towards mainframes and some of the well known ones are OS/360, OS/390 and OS/400. Apple Macintosh can also be considered as a separate class of OS rather than a variant of UNIX.

DOS was the first disk based OS which was developed for IBM PC by Microsoft. It was concerned about keeping OS, application programs and all user files on a disk and managing them through set of commands called DOS commands. There are two well known variants of DOS called PC-DOS (Personal Computer DOS) and MS-DOS (Microsoft DOS). PC-DOS was developed and sold with IBM PCs while MS-DOS was sold in open market. DOS was simple to use and learn therefore Microsoft was able to win a large market share among other PC OSs.

[23] Proving that a user is who he/she clams to be. This is mostly achieved by having a user name and a password.
[24] After authenticating a user the OS allow different roles to the user such as administrator, user, super user, etc. Based on the roles each user has different privileges while accessing the system.

Inspired by the user interface of Apple Lisa, Microsoft decided to give DOS a Graphical User Interface (GUI). Microsoft released their first version of GUI based OS called Windows 1.0 on 1985. However, only Windows 3.1 was commercially successful. Early versions of Windows were just an application running on top of DOS, behind the screen GUI was actually issuing DOS commands. With the success of Windows 3.1, Windows 95 was introduced. Then with the introduction of "Windows 98" Microsoft was able to purely escape from DOS and built a compute GUI based OS which does not depend on DOS. Still Windows 98 is considered to be the heavily used PC OS in the world.

Microsoft Windows supports multitasking (i.e. allowing users to have several applications running at the same time). It is mainly targeted for average "home" users who do not have much of a technical background. It supports most of the hardware and software components. Later version of Windows such as Windows 2000, Windows Millennium edition and Windows XP incorporates; user friendliness, better performance, efficient use of resources, security, stability, etc. However compared to some of the other OSs windows is not so stable, does not use resources in an effective way, lack some security features, etc. However due to its higher user friendliness Windows has a large market share compared most other commercial OSs.

One of the most important OSs to be discussed is the UNIX. It was developed in 1969 at Bell Labs by Ken Thompson and Dennis Ritchie. UNIX is still being used with large number of variants and versions. It was developed as a time-sharing system for *minicomputers* and mostly used by universities. The architecture of UNIX is so stable and so far it is the best in terms of security, reliability, robustness and performance.

It was developed by engineers for engineers. Therefore it was harder for an average person to effectively use UNIX or any of its variants such as Berkeley UNIX, Linux, MINIX, AIX, etc. This was an issue why UNIX was not so popular among the general public. However, modern variants such as; Fedora, Caldera and Suse Linux are much better in terms of user friendliness.

6.5.1 Linux

Linux was developed by Linus Torvalds in 1991. Linux is a UNIX like OS developed originally for home PCs. However today it runs on a many platforms including; PowerPC, Macintosh and Sun Sparc. It was developed with the intension of making it simple so that anyone can understand and improve. The most important thing about Linux is its totally free. You are even given the source code of the OS. Since its source code is freely available many people around the world had studied and improved it. Therefore Linux is a complete OS which is stable, reliable and efficient compared to most other OSs. It also supports excellent networking facilities.

If you compare Windows and Linux; Linux required lesser disk space, memory and processing power than Windows. Both OSs support multitasking and multiprocessing however Linux is better in terms of handling multiple users. Linux is free while Windows license is about LKR 20,000 to 30,000. If the user is not satisfied with the freely available software there is always wide variety of commercial software to be bought.

6.6 Application Software

Software refers to a program or set of instructions that instructs computer to perform some task. Software can be dived into two major categories called; *system software* and *application software.* Systems software includes operating systems and various device drivers. Application software is used to perform real-world tasks and solve specific problems.

6.6.1 Major Types of Software

Software can be categorised based on the intended customer:

- **Generic products** – are standalone systems or group of applications which are sold on the open market for any customer. MS Office, Photoshop, Flash are some of the examples.
- **Bespoke (customized) products** – are tailor made to handle a particular customer's requirement. A software system for Colombo Stock Exchange and a billing system for SLT are examples for products developed to a particular customer.

6.7 Types of Software

Software can be further classified as:

1. System software
2. Real time software
3. Business software
4. Embedded software
5. Engineering and scientific software
6. Personal computer software
7. Web based software
8. Artificial intelligence software

6.7.1 System Software

System software includes operating systems and device drivers.

6.7.2 Real Time Software

These are similar to real time operating systems. Only difference is these are applications run on top of some operating system. These are software solutions that monitor, analyze and control real time world events as they occur. Computerised heartbeat monitoring systems, rocket and nuclear power plant management systems belong to this category.

6.7.3 Business Software

Business information processing is the largest software application area. There are millions of application programs written for business purposes. Any computerized system related to any business matter can be considered as business software. Microsoft Money, Sage accounting and Lotus 123 are some of the well known applications.

6.7.4 Embedded Software

These software systems are used in modern consumer electronic devices such as mobile phones, PDAs, digital diaries, photocopiers, etc. These software runs on top of an embedded OS. Java games, Sinhala/Tamil SMS applications are some of the example embedded software runs on mobile phones.

6.7.5 Engineering and Scientific Software

These are mainly used for scientific and engineering applications. They have been characterized by number crunching algorithms. These software systems are used for various calculations, simulation purposes and testing. These software systems are used in CAD (Computer Aided Design) and CAM (Computer Aided Machining) applications such as AutoCAD, OrCAD and CircuitMaker.

6.7.6 Personal Computer Software

These are the tiny software applications used in PCs. PC software market has flourished over the past two decades. These sorts of systems are used in areas such as; word processing, spreadsheets, computer graphic, multimedia, database management, etc. MS Office, PageMaker, Photoshop, Winamp and Jet Audio are some of the examples.

6.7.7 Web Based Software

Any form of software that run on top of a network or Internet belongs to this category. Software systems such as; web browsers, messaging tools (Yahoo/MSN messenger), streaming media and video conferencing belong to this category.

6.7.8 Artificial Intelligence Software

These are also referred as *expert systems*. Artificial intelligence (AI) software makes use of non-numerical algorithms to solve complex problems. They try to simulate human like thinking and behaviour. Achieving this is not an easy task therefore most of these systems are still under research. Shuttles send to outer space, image processing and game playing applications make use of AI.

7 – Introduction to Networking

When two or more computers are connected together it forms a computer network. A computer network can rage from a small network in a house (where two computers are connected) to the Internet (where millions of computers are interconnected) that spans the entire world. Although networking is a broader subject we would focus only on basics of networking. Throughout the chapter the term network refers to a computer network.

7.1 Definition

Some form of an association between set of points or locations is called a network. It can be a network of users, resources, organisations or even computers. The courier services such as DHL or FedEx has branches, offices and collection points all around the world which are linked to each other with the objective of transferring documents and goods all around the world. Such a connection of points (locations) can be considered as a network of offices.

When two or more computers are connected using some communication medium it forms a computer network. Exchange of information is the main objective of a network. Figure 8.1 show a simple network that connects four computers and a printer.

7.2 The Need of a Network

Networks are useful due to two major reasons namely; resource sharing and data sharing.

Resource sharing allows resources such as printers (figure 7.1), scanners, centralised file storage and CD/DVD Writers to be shared among multiple workstations within the same network.

Data sharing allows exchange of information among users. The exchange of information can be in the form of e-Mails, video conferencing, web pages, text and voice messaging, etc. In order to exchange information both the source and destination should be connected and the source may route packets from one user to another. One of the best examples for sharing data is the publishing of O/L and A/L results (i.e. data) over the Internet by the Department of Examinations.

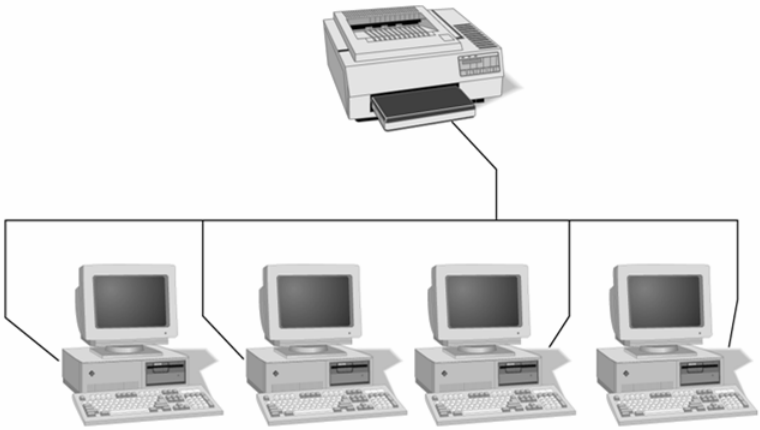


Figure 7.1 – A simple network

7.3 Components of a Network

A network consists of several components namely; computers, connections, applications, data and people. *Computers* are connected to each other using some type of *connections*. These connections are referred as communication medium. *Applications* make use of these *connections* to transmit/receive *data* with each other. e-Mail is one such application. All these *applications* are used by *people* for communication and resource sharing. When all these components are combined in an effective manner it forms a usable computer network.

7.4 Transmission Medium

Transmission medium provides the path for data communication. It allows a stream of bits to be transmitted (for e.g. as a variation of voltage) from one data communication device to another. The transmission medium can be categorised as guided media and unguided media. The signal path is guided by the guided media and unguided media does not provide specific signal path. Guide media includes transmission mediums such as copper wires and fibre optics while radio waves, micro waves and infrared belongs to the category of unguided media. A more specific categorisation of transmission media is given below:

- Guided media
  - Twisted pair
    - UTP – Unshielded Twisted Pair
    - STP – Shielded Twisted Pair
  - Coaxial cables
  - Fibre optics
- Unguided media
  - Radio waves
  - Microwave
  - Infrared

7.4.1 Twisted Pair

Twisted pair is formed by twisting pair of copper wires together (see figure 7.2). When 2 wires are kept in parallel they form an antenna, such wires are susceptible to Electro Magnetic Interference (EMI). In order to reduce the EMI wires are twisted around each other. The higher the number of twists per length the lower the EMI. Therefore a cable with large number of twists is better for data communication however the cost of such a cable can be higher.

Some of these twisted pairs are covered by an insulation jacket (figure 7.2) to further reduce the EMI. Such cables are suitable to transmitting data over a long distance. Those cables are referred as Shielded Twisted Pair or STP. The cables without such an insulation jackets is called an Unshielded Twisted Pair or UTP. Cost of STP is higher than UTP however STP can transmit signals to a long distance without reducing that much of signal power compared to UTP.

In data communication UTP cables are further categorised as:

- Category 1 – Voicegrade telephone cabling
- Category 2 – up to 4 Mbps 4 pairs
- Category 3 – up to 10 Mbps 4 pairs
- Category 4 – up to 16 Mbps 4 pairs
- Category 5 – up to 100 Mbps 4 pairs
- Category 5e – up to 1 Gbps 4 pairs

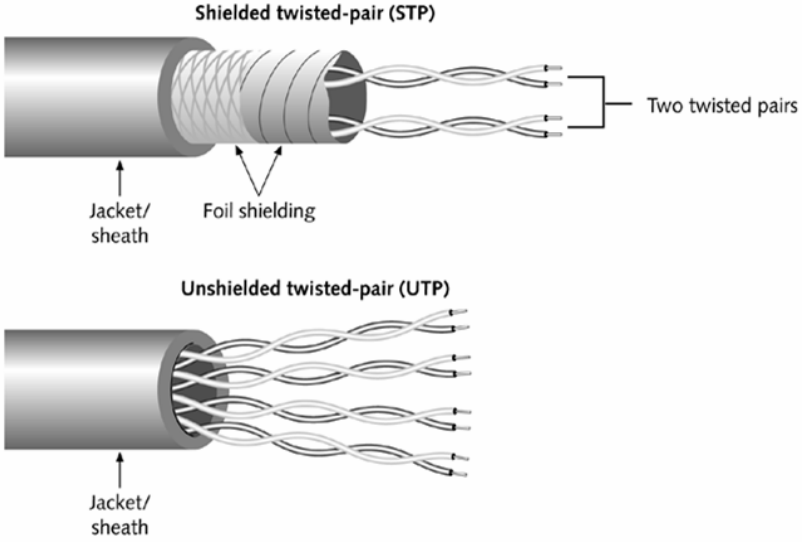


Figure 7.2 - Shielded Twisted Pair and Unshielded Twisted Pair

Category 1 (also referred as Cat 1) is used in domestic telephone systems while Category 5 (generally referred to as Cat 5) is mostly used in Local Area Networks (LAN). An enhanced version of Category 5 is also available and it is referred to as Category 5e or Cat 5e.

The most important advantage of both UTP and STP compared to other media is the lower cost. Those cables are easier to work with (they bend easily without getting damaged). STP can run several kilometres without amplification and for long distances repeaters are needed. However with distance bandwidth reduces and as a result the data transmission speed reduces.

7.4.2 Coaxial Cables

In early days of networking these types of cables were heavily used. A copper conducting core is covered by an insulation material (mostly PVC or Teflon) which is referred as the Insulation (figure 7.3). The Insulation is covered by a Braided shielding which is built either using a copper mesh or an aluminium foil. The braided shielding is covered by a plastic outer covering referred as the sheath.

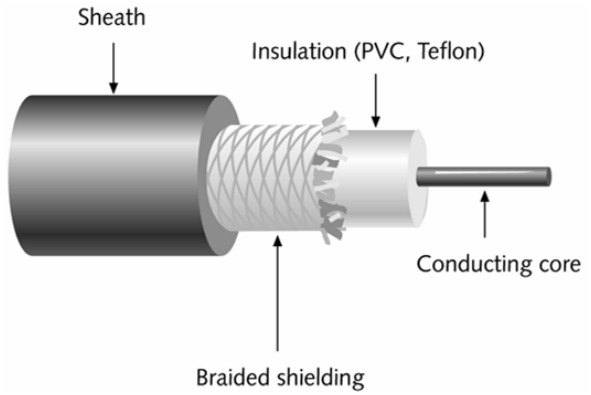


Figure 7.3 – Parts of a Coaxial cable

In coaxial cables, the braided shielding provides much better resistance to electro magnetic interferences. Therefore these cables can span much longer distances than STP.

These cables can be categorized as Baseband Coaxial Cables and Broadband Coaxial Cables. Baseband Coaxial Cables typically have a resistance of 50 O and used in digital transmission. Broadband Coaxial Cables typically have a resistance of 75 O and used in analog transmission such as in cable TV and as TV antenna cables.

Coaxial cables are harder to install and can easily get damaged than twisted pair. However, they are easier to install than fibre optics. In terms of cost these cables are cheaper than fibre optics but expensive than twisted pair. Due to the extra shielding these cables can span higher distances at higher data rates.

7.4.3 Optical Fibre

Use of fibre in data and voice communication is rapidly expanding. With mass production, fibre and associated devices are getting affordable. As a result more and more fibre based networks are being setup.

Fibre optics are formed by bundle of tiny glass or plastic fibres. Signals are transmitted as pulses of light. The optical fibre which is referred as the core is covered with a glass cladding (figure 7.4) and protective outer sheath which is known as the jacket.

An optical fibre based system includes 3 major components, namely; the light source, the transmission medium and the light detector (figure 7.5). The light source converts electrical signals to optical signals. It is mostly a special LED which produces light based on the applied voltage. Logical ‘1’ is indicated by switching on the LED while logical ‘0’ is indicated by switching it off. Transmission medium is fibre which provides the path for light transmission. Light travels through the fibre based on the principle of total internal reflection. The light detector is a light sensitive device such as a photo diode that converts the detected light to an electrical pulses.

Since there is no direct involvement of electric pulses optical fibre are immune to electro magnetic and other interferences. They have a much higher bandwidth and can carry data at very high speeds (in Gbps) over a long distance.

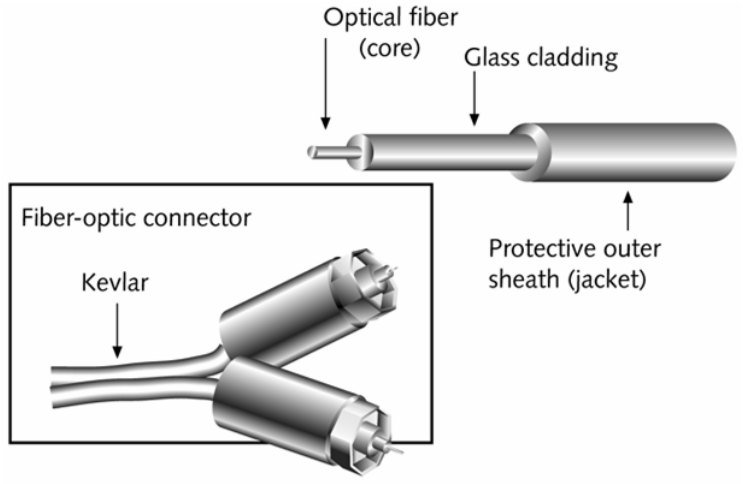


Figure 7.4 – Parts of a Fibre Optic cable

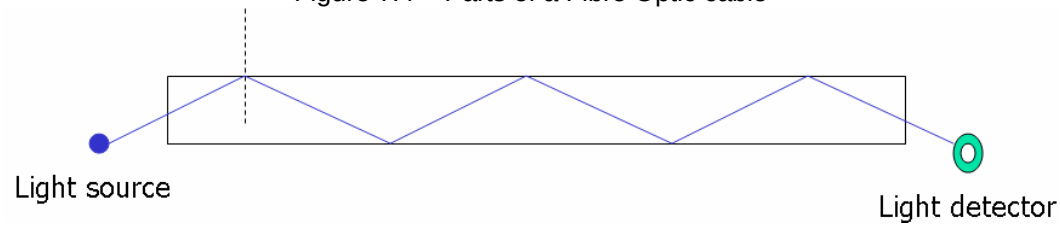


Figure 7.5 – Components of a fibre system

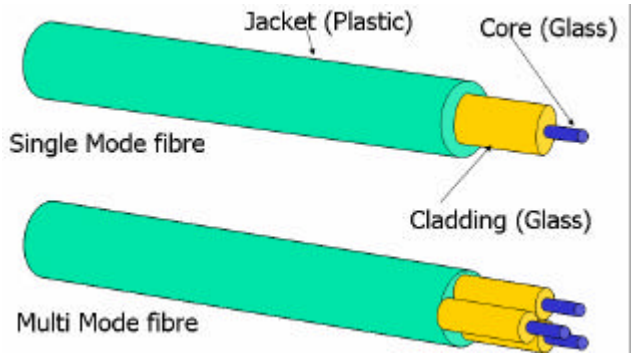


Figure 7.6 – Single mode and multimode fibre

There are 2 types of fibres namely single mode fibre and multi mode fibre (figure 7.6). Single mode fibre uses a single core to transmit light signals while multimode fibre uses multiple cores (fibres) within the jacket. Single mode fibre is so small about 25 µm in diameter. Therefore it is harder to manufacture and as a result the cost is high. Single mode fibre can transmit data at a very high speed over a distance of up to 100km. Multimode fibre cannot span a larger distance as single mode fibre. The data transmission speed of multimode fibre is also lower than single mode fibre but it is cost effective than single mode.

7.4.4 Radio Transmission

Radio transmission belongs to the category of unguided media. They can be easily generated and signals propagate in all directions (omni directional) over a long distance. Radio waves can easily penetrate buildings and other obstacles. Low frequency radio waves such as Short Wave (SW) are reflected by the ionosphere, therefore you could easy send a signal over the entire earth from a single location. High frequency radio waves such as microwaves tend to travel in a straight line and their ability to penetrate buildings is poor. In addition, such high frequency signals are absorbed by rain.

7.4.5 Microwave Transmission

Radio waves above 100MHz are referred as microwaves and they travel in straight lines. Microwave signals can be narrowly focused and can be concentrated to a small beam using a parabolic antenna. Antennas at both ends (transmitter and receiver) must be aligned to each other (referred as Line of Sight). In order to transmit signals using microwaves over a long distance repeaters are needed

because of the line of sight behaviour and the curvature of the earth. As mentioned earlier, microwaves do not penetrate through objects.

7.4.6 Satellite Communication

Satellite communication makes use of the geosynchronous (geo-stationary) satellites which are orbiting around the earth at a distance of 36000 km above the equator. Period of such satellites are 24 hours. Satellites use different signal bands for different purposes such as telecommunication and satellite TV.

7.4.7 Infra-Red (IR) Communication

IR transmission was so popular few years ago and is still being used with devices such as remote controllers for domestic appliances. It makes use of an IR beam for transmitting signals. IR transmission does not work accurately when natural sunlight is around which tends to interfere with IR signals. Therefore IR is used only for indoor devices. IR cannot span over of a long distance and hence ideal for short range communication. These signals do not pass through solid objects either. However, development of IR based communication systems are really simple and cost effective. IR communication does not support faster data transmission. IR devices are used in remote controllers, hand held devices, laptops, etc.

7.4.8 Light Wave Transmission

Data can also be transmitted using light waves (visible light) as in old days. However, both the transmitter and the receiver should be tightly focused, coherent and require very high degree of alignment. It can be used to interconnect LANs in 2 different buildings, yet accuracy may depend on weather conditions. Light waves cannot penetrate rain or thick fog. Further, they may be distorted during day time due to sunlight

7.5 Type of Networks

Depending on the geographical area that a network spans, they can be categorised as LAN, WAN, MAN and PAN.

7.5.1 Local Area Network (LAN)

A network within a building or a campus is called a Local Area Network (LAN). Such a network spans a limited distance (typically about 100m) and range from connecting few computers to a several hundred computers. LANs provide faster data communication and speeds are in the range of 10/100/1000 Mbps (Mega Bits per second). Generally one or more computers in such a network acts as a server (file server, print server, database server, etc.) for sharing peripherals and data. UTP Cat5 and Cat5e are the most commonly used type of LAN cabling. The first year lab is an example for a LAN and all your profiles are stored in the ‘firstyearlab’ file server.

7.5.2 Wide Area Network (WAN)

As its’ name implies a Wide Area Network (WAN) spans a much wider area than a LAN. These networks can span a city, country (a nation-wide network) or even the entire world (a world wide network). Certain corporations/industries such as banks, courier services and military may set up these type of networks to interconnect their branches or regional offices. Such networks are used only for a specific purpose and mostly privately owned. On the other hand the Internet which is the worlds biggest WAN, is a general purpose network and there is no specific owner. These networks are connected either using coaxial cables, fibre or microwave links. Most of these WANs have lower data transfer rates than LANs (except in fibre). A typical example for a WAN network is the ATM (Automatic Teller Machine) systems used by most of the commercial banks in Sri Lanka.

7.5.3 Metropolitan Area Network (MAN)

There is another network classification called Metropolitan Area Network (MAN) which spans larger area than a typical LAN and smaller area than a typical WAN. A MAN is a specialised network that covers a campus, an apartment block, a street or sometimes a suburb or a city. Today though, MAN and WAN technologies are converging. Even then, new technologies such as Metro Ethernet and Wireless MANs (WiMAX) have stopped MANs from becoming obsolete.

7.5.4 Personal Area Network (PAN)

Personal Area Network (PAN) is the latest addition to the network classification. A PAN defines a network that is much smaller than a LAN and that spans around 20-30 feet. PANs are becoming popular with the introduction of more and more wireless devices and peripherals. PANs use technologies such as radio wave, Bluetooth™ and Infra-Red (IR) as the communication medium and interconnects devices such as smart phones, pocket PCs, PDAs, iPods, wireless printers, wireless hands free kits, etc. PANs support communication among similar devices (for the purpose of sharing data such as appointments, electronic business cards) as well as peripherals (sharing resources such as wireless printers).

When multiple networks are connected to each other forming a one large network it is called an 'internetwork'.

7.5.5 Wireless Local Are Networks (WLAN)

A wireless network that spans a building or a campus is called a Wireless Local Are Network (WLAN). WLANs are becoming increasingly popular with the introduction of wireless enabled laptops, handheld devices, peripherals and expansion cards for PCs. Data is transmitted from one computer to another using RF (Radio Frequency) signals. These radio waves travel in all directions and any one within the signal range can access these networks.

The main advantage of WLAN is the mobility, it allows users to move within the organization (within signal range) while being connected to the network. A typical wireless system can span up to 100m and it can also penetrate obstacles such as walls.

7.6 Network Topologies

The network topology defines how computers in a network are connected to each other. There are four major topologies namely; bus, star, ring and mesh networks.

7.6.1 A Bus Network

A bus topology is a network where all the network nodes (computers and network printer) are connected to the same segment of cable in the logical shape of a line, with terminators at each end (figure 7.7). Coaxial cables are used as the transmission medium and network nodes are connected to it using a special type of a connector called the BNC connector. Two resisters are attached to the end of the wire and these are called terminators.

The bus topology is suitable for smaller networks having few machines. It is inexpensive to implement on a small scale since the cable length is small. Another workstation can be added to the network just by cutting the wire and attaching a new BNC connector.

However management costs of such a network is often higher and working with coaxial cables can be harder as well. In case of a problem it would be difficult to isolate a malfunctioning node or cable segment and associated connectors. Since every one is using the same shared bus only two devices can communicate at a time, all others have to wait until the bus is available which could result in congestion.

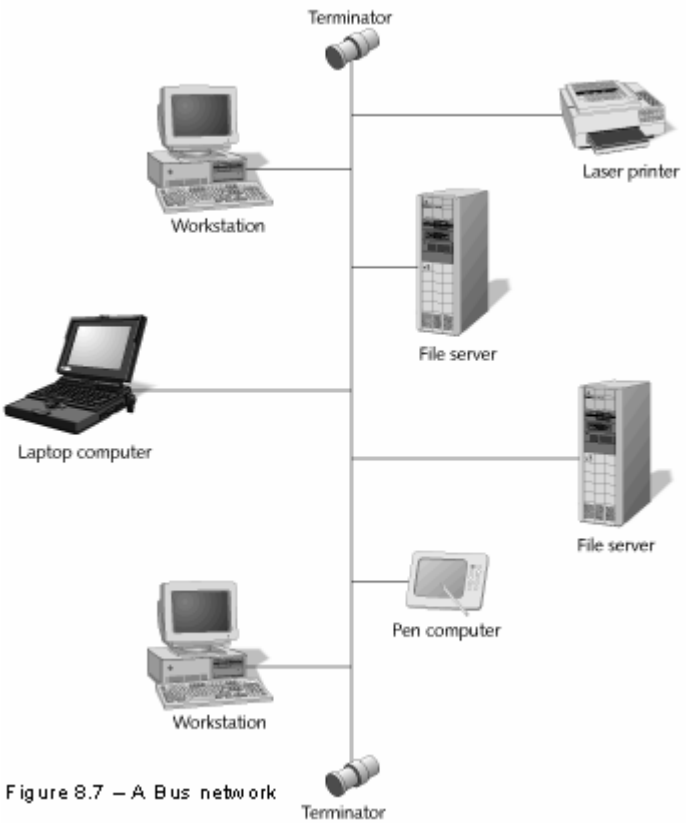
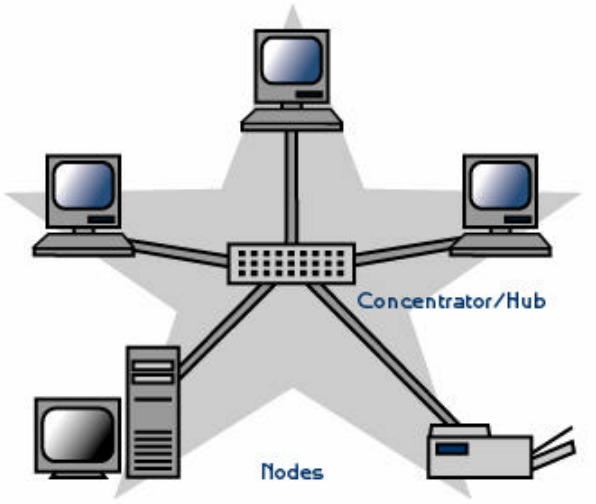


Figure 8.7 – A Bus network

7.6.2 Star Topology

The star topology is a network configured with a central hub and individual cable

segments are connected to the hub, like the shape of a star (figure 7.8). A hub is a central device used in the star topology that joins single cable segments or individual LANs into one network. The Ethernet (also referred as IEEE 802.3) is a dominant LAN network which is based on the star topology. Ethernet networks are built using UTP Cat5 or CAT 5e cables and in modern Ethernet networks you will find a device called a Switch instead of a Hub.



7.8 – A Star network

Managing and locating cable or node problems are easier in star networks. A star network can be easily expanded by having a hub/switch with larger number of ports. If number of ports in a switch is not enough more switches can be added and they can also be connected to the central switch as another set of nodes. Therefore these types of networks are suitable for enterprise networking where larger number of hosts resides within the same LAN. Higher data rates can be achieved by using switches, (allows concurrent connections) instead of Hubs. However, if the central hub or switch fails the entire network will become inaccessible. Since every node is directly connected to the hub/switch more cables are needed than a Bus network.

7.6.3 Ring Topology

A ring topology is a network in the shape of a ring or a circle, with nodes connected around the ring (figure 7.9a). Most of the ring networks are built using coaxial cables and BNC connectors. Fibre based ring networks are also being used.

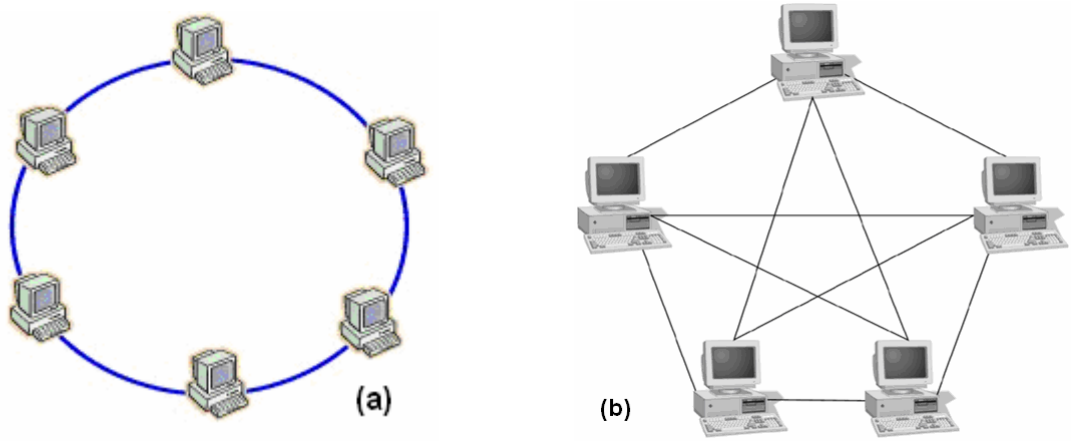


Figure 7.9 – (a) A Ring network, (b) A Mesh network

Locating node and cable errors in a ring network is easier compared to a bus network. Ring networks can handle high volume of traffic over a long distance. Since each node can be reached by two paths these networks are very reliable. On the other hand they require more network cables than a bus network. Use of ring networks are less compared to bus and star networks.

7.6.4 Mesh Topology

In a mesh network each and every network node is interconnected (figure 7.9b). As a result a node can be reached by multiple paths providing high level of fault tolerance. Due to use of large number of cables cost of such a network can be high. In addition, locating malfunctioning cables will be problematic. Internet is the best example for a mesh network.

Hybrid networks are built by combining multiple network topologies. Figure 7.10 illustrates such a Bus-Star hybrid network. The modern Ether networks that span more than one hub/switch are of this type.

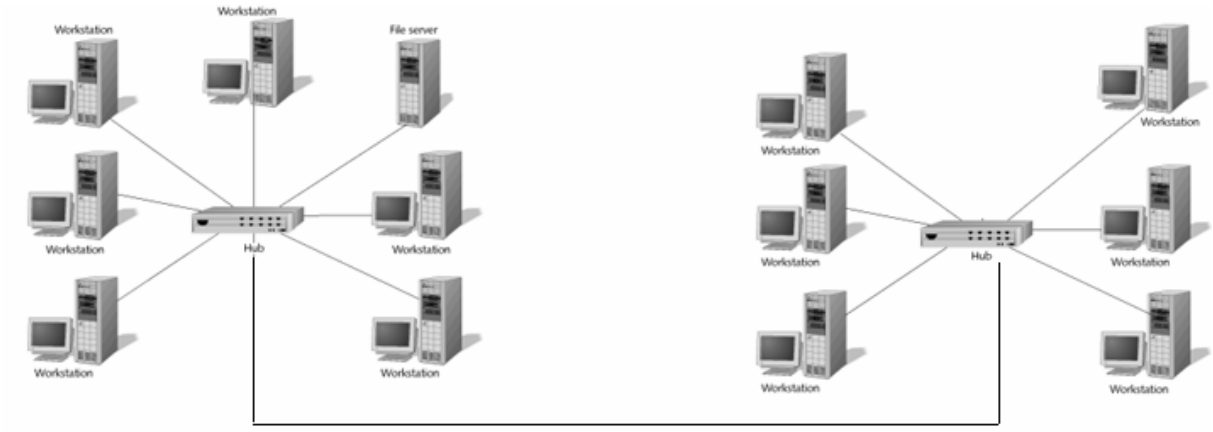


Figure 7.10 – Bus-Star topology

7.7 The Internet

The Internet is made up of millions of computers linked together around the world in such a way that information can be sent from any computer to any other 24 hours a day. These computers can be in homes, schools, universities, government departments, or businesses, small and large. They can be any type of computer, for example personal computers or workstations in a school or servers and mainframes in a company network. It is the largest information store in the world, holding vast amount of up-to-date information. The Internet is often described as 'a network of networks' because all the smaller networks of organisations are linked together into one gigantic network called the Internet.

7.7.1 The History

Internet was a successor of the network called ARPANET (Advanced Research Projects Agency NETwork) which was initiated as a research project and it was funded by US Defense Department. During the cold war the US military set up an automated missile guard system all over the country and they wanted to connect each of these missile stations through a network. Later this network was used to connect many research institutions around the country starting from University of California – Los Angeles, Stanford Research Institute, University of California – Santa Barbara and University of Utah.

With ARPANET users were able to log into a remote computer, print to a remote printer and transfer files between computers. Even with this limited set of capabilities, the network was an instant success and more and more institutions were connected over time. As a result of further research e-mail was introduced and it revolutionised world of communication.

Then the most significant development was the introduction of TCP/IP (Transmission Control Protocol/Internet Protocol). This set of network standards serve as the basis for the "network of networks" (i.e. the Internet) that was to follow and eventually it made ARPANET obsolete. Later on in early 1990's the Internet was expanded beyond USA and several other countries were added to it. Then it became more of a commercial network and customers were given access to the Internet through commercial Internet Service Providers (ISP). With the realisation of the commercial value of the Internet more and more commercial sites were added and further research was carried out to implement new services and protocols.

In 1995 Internet was introduced in Sri Lanka and it has had a significant impact in our country, academia and industry over the past decade. All Sri Lankan academic and research institutions are linked by a network called LEARN (Lanka Education And Research Network) which was established

in 1990. This is effectively part of the Internet. The web, email and other services can be easily exchanged within all national intuitions. LEARNmail is the first e-mail service in Sri Lanka, which was operated by the Department of Computer Science and Engineering, University of Moratuwa.

The key aspect about the Internet is, it is not owned by anyone. It is handled by mutual agreements between countries, ISPs and organisations like IETF (The Internet Engineering Task Force) and ICANN (Internet Corporation for Assigned Names and Numbers)

7.7.2 Internet Services

Internet is just an infrastructure which allows multiple services to run on top of them. Therefore the number of services that could be offered is limitless. Most of these services are generic while there are some very specific services which are used only by specific companies, research institutions or military. Some of the services are e-Mail, World Wide Web, file transfer, e-Leaning, Internet chat, discussion groups, Internet phone capabilities, video conferencing, news groups, interactive multimedia, games, and so on.

Electronic Mail (e-mail)

Electronic mail, sometimes called e-mail, is a computer based method of sending messages from one computer user to another. These messages usually consist of individual pieces of text that a user can send to another computer user even if the other user is not logged in at the time that the message is sent. The message can then be read at a later time. This procedure is analogous to sending and receiving a letter.

When e-mail is received on a computer system, it is usually stored in an electronic mailbox for the recipient to read later. Electronic mailboxes are usually special files on a computer (this computer is referred as the mail server) which can be accessed using various commands. Each user normally has their individual mailbox.

An e-mail message can be sent to a single recipient, a number of recipients or a mailing list[25]. Originally, e-mail messages were restricted to simple text, but now many systems can handle more complicated formats, such as graphics, video clips and word processed documents.

It is straight forward to send electronic mail between users of different computer systems which are connected to major networks. Almost all major academic and research institutions and companies throughout the world can now be reached by electronic mail. It has become a tool for business, research and academic communication. It is also widely being used for personal communication. It has become a key tool in communication because of following reasons:

- Its much faster and convenient than surface mail
- Can be delivered anywhere in the world in a few minutes (occasional delays are possible)
- The receiver does not need to be online while the e-mail is sent
- All received mails are kept in a mailbox until it is read by the recipient. However a mail box has a limited capacity.

In order to access e-mail you need the following components; a computer, communication equipment (modems, network cards) to connect to the mail server, an Internet/network connection and finally some specific software called an e-mail client. Some of the well known e-mail clients are Microsoft Outlook, Outlook Express, Eudora, Thunderbird, and Hena Kurulla (Sinhala version of Thunderbird).

Webmail is a web application that allows users to read and write e-mail using a web browser. You can read or send e-mail from anywhere in the world provided that you have access to the web. Most of these webmail systems provide advance features like grouping and organising mails, keeping an address book, calendar, spellchecking, small notepad, etc. Hotmail, Yahoo and Gmail are some of the well known webmail systems that provide free e-mail access to anybody who wants an e-mail account.

An e-mail address has two parts and each is separate by the @ symbol. You can only send an e-mail to a valid address, if not the e-mail will be returned to the sender. Consider the following example:

[25] A list of e-mail addresses identified by a single name, such as *firstyear@mrt.ac.lk.* When an e-mail message is sent to the mailing list name, it is automatically forwarded to all the addresses in the list.

A typical e-mail message has several header fields and a message body (figure 7.11). The meaning of each head is as follows:

- From: - E-mail address of the sender
- To: - E-mail address of the receiver(s)
- Cc: - Cc is an abbreviation for carbon copy. A copy of the message is sent to the recipient(s) give in Cc, and the recipient's name is visible to other recipients of the message.

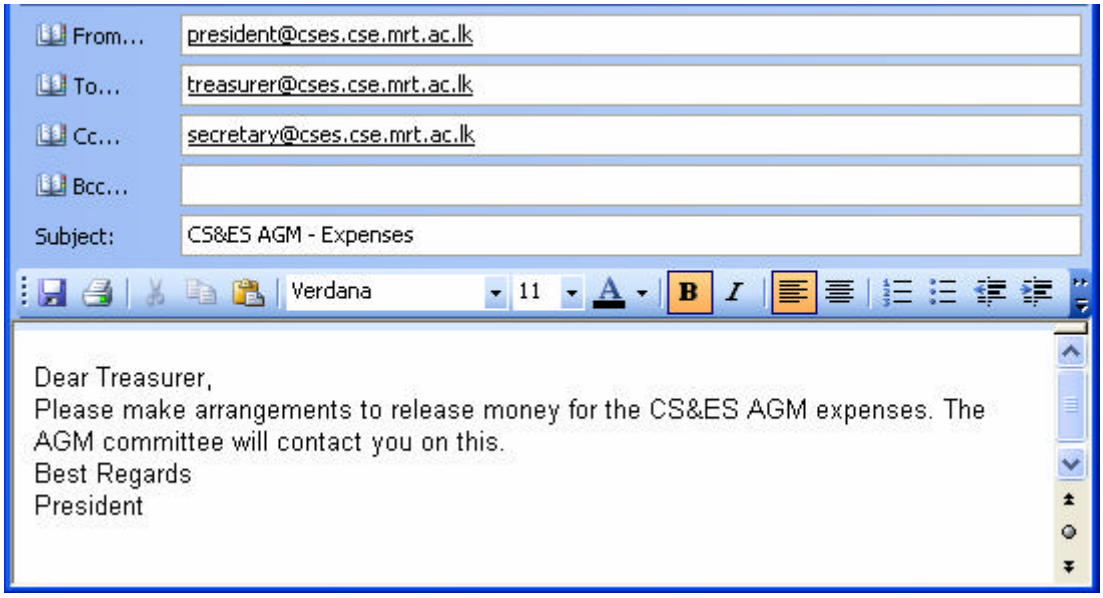


Figure 7.11 – A typical e-mail message

- Bcc: - Bcc is an abbreviation for blind carbon copy. A copy of the message is sent to that recipient, but the recipient's name is not visible to other recipients of the message.
- Subject: - Title given for the message

World Wide Web

The words Internet and World Wide Web (WWW, or just the Web) are used synonymously but they are different. The WWW is another service running on top of the Internet that presents information in a graphical interface. The WWW can be considered as the illustrated version of the Internet. It began in the late 1980's when physicist Dr. Berners-Lee wrote a small computer program for his own personal use. This program allowed pages, within his computer, to be linked together using keywords. It soon became possible to link documents in different computers, as long as they were connected to the Internet. The document formatting language used to link documents is called HTML (HyperText Markup Language) and a protocol called HTTP (HyperText Transfer Protocol) was used to transfer these HTML documents across the Internet.

The Web remained primarily text based until 1992. Two events occurred during this year that changed the way the Web looked forever. Marc Andersen developed a new computer program called the NCSA Mosaic and gave it away! The NCSA Mosaic was the first Web browser. The browser made it easier to access different Web sites that had started to appear.

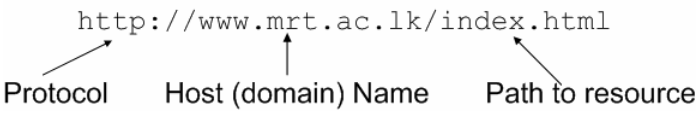
Soon Web sites contained more than just text. They also had images, sound and video. More and more commercial web sites were added with the introduction of e-Commerce (electronic Commerce). The development of the web has boosted the popularity of the Internet and is also the easiest part of the Internet to use.

In order to access web you need the following components; a computer, communication equipments (modems, network cards) to connect to a web server(s), an Internet/network connection and finally some specific software called a web browser. Some of the well known web browsers are Microsoft

Internet Explorer, Mozilla Firefox, Netscape, Konqueror and Opera. Following are some of the terms that you may come across while accessing the web:

- Web page - A single disk file with a single file name
- Web site - A collection of one or more web pages
- Home page - The first page of a web site
- Web server - A computer on the Internet containing one or more web sites
- Web browser - A special application that is used to see web pages
- Web development tools - Applications that are used to develop web pages and web sites

A web page can be accessed by giving its unique address over the Internet called the URL (Uniform Resource Locator). Format of a URL is:



Since Web contains vast amount of up-to-date information it would be difficult to locate the exact information that a user wants. In order to help this process another type of web sites called the search engines are available. An Internet search engine helps users to find web pages on a given subject. The search engines maintain databases of web sites and use programs (often referred to as "spiders" or "robots") to collect information, which is then indexed by the search engine. Similar services are provided by "directories," which maintain ordered lists of websites.

Content can be searched in two different ways. A keyword based search allows the user to search the content based on a specific word(s) that the user is interested in. In a directory based search engine contents can be found by browsing some set of categories. Google, Yahoo and AltaVista are some of the well known search engines.

File Transfer

Files can be transferred from one computer to another over the Internet. The File Transfer Protocol (FTP) is used for this purpose. It allows you to upload or download a file. For large file transmission users are required to have high speed and reliable network/Internet connections.

Instant Messaging

Instance messaging allows real time exchange of messages among two or more users. This is normally done as chat. These are text based systems and what ever is typed and sent by the sender is displayed to the recipient(s). When multiple users are involved in the same chat it is called a conference. In order to chat, users are required to have a chat client like Yahoo messenger, MSN messenger or Google talk or a web based chat system like kaputa.com or sinhalaya.com.

Multimedia over the Internet

Internet can be used for one way transmission of audio or video content. Data is streamed from a streaming server (output as it comes out) and millions of users can receive it at the same time (referred as multicast). Some of these audio or video broadcasts can be pre-recorded or live. A faster Internet/network connection is required to get a good quality sound or video.

Internet Phone

Two or more people can talk (voice) over the Internet. This is normally referred as IP telephony. It can either be from one PC to another PC or from one PC to a land or mobile phone. Today even phone to PC is also becoming popular. Sound quality of such a system is not so good. However, if you have a faster Internet/network connection the conversation will be much clear. These calls cost much lower than standard IDD calls. Skype and Yahoo Messenger are some of the examples.

Electronic Learning (e-Learning)

Education offered using electronic delivery methods such as CD-ROMs, video conferencing, websites and e-mail is referred as e-learning. Internet is full of vast amount of learning material that includes text, graphics, animations, audio, video and interactive activities. Some of this content may be public

while some web sites provide access only to registered users. Today you may even register with a foreign university and follow your entire diploma or degree online.

7.7.3 Connecting to the Internet

In order to connect to the Internet you need to get an Internet connection from a local ISP. These Internet connections can be post-paid or prepaid. Some of the well known ISPs operating in Sri Lanka are; SLT, Suntel, Lanka Bell, Lanka Internet, CeyCom and Dialog.

Then there should be a way that you can connect to your ISP. These connections can be full time connections such as leased lines, ISDN (Integrated Services Digital Network) or ADSL (Asymmetric Digital Subscriber Line) or part time connections such as Dial-up links. The mobile service providers provide facilities like WAP (Wireless Application Protocol) and GPRS (General Packet Radio Service) for accessing Internet over mobile phones. VSAT (Very Small Aperture Terminal) is a satellite based data communication system which you can use to access Internet for, any part of the world.

Business users need full time dedicated connections wile a part-time connection is acceptable for a home user. ADSL is a low cost full time connection that can be used by small business owners and home users. In Sri Lanka Dial-up connections can be used in every part of the country but connections such as ADSL is still only available around Colombo and some other cities.

7.7.4 Security on the Internet

Security over the Internet can be a major concern since it provides open connection to everyone and all forms of content. If you are connected to the Internet mean you are connected to the rest of the world. Thus there are various threats that you need to be aware of. The major problems would be spam mails and viruses.

Spam mails are indiscriminately sent unwelcome, unwanted, irrelevant, or inappropriate messages, specially used for commercial advertising in mass quantities. These are also referred to as junk mail. A computer virus is a software program capable of reproducing itself and usually capable of causing great harm to files or other programs on the same computer. To be safe from spam mails do not provide your e-mail address to various websites that you are not really keen on receiving any advertisements or news letters. They key to be safe from viruses is to install an antivirus software (also called a virus guard) on your machine and update it frequently. Also do not download any software, files or e-mail attachment that you are not sure about or from an unknown sender.

On the other hand there are people called hackers (the correct term is crackers) that would try to get into other peoples computers (mostly to machines in banks, military or research institutions) and who try to steal information such as credit card numbers, personal information , etc. These could affect the integrity or confidentiality of Internet users or customers. These problems can be prevented up to a certain extend by installing special software called a firewall.

People may post pornographic content over the web and in a way it has become one of the major businesses in the Internet. But having pop-up web pages on such content would be a problem and most people will not welcome those. It would be annoying and embarrassing when those automatically come up when other people are around. To overcome such problems you can configure your web browser to block any unwanted pop-up messages. This could be a major problem for parents with kids. Access to such web sites or content can be controlled by installing some content filtering programs.

Internet is vast, up-to-date and open; it is your responsibility to make use of it for the betterment of yourself as well as for the rest of the world.

References

3 – Data Representation

- Figure 3.2 - The Unicode website – www.unicode.org

5 – Introduction to Computer Hardware

- Scott Mueller's "Upgrading and Repairing PCs", 13th edition and "Microsoft Encarta Encyclopaedia Deluxe 2001".
- Figure 5.1 – Intel Corporation
- Figure 5.2 – Microsoft Encarta Encyclopaedia
- Figure 5.3 – Scott Mueller's "Upgrading and Repairing PCs"
- Figure 5.7 – Scott Mueller's "Upgrading and Repairing PCs"
- Figure 5.12 – Scott Mueller's "Upgrading and Repairing PCs"
- Figure 5.17 – Scott Mueller's "Upgrading and Repairing PCs"
- Figure 5.29 – Wikipedia, 27/07/06, – http://en.wikipedia.org/wiki/Multicore
- Figure 5.31 – Scott Mueller's "Upgrading and Repairing PCs"
- Figure 5.33 – Scott Mueller's "Upgrading and Repairing PCs"
- Figure 5.35 – Scott Mueller's "Upgrading and Repairing PCs"
- Figure 5.38 – Find out what is inside a floppy disk, 13/03/06, http://www.exploratorium.edu/science_explorer/dissect_disk.html
- Figure 5.41 – Scott Mueller's "Upgrading and Repairing PCs"
- Figure 5.43 – Scott Mueller's "Upgrading and Repairing PCs"
- Figure 5.47 – How stuff work, 13/03/06, http://computer.howstuffworks.com/laser-printer.htm

6 – Operating Systems and Application Software

- Figure 6.2 – Andrew S. Tanenbaum, Modern Operating Systems, 2nd Edition
- Figure 6.4 – Smart Card Supply, 27/07/06, http://www.smartcardsupply.com/

7 – Introduction to Networking

- Figure 7.8 – An educator's Guide to School Networks, 27/07/06, http://fcit.usf.edu/network/
- Figure 7.9 – Free Network+ 2005 Training , 27/07/06, http://www.learnthat.com/certification/

Note: For most of the other pictures/diagrams original author cannot be found.