1.1 Introduction

An operating system is a program that manages the computer hardware. It also provides a basis for application programs and acts as an intermediary between a user of a computer and the computer hardware. An operating system is an important part of almost every computer system. A computer system can be divided roughly into four components: the h*ardware*, the *operating system*, the *application programs*, and the *users* (Figure 1.1)

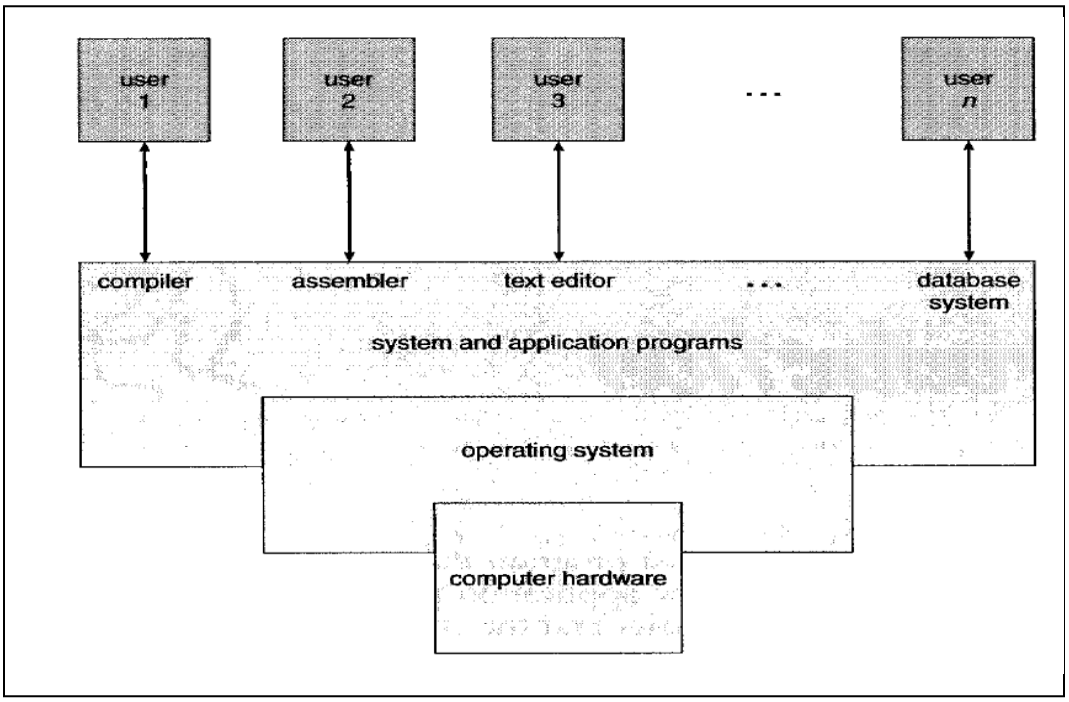

Fig.1.1 Abstract view of the components of a computer system.

The components of a computer system are its hardware, software, and data. The operating system provides the means for the proper use of these resources in the operation of the computer system. **An** operating system is similar to a *government.*

Like a government, it performs no useful function by itself. It simply provides an *environment* within which other programs can do useful work. Operating systems can be explored from two viewpoints: the user and the system.

1.1.1 User View

The user view of the computer varies by the interface being used. Most computer users sit in front of a PC, consisting of a monitor, keyboard, mouse, and system unit. Such a system is designed for one user to monopolize its resources, to maximize the work (or play) that the user is performing. In this case, the operating system is designed mostly for ease of use, with some attention paid to performance, and none paid to resource utilization. Performance is important to the user, but it does not matter if most of the system is sitting idle. Some users sit at a terminal connected to a mainframe or minicomputer. Other users are accessing the same computer through other terminals. These users share resources and may exchange information. The operating system is designed to maximize resource utilization-to assure that all available CPU time, memory, and **I/O** are used efficiently, and that no individual user takes more than her fair share. Other users sit at workstations, connected to networks of other workstations and servers. These users have dedicated resources at their disposal, but they also share resources such as networking and servers-file, compute and print servers. Therefore, their operating system is designed to compromise between individual usability and resource utilization.

1.1.2 System View

From the computer's point of view, the operating system is the program that is most intimate with the hardware. We can view an operating system as a resource allocator. A computer system has many resources-hardware and software-that may

be required to solve a problem: CPU time, memory space, file-storage space, **I/O** devices, and so on. The operating system acts as the manager of these resources and must decide how to allocate them to specific programs and users so that it can operate the computer system efficiently and fairly. A slightly different view of an operating system emphasizes the need to control the various **I/O** devices and user programs. An operating system is a control program. A control program manages the execution of user programs to prevent errors and improper use of the computer. Operating systems exist because they are a reasonable way to solve the problem of creating a usable computing system. The fundamental goal of computer systems is to execute user programs and to make solving user problems easier. Toward this goal, computer hardware is constructed. Since hardware alone is not particularly easy to use, the common functions of controlling and allocating resources are then brought together into one piece of software: the operating system.

A more common definition is that the operating system is the one program running at all times on the computer (usually called the kernel), with all else being application programs.

1.1.3 System Goals

The primary goal of some operating system is *convenience for the user*. Operating systems exist because they are supposed to make it easier to compute with them than without them. This view is particularly clear when you look at operating systems for small PCs.

The primary goal of other operating systems is *efficient* operation of the computer system. This is the case for large, shared, multiuser systems. These systems are expensive, so it is desirable to make them as efficient as possible. These two goals-convenience and efficiency-are sometimes contradictory.

1.2 Mainframe Systems

Mainframe computer systems were the first computers used to tackle many commercial and scientific applications. In this section, we trace the growth of mainframe systems from simple batch systems, where the computer runs one -and only one-application, to time-shared systems, which allow for user interaction with the computer system.

1.2.1 Batch Systems

Early computers were physically enormous machines. The common input devices were card readers and tape drives. The common output devices were line printers, tape drives, and card punches. The user did not interact directly with the computer systems. Rather, the user prepared a job which consisted of the program, the data, and some control information about the nature of the job (control cards) and submitted it to the computer operator. The job was usually in the form of punch cards. At some later time (after minutes, hours, or days), the output appeared. The output consisted of the result of the program, as well as a dump of the final memory and register contents for debugging.The operating system in these early computers was fairly simple. Its major task was to transfer control automatically from one job to the next. The operating system was always resident in memory (Figure 1.2).

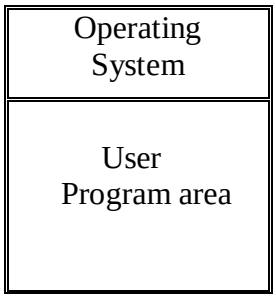

Figure 1.2 Memory layouts for a simple batch system.

To speed up processing, operators batched together jobs with similar needs and ran them through the computer as a group. In this execution environment, the CPU is often idle, because the speeds of the mechanical I/O devices are slower than electronic devices.

The introduction of disk technology allowed the operating system to keep all jobs on a disk, rather than in a serial card reader. With direct access to several jobs, the operating system could perform job scheduling, to use resources and perform tasks efficiently.