1 Basic concepts of computer systems

Modern computer systems are constructed from various different component parts interconnected to produce machines of various different characteristics. These computer systems are used for a wide variety of tasks from applications such as word processing, spreadsheet and database, to controlling a mobile phone or a passenger jet. This chapter will introduce the features that are common to all these different systems.

1.1 COMPUTERS AND COMPUTER SYSTEMS

Electronic computer systems have come a long way since their initial development in the late 1940s. Systems are now capable of storing vast amounts of information and processing that information at great speed. The Cray TE-350 is capable of performing over 3,000,000,000,000 floating point calculations per second (www.cray.com). Considerable advances have also been made in the area of miniaturisation: for example a web server has been developed at Stanford Wearables Lab that can be fitted into a matchbox (wearables.stanford.edu). Figure 1.1 shows pictures of these systems, the web server is shown standing on a computer keyboard.

These advances in computer system performance have also been matched by significant developments in software technology, which allows systems to perform more complex and diverse tasks than ever before. These new power-hungry applications have created an unceasing demand for increased system performance. The development of faster processors, quicker memory with increased storage capacity and more complex and demanding software continues unabated.

This demand for speed is particularly an issue where systems are expected to display some degree of intelligence, for example, speech and video processing and the understanding of text. One of the ultimate aims of a lot of research in this area is to develop computer systems that will communicate like a human being (i.e., understand and reproduce natural spoken language). Another area where the demand for computing power is practically limitless is scientific modelling, simulation and prediction. For example, weather-forecasting systems are required to run models using vast amounts of data to produce results within reasonable time scales.

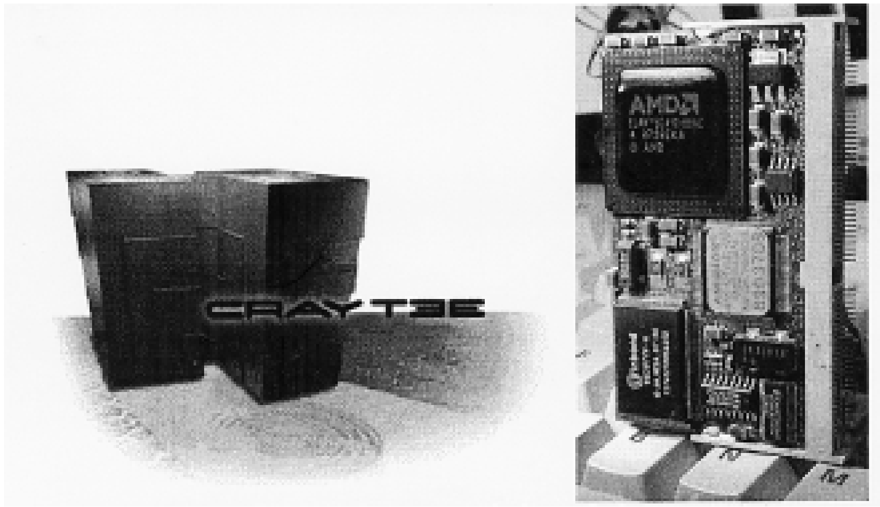

Figure 1.1 Large and small computer systems *left* Cray computer; *right* miniature web server

With all the advances in hardware performance it is important to note that software applications have not always increased in performance (efficiency) and a lot of new applications in fact run slower (for a given hardware platform) than their predecessors. This is due to a number of factors including more complex software requirements (e.g., word-processors that support embedded graphics), less emphasis on optimisation, greater memory requirements (particularly for the operating system) and introduction of new slower technologies such as interpreted languages like Java and Javascript. This means that the importance of improving hardware performance is as important as ever and will be for the foreseeable future.

1.1.1 WHAT IS A COMPUTER?

A computer is a device which given a set of instructions can be used to perform a given task or tasks. The computer reads in information (instructions or data), does some processing and generally stores or outputs some result. The computer has internal storage (for its instructions and data) and also has mechanisms to communicate with the outside world, input for reading data into the computer and output for writing data out.

1.1.2 COMPUTER SYSTEMS

A computer system is the combination of a computer plus software plus support hardware, which work together to perform a task. Computer systems can be linked together with other computer systems using a computer network. The computer network transfers data between the systems allowing them to communicate. Most computer systems nowadays provide connectivity to a network, more often than not to the Internet (see 1.3.3).

A set of computer systems connected to a network, and configured to work together on a given task is referred to as a distributed system. One of the advantages of a distributed system is that many computers can be linked together to provide a level of performance and/or storage capability that would be beyond any of the computers working alone. An example of a distributed computer system is the World Wide Web; this uses millions of computers, which work together storing and delivering data across the world using the Internet to communicate.

Computer systems can be classified as general purpose or dedicated. An example of a general-purpose computer is the PC. It can be used for a wide variety of tasks – for example word-processing (which was used to write this book), spreadsheets and database applications. The computer system is loaded up with different software depending on the task required.

A dedicated computer system is designed for a single purpose only, with the software being loaded up at manufacture. For dedicated systems designed to control a piece of electronic equipment the software is usually stored on a chip. These are referred to as embedded systems as the computer is 'embedded' within another product (e.g. a videocassette recorder, washing machine). An example of an embedded system is the cell phone. The system needs to read the keypad, control the display, set up calls, update the phone's SIM card and many other functions. Both dedicated and general purpose computer systems are similar in their internal design but they do differ somewhat in their features and physical design (imagine carrying a desktop PC around to make your phone calls).

1.2 BASIC COMPONENTS OF A COMPUTER SYSTEM

Figure 1.2 shows a block diagram of a basic computer system. This is a highly simplified representation of the computer system displaying the most important components.

1.2.1 CENTRAL PROCESSING UNIT

The Central Processing Unit (CPU) controls the operation of the computer. It is responsible for carrying out (executing) lists of instructions (software)

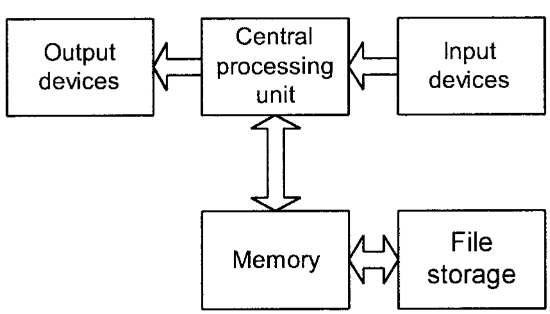

Figure 1.2 Computer system block diagram

stored in the computer's memory. The CPU can also write data to output devices, read data from input devices and store and retrieve data from the system's memory and file storage.

Modern CPUs can execute multiple instructions at the same time (concurrent processing). Another method used is to split the execution of individual instructions into more than one process allowing more than one instruction to be worked on at the same time. This technique, called pipelining, is similar to how a production line is used to work on multiple products at the same time (e.g., in a car production line). Some systems use a number of processors connected together on the same circuit board in what is termed Multi Process Architecture. Many network servers use this technique to allow them to provide data for many client computers simultaneously

In this book examples will be provided using both the Motorola PowerPC® and Intel Pentium® processors. These are two of the most commonly used processors in general purpose computing and, in the case of the Intel Pentium®, the most commonly used processor for personal computing.

1.2.2 MEMORY

The computer's memory (or primary storage) is used to store two types of information, instructions and data. The instructions are read from the memory by the CPU, which then carries them out in sequence. The list of instructions that is stored in memory for a given task is called a computer program. The program that the CPU executes processes a set of data that is stored in memory. For example with a word-processing program the data would be the document that is currently being edited.

Both the instructions and data are stored in memory cells called memory locations, see Figure 1.3. Each of the locations is referred to using a number called its address. The first address is 0, the second 1, etc. The

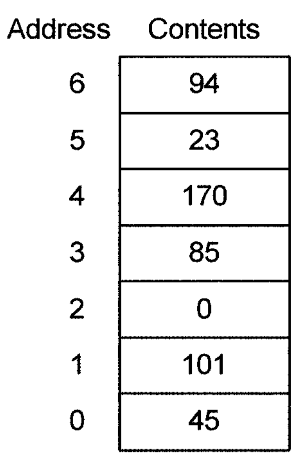

Figure 1.3 Main memory organisation

number stored in the memory location is referred to as the location's content. In Figure 1.3 the location with address 5 contains the number 23.

Each memory location is only capable of storing numbers. If characters or other information have to be stored then this must be converted to a numeric form first. Each of the instructions that the CPU is capable of executing is allocated a number (called its opcode). This allows the computer's program to be stored in the memory. CPUs cannot distinguish between locations containing data, code or just random garbage. This means that it is crucial for memory to be loaded with the correct data and instructions before the CPU can do any useful work.

Memory can provide data to the CPU very quickly. Typically access time is measured in nanoseconds (one nanosecond = 1/1,000,000,000 seconds). The Motorola MCM63P737A, for example provides access times as short as 2.5 nanoseconds. The faster the memory chip access time the faster the CPU can operate. One problem with fast memory is that is relatively expensive so most systems use a mixture of fast and slower memory employing a technique called caching.

1.2.3 FILE STORAGE

File (or secondary) storage provides bulk storage for the computer system. The storage capacity of file storage is typically a hundred times the capacity of the computer's memory and is also cheaper per unit of storage than memory by at least ten fold.

The programs and data are stored in file storage in logical units called files. These files are organised and grouped together to form a file system. The file system on most computer systems is controlled in such a way that users cannot access certain files that they do not have the correct permissions for. This allows the secure storage of sensitive data for multi-user systems.

File storage has much slower access times than the memory (typically at least a thousand times slower). For this reason the programs must be loaded up into memory before being executed by the processor. Since main memory is relatively expensive, the computer system is not provided with enough to store all the programs the user might want to execute. File storage is used to keep copies of data and programs not currently being used by the computer system. When required the computer loads the program into memory. File storage (unlike the memory) does not lose data when the computer system is switched off making it suitable for offline storage of programs and data.

1.2.4 PERIPHERALS

The computer system needs a way of communicating with the outside world. Peripherals are classified into three groups: input devices (e.g., keyboard or mouse), output devices (e.g., VDU or printer) and I/O devices that can do both functions (e.g., a network card or modem). Peripheral technology has advanced at a very rapid pace with headsets that allow the projection of 3D graphics (for Virtual reality) and colour printing technology that can produce images of photographic quality.

The correct selection and connection of peripherals for the computer system is critical to its functionality. For example, a system that needs to produce photo images would require a very high-resolution colour printer. Some peripherals such as a radio transceiver or keypad would be only relevant to dedicated systems such as cellular phone. The peripherals are connected to the main system via a standard connection called an interface. The selection of an appropriate interface for each peripheral is an important design decision when constructing the system.

1.3 PUTTING IT ALL TOGETHER

So far we have discussed each of the component parts of the computer system in isolation. When designing a computer system choices have to made according to budget, purpose of the system, required processing power etc. Each component must be chosen appropriately and decisions made about component interconnection, interfacing and software. In general this involves a trade off between budget and processing power (faster computer hardware is more expensive) but a judicious choice of software and well-optimised hardware can also produce a lot of gains in system performance with no increase in cost. We will now discuss in brief some of the design decisions necessary when building a system.

1.3.1 INTERCONNECTION

The way that the system components are interconnected can affect overall performance considerably. By providing dedicated links (mesh interconnection) between each pair of components the speed of data transfer can be increased substantially. The problem with this approach is that it increases the system cost and complexity, since each component will require interface electronics to every other component. Also it will be difficult to manufacture the hardware since the number of interconnecting wires will be very great.

At the other extreme all the components can be connected to a common channel called a bus. This is cheaper than the mesh approach but slower. This slowness is due to the fact that each component has to share the bus and wait for others to finish access before they can start their transfer. Figure 1.4 shows the two configurations. In fact, a compromise between these two extremes is commonly used with systems providing multiple buses for component interconnection.

1.3.2 OPERATING SYSTEM

Most computer systems when delivered come complete with a basic set of software called the operating system. The operating system is responsible for interfacing at a low level with the system's hardware and controlling the execution of programs within the computer. Operating systems can be classified in various ways; for example, single tasking (capable of executing only one program at a time) and multitasking (capable of executing many programs simultaneously). Another distinction is between single and multi-

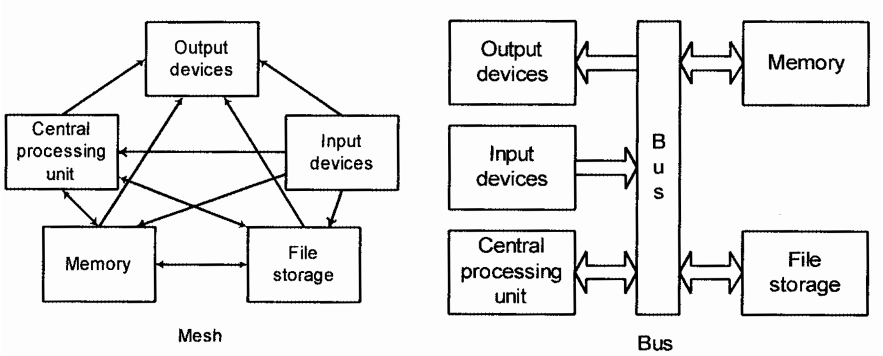

Figure 1.4 System interconnection

user operating systems (multi-user systems differentiate between different users of the system; for example stopping one user accessing or deleting another user's files).

Operating system functions include low level control of hardware, sharing memory securely between programs, scheduling the execution of different tasks (multitasking systems only) and facilitating communication across a network. The correct choice of operating system can make a great deal of difference to overall system performance in terms of the optimal use of memory, speed of execution and system reliability. A vigorous debate has been raging for some time between advocates of both the Windows and Linux operating systems. This book does not intend to add to that discussion.

Some operating systems are designed to run on particular hardware platforms while others are available for a wider range of CPUs and different configurations. The latter are termed portable. If a single operating system is required for a set of systems that will be using many different hardware platforms, finding a portable operating system will be of critical concern.

1.3.3 NETWORKS AND DISTRIBUTED SYSTEMS

These days most computer systems do not work alone but are connected to a network allowing them to communicate with others. Reliable and efficient intercommunication between systems which are using different hardware and software and may be located far apart, is a non-trivial task. A lot of work has been done in this area, particularly in producing standards which can be used to facilitate communication between heterogeneous systems. Another important development in networking is the increased performance and use of the Internet, a global network, which provides connectivity for millions of computer users.

The communication between systems is controlled by a set of rules and standards called a protocol. For two machines to communicate they must both 'talk' the same protocol. The most commonly used family of protocols is TCP/IP. This is used by the Internet and many other networks to provide the reliable transfer of data.

As explained before, in a distributed system many computers are interconnected and co-operate with each other to perform a particular task. An example of such a system would be a distributed database in which data is stored in various different locations across the network. This has a number of advantages;, for example, data may be accessible quickly from many different locations and the system may still work if one or more computers fail. That given, distributed database systems are more complex to maintain and require complex protocols to ensure all copies of the data is kept up to date. Distributed systems can also have their performance affected by the speed (bandwidth) limitations of the network they are connected to. This is particularly a problem with global networks (such as the Internet) as many systems may be sharing the same long-distance connection.

1.4 SUMMARY

In this chapter we have discussed what the computer system can do, how it is constructed and have described in brief some of its component parts. We have also seen how the computer system consists of both hardware and software and how both must work together to achieve optimal performance. In Chapters 2 to 11 we will examine in detail the function of the systems component parts and how they are integrated to produce a functioning whole. Chapters 12–15 cover distributed systems and networking, explaining how computers are interconnected and can be made to work together.

1.5 EXERCISES

1 Find out in what ways computer system automation have been used to improve the following public services:
health provision
education
policing
emergency services
a library

2 Find out how distributed computer systems have been used to solve computationally difficult problems. (*Hint*: look for SETI on the World Wide Web.) Try and find three different applications and describe the distributed systems used.

3 What are the forces that drive the search for faster and more efficient computer hardware?