Student notes 2.3

Memory

We saw in Section 2.2 that the processor fetches instructions and data in binary format. This means that the instructions and data need to be stored ready for the processor to fetch. This storage is called the primary memory. The processor can also store the result of the processing in primary memory.

In this section, we study the different kinds of primary memory. The primary memory, is the computer storage that the processor can access directly. Most computer systems also have a certain amount of secondary storage that is used to store programs and data for later use, but cannot be accessed directly by the processor. We will look at this type of storage in Section 2.5.

RAM and ROM

The two main types of primary memory are RAM and ROM.

RAM is Random Access Memory. This means that you can access any item you want in the memory.

ROM is Read-Only Memory. This means that you cannot change what is stored in the memory.

It is important to realise that these names only describe a small element of what they are, and they are not a definition. There are, for example, many different types of memory which have random access, but are not RAM.

Instead, we should consider what the memory is used for.

RAM

RAM is used to store the programs and data that the computer is currently working on. It is the working memory of the computer.

Most computers use an operating system. The operating system needs to be in the RAM so that it can run. If you want to run a program or listen to a piece of music, it will need to be loaded into the RAM first. If you type into a text file, the text file will also be in the RAM. Any changes you make to the file will be made in the RAM until you save the file.

RAM is volatile. This means that when the power is turned off, everything that is stored in the RAM is lost. (This is why, if you turn off your computer before saving a file, you may lose the data in that file.) RAM is volatile because the memory is stored using electronic circuits which need electricity to operate.

If a computer has a lot of RAM, this means that it can deal with larger programs and larger data files. More usually, this also means that the computer can multi-task more. This means the computer is dealing with several programs and files at the same time.

Continued

For example, suppose you are listening to music while researching something on a website and writing your homework into a word processor. For this to happen, the RAM needs to contain:

- the operating system of your computer
- your music player software
- the music file you are listening to
- your web browser software
- the website you are currently browsing
- your word-processor software
- the word-processed file of your homework.

Typically, your RAM will also need to contain additional items such as your anti-virus software to make sure that you do not accidentally download a virus from the internet.

We can see that the more RAM you have, the more the computer will be capable of. Therefore, one way of improving the performance of a computer is to incease the amount of RAM. In a standard desktop computer this is usually done by adding or replacing memory modules on the motherboard. This allows some computers to have up to 8 GB or more of RAM.

ROM

When you turn on a computer, the processor needs to begin fetching and executing instructions. However, the RAM is empty because it is volatile, and any instructions which were stored in the RAM were lost when the computer was turned off. This is where the ROM comes in.

The ROM contains the program that is used to start up the computer. This is also called the boot program because the process of starting up the computer is also called 'booting up' the computer. In many desktop computer systems, the boot program is also called the BIOS or Basic Input Output System.

The circuits inside the computer are designed so that when the computer is turned on the processor first runs the program in the ROM. ROM is not volatile and the program is preserved, even when the computer is turned off. Also, the program inside the ROM is not easily deleted or changed. This means the boot program will always be there when the computer is started. This is why any computer system needs some ROM.

On a standard personal computer, the boot program does the following things:

- It performs some basic checks.
- It finds the operating system and loads it into the RAM.
- It then hands control over to the operating system.

Continued

A program which does just these things does not need to be very big. In 2010, the typical amount of ROM in a personal computer was 1 or 2 MB.

After the operating system kicks in, the boot program closes and the ROM is no longer needed until the next time the computer is started again.

Other computer-based devices, such as MP3 players and mobile phones, also need some ROM to boot up the device. However, since these systems are designed for a particular use, the whole operating system is usually embedded into the ROM. This means that the ROM is used all the time, not just when the device is turned on.

ROM is called Read-Only Memory, because when used normally, the computer does not change the program stored in it. Many modern computers use flash memory for the ROM. This allows the contents to be updated using a special process known as 'flashing'. We will discuss flash memory later.

Differences between RAM and ROM

The table below summarises the differences between RAM and ROM.

	RAM	ROM
What does it contain?	Operating system, programs and data which are currently being worked on by the processor.	A program used to start the computer called the 'boot program'.
Is it volatile?	Yes	No
How big is it?	Typically quite large and measured in gigabytes. The larger the better because this means that the computer can work on more things at the same time.	Usually small because it only needs to store the boot program. It is usually measured in megabytes.
Can the contents be changed?	Yes, the contents of the RAM are changed all the time, while the computer is running.	The contents of ROM cannot normally be changed except by a special 'flashing' operation.

Virtual memory

Virtual memory is a part of the hard drive (which is secondary storage) that is reserved to be used as though it were an extention of the RAM. This is useful when the RAM in the computer is not enough to contain all the programs and data that a user wants to use.

This can happen, for example, when you have too many programs and files open. Although, there are many programs and files open, the computer is only using a few of

Continued

these at a time. Some of the files and programs are just in a waiting state, until the user activates that program, for example by clicking on it.

The computer can take these 'waiting' programs out of the RAM and store them in a specially reserved part of the hard drive. This then frees some RAM which can be now be used to open more programs.

When any of the 'waiting' programs that have been transfered to the hard disk are needed again, they need to be transfered back into the RAM first before they can be used. If it is necessary to make space for it in the RAM, some other 'waiting' program is transfered to the hard disk at the same time.

Other types in memory in modern computer systems

Cache memory

Cache memory is memory that is located on the processor itself and is used to make the computer faster.

The cache memory acts as a buffer between the processor and the RAM. When the processor fetches data from the RAM, it is first copied onto the cache memory, and then from the cache memory into the main processor itself. The transfer between the cache memory and the main processor is much faster than the transfer between the RAM and the cache memory.

When the processor tries to fetch the data again, it first checks whether the data is already in the cache memory. If it is, then the data is fetched from the cache memory and not from the RAM.

This way, by storing the data that is fetched most often so that it does not need to fetched from the RAM, the cache memory makes the whole system a lot faster. The larger the cache memory, the more the potential improvement in performance will be. For this reason, modern processors include up to 8 MB or more of cache memory.

Flash memory

Flash memory is a type of memory where the data is stored using transistors in electrical circuits in a non-volatile way. It is a form of EEPROM or 'Electronically Erasable Programmable Read-Only Memory'. This means that while in normal use, you can read any data from it, and, unlike older types of ROM, you can also erase the data. This is usually done by applying an electric current to a large block of data at a time. New data can then be programmed into the memory.

In modern computers, the ROM is usually implemented using flash memory. This allows the boot program which is stored in the ROM to be erased and replaced with a new version when necessary. This is what happens when you upgrade the BIOS of a computer, or when you upgrade the software of a phone. The ROM in the computer is erased and replaced with the new version. This must be done with care because if the boot program is

Continued

stored incorrectly, the computer cannot start. This is why it is usually done by expert users.

Recent developments in flash memory have greatly improved the speed with which the contents can be erased and written, and this has made it possible to use flash memory in many new ways. For example, many portable devices use internal flash memory to store the operating system and store user data such as music files and contact details. In this way, the flash memory is acting like secondary storage (which we discuss in the next section), but at the same time it has the advantage of ROM because the operating system can always be found at the beginning of the flash memory, when the device is turned on.