3.3 Data storage

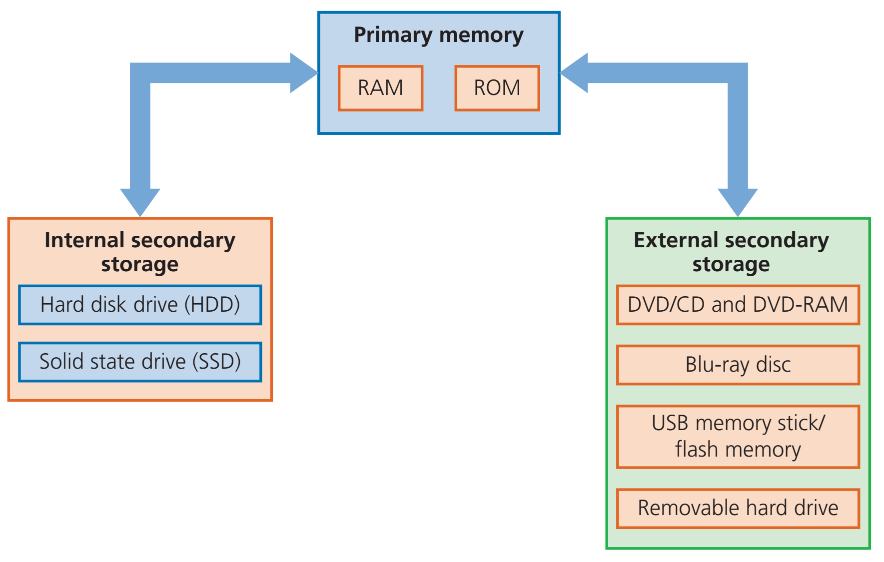
All computers require some form of memory and storage. Memory is usually referred to as the internal devices used to store data that the computer can access directly. This is also known as primary memory. This memory can be the user's workspace, temporary data or data that is key to running the computer.

Storage devices allow users to store applications, data and files. The user's data is stored permanently and they can change it or read it as they wish. Storage needs to be larger than internal memory since the user may wish to store large files (such as music files or videos). Storage devices can also be removable to allow data, for example, to be transferred between computers. Removable devices allow a user to store important data in a different location in case of data loss.

However, all of this removeable storage has become less important with the advent of technology such as 'data drop' (which uses Bluetooth) and cloud storage. Figure 3.57 summarises the types of memory and storage devices covered in this section.

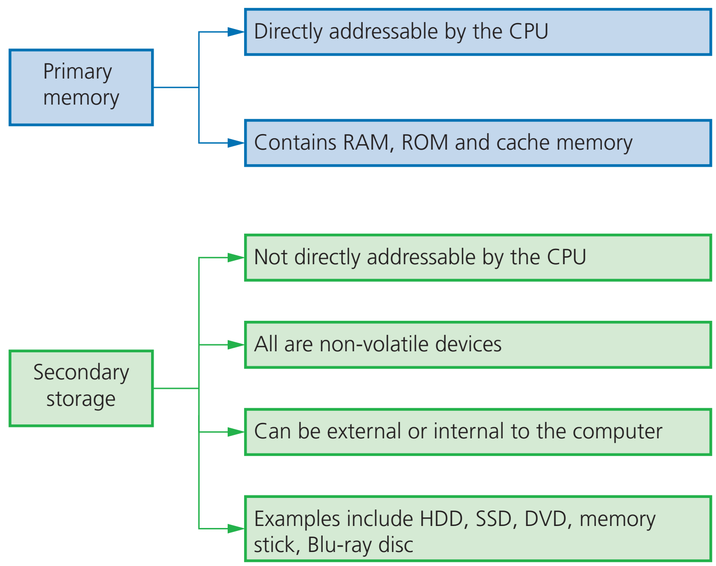
Memory and storage devices can be split up into two distinct groups:

- primary memory
- secondary storage.



▲ **Figure 3.57** Typical memory and storage devices

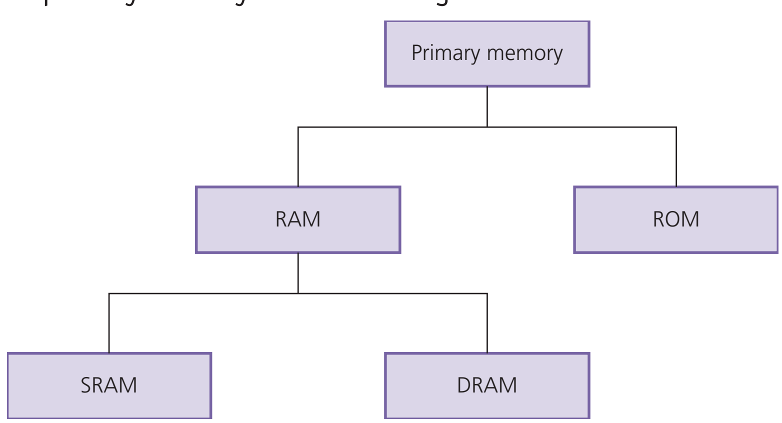
Here is a summary of the differences between primary memory and storage devices:



▲ **Figure 3.58** Summary of primary, secondary and off-line devices

3.3.1 Primary memory

Primary memory is the part of the computer memory which can be accessed directly from the CPU; this includes **random access memory (RAM)** and **read-only memory (ROM)** memory chips. Primary memory allows the CPU to access applications and services temporarily stored in memory locations. The structure of primary memory is shown in Figure 3.59.



▲ **Figure 3.59** Primary memory

Random access memory (RAM)

All computer systems come with some form of RAM. These memory devices are not really random; this refers to the fact that any memory location in RAM can be accessed independent of which memory location was last used. When you run an application or program, data is retrieved from secondary storage and placed temporarily into RAM. Access time to locate data is much faster in RAM than in secondary or off-line devices. Features of RAM include:

- can be written to or read from, and the data can be changed by the user or the computer (i.e. it is a temporary memory)

- used to store data, files, part of an application or part of the operating system **currently in use**
- it is **volatile**, which means memory contents are lost when powering off the computer.

In general, the larger the size of RAM the faster the computer will operate. In reality, RAM never runs out of memory; it continues to operate but just becomes slower and slower as more data is stored. As RAM becomes 'full', the CPU has to continually access the secondary data storage devices to overwrite **old** data on RAM with **new** data. By increasing the RAM size, the number of times this has to be done is considerably reduced; thus making the computer operate more quickly.

There are currently two types of RAM technology:

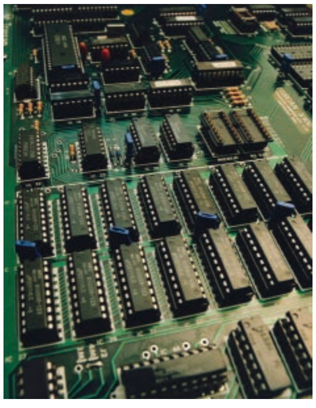
- dynamic RAM (DRAM)
- static RAM (SRAM).

Dynamic RAM (DRAM)

▲ **Figure 3.60** DRAM

Each DRAM chip consists of transistors and capacitors. Each of these parts is tiny since a single RAM chip will contain millions of transistors and capacitors. The function of each part is:

- capacitor – this holds the bits of information (0 or 1)
- transistor – this acts like a switch; it allows the chip control circuitry to read the capacitor or change the capacitor's value.



▲ **Figure 3.61** SRAM

This type of RAM needs to be constantly **refreshed** (that is, the capacitor needs to be re-charged every 15 microseconds otherwise it would lose its value). If it wasn't refreshed, the capacitor's charge would leak away very quickly leaving every capacitor with the value 0.

DRAMs have a number of advantages over SRAMs:

- they are much less expensive to manufacture than SRAM
- they consume less power than SRAM
- they have a higher memory capacity than SRAM.

Static RAM (SRAM)

A major difference between SRAM and DRAM is that SRAM doesn't need to be constantly refreshed.

It makes use of **flip flops**, which hold each bit of memory.

SRAM is much faster than DRAM when it comes to data access (typically, access time for SRAM is 25 nanoseconds and for DRAM is 60 nanoseconds).

DRAM is the most common type of RAM used in computers, but where absolute speed is essential, for example, in the CPU's memory cache, SRAM is the preferred technology. Memory cache is a high-speed portion of the memory; it is effective because most programs access the same data or instructions many times. By keeping as much of this information as possible in SRAM, the computer avoids having to access the slower DRAM.

Table 3.9 summarises the differences between DRAM and SRAM.

▼ **Table 3.9** Differences between DRAM and SRAM

DRAM	SRAM
consists of a number of transistors and capacitors	uses flip flops to hold each bit of memory
needs to be constantly refreshed	doesn't need to be constantly refreshed
less expensive to manufacture than SRAM	has a faster data access time than DRAM
has a higher memory capacity than SRAM	CPU memory cache makes use of SRAM
main memory is constructed from DRAM	
consumes less power than SRAM	

Read-only memory (ROM)

Another form of primary memory is read-only memory (ROM). This is similar to RAM in that it shares some of its properties, but the main difference is that it cannot be changed or written to. ROM chips have the following features:

- they are non-volatile (the contents are not lost after powering off the computer)
- they are permanent memories (the contents cannot be changed or written to by the user, the computer or any application/program)
- the contents can only be read
- they are often used to store data that the computer needs to access when powering up for the first time (the basic input/output system (BIOS)); these are known as the start-up instructions (or bootstrap)

Here is a summary of the main differences between RAM and ROM:

▼ **Table 3.10** RAM and ROM features

RAM	ROM
temporary memory device	permanent memory device
volatile memory	non-volatile memory device
can be written to and read from	data stored cannot be altered
used to store data, files, programs, part of OS **currently** in use	always used to store BIOS and other data needed at start up
can be increased in size to improve operational speed of a computer	

Example of an application

We will now consider an application, other than a computer, where both RAM and ROM chips are used:

A remote-controlled toy car has circuitry which contains both RAM and ROM chips. The remote control is a hand-held device. Explain the function of the RAM and ROM chip in this application.

We will consider the function of each type of memory independently:

ROM

- storing the factory settings such as remote control frequencies
- storing the 'start-up' routines when the toy car is first switched on
- storing of the set routines; for example, how the buttons on the hand-held device control turning left, acceleration, stopping, and so on.

RAM

- the user may wish to program in their own routines; these new instructions would be stored in the RAM chip
- the RAM chip will store the data/instructions received from the remote control unit.

Activity 3.6

1 Describe how ROM and RAM chips could be used in the following devices:
   a a microwave oven
   b a refrigerator
   c a remote-controlled model aeroplane; the movement of the aeroplane is controlled by a hand-held device.

3.3.2 Secondary and off-line storage

Secondary (and off-line) storage includes storage devices that are not directly addressable by the CPU. They are non-volatile devices that allow data to be stored as long as required by the user. This type of storage can store more data than primary memory, but data access time is considerably longer than with RAM or ROM. All applications, the operating system, device drivers and general files (for example, documents, photos and music) are stored on secondary storage. The following section discusses the various types of secondary storage that can be found on the majority of computers.

3.3.3 Magnetic, optical and solid-state storage

Secondary (and off-line) storage falls into three categories according to the technology used:

- magnetic
- solid state
- optical.

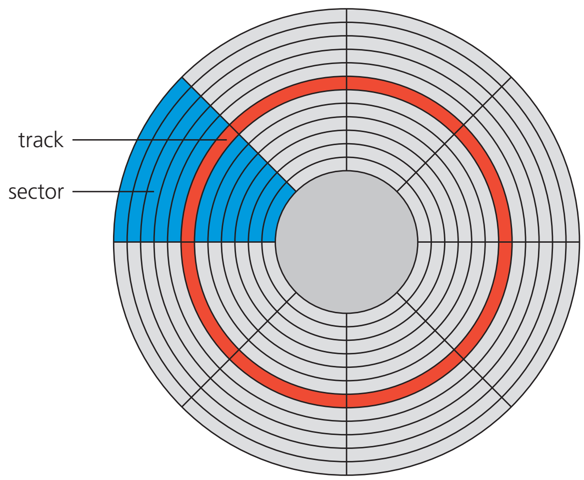
▲ **Figure 3.62** HDD

Magnetic storage

Hard Disk Drives (HDD)

Hard disk drives (HDD) are still one of the most common methods used to store data on a computer.

Data is stored in a digital format on the magnetic surfaces of the disks (or platters, as they are frequently called). The hard disk drive will have a number of platters that can spin at about 7000 times a second. Read-write heads consist of electromagnets that are used to read data from or write data to the platters. Platters can be made from aluminium, glass or a ceramic material. A number of read-write heads can access all of the surfaces of the platters in the disk drive. Normally each platter will have two surfaces which can be used to store data. These read-write heads can move very quickly – typically they can move from the centre of the disk to the edge of the disk (and back again) 50 times a second.



▲ **Figure 3.63** Tracks and sectors

Data is stored on the surface in sectors and tracks. A sector on a given track will contain a fixed number of bytes. Unfortunately, hard disk drives have very slow data access when compared to, for example, RAM. Many applications require the read-write heads to constantly look for the correct blocks of data; this means a large number of head movements. The effects of **latency** then become very significant. Latency is defined as the time it takes for a specific block of data on a data track to rotate around to the read-write head.

Users will sometimes notice the effect of latency when they see messages such as 'Please wait' or, at its worst, 'not responding'.

When a file or data is stored on a HDD, the required number of sectors needed to store the data will be allocated. However, the sectors allocated may not be adjacent to each other. Through time, the HDD will undergo numerous deletions and editing which leads to sectors becoming increasingly **fragmented** resulting in a gradual deterioration of the HDD performance (in other words, it takes longer and longer to access data). Defragmentation software can improve on this situation by 'tidying up' the disk sectors.

Link

See Section 4.1.1 for more on defragmentation.

All data in a given sector on a HDD will be read in order (that is, sequentially); however, access to the sector itself will be by a direct read/write head movement.

Removable hard disk drives are essentially HDDs external to the computer that can be connected to the computer using one of the USB ports. In this way, they can be used as a back-up device or another way of transferring files between computers.

Solid state drives (SSD)

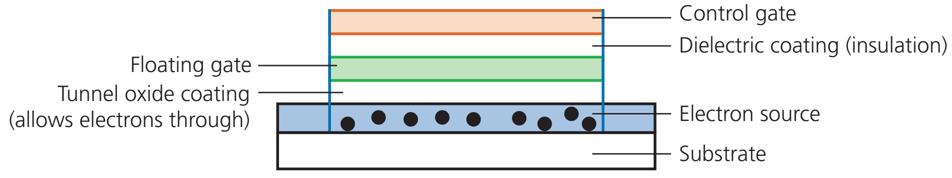
Latency is an issue in HDDs as described earlier. Solid state drives (SSD) remove this issue considerably since they have no moving parts and all data is retrieved at the same rate. They don't rely on magnetic properties; the most common type of solid state storage devices store data by controlling the movement of electrons within NAND or NOR chips. The data is stored as 0s and 1s in millions of tiny transistors (at each junction one transistor is called a floating gate and the other is called a control gate) within the chip. This effectively produces a non-volatile rewritable memory.

Link

For more on NAND and NOR gates see Chapter 10.

Floating gate and control gate transistors

Floating gate and control gate transistors use CMOS (complementary metal oxide semi-conductor) NAND technology. Flash memories make use of a matrix; at each intersection on the matrix there is a floating gate and a control gate arranged as follows:



▲ **Figure 3.64** Flash memory

A dielectric coating separates the two transistors, which allows the floating gate transistor to retain its charge (which is why the memory is non-volatile). The floating gate transistor has a value of 1 when it is charged and a value of 0 when it isn't. To program one of these 'intersection cells' a voltage is applied to the control gate and electrons from the electron source are attracted to it. But due to the dielectric coating, the electrons become trapped in the floating gate. Hence, we have control over the bit value stored at each intersection. (**Note:** After about 12 months, this charge can leak away, which is why a solid state device should be used at least once a year to be certain it will retain its memory.)

The main benefits of this newer solid state technology over hard disk drives are:

- they are more reliable (no moving parts to go wrong)
- they are considerably lighter (which makes them suitable for laptops)
- they don't have to 'get up to speed' before they work properly
- they have a lower power consumption
- they run much cooler than HDDs (both these points again make them very suitable for laptop computers)
- because of no moving parts, they are very thin
- data access is considerably faster than HDD.

The main drawback of SSD is still the longevity of the technology (although this is becoming less of an issue). Most solid state storage devices are conservatively rated at only 20 GB of write operations per day over a three year period – this is known as **SSD endurance**. For this reason, SSD technology is still not used in all servers, for example, where a huge number of write operations take place every day. However, the durability of these solid state systems is being improved by a number of manufacturers and they are rapidly becoming more common in applications such as servers and **cloud storage** devices.

Note: It is also not possible to over-write existing data on a flash memory device; it is necessary to first erase the old data and then write the new data at the same location.

Memory sticks/flash memories

Memory sticks/flash memories (also known as pen drives) use solid state technology.

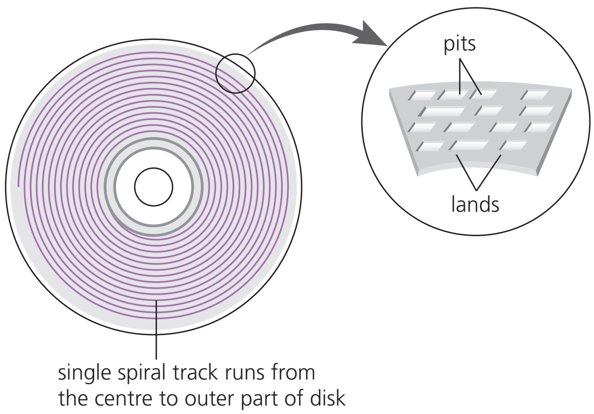
They usually connect to the computer through the USB port. Their main advantage is that they are very small, lightweight devices, which make them very suitable as a method for transferring files between computers. They can also be used as small back-up devices for music or photo files, for example.

Complex or expensive software, such as financial planning software, often uses memory sticks as a dongle. The dongle contains additional files that are needed to run the software. Without this dongle, the software won't work properly. It therefore prevents illegal or unauthorised use of the software, and also prevents copying of the software since, without the dongle, it is useless.

Optical media

CD/DVD disks

CDs and **DVDs** are described as **optical storage devices**. Laser light is used to read and write data to and from the surface of the disk.

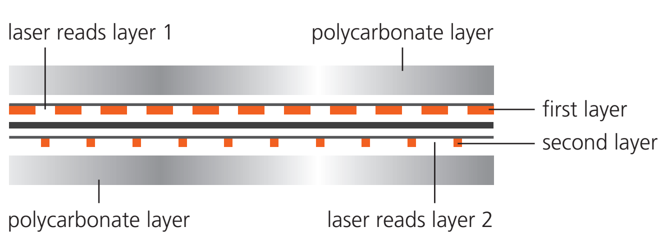


▲ **Figure 3.65** Optical media

Both CDs and DVDs use a thin layer of metal alloy or light-sensitive organic dye to store the data. As can be seen from the diagram in Figure 3.65, both systems use a single, spiral track which runs from the centre of the disk to the edge. When a disk spins, the optical head moves to the point where the laser beam 'contacts' the disk surface and follows the spiral track from the centre outwards. As with a HDD, a CD/DVD is divided into sectors allowing direct access to data. Also, as in the case of HDD, the outer part of the disk runs faster than the inner part of the disk.

The data is stored in 'pits' and 'lands' on the spiral track. A red laser is used to read and write the data. CDs and DVDs can be designated 'R' (write once only) or 'RW' (can be written to or read from many times).

DVD technology is slightly different to that used in CDs. One of the main differences is the potential for **dual-layering**, which considerably increases the storage capacity. Basically, this means that there are two individual recording layers. Two layers of a standard DVD are joined together with a transparent (polycarbonate) spacer, and a very thin reflector is also sandwiched between the two layers. Reading and writing of the second layer is done by a red laser focusing at a fraction of a millimetre difference compared to the first layer.



▲ **Figure 3.66** Dual-layering on a DVD

Standard, single layer DVDs still have a larger storage capacity than CDs because the 'pit' size and track width are both smaller. This means that more data can be stored on the DVD surface. DVDs use lasers with a wavelength of 650 nanometres; CDs use lasers with a wavelength of 780 nanometres. The shorter the wavelength of the laser light, the greater the storage capacity of the medium.

Blu-ray discs

Blu-ray discs are another example of optical storage media. However, they are fundamentally different to DVDs in their construction and in the way they carry out read-write operations.

Note: it is probably worth mentioning why they are called Blu-ray rather than Blue-ray; the simple reason is it was impossible to copyright the word 'Blue' and hence the use of the word 'Blu'.

The main differences between DVD and Blu-ray are:

▲ **Figure 3.67** Blu-ray disc

- a blue laser, rather than a red laser, is used to carry out read and write operations; the wavelength of blue light is only 405 nanometres (compared to 650 nm for red light)
- using blue laser light means that the 'pits' and 'lands' can be much smaller; consequently, Blu-ray can store up to five times more data than normal DVD
- single-layer Blu-ray discs use a 1.2 mm thick polycarbonate disk; however, dual-layer Blu-ray and normal DVDs both use a sandwich of two 0.6 mm thick disks (i.e. 1.2 mm thick)
- Blu-ray disks automatically come with a secure encryption system that helps to prevent piracy and copyright infringement
- the data transfer rate for a DVD is 10 Mbps and for a Blu-ray disc it is 36 Mbps (this equates to 1.5 hours to transfer 25 GiB of data).

Since Blu-ray discs can come in single layer or dual-layer format (unlike DVD, which is always dual-layer), it is probably worth also comparing the differences in capacity and interactivity of the two technologies.

Comparison of the capacity and interactivity of DVDs and Blu-ray discs

- A standard dual-layer DVD has a storage capacity of 4.7 GB (enough to store a 2-hour standard definition movie)
- A single-layer Blu-ray disc has a storage capacity of 27 GB (enough to store a 2-hour high definition movie or 13 hours of standard definition movies)
- A dual-layer Blu-ray disc has a storage capacity of 50 GB (enough to store 4.5 hours of high definition movies or 20 hours of standard definition movies).

Blu-ray allows greater interactivity than DVDs. For example, with Blu-ray, it is possible to:

- record high definition television programs
- skip quickly to any part of the disc
- create playlists of recorded movies and television programmes
- edit or re-order programmes recorded on the disc
- automatically search for empty space on the disc to avoid over-recording
- access websites and download subtitles and other interesting features.

Finally, Table 3.11 summarises the main differences between CDs, DVDs and Blu-ray.

All these optical storage media are used as back-up systems (for photos, music and multimedia files). This also means that CDs and DVDs can be used to transfer files between computers. Manufacturers sometimes supply their software (e.g. printer drivers) using CDs and DVDs. When the software is supplied in this way, the disk is usually in a read-only format.

▼ **Table 3.11** Comparison of CD, DVD and Blu-ray (Note: nm = 10^{-9} metres and μm = 10^{-6} metres)

Disk type	Laser colour	Wavelength of laser light	Disk construction	Track pitch (distance between tracks)
CD	Red	780 nm	single 1.2 mm polycarbonate layer	1.60 μm
DVD (dual-layer)	Red	650 nm	two 0.6 mm polycarbonate layers	0.74 μm
Blu-ray (single layer)	Blue	405 mm	single 1.2 mm polycarbonate layer	0.30 μm
Blu-ray (dual-layer)	Blue	405 nm	two 0.6 mm polycarbonate layers	0.30 μm

The most common use of DVD and Blu-ray is the supply of movies or games. The memory capacity of CDs isn't big enough to store most movies (see earlier comparison notes).

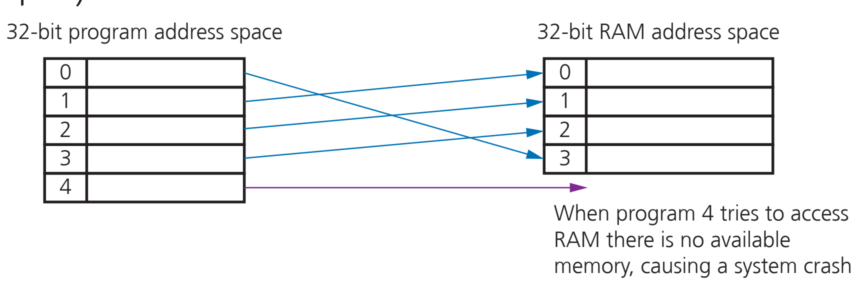
3.3.4 Virtual memory

One of the problems associated with memory management is the case when processes run out of RAM. If the amount of available RAM is exceeded due to multiple programs running, it is likely to cause a system crash. This can be solved by utilising the hard disk drive (or SSD) if we need more memory. This is the basis behind **virtual memory**. Essentially RAM is the ***physical memory***, while ***virtual memory*** is RAM + swap space on the hard disk or SSD.

To execute a program, data is loaded into memory from HDD (or SSD) whenever required. It is possible to show the differences between using normal memory management and virtual memory management in two simple diagrams.

Without virtual memory

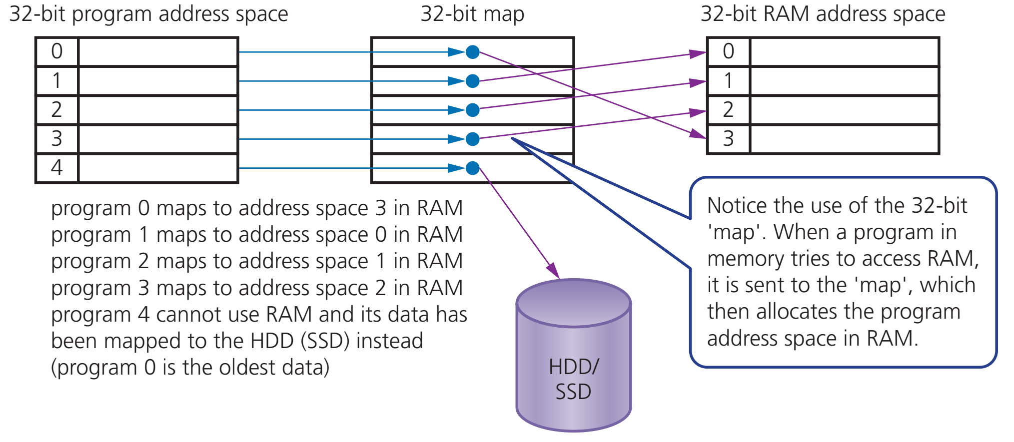
Suppose we have five programs (numbered 0 to 4) that are in memory, all requiring access to RAM. The first diagram shows what would happen without virtual memory being used (i.e. the computer would run out of RAM memory space):



▲ **Figure 3.68** Normal memory management

With virtual memory

We will now consider what happens if the CPU uses virtual memory to allow all five programs to access RAM as required. This will require moving data out of RAM into HDD/SSD and then allowing other data to be moved out of HDD/SSD into RAM:

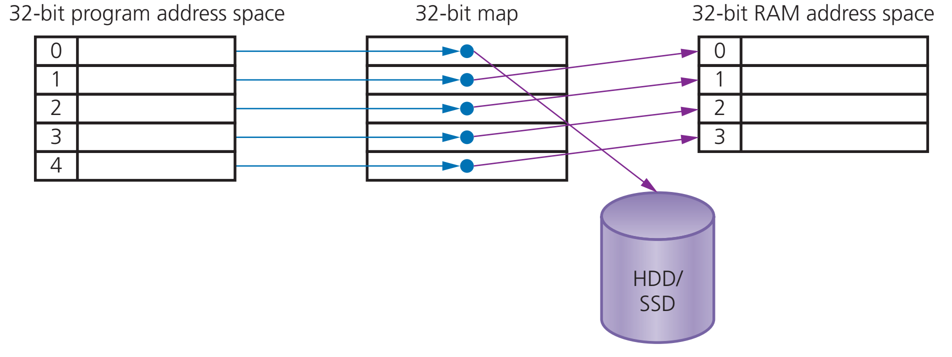


▲ **Figure 3.69** Status just before program 4 is given RAM space

Virtual memory now moves the oldest data out of RAM into the HDD/SSD to allow program 4 to gain access to RAM. The 32-bit 'map' is now updated to reflect this new situation:

- data from program 0 (which was using RAM address space 3 – the oldest data) is now mapped to the HDD/SSD instead, leaving address space 3 free for use by program 4
- program 4 now maps to address space 3 in RAM, which means program 4 now has access to RAM.

Our diagram now changes to:



▲ **Figure 3.70** Status with program 0 now mapped to HDD and program 4 has access to RAM

All of this will continue to occur until RAM is no longer being over-utilised by the competing programs running in memory. Virtual memory gives the illusion of unlimited memory being available. Even though RAM is full, data can be moved in and out of the HDD/SSD to give the illusion that there is still memory available. In computer operating systems, **paging** is used by memory management to store and retrieve data from HDD/SSD and copy it into RAM. A **page** is a fixed-length consecutive (or contiguous) block of data utilised in virtual memory systems.

This is a key part of how virtual memory works allowing **data blocks** (pages) to be moved in and out of a HDD/SSD. However, accessing data in virtual memory is slower so, as mentioned earlier on in this chapter, the larger the RAM the faster the CPU can operate. This is one of the benefits of increasing RAM size as far as possible.

The main benefits of virtual memory are:

- programs can be larger than physical memory and still be executed
- there is no need to waste memory with data that isn't being used (e.g. during error handling)
- it reduces the need to buy and install more expensive RAM memory (although as mentioned earlier there are limits to the value of doing this).

When using HDD for virtual memory the main drawback is **disk thrashing**. As main memory fills, more and more data needs to be swapped in and out of virtual memory leading to a very high rate of hard disk read/write head movements; this is known as disk thrashing. If more and more time is spent on moving data in and out of memory than actually doing any processing, then the processing speed of the computer will be considerably reduced. A point can be reached when the execution of a process comes to a halt since the system is so busy moving data in and out of memory rather than doing any actual execution – this is known as the **thrash point**. Due to large numbers of head movements, this can also lead to premature failure of a hard disk drive. Thrashing can be reduced by installing more RAM, reducing the number of programs running at a time or reducing the size of the swap file. Another way of reducing this problem is to make use of a solid state drive (SSD) rather than using HDD.

3.3.5 Cloud storage

Public and private cloud computing

Cloud storage is a method of data storage where data is stored on remote servers. The same data is stored on more than one server in case of maintenance or repair, allowing clients to access data at any time. This is known as **data redundancy**. The physical environment is owned and managed by a hosting company and may include hundreds of servers in many locations.

There are three common systems:

- Public cloud – this is a storage environment where the customer/client and cloud storage provider are different companies
- Private cloud – this is storage provided by a dedicated environment behind a company firewall; customer/client and cloud storage provider are integrated and operate as a single entity
- Hybrid cloud – this is a combination of the two above environments; some data resides in the private cloud and less sensitive/less commercial data can be accessed from a public cloud storage provider.

Instead of saving data on a local hard disk or other storage device, a user can save their data 'in the cloud'. The benefits and drawbacks of using cloud storage are shown in Table 3.12.

▼ **Table 3.12** Benefits and drawbacks of cloud storage

Benefits of using cloud storage	Drawbacks of using cloud storage
customer/client files stored on the cloud can be accessed at any time from any device anywhere in the world provided internet access is available	if the customer/client has a slow or unstable internet connection, they would have many problems accessing or downloading their data/files
there is no need for a customer/client to carry an external storage device with them, or even use the same computer to store and retrieve information	costs can be high if large storage capacity is required; it can also be expensive to pay for high download/upload data transfer limits with the customer/client internet service provider (ISP)
the cloud provides the user with remote back-up of data with obvious benefits to alleviate data loss/disaster recovery	the potential failure of the cloud storage company is always possible – this poses a risk of loss of all back-up data
if a customer/client has a failure of their hard disk or back-up device, cloud storage will allow recovery of their data	
the cloud system offers almost unlimited storage capacity	

Data security when using cloud storage

Companies that transfer vast amounts of confidential data from their own systems to a cloud service provider are effectively relinquishing control of their own data security. This raises a number of questions:

- what physical security exists regarding the building where the data is housed?
- how good is the cloud service provider's resistance to natural disasters or power cuts?
- what safeguards exist regarding personnel who work for the cloud service company; can they use their authorisation codes to access confidential data for monetary purposes?

Potential data loss when using cloud storage

There is a risk that important and irreplaceable data could be lost from the cloud storage facilities. Actions from hackers (gaining access to accounts or pharming attacks, for example) could lead to loss or corruption of data. Users need to be certain that sufficient safeguards exist to overcome these risks.

The following breaches of security involving some of the largest cloud service providers suggest why some people are nervous of using cloud storage for important files:

- The XEN security threat, which forced several cloud operators to reboot all their cloud servers, was caused by a problem in the XEN hypervisor (a hypervisor is a piece of computer software, firmware or hardware that creates and runs virtual machines).
- A large cloud service provider permanently lost data during a routine back-up procedure.
- The celebrity photos cloud hacking scandal, in which more than 100 private photos of celebrities were leaked. Hackers had gained access to a number of cloud accounts, which enabled them to publish the photos on social networks and sell them to publishing companies.
- In 2016, the National Electoral Institute of Mexico suffered a cloud security breach in which 93 million voter registrations, stored on a central database, were compromised and became publicly available to everyone. To make matters worse, much of the information on this database also linked to a cloud server outside Mexico.

Activity 3.7

1 Name two types of memory used in a mobile phone. For each named memory, describe its purpose in the mobile phone.

2 a Explain what is meant by ***virtual memory***.

b Five programs are currently being run in a computer. Program 1 is using 10 GiB of RAM, program 2 is using 5 GiB of RAM, program 3 is using 12 GiB of RAM and program 4 is using 4 GiB of RAM. The programs are at the stage where program 5 now needs to access RAM, but RAM is presently full (RAM has a 32 GiB maximum capacity). Explain how virtual memory could be used to allow program 5 to access RAM without any of the data from the other four programs being lost.

3 Five descriptions of computer terms are shown on the left and five terms are shown on the right. Draw lines to connect each description to the correct computer term.

storage environment where the client and remote storage provider are different companies	thrashing
high rate of HDD read/write operations causing large amount of head movement	swap space
space on HDD or SSD reserved for data used in virtual memory management	cloud storage
situation where a HDD is so busy doing read/write operations that execution of a process is halted	thrash point
method of data storage where the data is stored on hundreds of off-site servers	public cloud

4 a Give four differences between RAM and ROM chips.

b Give an example of the use of each type of memory.

5 a Use the following words to complete the paragraph below which describes how solid state memories work (each word may be used once, more than once or not at all).

Word list:		
	control gate	NAND
	electrons	negative
	floating gate	positive
	intersection	transistor
	matrix	volatile

Solid state devices control the movement of .... within a .... chip. The device is made up of a .... and at each .... there is a .... and a .... transistor. .... are attracted towards .... when a voltage is applied.

b Give three advantages of using SSD, rather than HDD, which make SSD technology particularly suitable for use in laptop computers.

c Describe one disadvantage of solid state technology.