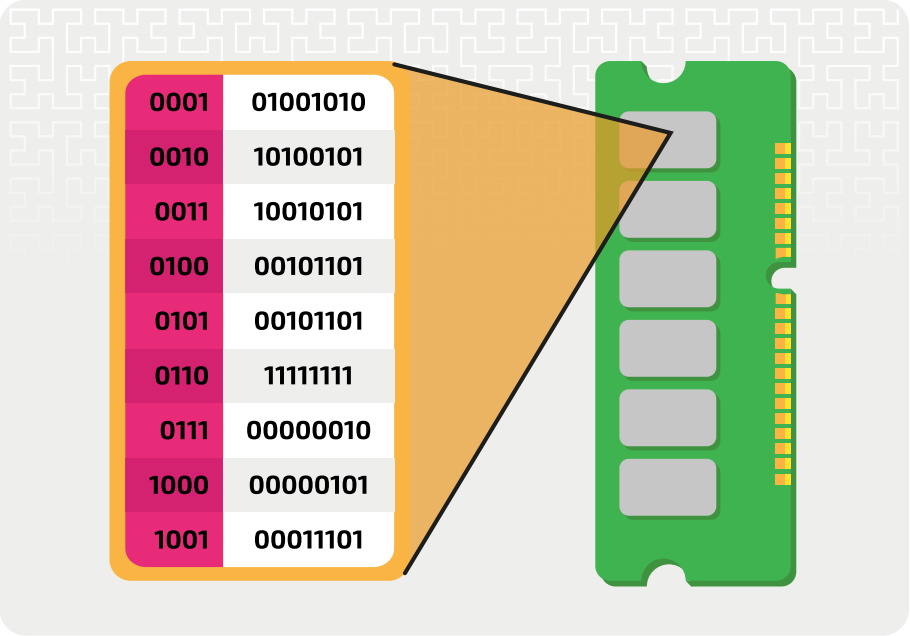
isaac computer science

Main memory

Main memory (sometimes called *primary storage*) refers to storage locations that are directly accessible by the processor. Main memory offers very fast read/write speeds but is typically much lower capacity than secondary storage devices.

GCSE **Random access memory (RAM)**



RAM "sticks" are plugged into the motherboard

The RAM inside a computer holds **all of the data and instructions that are currently being processed**. In order for any program to be executed, it must first be loaded into the RAM from Secondary storage. This is because RAM has quicker read and write speeds than secondary storage devices. The processor will fetch instructions (and any necessary data) from the RAM directly before decoding and executing them.

How does the increased access speed of RAM affect the execution of instructions?

Click a button to show the answer

What is your level of confidence that your own answer is correct?

Low Medium High

Within the RAM, the memory is split up into separate locations, each of which is given a unique **memory address**. The processor will use these addresses to access the data stored in RAM. Any of the locations in RAM can be accessed in the same amount of time, unlike other storage devices where a sensor (such as a read/write head) needs to be moved into place. This is why it is called **random access**: any location can be accessed "at random" with no impact on speed.

RAM is **volatile**, meaning power is required to hold the data, and if the power is switched off all the data stored in RAM is lost. Any data that needs to be stored permanently must be moved out of RAM before a computer is switched off. If you have ever lost a document or piece of code you were working on because you hadn't saved it and your computer switched off, you have seen the impact of RAM's volatility.

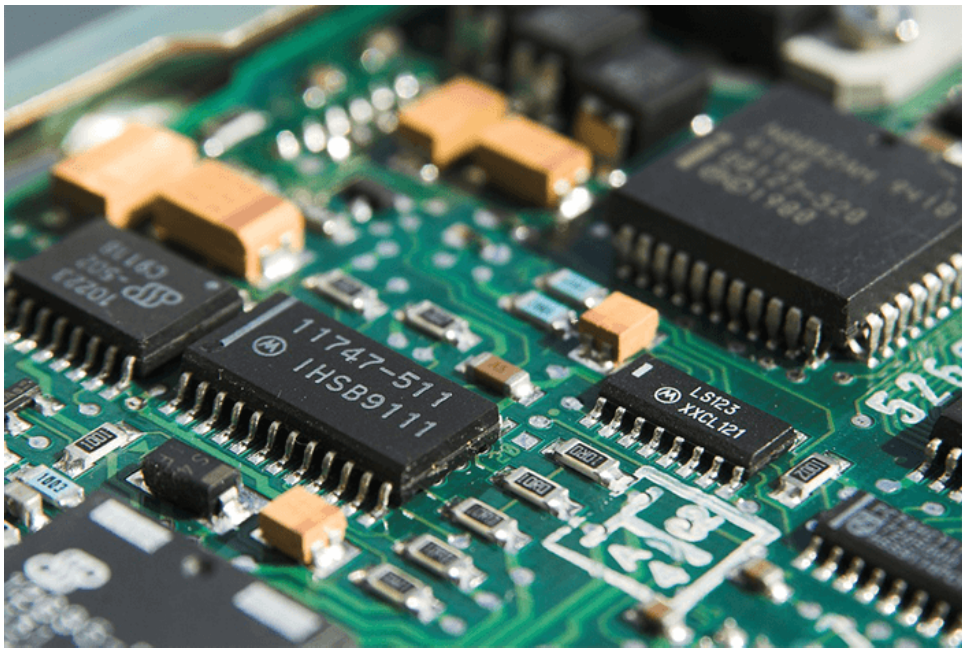
Although the capacity of RAM is typically lower than secondary storage devices, it is often the highest capacity form of main memory.

Summary

The characteristics of RAM are:

- It is volatile
- You can read and write to it
- It is quicker to access than secondary storage
- It has the largest capacity of all main memory

GCSE **Read-only memory (ROM)**



The ROM chip on your computer is attached to the motherboard
Pixabay

ROM is typically used to store the boot sequence (BIOS) for a computer. Unlike RAM, it is **non-volatile** and will not lose its data when the power is switched off. The contents of ROM are set by the computer manufacturer and, as the name implies, your computer usually **only reads** from ROM. Which is good because it is crucial to your computer system and any changes you make **could** stop the computer booting up properly.

The BIOS (Basic Input/Output System) stored on ROM is a very limited sequence of instructions that checks that the core components of the computer system (RAM, fundamental input/output devices, secondary storage) are connected and responding correctly.

Once done, the boot sequence loads the essential parts of the main operating system from secondary storage into RAM. From this point, the operating system will oversee the operation of the computer, managing memory, storage, and requests for input/output.

Which of these components do you think the BIOS checks first?

A - The screen
B - The keyboard
C - RAM
D - The CPU

A - The screen

Working out: The BIOS does a few things before checking on any components, but the screen is typically first. This is so the BIOS can display any errors or broken components to the user if the boot-up sequence fails. If a screen is not found, the BIOS can save resources by skipping the display commands.

Summary

The characteristics of ROM are:

- It is non-volatile
- It is written by the computer manufacturer
- Usually stores the BIOS
- Smaller capacity than RAM

GCSE

Comparing RAM and ROM

Whilst RAM and ROM are both types of main memory, they are different in a number of key ways, and it is useful to be able to compare them.

RAM has a much, much greater capacity than ROM. RAM is used to hold the data and instructions that are being executed, such as the operating system and any programs you are using. When you insert a disc into a games console, the code on the disc won't be executed until it has been copied from the disc into the system RAM, which is why you see a loading screen.

As ROM is read-only memory, it tends to store core software instructions that don’t need to be changed, such as the code needed to load the operating system into RAM (known as bootstrapping) or the BIOS.

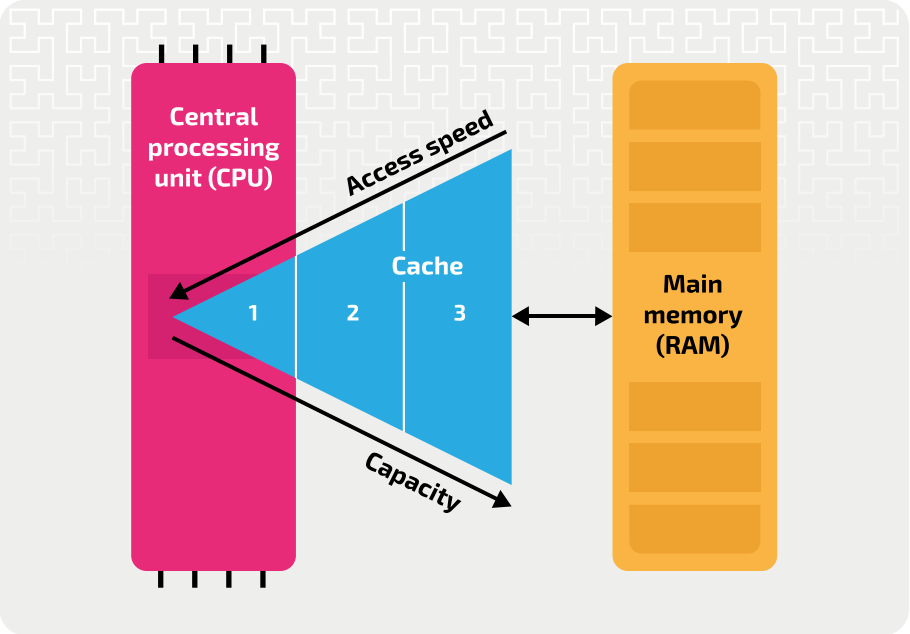
The table below summarises the comparisons of these two memory devices.

Characteristic	RAM	ROM
Purpose	Stores data and instructions during processing	Stores boot-up instructions set by the manufacturer
Volatility	Volatile, data lost without power	Non-volatile, data remains after power switched off
Read/Write	Read and write	Read only
Capacity	Usually several gigabytes	A few megabytes in size

GCSE **Cache**

Whilst a program is being executed, the processor stores and retrieves data from the RAM repeatedly, and although RAM is quick to access, there is another type of memory designed to make access to frequently used data much faster: **cache memory** (pronounced cash).

You may have heard the term cache used in relation to a web browser; that is a different sort of cache, but the role it plays is the same. Your browser will save frequently accessed web pages, images, and files so the next time you access them they are quickly available. The cache used by your processor works the same way, but stores instructions and data.



Cache sits between the processor and RAM

There are three levels of cache used by the processor, each level has a different capacity and access speed. The closer the cache is to the CPU, the faster it can read and write data, but the less it can store.

The level one cache is usually a part of the processor itself. This cache has the lowest capacity, but the quickest access speed.

In some processors, the level two cache might be built into the processor along with level one. Level two is slightly slower to read, but has a higher capacity than level one.

The level three cache might be the slowest at reading and writing, but it is still roughly twice as fast as RAM.

In a multi core processor, each CPU core will have its own level one cache but might share levels two and three with the other cores in the processor.

The levels of cache are designed to make the most frequently used instructions and data the quickest to access.

Which level of cache would a processor store the most frequently used data in?

1

Working out: The level one cache is the quickest to access, so it will be used to store the most frequently used instructions and data.

You can find out more about how the cache helps speed up processing on the system architecture page.

GCSE

Registers

Onboard the CPU there are very small memory devices called **registers**. Registers provide faster access than both the RAM and cache, but are only able to hold a few bytes at most.

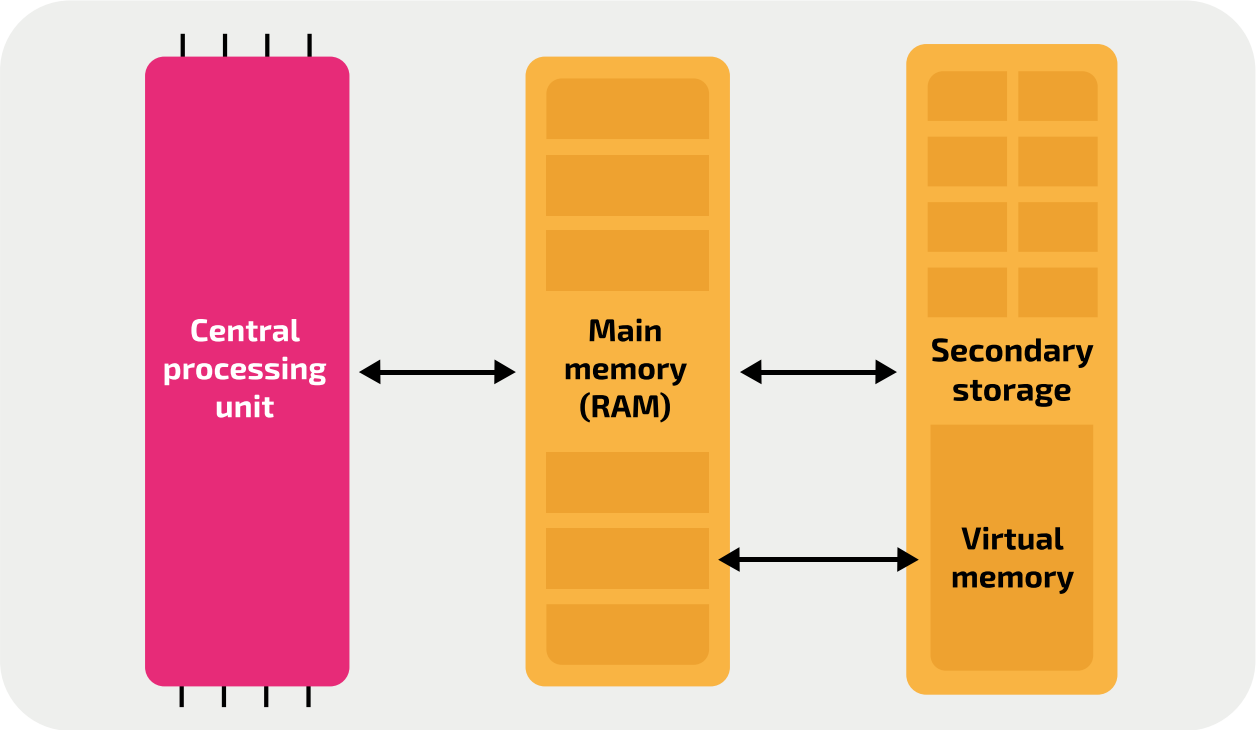
Registers are split into two categories: **general purpose** and **dedicated** (or special purpose).

- General purpose registers are used to temporarily store values generated by instructions, if they need to used again in the next instruction. The number of general purpose registers will differ between models of CPU.
- Dedicated registers serve a specific purpose in the Fetch-decode-execute cycle. To find out more about the cycle and the registers used, visit the system architecture page.

GCSE **Virtual memory**

Virtual memory is a way a computer system can compensate for a shortage of RAM. Older computers often have small main memories, but modern applications often have high main memory requirements. Running such applications on systems with minimal main memory will lead to issues as applications compete with each other for access to RAM.

In systems with a small amount of memory installed and/or multiple applications open, the effect is typically a slowdown in performance, known as a lag. In extreme cases, the system may freeze or crash. To reduce the likelihood of this happening, computer systems allocate a portion of secondary storage as virtual memory. This process allocates a portion of the HDD / SSD for storage of applications and files that are currently open.



A part of secondary storage is assigned to virtual memory

When a system requires virtual memory, the operating system creates a set of virtual addresses (RAM is separated into a set of physical addresses). Data moved to virtual memory is stored as pages. When the processor (CPU) is ready to accept instructions from the files in virtual memory, it moves the page files into RAM, allocating a physical address to the data. When page files are moved from virtual to real memory, other data is moved from RAM to either virtual memory or back to the secondary storage.

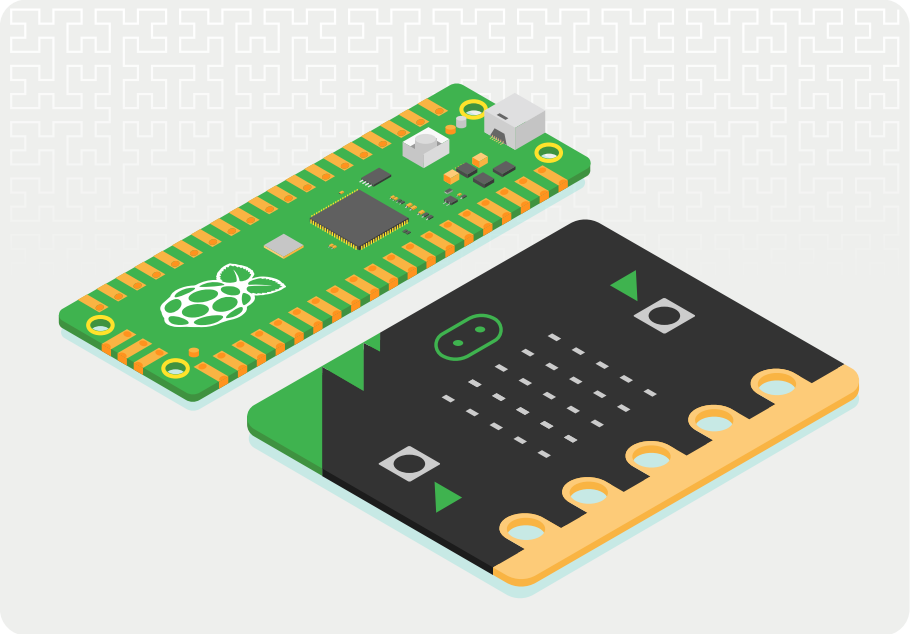
Virtual memory shouldn't be confused with **virtual storage** — a term used to refer to secondary storage located away from the computer, such as cloud storage and other network storage services.

Both HDDs and SSDs are used for virtual memory. SSDs have an advantage over HDDs in that access speed is closer to that of RAM. However, a system that uses SSDs regularly for virtual memory reduces the lifespan of the SSDs, as a result of their finite read/write lifecycle.

GCSE

Flash Memory

Some microcontrollers use a memory chip to store instructions, called **flash memory**. Flash memory is **non-volatile** and can be both written to and read from.



The Raspberry Pi Pico and BBC micro:bit both use flash memory.

Microcontroller and microproccessors are used to perform very specific tasks and will often have only a single program to run. These programs can be stored and fetched from the same place as there is no need for multiple programs to share the same main memory, which would require one of them to be removed to make room (as with RAM on larger PCs). For this reason, flash memory is used as both storage and main memory.

A Level

Memory locations

The basic building block of RAM is a memory cell. A cell can store a single binary digit (1 or 0). A group of cells is called a **memory location**.

Therefore, RAM consists of a number of memory locations which are used to store data (that is data values or instructions) in the form of bit patterns. RAM is addressable; each memory location has a physical **unique address**, a **natural number** that is used to find the memory location and access its contents to retrieve or store an instruction or data. For example, if a main memory has 64 memory locations, then the addresses of these locations will be the numbers from 0 up to 63.

Each memory location of RAM can be **accessed directly** by the processor (while the contents of secondary storage need to be transferred to the main memory first). This means that **it takes the same time to access any memory location**, i.e. the time required to access a memory location doesn't depend on the address or position of that memory location. As a result, memory locations do **not** have to be accessed in a specific sequence, instead they can be accessed in any order — hence the name random-access memory.

The processor can **read** a memory location, i.e. load the contents of that memory location to one of its registers. In this case, the contents are copied over and remain unchanged to that memory location.

The processor can **write** to a memory location, i.e. store the contents of one of its registers to that memory location. In this case, the contents of that memory location are overwritten.

A Level

RAM capacity

The maximum amount of RAM is determined by the width of the address bus. The overall capacity has an effect on the performance of the computer system. In general, the larger the size of RAM, the more efficiently a computer will run. Not as much memory means that data is moved between secondary storage and RAM more frequently. These operations increase the amount of time an operating system spends on management tasks and is witnessed through a slowing of operations on a computer.