Main Memory, Caches and Registers

Learn this definition

Main memory is any form of computer memory that the CPU can access **directly**.

Main memory is the memory part of the von Neumann architecture

This is the memory that the computer uses when it carries out tasks. It doesn't include the hard drives or removable media – these are called **secondary storage**.

How does main memory work?

This is a really simplified version of what happens

When a document is loaded, it first needs to be moved from secondary storage into the computer's working memory. This is the RAM chips that fill the DIMM slots on the motherboard.

This often takes some time – which is why it often seems to take so long for programs to start up. This is just because a lot of data needs to be moved into RAM.

When a specific piece of data from the document is processed, the **Fetch–Decode–Execute cycle** fetches the data from the RAM chip into registers in the CPU where it is decoded and executed.

Types of Memory

The exam board doesn't count registers and caches as "main memory". Think of main memory as just the RAM and ROM

There are four types of memory you need to know about:

1. RAM
2. ROM
3. Registers – not part of "main memory"
4. Caches – not part of "main memory"

You need to know what each of these are used for and how they work.

1. RAM

RAM stands for **Random Access Memory**. It is used to hold the programs that are being run by the computer and the data those programs are using – including the documents that are being worked on.

DIMM stands for Dual Inline Memory Module

Think about this as the memory chips that fill the **DIMM slots** on the motherboard.

Volatile is **really important**. Basically if the power is turned off the data simply disappears.

The really important thing about this memory is that when the power to the computer is turned off, the data and instructions stored in RAM is lost. RAM can only hold data when it is powered up. We say that this means that RAM is **volatile**.

The flexibility of RAM is what makes modern general purpose computers possible. The ability to reuse memory for lots different purposes makes so many things possible.

The contents of any RAM memory address can be changed as required. The CPU may start by storing an instruction at one location before overwriting the memory address with an item of data. This means that RAM is really flexible and can be reused for lots of different purposes.

Data is moved from RAM to the CPU using the **Bus** – the set of connections which link the elements of the motherboard together. This is often a relatively slow process.

2. ROM

ROM stands for **Read Only Memory**. ROM is used to store data that is never changed – it is **Read Only**. The data stored in ROM is **not** lost when the power is switched off – so it is **non-volatile**.

ROM is used to store the instructions required to **boot** the computer. A small program in ROM starts the process of loading the operating system into RAM – moving it from secondary storage to the RAM slots. Once the process is started, the OS can take over, running its own loading programs from the data initially moved into RAM.

The program used to start to load the operating system is called the **bootstrap loader**. ROM also includes the **BIOS** – this tests the hardware systems and initiates the bootstrap loader.

3. Registers

Registers are small, quick memory locations **within the CPU**. They are the memory the computer uses to process data during the Fetch–Decode–Execute cycle.

Registers are at the heart of the Fetch–Decode–Execute cycle

Data is copied into registers during the Fetch part of the cycle and can then be processed. Data can then be copied back into RAM if it has been changed.

Registers are expensive because they are powerful and quick. Most of the registers within the CPU can be used for a variety of tasks, but some, such as the Program Counter, are only used for specific jobs.

Registers and Caches are also **volatile**. When the power is lost, they are emptied

4. Caches

Caches are memory locations **close to the CPU**. They are used to hold data and instructions which have been recently used by the CPU.

Caches are quicker areas of memory, often with a quicker bus. This means the data and instructions stored in them can be retrieved more quickly, improving performance.

Caches are quicker types of memory, which means they are more expensive

When data is used by the CPU it gets dumped at the top of the cache. If the cache is full, the data at the bottom falls out. The next time the CPU needs to retrieve data or instructions, it looks in the cache for it first. If it's in the cache, it retrieves it and after it's finished with it, dumps it back at the top of the cache. This means that data the CPU is using a lot stays at the top of the cache and never falls out – so it's always quicker to retrieve it.
The bigger the cache, the more data can be stored in it.

Cache size is one of the three ways to improve CPU performance make it "quicker"

Activities:

a) What is the definition of **main memory**?
b) List the **four** types of memory. Which ones are included in main memory?
c) Define the terms **volatile** and **non-volatile** in terms of main memory
d) What is **RAM**? What is it used for? Where would you find it?
e) What is **ROM**? What is it used for? Where would you find it?
f) Summarise the four types of memory – what they do, where they are and their differences
g) Explain the relationship between **main memory** and **secondary storage**