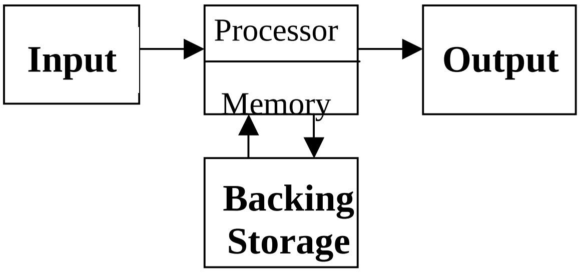
The Central Processing Unit



All computers derive from the same basic design, usually referred to as the **von Neumann architecture**. This concept involves solving a problem by defining a sequence of commands or instructions which are input, processed and then output. The **Central Processing Unit**, (CPU) is the part of the computer which is responsible for executing the instructions of the program. The CPU has 2 main parts, the processor and main memory. Buses are used to transfer data to and from the processor and memory.

INTERNAL PROCESSOR STRUCTURE

A processor can be considered to be made up of three components:

* Arithmetic and Logic Unit (ALU)
* Control Unit
* Registers

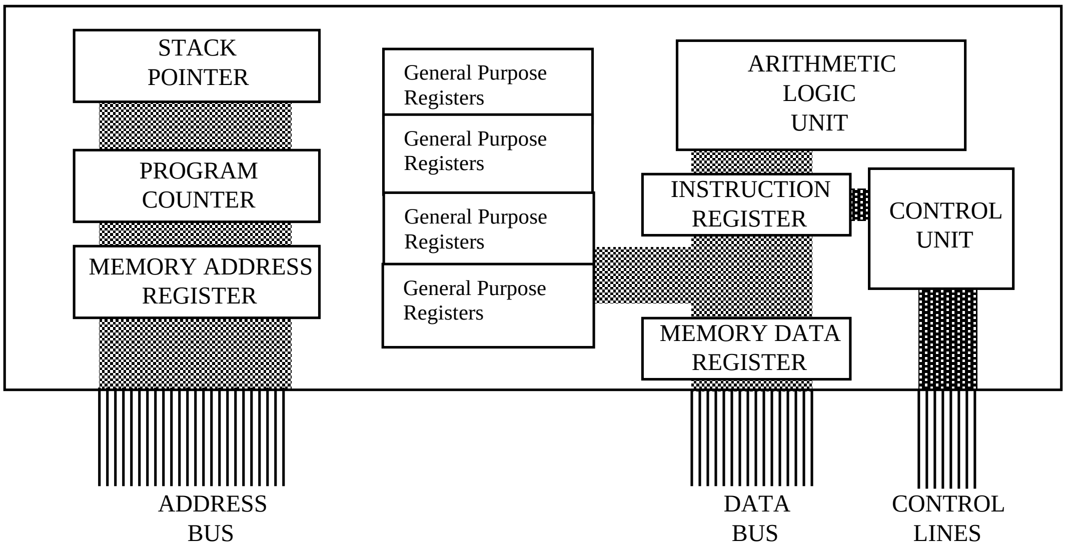
The **ALU** is the part of the processor where data is processed and manipulated. These consist of arithmetic operations or logical comparisons allowing a program to make decisions.

The **Control Unit** is the part of the processor which manages the execution of instructions. The control unit is responsible for the fetching, decoding and execution of instructions by sending control signals to other parts of the computer.

The other main component of a processor is the **registers**. A register is a very fast temporary storage location inside the processor itself. There are many registers including the memory address register (MAR), memory data register (MDR), program counter (PC), stack pointer (SP), accumulator (A), general purpose registers and the status register.

A register can be used for tasks like holding data for a calculation, storing the address of the next instruction to be executed or holding the instruction as it is being decoded and executed. Internal buses are used to transfer data from one register to another.

As fetches to memory take time it makes sense to keep as much data on the processor as possible. This means that it can be accessed immediately rather than waiting while it is being fetched from memory. Older computers might only have those registers mentioned above (or similar) but modern processors will have many more registers than that including some with over 100 general purpose registers.



Example of the structure of a processor

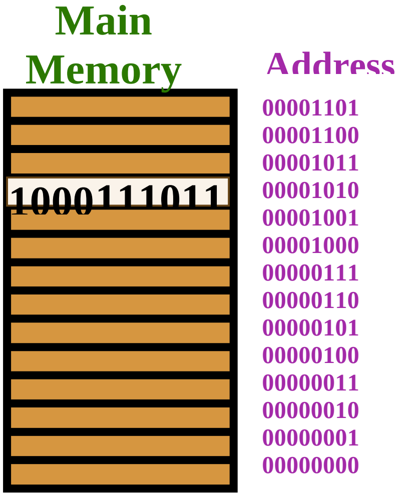
Produced by S Lambert, R Simpson and HSDU for The City of Edinburgh Council, 2004

The Processor And Memory

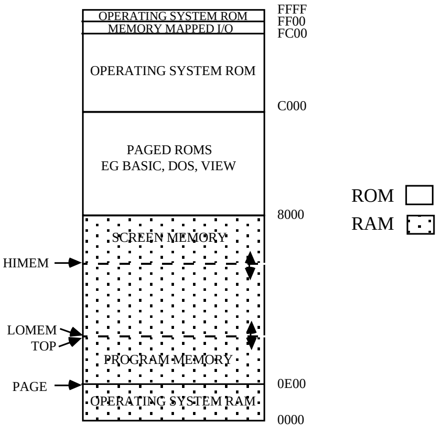
Random Access Memory

To execute a program you must first load the program and any relevant data into the computer's internal memory (**RAM**) from disk. The program and data is stored in memory until needed by the processor. This is called the **stored program** concept.

A program may contain thousands of instructions but the processor can only execute one instruction at a time. Obviously the speed at which a computer operates must be very fast. The first instruction is fetched from memory into the processor where it is decoded and executed. Then the second instruction is fetched, decoded and then executed and so on until the program ends.
This is known as the **FETCH - EXECUTE - CYCLE**.



RAM is split up into **memory locations**. The processor has to be able to pinpoint any memory location needed so each memory location is assigned an **address**. This is a unique binary number from zero up to the number of locations – 1.



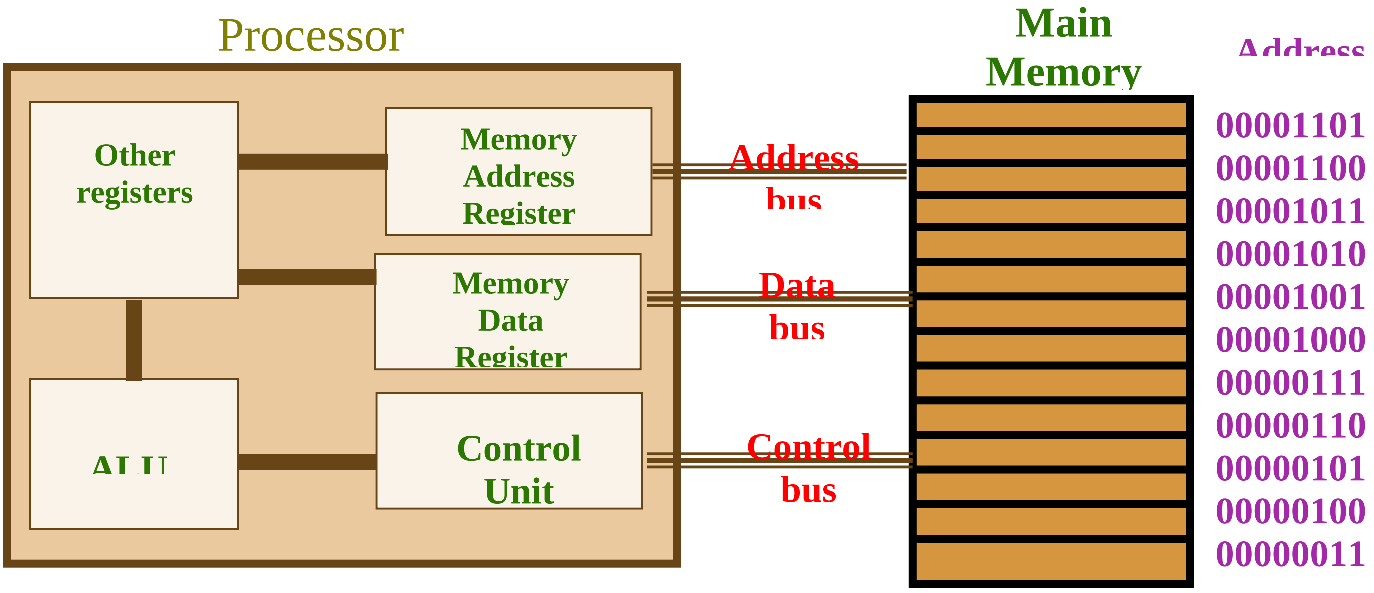
Memory Maps

Memory maps are often included with the sale of the computer. These describe the way memory locations are organised in the computer. For example, parts of the memory may be allocated to the screen display, program code, variables etc. This information can be useful for the programmer to know.

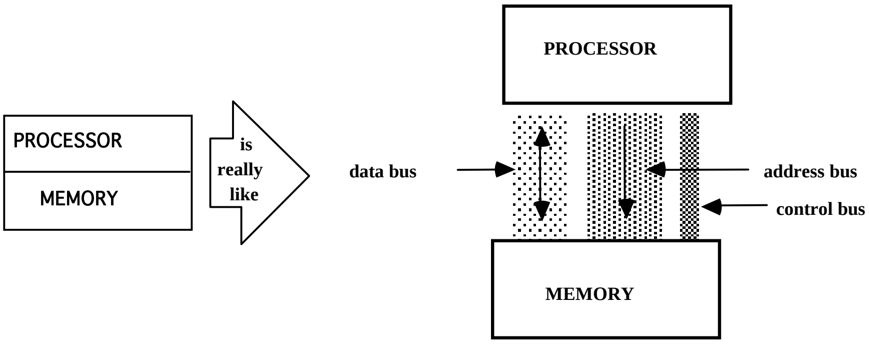
Produced by S Lambert, R Simpson and HSDU for The City of Edinburgh Council 2004

The Processor And Memory

The processor and memory are linked up like so:



Obviously information has to be transferred backwards and forwards between memory and the processor. Buses are used to do this. A bus is an electronic highway or a collection of cables where each cable can transmit a bit.



Produced by S Lambert, R Simpson and HSDU for The City of Edinburgh Council 2004

Buses

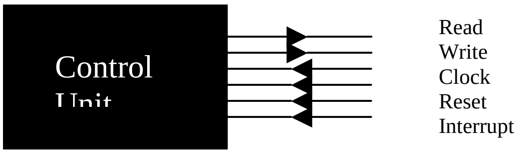
There are three buses:

The address bus - is used by the processor to pinpoint the memory location needed. This is a one way (uni-directional) bus.

The data bus - is used to transfer the data from processor to memory and vice-versa. A two way (bi-directional) bus

The control bus - is used to tell memory (and other parts of the system) what is happening. The control bus is not a true bus because each line is used independently and it does not send or receive data.

The control bus usually consists of the following control lines:



read - A signal from the processor that initiates a memory read operation (sends data from memory to processor) once the address bus and the data bus have been set up.

write - A signal from the processor that initiates a memory write operation (sends data from procesor to memory) once the address bus and the data bus have been set up.

clock - Every processor has a clock which ticks continuously and is used by the control unit to tell all the components when certain tasks have to be done. A current (2004) processor might have a clock of 1GHz (Gigahertz) which means that the clock ticks 1 000 000 000 times a second. So a processor with a 1 GHz clock will be faster than a similar processor with a 500 MHz clock.

reset - This clears all internal processor registers and starts fetching instructions from a predefined place.

Interrupt - This causes the current state of processing to be saved in a temporary area (called the stack). The processor then deals with the device that made the interrupt, returning to the previous operation when the interrupt is complete. Instructions can tell the processor to ignore or ‘mask’ the interrupt system.

NMI (Non Maskable Interrupt) - This operates in the same way as the ‘interrupt’ signal except that it cannot be ‘masked’ (ignored). The processor must deal with this interrupt.

A processor such as the PowerPC might have a 36 bit address bus and a 32 bit data bus whereas an older 6502 processor for a BBC computer would have 16 bit address and 8 bit data buses.

Produced by S Lambert, R Simpson and HSDU for The City of Edinburgh Council 2004

Fetch – Execute Cycle

Read From memory (Fetch - Execute)

A processor needs an instruction from memory or requires some data to perform a calculation. Here are the steps and buses needed to fetch that information from memory.

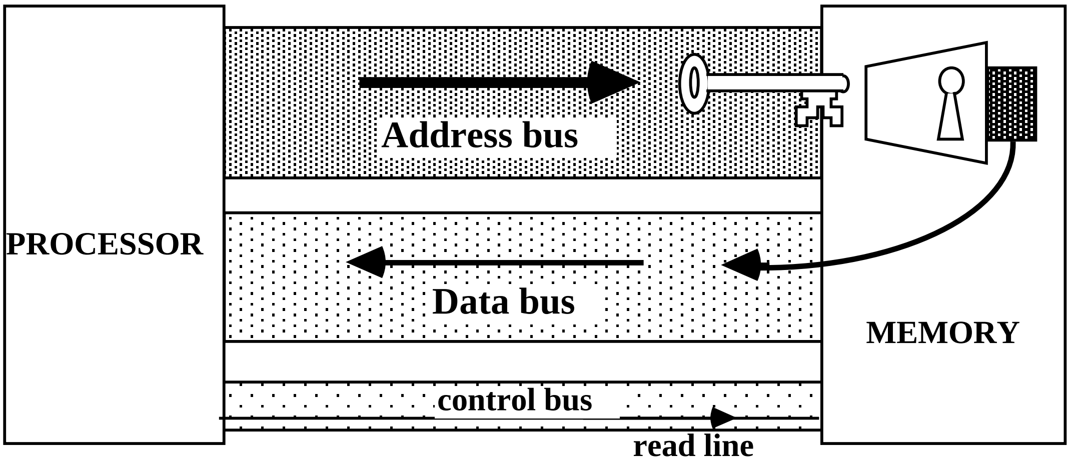
1. Processor sets up *address bus* with the appropriate address.
   This pinpoints the desired location.

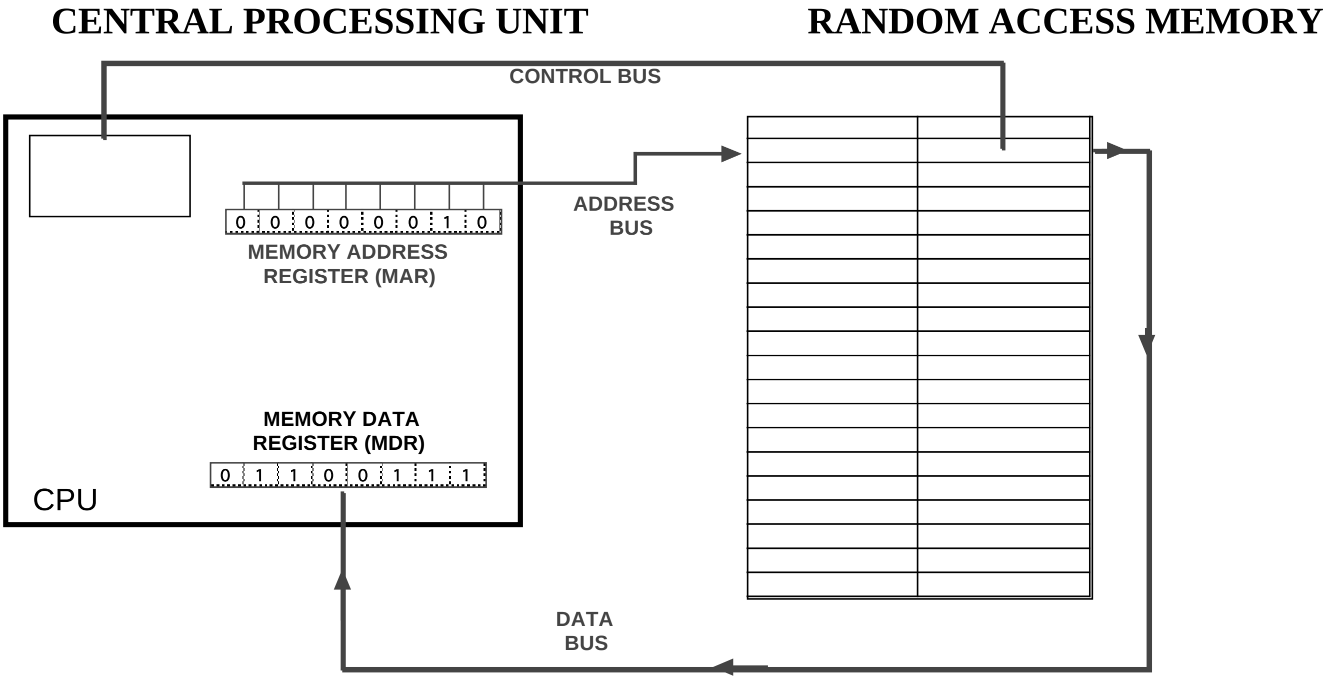
2. The address bus then opens the appropriate memory location of this address.

3. Activate the read line on the *control bus*.
   This tells the location that it is to be read from.

4. Memory location releases its data on to the *data bus*.

5. Data is transferred to the processor where it is decoded and executed.





Fetch – Execute Cycle

Write to Memory

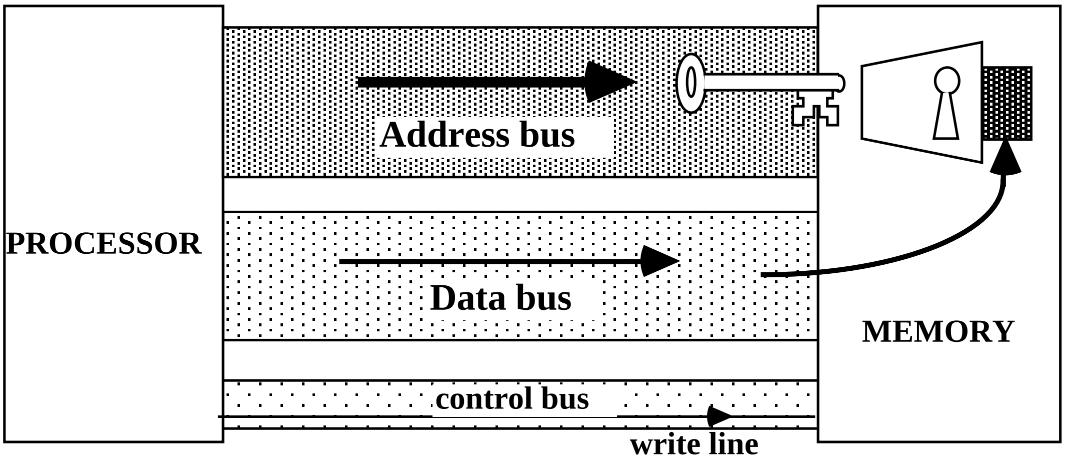
After a calculation or other task has been completed the processor has to store the information into a memory location. Here are the steps and buses needed to send that information from the processor to memory.

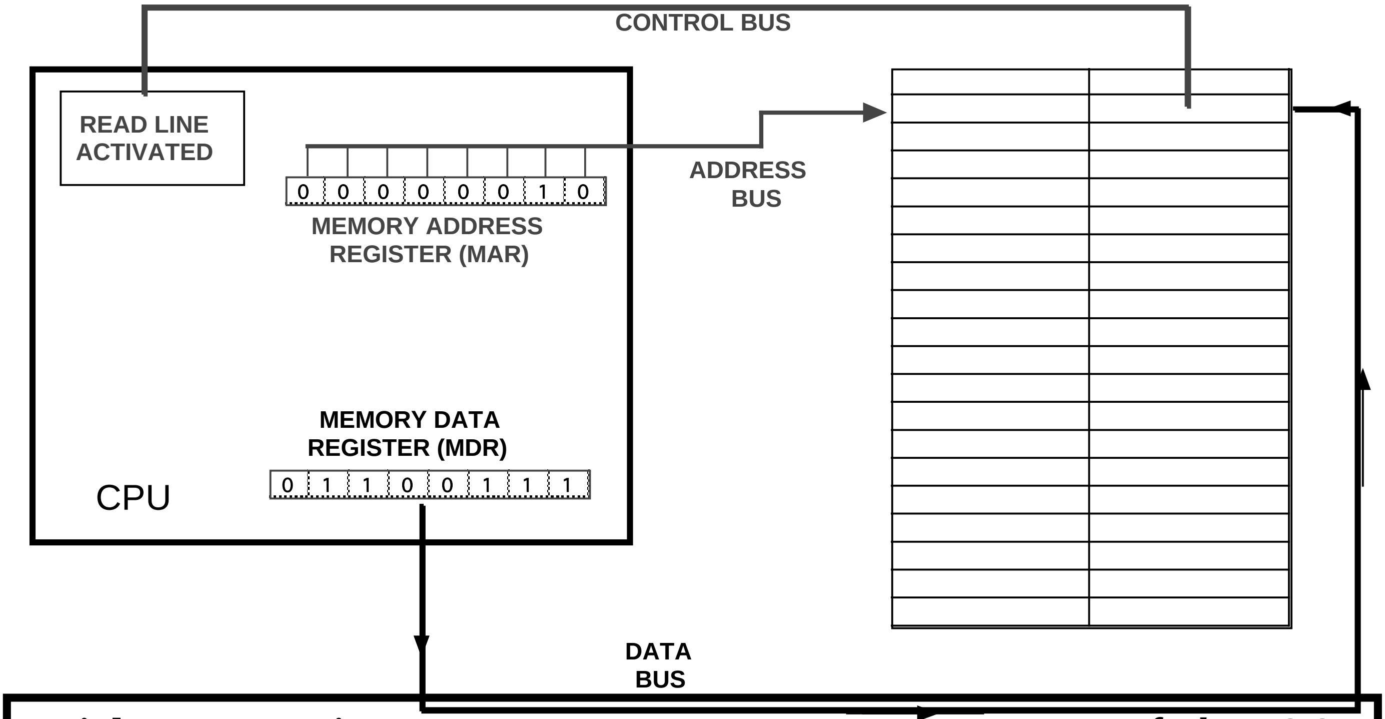
1 Processor sets up *address bus* with the appropriate address.
This pinpoints the desired location.

2 The address bus then opens the appropriate memory location of this address.

3 Activate the write line on the *control bus*.
This tells the location that it is to be written to.

4 Data is transferred from processor to the memory location using the *data bus*.





Produced by S Lambert, R Simpson and HSDU for The City of Edinburgh Council 2004

Computer Memory

Main Memory

The main purpose of main memory or RAM is to store data given by the processor. The computer can only manipulate data that is in main memory, therefore, every program executed and every open file must be in main memory before it can be used by the processor. The amount of main memory on a computer determines how many programs can be executed at one time and how much data can be readily available to a program.

Static And Dynamic RAM

Static RAM (SRAM) is fast access memory and its chips hold their contents as long as power is applied to the chip. Dynamic RAM (DRAM) chips not only need the power applied but also need a continuous signal to refresh the contents of the chip (hence the term dynamic). These are slower to access than SRAM. Dynamic RAM is more widely used than static RAM because it needs less power and its circuitry is simpler.

Read Only Memory (ROM)

This is where data can only be read from. Its contents are **not** lost when power is removed and is therefore permanent or non-volatile. The software and data required on the ROM are normally fixed in at manufacture.

Programmable Read Only Memory (PROM)

These are types of ROMs that come with every bit set to 1 and the user (or manufacturer) can program these themselves. This process, which is irreversible, is usually called 'blowing' since each bit is a fusible link which becomes a zero when destroyed. The difference between a PROM and a ROM is that a PROM is manufactured as blank memory, whereas a ROM is programmed during the manufacturing process. To write data onto a PROM chip, you need a special device called a PROM burner.

Erasable Programmable Read Only Memory (EPROM)

These are a family of ROMs which can also be programmed by the user, but can additionally be erased (usually by ultra violet light). These work on a stored charge principle and are often used for the development and testing of ROM based software before the ROMs are manufactured.

Produced by S Lambert, R Simpson and HSDU for The City of Edinburgh Council 2004

Computer Memory

Memory Modules

SIMMs (single in-line memory modules) or DIMMs (dual in-line memory modules) are small circuit boards that can hold a group of memory chips. These can easily be installed into slots on the mother board of a computer so that extra memory can be installed.

Cache Memory

Cache memory is high speed memory (usually SRAM). It allows instructions that are accessed by the processor on a regular basis to be stored. By keeping as much of this data as possible in cache memory, the processor avoids accessing the slower main memory (usually DRAM). Caching is effective because most programs access the same data or instructions over and over.

Registers

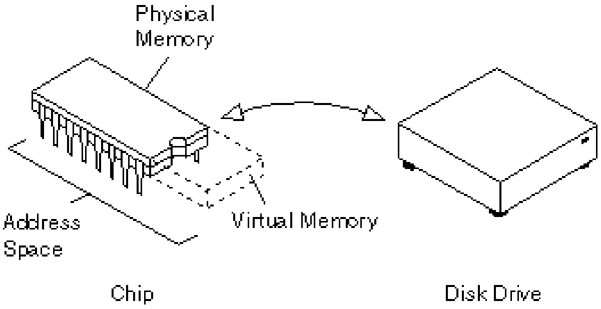
Registers are fast temporary storage areas within the processor.

Backing Storage

Backing storage is classified as **external memory**. Any data that the user wishes to keep and return to must be saved onto backing storage as RAM loses its contents when power is switched off. Typical examples of backing storage devices are hard disk, CD-RW, DVD-RAM etc.

Virtual Memory

Part of the backing storage (hard disk) is set aside for **virtual memory**. This enables virtual addresses to be set up on the hard disk and then are ‘mapped’ to real addresses in main memory when required. This is known as **paging**. It enables the processor to ‘pretend ‘ it has more addressable memory than it actually has access to**.**



Produced by S Lambert, R Simpson and HSDU for The City of Edinburgh Council 2004

Computer Memory

Speed Of Access From Memory

Increasing the system performance of a computer could depend upon what type of memory is being used and for what.

Having a processor with many registers will improve system performance as this means that data can be accessed immediately rather than waiting for it to be fetched from main or cache memory.

Using cache memory will improve system performance as this means that data can be accessed at a faster rate than from main memory.

If the amount of RAM is small then that usually means that large applications and their data cannot be opened at the same time. This means that more fetches must be made to backing store and this will slow down system performance. However most computers can be upgraded by allowing additional RAM to be installed. Single In-line Memory Modules (SIMMs) plug into a SIMM socket on the motherboard allowing easy installation of increased RAM and a minimal amount of space used on the motherboard. Each SIMM contains a number of DRAM chips and varies depending on the type of computer and the amount of RAM you require. If a computer is struggling to run some software or all the software can’t load at the one time, then adding extra memory can improve system performance. This is particularly useful when dealing with multimedia software and files (graphics, video and sound). Latest machines come with 512 Mb (or more) RAM which can be upgraded to 1 Gb (or more).

Produced by S Lambert, R Simpson and HSDU for The City of Edinburgh Council 2004

Measures Of Processing Speed

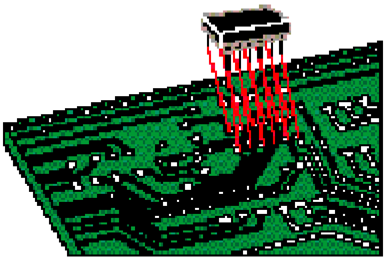
Clock Speed

When comparing one computer's performance against another, one of the main criteria is the **clock speed** of the processor. This indicates how many **clock pulses** the processor can deal with per second. Normally each instruction the processor carries out can take many clock pulses. The faster the clock speed the faster the processor will perform. However if there is not much difference between clock speeds and different manufacturers are involved, then other factors must be taken into consideration.

> ***Example***
> Intel Celeron 2.4 GHz against a Intel Celeron 2.6 GHz - obviously the second processor is faster
>
> Pentium 2.8 GHz against a Intel Celeron 2.6 GHz - no indication of which is the faster

Instructions Per Second

Another measure is the **number of instructions** that the processor executes per second (**mips**, millions of instructions per second). Again if this number is quite close when comparing processors and different manufacturers are involved, a problem still exists. What sort of instructions are being carried out? There is no standard set and so some manufacturers could use simpler and faster instructions than others.



Floating Point Instructions

An even better measure of processor performance is the **Flop** (Floating point operations per second). The procedures involved in doing a floating point multiplication are basically the same for every processor and these kinds of operations are used in most software so they provide the basis of a reasonable comparison of system performance.

Application Based Tests

Ever been frustrated trying to use a software package like a Desk Top Publisher or Multimedia Authoring as the computer seems to perform really slowly when you are trying to do even a basic task? Application based tests are based on such powerful applications. Basically system performance is measured by how well the processor can deal with requests from large applications. These applications are usually memory intensive and involve data types such as graphics, sound and video. These are also known as benchmarks for computer systems. a standard set of computer tasks designed to allow a computer's performance to be measured.

The type of the tests that can be carried out could be how well the processor performs when:

- file access is required. How fast can the file be saved, edited etc;
- many memory fetches are necessary.

If a processor can cope with these large applications and their related tasks well, then that is a good indication of high system performance.

Produced by S Lambert, R Simpson and HSDU for The City of Edinburgh Council 2004

Factors Which Affect System Performance

The **word** length of a computer is the size of the data, in bits, which can be manipulated as a single unit by the processor. This size is determined by the width of the data bus within the computer. In an ideal computer the width of the data bus and the size of the memory locations will match. Thus if the capacity of the memory locations and the word length is 16 bits then every fetch will bring the contents of one memory location (2 bytes) into the processor.

Early computers had a word size of a byte so a four byte integer would take four fetches before it could be used in a calculation. Most modern computers have a word size of four bytes (32 bit) so the same integer can be brought into the processor with one fetch. This also means that a 32-bit data bus can carry twice as much as a 16-bit data bus. However this does **not** mean it can go twice as fast as other factors come into consideration.

The width of the address bus defines the maximum amount of **memory locations** available to the processor. A single line can address locations with addresses 0 and 1 ie **2** (2^1) locations. If 2 lines were used then **4** (2^2) addresses would be available 00, 01, 10 and 11.

In general **number of memory locations (addresses) =** $\mathbf{2}^{\text{number of bits for address bus}}$

Combining the two buses it is then possible to work out the actual **memory size**.

Example

To find the maximum memory size which a 24-bit address bus can support, assuming 16-bit data bus:

Maximum memory size = 2^{24} x 16 bits (each location can hold 16 bits)
= 16 777 216 x 2 bytes
= 33 554 432 ÷ 1024 Kb
= 32 768 ÷ 1024 Mb
= 32 Mb

In general **Maximum Memory Size = number of memory locations * bus width**

Cache Memory

Using cache memory implies that the memory can be accessed very rapidly. A cache could be used inside a processor or between processor and memory. The effect is to speed up computing actions by reading from and writing to fast cache memory instead of RAM.

Example

In one particular computer it takes the CPU as much as 180 nanoseconds (0.000000180 s) to obtain information from main memory, compared to just 45 nanoseconds (0.000000045 s) from cache memory.

Therefore, the more instructions and data the CPU can access directly from cache memory the faster the computer can run.

Peripherals

Peripherals generally perform at much slower speeds than the CPU and this will **reduce** system performance.
Buffers and spoolers can be used to help this situation (see later).
However most peripheral manufacturers are increasing the power of their interface.

Example

Sound cards now have their own processor, RAM and ROM.

Produced by S Lambert, R Simpson and HSDU for City of Edinburgh Council 2004

Current Trends In Hardware

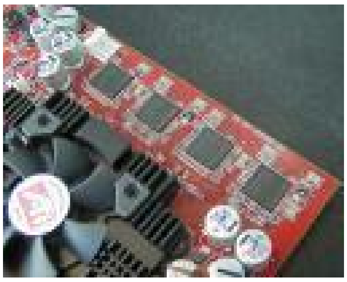
Over the last 20 years there has been a phenomenal increase in the performance of computer systems. There are many reasons for such changes. Three of the reasons are down to clock speeds, memory and backing storage capacity.

Clock Speeds

The processor's clock "ticks" so many millions of times a second. During these pulses instructions are carried out. The greater the number of pulses per second the faster the processor should perform (obviously depending upon comparison of similar manufacturers). Below is an indication of how processing clock speed has evolved over the past 20 years.

Processor Type	**Year**	**Clock speed**
8088	1979	4.77 - 8 MHz
8086	1978	4.77 - 8 MHz
80286	1982	6 - 20 MHz
80386DX	1985	16 - 33 MHz
Pentium	1993	60 - 200 MHz
MMX	1997	166 - 233 MHz
Pentium	2000	1.4 GHz - 3.2 GHz
AMD Athlon	2003	2 GHz

Memory



Random Access Memory (RAM) is where programs and data that are currently used are being stored. The more data it is possible to have available in RAM the faster the PC will run. This is due to less fetches required from backing storage.

The width of the address bus dictates how many different memory locations can be accessed, and the width of the data bus dictates how much information is stored at each location. Every time a bit is added to the width of the address bus, the address range doubles. In 1985, Intel's 386 processor had a 32-bit address bus, enabling it to access up to 4GB of memory. The Pentium processor, introduced in 1993, increased the data bus width to 64-bits, enabling it to access 8 bytes of data at a time.

Backing Storage

External memory areas within a computer system allow data to be saved so that it could be retrieved again for later use. This is due to RAM being volatile and losing its contents when no power is available.

The first home computers did not have any type of backing storage and all data had to be dealt with whilst the computer was running. All programs used were stored in ROM. The floppy disk drive was the first backing storage device to be manufactured. Dedicated word processors used these disks to save data onto.

The hard disk drive revolutionised computers as software and data could be changed and stored on the computer itself. Back in 1954, when IBM first invented the hard disk, the capacity was 5 Mb. Nowadays we would expect the hard drive's capacity to be at least 20 Gb. The larger the hard drive, the more applications and data can be stored.

Other backing storage devices offer portability of data etc. These are commonly the CD-R, CD-RW, DVD-RAM, DVD-RW to name but a few. These offer storage from 850 Mb upwards and allow multimedia elements like video, graphics and sound to be stored. The floppy disk drive is almost obsolete due to its poor storage capacity of its disks - 1.4 Mb. Good for text files using the dedicated word processor, but useless for multimedia storage

Produced by S Lambert, R Simpson and HSDU for City of Edinburgh Council 2004