Computer Performance

Prerequisite knowledge

Before studying this topic you should be able to:

- Describe the uses and compare the features of embedded, palmtop, laptop, desktop and mainframe computers
- Make comparisons in terms of processor speed, main memory, backing store and peripherals
- Describe clock speed as an indicator of system performance.

Learning Objectives

By the end of this topic you will be able to:

- Describe and evaluate measures of computer performance including clock speed, MIPS, FLOPS, and application benchmark tests
- Describe the factors affecting system performance including data bus width, cache memory and data transfer rates between peripherals
- Describe the effects of increases in clock speed, memory and storage capacity.

Clock Speed and Memory Usage

The System Clock

All of the **events** which take place **inside** the computer system **have** to be timed/synchronised. This ensures, e.g. that instructions are carried out in the correct sequence.

Each computer system, then, has a **built-in clock** This clock is regulated by a **crystal** which oscillates several million times per second, e.g. a **100 MHz (MegaHertz)** clock oscillates 100 million times per second.

The accuracy of the physical switching in the computer system is built in to its architecture. The **speed** of the **switching** is down to the **speed** of the **system clock** - the faster the clock, the faster your computer will process its data. Each 'tick' of the system clock is referred to as a **machine cycle.**

In theory, a 100 MHz system should be able to carry out a **hundred million instructions per second**. However, some instructions take up **several** machine cycles, i.e. they involve **several** memory **fetches** or separate **read/write** operations.

MIPS

A more **accurate** description of a computer's speed is the **MIPS** measure, **MIPS** standing for **millions of instructions per second**. However, some manufacturers use different types of instructions (simple/complex) when testing, which throws the **reliability** of this measure off.

The MIPS rate can be influenced by a number of factors:

- the clock speed of the processor
- the speed of communication lines (buses)
- the speed of memory access
- the speed of execution of instructions.

Another aspect of the computer system which may affect its performance in terms of speed and storage capacity is the **type** of **RAM** (Random Access Memory) chips used for data storage.

FLOPS (Floating point operations per second)

You should be aware that using mip rate as a comparison factor also has problems. 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. A better measure of 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. As these kinds of operations are used in most software they provide the basis of a reasonable comparison of system performance.

Benchmark tests

A benchmark is a well defined standardised routine used to test the performance of computer systems. (Benchmark testing is also used to test software performance). They consist of standard operations that measure the speed of processing in terms of floating point operations per second (FLOPS) and in some cases the number of instructions performed per second (MIPS).

Examples of benchmarks are the Dhrystone and Whetstone tests. The Dhrystone test measures the processor's performance in executing frequently used statements and string comparisons. The Whetstone test measures the processor's performance in executing arithmetic functions.

Types of RAM Chips Available

SRAM (Static RAM)

· Each bit represented in SRAM requires two transistors, one to maintain the charge and the other to store the binary state (1 or 0).

· This means that Static RAM maintains the binary values until the power is switched off.

DRAM (Dynamic RAM)

· uses only one transistor for each bit represented in memory and, potentially, is capable of double the storage capacity of SRAM.

· the circuitry for DRAM is more involved since it requires to refresh the RAM store at regular intervals to maintain the bit states.

· is cheaper than SRAM and it is the commonest type of RAM used in PCs.

Although DRAM storage capacity has increased dramatically, its memory access times have been only marginally improved.

How can **DRAM** performance be improved?

EDO (Extended Data Output**) RAM**

The data is left active on the output of the memory chip until the next read cycle starts. Since there is less 'dead' time, EDO RAM improves memory access by 30%. Single-cycle EDO performs a complete memory operation in one clock cycle instead of two, increasing memory access speed by 100%.

CDRAM (Cache DRAM)

This uses SRAM & DRAM on one chip & has a very fast bus between the SRAM & DRAM components, which are separate, so that they can operate in parallel. The SRAM provides the '**cache**' element.

SDRAM (Synchronous DRAM)

This type of chip **synchronises** the memory data transfer operation with the system clock. Once a transfer has been initiated, the chip sends or receives data at the clock frequency, relying on the fact that other components are using the system clock and that the transfer will take place with **precision**.

Rambus

This technique uses two banks per chip and a very fast synchronous Local bus. The operation of the Local bus is strictly controlled and there is an acknowledgement or 'handshaking' protocol involved.

VRAM (Video RAM)

Video data needs to be streamed at a precise rate. VRAM has to be able to cope with requests from the CPU and the video accelerator chip for this constant data stream. If it pauses the CPU operation, the video card's operation slows down. If it holds up the video card's operation, the picture is disrupted.

WRAM (Windows RAM)

This is a memory chip which is optimised for Windows OS video operations. It uses its own 256 bit bus to perform video display operations.

Multi-Banked DRAM

A cheap and cheerful solution which allows video data to be read in parallel from discrete chips, so accelerating the updating of the screen display.

Standard functions of an interface

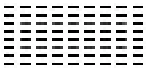
Every interface will need to carry out the following:

- convert data from the format understood by the processor to the format understood by the peripheral
- hold data in a Buffer as it is transferred from the processor to the peripheral and vice versa
- transmit/receive control signals to/from the CPU
- maintain status information that informs the processor whether the peripheral is ready to send or to receive data.

Serial and Parallel - these refer to different ways that **communication** (**data transfer**) takes place between devices.

•**Serial** communication means a **stream of bits** travelling along a **single** channel. (one bit after another in single file) e.g.

•**Parallel** communication means a group of **bits** (bytes) travelling along **several** channels **at the same time**, e.g.



•**Serial** communications are therefore **slow** (one bit at a time) **compared** to **Parallel** communication.

Serial interfaces have **simpler** electronics than **parallel** interfaces because the interface only has to deal with **a single stream** of bits whereas the **parallel interface** has to handle **multiple bits** simultaneously, and as such, the parallel interface has more complex electronics.

Because of the complexity of parallel interfaces and the problems that can arise through signal degradation, parallel communication tends to be used over very short distances. Serial communication is used for long distances such as telephone networks and the Internet.

Main memory

There are various aspects of main memory that can affect system performance. These include:

- speed of access
- word size
- amount of memory
- cache memory.

The effect of data bus width on performance (Word size)

This is the basic number of bits that the processor can handle in a single operation. Thus a 32-bit processor can handle 32 bits in a single operation.
An 8-bit processor can add together two 32-bit numbers but this would take quite a few operations, whereas a 32-bit processor could perform the same task in a single operation.
If the word size of the computer and the data bus width are the same, this allows data transfers to and from main memory to be carried out in a single operation. However, computers are not always designed like this and often compromises are made due to chip fabrication and manufacturing costs.

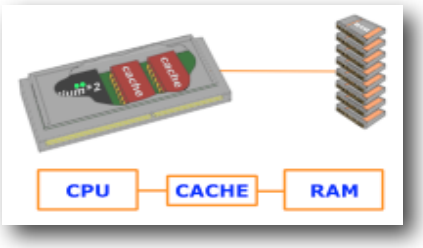
Amount of memory

Main memory is a mixture of random access memory (RAM), read only memory (ROM) and empty space. Empty space means there is less physical memory present than can be directly addressed. Physical memory can therefore be expanded by adding more memory modules as and when required (memory upgrade). If you cannot load all the software you want at the one time, then adding extra memory will improve your system's performance.

Cache Memory

Cache memory is a dedicated, small, fast RAM set (usually **SRAM**) used to store data in transit between the CPU and memory. This is usually SRAM which is faster than DRAM, and although this is much smaller than RAM there is a benefit from the fact that it is always faster to access a small memory segment.
With modern, fast operating systems handling large data sets, main memory access times again become important.

Cache Memory is an **extremely fast** storage area used as a **temporary** store for data in **transit** between the **processor** and main **memory**.
It is often part of the processor itself. so that there is infinitesimal delay in, **e.g.** transferring data from a register to **cache** memory. This allows the processor to proceed with some other task.



Virtual storage

A typical processor today may have a 32-bit address space allowing it to address 4Gb of memory; however, it is rare to find a machine equipped with a full 4Gb of RAM. In contrast, it is very common to find machines equipped with large, cheap amounts of hard disk – which can easily be in excess of 40Gb. It therefore seems quite apparent that using some of the hard disk as slow memory would allow us to utilise the full address space of the processor.