From the specification

**OCR GCSE Computing**
- Explain how common characteristics of CPUs such as … **cache size** and …
- Explain the need for the following functions of an **operating system**: user interface, memory management, peripheral management, multi-tasking and security

**AQA GCSE Computer Science**
- Be able to explain the effect of common CPU characteristics on the performance of the processor. These should include … **cache size/types** …
Outline

• Memory overview

• Understanding performance
  • Latency versus Bandwidth
  • How a cache works

• What is an Operating System for?
TYPES OF MEMORY
Semiconductor

- RAM
  - Volatile
  - Dynamic or static
- ROM
  - Non-volatile
  - Maybe eraseable
- Flash
  - Non-volatile
  - Limited life

What is ROM used for?
Disk

- Capacity: GBytes
- RPM: how fast it spins (RPM)
- Size (diameter – in): how big?
- Interface: will it work in my PC?
- Buffer size (*it's a cache*): MBytes
- Bandwidth (peak, sustained): MByte/second
- **Performance?**
Optical and Tape

- Tape
  - Magnetic
  - Re-writeable
  - Serial access

- CD-ROM
- DVD
  - Distribution
  - Backup
  - Read-only or read/write
Trends

• Tape is on the way out
  • Capacity no longer exceeds disk
  • Price / byte no longer less than disk

• Disk being replaced by Flash
  • Flash only devices: e.g. iPad, RPi, most phones
  • Solid state – 'faster'
Storage Over a Network

- Shared storage
  - Local network – e.g. around office
  - Internet – e.g. dropbox
- Bandwidth: network must be
  - ~ as fast as disk
  - Not too far away
Storage Characteristics

- Capacity – Bytes
  - Cost per byte (disk is cheapest)
- Volatile / non-volatile (permanent)
  - RAM is volatile
- Access
  - Random – anywhere (RAM)
  - Sequential – only in sequence (tape)
  - Block – hard disk
- Speed – latency and/or bandwidth
Understanding Performance

- Latency versus bandwidth
- How does a cache work?
Latency versus Bandwidth

• Wait for the bus *versus* how quickly it goes
• Bandwidth: is rate of data – Bytes / sec
  • Can be increased
• Latency: is **delay** – seconds
  • Speed of light (c) – $3 \times 10^8$ m/s
  • Fibre-optic / electrical signals go at $\sim 1/2$-1/3 c

<table>
<thead>
<tr>
<th>Distance</th>
<th>Time</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>30cm</td>
<td>1ns</td>
<td>1 tick of the CPU clock</td>
</tr>
<tr>
<td>30m-300m</td>
<td>100ns-1μs</td>
<td>Distance to school file server</td>
</tr>
<tr>
<td>6000Km</td>
<td>20ms</td>
<td>Distance to USA</td>
</tr>
<tr>
<td>30,000Km</td>
<td>100ms</td>
<td>Distance to a geostationary satellite</td>
</tr>
</tbody>
</table>
How Does a Disk Work?

- Disk spins e.g. 100 times/sec (6000rpm)
- Reading head moves
  - Track: ring around the disk surface
Disk and Memory Latency

- Latency of the disk
  - 1 rotation every 10ms (100x a second)
  - Average wait 5ms (= 5,000,000 cpu instructions)
  - ... also arm movement

<table>
<thead>
<tr>
<th>Memory Type</th>
<th>Delay (CPU cycles)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAM</td>
<td>10 or more</td>
</tr>
<tr>
<td>SSD (Flash)</td>
<td>1,000s</td>
</tr>
<tr>
<td>Disk</td>
<td>1,000,000s</td>
</tr>
</tbody>
</table>
What's a Cache?

- Small fast memory
- Copy of part of a larger slower memory

Can keep up with CPU

10-20x slower than CPU
Lots of Caches in a Computer

- Main memory cache
  - Hierarchy: L1 (small & fast), L2, L3 (larger & less fast)
- Disk cache
  - RAM in the disk drive
- File buffer
  - Read block of file into memory
  - *Buffer writes too*
Memory Hierarchy

• Trade-off cost, capacity and speed
What’s an Operating System For?

- Input and output
- Organising files
- Managing processes and memory
Manage I/O Buses and Drivers

- I/O cards on buses
  - How to move data?
- Interface standards
  - PCI, USB, SATA
- Device drivers
File Systems

- Files organised into directories
- Hierarchy
- Ownership and protection
- Attributes
  - Size
  - Dates

- Illusion: disk is organised into file
History

• Mainframes
  • Multi-programming
  • Time sharing
• Personal computer
  • *Now with Mainframe OS*

Large computers  One Program  Many Programs  ... & many users

Small computers
Processes

• It’s Quicker to Run Several Programs at Once
  • A single program must wait for the disk
  • Several programs improves throughput

• Multiple processes on one processor
  • An illusion ...
  • ... created by juggling

• But there are consequences ...
Memory

- Lots of process ➔ sharing memory

- Where in memory is my code and data?
  - No reserved place

- Who get the memory?

- Virtual memory
  - Illusion: your program can use all the memory
  - Illusion: the computer has as much memory as you want
Memory – Where?

- Is your data in cache or main memory?
- What happens when the programs overfill memory?
- The OS ‘hides’ this from the programmer
Interrupts

• I/O device speed varies
• How many CPU clock cycles per keystroke?

• How to avoid CPU waiting for I/O?
• Interrupt
  • Signal from outside the CPU
  • ... changes the program
OS is about Illusions

- Several programs are running simultaneously
  - The computer switches from one to another
- My program has all the memory
  - The memory is shared between programs
- Disk is organised in files
  - The disk has blocks; the OS maps file names to blocks
- Storage devices work the same
  - Files may be arranged differently on magnetic, flash and optical drives
- OS creates an ‘ideal’ computer from a real one
Computer Layers

- Each layer solves a problem

<table>
<thead>
<tr>
<th>GUI</th>
<th>Shell</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processes</td>
<td></td>
<td>Files</td>
</tr>
<tr>
<td>Operating System</td>
<td></td>
<td>Networks</td>
</tr>
<tr>
<td>Binary data</td>
<td>Binary code</td>
<td></td>
</tr>
<tr>
<td>CPU</td>
<td>RAM</td>
<td>Disk, I/O</td>
</tr>
<tr>
<td>Logic Gates</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Summary

• Storage characteristics
  • Capacity and cost
  • Volatile / permanent
  • Access: random, block, sequential

• Performance
  • Bandwidth – rate of data
  • Latency – delay for data
  • A cache reduces (effective) latency

• Operating System creates illusions
  • Standard I/O interfaces
  • Process and memory management