SCHOLAR Study Guide

SQA Higher 2004-5

Computing Unit 1
Computer Systems

Approved for use in the SCHOLAR Forum schools in 2004-5.

David Bethune
Heriot-Watt University

Andy Cochrane
Heriot-Watt University

Tom Kelly
Heriot-Watt University

Ian King
Heriot-Watt University

Richard Scott
Heriot-Watt University

Interactive University

Edinburgh EH12 9QQ, United Kingdom.

First published 2004 by Interactive University

Interactive University publishes and delivers, through Local Learning Partners around the world, a range of programmes from universities and colleges in Scotland.

British Library Cataloguing in Publication Data

Interactive University

SCHOLAR Study Guide Unit 1: Computing

1. Computing

ISBN 1 904647 33 2

Typeset by: Interactive University, Wallace House, 1 Lochside Avenue, Edinburgh, EH12 9QQ.

Printed and bound in Great Britain by Graphic and Printing Services, Heriot-Watt University, Edinburgh.

Part Number 2004-1111

Contents

Acknowledgements

Thanks are due to the members of Heriot-Watt University's SCHOLAR team who planned and created these materials, and to the many colleagues who reviewed the content.

Programme Director: Professor R R Leitch

Series Editor: Professor J Cowan

Subject Directors: Professor P John (Chemistry), Professor C E Beevers (Mathematics), Dr P J B King (Computing), Dr P G Meaden (Biology), Dr M R Steel (Physics), Dr C G Tinker (French)

Subject Authors:

Biology: Dr J M Burt, Ms E Humphrey, Ms L Knight, Mr J B McCann, Mr D Millar, Ms N Randle, Ms S Ross, Ms Y Stahl, Ms S Steen, Ms N Tweedie

Chemistry: Mr M Anderson, Mr B Bease, Dr J H Cameron, Dr P Johnson, Mr B T McKerchar, Dr A A Sandison

Computing: Mr I E Aitchison, Dr P O B Holt, Mr S McMorris, Mr B Palmer, Ms J Swanson, Mr A Weddle

Engineering: Mr J Hill, Ms H L Jackson, Mr H Laidlaw, Professor W H Müller

French: Mr M Fermin, Ms B Guenier, Ms C Hastie, Ms S C E Thoday

Mathematics: Mr J Dowman, Ms A Johnstone, Ms O Khaled, Mr C McGuire, Ms J S Paterson, Mr S Rogers, Ms D A Watson

Physics: Mr J McCabe, Mr C Milne, Dr A Tookey, Mr C White

Learning Technology: Dr W Austin, Ms N Beasley, Ms J Benzie, Dr D Cole, Mr A Crofts, Ms S Davies, Mr A Dunn, Mr M Holligan, Dr J Liddle, Ms S McConnell, Mr N Miller, Mr N Morris, Ms E Mowat, Mr S Nicol, Dr W Nightingale, Mr R Pointon, Mr D Reid, Dr R Thomas, Dr N Tomes, Ms J Wang, Mr P Whitton

Cue Assessment Group: Ms F Costigan, Mr D J Fiddes, Dr D H Jackson, Mr S G Marshall

SCHOLAR Unit: Mr G Toner, M G Cruse, Ms A Hay, Ms C Keogh, Ms B Laidlaw, Mr J Walsh

Media: Mr D Hurst, Mr P Booth, Mr G Cowper, Mr C Gruber, Mr D S Marsland, Mr C Nicol, Mr C Wilson

Administration: Ms L El-Ghorr, Dr M King, Dr R Rist,

We would like to acknowledge the assistance of the education authorities, colleges, teachers and students who helped to plan the SCHOLAR programme and who evaluated these materials.

Grateful acknowledgement is made for permission to use the following material in the SCHOLAR programme:

To the Scottish Qualifications Authority for permission to use Past Papers assessments.

The financial support from the Scottish Executive is gratefully acknowledged.

Topic 1

Data Representation

Contents

Prerequisite knowledge

Before studying this topic you should be able to:

- *Represent positive numbers in binary up to 8 bits;*

- *Understand the advantages of using binary numbers;*
- *Describe the use of floating point representation of real numbers using the terms mantissa and exponent;*
- *Describe file sizes, backing storage and main memory capacities using the terms: bit, byte, kilobyte, Megabyte, Gigabyte, Terabyte;*
- *Understand the use of the ASCII code and character sets;*
- *Describe the bit map method of graphic representation of black and white;*
- *Calculate storage requirements.*

Learning Objectives

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

- *Represent positive and negative numbers in binary;*
- *Convert between binary and decimal;*
- *Describe the use of floating point numbers;*
- *Convert to and from bit, byte, kilobyte, Megabyte, Gigabyte, Terabyte;*
- *Describe the need for Unicode;*
- *Describe the bit map method of graphic representation;*
- *Understand the relationship of bit depth and colour;*
- *Describe the vector method of graphic representation;*
- *Discuss relative advantages and disadvantages of bit mapped and vector graphics;*
- *Understand the relationship between bit depth and file size;*
- *Realise the need for data compression.*

Revision

The following exercise tests the prerequisites for this topic. Ensure that you are happy with your responses before progressing.

Q1: The decimal number 73 is expressed in binary as

a) 01001001
b) 01111100
c) 11111
d) 01110011

Q2: Use the ASCII character set to translate the following bytes:

01000011 0110111 1 01100100 01100101

Q3: A 1 bit/pixel monochrome display has 200 vertical lines. Each vertical line has 640 horizontal dots. Calculate the memory required to store a single screen shot. Express your answer in bytes and Kilobytes.

1.1 Introduction

This unit on Data Representation considers how computers represent data by using binary numbers. At the lowest level in the computer, only binary numbers can be understood. As humans we are much more familiar with the decimal system and think of numbers in terms of base 10. So in order to understand how the computer processes data, we must be comfortable with both binary and decimal numbers. Like us, the computer must also be able to process positive and negative numbers that can also be very large or very small. This unit considers all such numbers and how they can be represented.

Then we can consider how the computer uses numbers to represent other forms of data, such as text and graphics by using codes and other numeric techniques. The unit is completed by considering some of the practicalities of representing data in its various forms and how such data can be stored within a computer system.

1.2 Numbers

We use the base 10 number system to represent *whole numbers*, *integers* and *fractional* numbers. This number system uses the 10 digits 0,1,2,3,4,5,6,7,8,9 to represent numbers, which explains why it is called base 10. The *value* of a decimal digit is given by its *position* within the base 10 number system.

When expressing large numbers in terms of powers of 10 the following abbreviations are used:

- $10^1 = 10$
- $10^2 = 100$

- 10^3 = 1000 = **1 kilo**
- 10^6 = 1,000,000 = **1 Mega**
- 10^9 = 1,000,000,000 = **1 Giga**
- 10^{12} = 1,000,000,000,000 = **1 Tera**

Example : Deriving the value of 345

The whole number 345 is represented in base 10 as:

10^2	10^1	10^0
3	4	5

This gives the value of:

	3 x 100
+	4 x 10
+	5 x 1
=	345

Example : Deriving the value of 23.75

The number 23.75 is represented in base 10 as:

10^1	10^0	10^{-1}	10^{-2}
2	3	7	5

This gives the value of:

	2 X 10
+	3 X 1
+	7 X 0.1
+	5 X 0.01
	23.75

1.2.1 The binary number system

When numbers are represented electronically, the most convenient base is 2, where each column, reading from the right is a power of two. The base 2 number system uses 2 symbols, 0 and 1 to represent a value.

In computing systems, large numbers are expressed in terms of powers of 2 and use the following abbreviations:

- 2^1 has a decimal equivalent of 2
- 2^2 has a decimal equivalent of 4
- 2^3 has a decimal equivalent of 8
- 2^4 has a decimal equivalent of 16
- 2^5 has a decimal equivalent of 32
- 2^6 has a decimal equivalent of 64

- 2^7 has a decimal equivalent of 128
- 2^8 has a decimal equivalent of 256
- 2^9 has a decimal equivalent of 512
- 2^{10} has a decimal equivalent of 1024 and is abbreviated to **1 kilo**
- 2^{20} has a decimal equivalent of 1,048,576 and is abbreviated to **1 Mega**
- 2^{30} has a decimal equivalent of 1,073,741,824 and is abbreviated to **1 Giga**
- 2^{40} has a decimal equivalent of 1,099,511,627,776 and is abbreviated to **1 Tera**

Example : Converting 00111101_2 to decimal

The number represented by 00111101_2 is equal to 61_{10}. The value 61 is obtained using the method illustrated below:

	2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
	0	0	1	1	1	1	0	1
=	(0*128) +	(0*64) +	(1*32) +	(1*16) +	(1*8) +	(1*4) +	(0*2) +	(1*1)
=			32 +	16 +	8 +	4 +		1
=	61_{10}							

= (0*128) + (0*64) + (1*32) + (1*16) + (1*8) + (1*4) + (0*2) + (1*1)

= 32 + 16 + 8 + 4 + 1

= 61_{10}.

The binary number system has the huge advantage that only two symbols are required, 0 and 1. These can easily be represented in a computer system by a switch or transistor being on or off, or by a high or low voltage level. Imagine how difficult it would be to represent 10 discrete logic values, as required in the base 10 number system.

Binary representation also simplifies the number of arithmetic rules that need to be applied in calculations. Binary arithmetic has fewer rules. Try adding two binary numbers together to find out how many rules are required.

1.2.2 Converting decimal representation to binary

There are two ways to convert decimal numbers into binary.

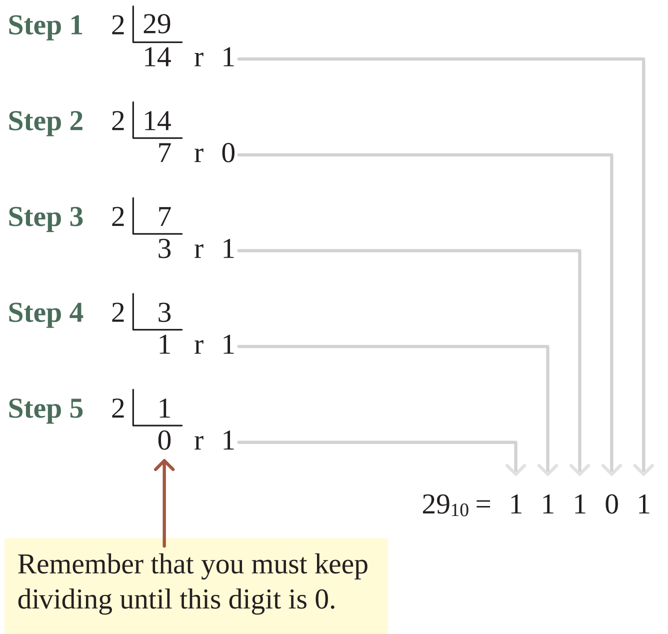
Method 1.

If you were writing a conversion routine in a computer program this is the way you would do it.

To convert 29 into binary:

Decimal to Binary Conversion Using Division by 2

Convert 29_{10} to binary



Notice that you have to keep dividing until the dividend is 0.

Simulation of decimal to binary conversion

On the web is a simulation of the conversion of a decimal number to binary. You should now look at this simulation.

Method 2.

Convert 29 into binary. If you're doing this conversion on paper this is probably the easier way.

1. Write down the position values at the top of the page.

2^4	2^3	2^2	2^1	2^0
16	8	4	2	1

2. Select the biggest position value that is not greater than your chosen number (in this case 16 is the largest value $<$ 29).

3. Put a 1 in the 16s column.

2^4	2^3	2^2	2^1	2^0
16	8	4	2	1
1				

4. Subtract 16 from 29, $29 - 16 = 13$.

5. Start again with the number 13.

6. Select the biggest position value that is not greater than 13 (in this case 8).

7. Put a 1 in the 8s column.

2^4	2^3	2^2	2^1	2^0
16	8	4	2	1
1	**1**			

8. Subtract 8 from 13, $13 - 8 = 5$.

9. Start again with the number 5.

10. Select the biggest number that is not greater than 5 (in this case 4).

11. Put a 1 in the 4s column.

2^4	2^3	2^2	2^1	2^0
16	8	4	2	1
1	**1**	**1**		

12. Subtract 4 from 5, $5 - 4 = 1$.

13. Start again with 1.

14. Select the biggest position value that is not greater than 1 (in this case 1).

15. Put a 1 in the 1s column.

2^4	2^3	2^2	2^1	2^0
16	8	4	2	1
1	**1**	**1**		**1**

16. Subtract 1 from 1, $1 - 1 = 0$.

17. Put a 0 in the empty columns.

2^4	2^3	2^2	2^1	2^0
16	8	4	2	1
1	**1**	**1**	**0**	**1**

$29_{10} = 11101_2$

Decimal to binary conversion - a ready reckoner

On the Web is a simulation of the repeated division method used to convert a decimal number to binary. Using this interaction you can enter your own numbers and test the conversions. You should now look at how this method works and then use it to convert from decimal to binary the list of numbers which follow.

Q4: Convert 134_{10} to binary

Q5: Convert 148_{10} to binary

Q6: Convert 394_{10} to binary

Using both methods to convert decimal to binary

Convert the following numbers to binary. Do each calculation using both conversion methods.

Q7: 29_{10}

Q8: 18_{10}

Q9: 79_{10}

Q10: 273_{10}

Q11: 127_{10}

Q12: 742_{10}

Q13: 4023_{10}

Q14: 9755_{10}

1.2.3 Converting binary representation to decimal

Again there are two ways to do this.

Method 1

1 1 0 1 0

1*2
=2
2 + 1
=3

3*2
=6
6 + 0
=6

6*2
=12
12 + 1
=13

13*2
=26
26 + 0
=**26**

Note that the last line is not multiplied by 2. Again this is the method you would use if this conversion were being coded for a computer.

Method 2.

The second conversion method is easier to calculate on paper.

2^4	2^3	2^2	2^1	2^0
16	8	4	2	1
1	**1**	**0**	**1**	**0**
1*16				
	+1*8			
		+0*4		
			+1*2	
				+0*1

= 16 + 8 + 0 + 2 + 0

= $\mathbf{26}_{10}$

Simulation of binary to decimal conversion

On the Web is a simulation that shows you how binary numbers are converted to decimal. You should now look at this simulation. Once you are confident in your understanding of how it works, you should use it to convert the following binary numbers to decimal.

Binary to decimal conversion - practice

Try these examples, doing each both ways.

Q15: 1011_2

Q16: $1\ 1010\ 1101_2$

Q17: $1\ 0001\ 0011_2$

Q18: $111\ 1111_2$

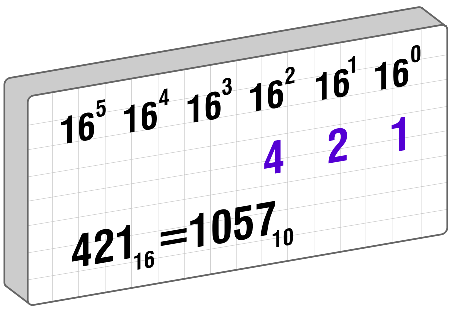
Q19: $1000\ 0000_2$

Q20: $101\ 1101\ 0001_2$

Q21: $11\ 1001\ 0110_2$

Q22: $110\ 0010\ 0101_2$

1.2.4 Hexadecimal



Values can be represented in any base. Long binary numbers can be difficult to read correctly. A memory address position that is 4 bytes long might look like this:

11010100010110010011001010010110

It is very easy to make mistakes when you work with a number like this. To solve this problem numbers are often represented in base 16 which is called hexadecimal. The numbers are much easier to work with and the conversion between base 2 and base 16 is very simple.

Numbers in base 16 need 16 symbols. The digits 0-9 and the letters A-F are used to give 16 symbols.

Decimal	**Hexadecimal**
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	A
11	B
12	C
13	D
14	E
15	F

The number positions in hexadecimal are:

position 4	position 3	position 2	position 1
16^3	16^2	16^1	16^0
4096	256	16	1
		2	**C**

= 2*16 + 12*1

= $\mathbf{44}_{10}$

Decimal, binary and hexadecimal representations of the numbers 1 to 24 are shown below.

Decimal	Binary	Hexadecimal	Decimal	Binary	Hexadecimal
1	1	1	13	1101	D
2	10	2	14	1110	E
3	11	3	15	1111	F
4	100	4	16	1 0000	10
5	101	5	17	1 0001	11
6	110	6	18	1 0010	12
7	111	7	19	1 0011	13
8	1000	8	20	1 0100	14
9	1001	9	21	1 0101	15
10	1010	A	22	1 0110	16
11	1011	B	23	1 0111	17
12	1100	C	24	1 1000	18

1.2.5 Why binary?

All the logic circuits used in digital computers are based upon two-state logic. That is, quantities can only take one of two values, typically 0 or 1. These quantities will be represented internally by voltages on lines, zero voltage representing 0 and the operating voltage of the device representing 1. The reason two-state logic is used is because it is easy and economic to produce such devices.

Modern processors contain many millions of **logic gates**. A logic gate performs a logical operation upon its input signals to produce its output signals. Logic gates are implemented nowadays using transistor circuits. The logical operations used are those of **Boolean** algebra developed by George Boole. The main logical operations that logic gates must implement are: AND, OR, NOT and NAND (not AND). These are defined by:

- **A AND B** - outputs a 1 if the inputs A and B are both 1, otherwise outputs 0.
- **A OR B** - Outputs a 1 if either (or both) of the inputs are 1, otherwise outputs 0.
- **NOT A** - Outputs a 1 if A is 0 and a 0 if A is 1.
- **A NAND B** - Outputs a 0 if A and B are both 1, otherwise outputs a 1.

The computer represents all information internally, both instructions and data, by patterns of 0s and 1s. Each individual 0 or 1 is held in a **bit**, and memory and registers within the computer are implemented as collections of bits. In memory the smallest addressable quantity is the **byte**, which is eight bits. While the bits are often interpreted as truth values (true or false), collections of bits are used to represent numbers and characters. Numbers are represented by some version of the **binary number system**, which only has the two digits 0 and 1. Characters are represented by specific patterns of 0s and 1s in a byte, using some encoding such as ASCII (American Standard Code for Information Interchange) or in two bytes using the Unicode standard.

1.2.5.1 Review questions 1

Q23: Why do computers use a binary representation to store information?

Q24: A truth table for the logical AND operation is given as follows:

A	B	A AND B
0	0	0
0	1	0
1	0	0
1	1	1

Write down similar truth tables for the logical operations OR and NAND.

Q25: How would a character in the ASCII character set be represented inside a computer?

1.2.6 Range of positive integers

The range of positive integers that can be represented using n bits is **0..2^n-1**. You can see a pattern leading to this general formula in Table 1.1

Table 1.1

No. of bits	Binary patterns	Decimal values	Range
1	0, 1	0, 1	$0..1 = 0..2^1 - 1$
2	00, 01, 10, 11	0, 1, 2, 3	$0..3 = 0..2^2 - 1$
3	000, 001, 010, 011, 100, 101, 110, 111	0,1, 2, 3, 4, 5, 6, 7	$0..7 = 0..2^3 - 1$

Q26: What is the range of positive integers that can be represented using 8 bits?

a) 0..16
b) 0..255
c) 0..256

Q27: What is the range of positive integers that can be represented using 12 bits?

a) 0..24
b) 0..144
c) 0..4095

Q28: The general formula for the range of positive integers for n bits is:

a) $0..2^n-1$
b) $0..2^n$
c) 0..2n-1

1.2.7 Negative integers - Sign and Magnitude

Sign and magnitude is one way of representing negative integers.

Let **0**, indicate a positive number and **1** indicate a negative number. Using 3 bits we can represent the following range of numbers:

Binary Pattern	Number Value
011	+3
010	+2
001	+1
000	+0
100	-0
101	-1
110	-2
111	-3

There are some difficulties in using sign and magnitude representation. These are:

- carrying out addition and subtraction requires consideration of the signs of the numbers and their relative magnitudes;
- there are two representations of zero, testing for a zero result is more complex than it need be.

For these reasons, sign and magnitude representation is not used for the representation of negative integers in modern computers.

1.2.8 Negative integers - Two's Complement

Two's Complement is another method of representing negative integers.

Let us assume we have a 3-bit register and we use the most significant bit (MSB) as a sign bit. This bit also contributes to the size of the number.

The range of positive and negative numbers that can be represented using 3 bits is shown below.

Register Content	Value	Register Content	Value		Comments
0 0 0	0	**1** 0 0	-4	= -4 + 0	notice the sign bit
0 0 1	+1	**1** 0 1	-3	= -4 + 1	0 for positive
0 1 0	+2	**1** 1 0	-2	= -4 + 2	1 for negative
0 1 1	+3	**1** 1 1	-1	= -4 + 3	

Notice the following **three** facts:

- the largest negative number is -4 = -2^{n-1}
- the largest positive number is 3 = $+2^{n-1}-1$
- 1-1 = (1) + (-1) = 0. So the arithmetic is fine and produces the expected results.

Converting an integer to Two's Complement is not easy. There are two methods that you can apply.

Method 1: Converting a number to Two's Complement.

- Step1. For positive numbers, treat all the bits as usual - e.g. +3 = 0011.
- Step 2. For a negative number, set the sign bit (MSB) and then:
  - set the remaining bits such that the value of the sign bit plus the value of the remaining bits = the value of the negative number i.e.
  - **-5 = 1000 (sign bit value -8) + (0011) (value +3) = 1011**

Q29: Using this method, what is the Two's Complement representation of -10?

Method 2: Two's complement representation

On the Web is a simulation that illustrates a second method of converting a negative integer to binary. You should now look at this simulation.

5 min

Using either of the two methods you have learned, represent the following numbers in Two's Complement form:

Q30: -35

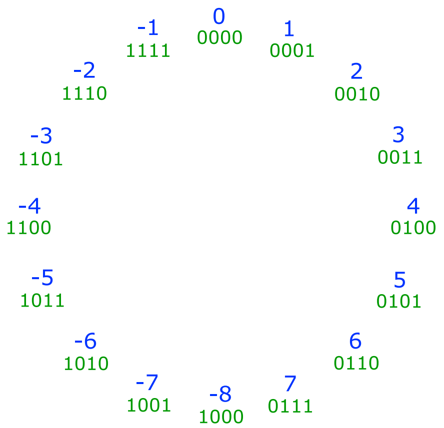
Q31: -9

Q32: 23

Deriving the range of numbers in Two's Complement

5 min

Construct a table showing the range of positive and negative numbers that can be represented in Two's Complement using 8 bits. How many unique numbers are there? What can you say about the range? Can you construct a general formula?



Decimal arithmetic uses different operations for addition and subtraction. Using Two's Complement, subtraction is carried out using the machine operation for addition. This system works for bit strings of any length. The examples given will use 8 or 4-bit strings for simplicity but a working PC is likely to use 32- or 64-bit strings. The first bit is again used to indicate a negative value but it also bears the position value.

The Two's Complement number 1000 1010 is evaluated as follows:

1000 1010

= -128 +0+0+0+8+0+2+0

= -118

Converting a number to its negative form is a two stage process:

1. complement (invert) all the bits (the result at this stage is known as **One's Complement**).
2. add 1.

If you're working with 8-bit strings all values must appear as 8 bits and the possible values will be in the range -128 to 127.

Example : Using a 4-bit binary string

To convert 1101 (13_{10}) to its negative:

write the number in 8 bit format	0000 1101
invert the bits	1111 0010
add 1	0000 0001
	1111 0011

It is easy to check the answer:

1111 0011

= -128 +64 +32 +16 +0 +0 +2 +1

= -128 + 115

= -13

Simulation of 4-bit number conversion to Two's Complement

On the web is a simulation of the conversion a 4-bit binary number to Two's Complement. You should now look at this simulation.

The table below shows the Two's Complement representation of the range -8.. + 7

Decimal	Binary Two's Complement representation
7	0111
6	0110
5	0101
4	0100
3	0011
2	0010
1	0001
0	0000
-1	1111
-2	1110
-3	1101
-4	1100
-5	1011
-6	1010
-7	1001
-8	1000

When 1 is added to the binary representation for -1 the result is a 5 bit number. Because the machine has only allocated 4 bits, the 5th bit is ignored as overflow. Adding one to any number leads to a continuous cycle of 4 bit numbers. The mileometer in a car works in the same way.

	1111
+	0001
1	0000

Car mileometer

On the Web is a simulation which shows you how the values in a car mileometer change as you move the car forwards and backwards. You should now look at this animation.

This same process can be used to convert negative numbers to positive.

This ability to treat positive and negative numbers in the same way is the reason that Two's Complement is the most commonly used machine representation for negative numbers.

Simulation of Two's Complement representation - a ready reckoner

On the Web is a simulation which illustrates how any number in the range -1..-100 is converted to Two's Complement. You should now look at this simulation.

Converting 8-bit binary numbers to Two's Complement - practice

Convert the following 8 bit binary numbers to their negative two's complement equivalents. You should check your answers by converting all binary numbers to decimal.

Q33: 0010 1010

Q34: 0111 1000

Q35: 0000 1000

Q36: 0011 0111

Q37: 1010 1001

Q38: 1111 1111

Q39: 0100 0110

Q40: 1001 0010

1.2.9 Floating point numbers

Floating-point representation represents decimal fractions in *scientific notation*. This notation represents numbers as a *base* number and an *exponent*. For example, 234.567 in decimal could be represented as 2.34567×10^2, with base = 10 and exponent = 2.

Q41: Using scientific notation, which of the following is the correct representation of the decimal value 2567.348?

a) 256.7348×10^1
b) 25.67348×10^2
c) 2.567348×10^3
d) all of the above
e) none of the above

Floating-point has the main advantage over fixed-point of being able to represent very large or very small numbers. Furthermore, when dividing two large numbers, floating point representation does not suffer from any loss of precision, as fixed point representation can.

To represent a decimal fraction in a computer, floating point representation uses a non-zero fractional part, the **mantissa**, and an integer part, the **exponent**, to represent a decimal number. The mantissa can be represented in Two's Complement while the exponent uses Sign and Magnitude.

Examples of floating point representation in decimal and binary are shown in Table 1.2 and Table 1.3

Table 1.2: Floating point representation in decimal

Decimal	Sign	Mantissa	Exponent	Equivalence
346.789	+	.346789	+3	0.346789×10^3
-0.002639	-	.2639	-2	-0.2639×10^{-2}

Table 1.3: Floating point representation in binary

Binary	Sign	Mantissa	Exponent	Equivalence
+110.011	+	.110011	+11	0.110011×2^3
-0.001011	-	.1011	-110	-0.1011×2^{-2}

The *range* of the number is represented by the exponent, while the *precision* is represented by the mantissa.

Range vs precision

Given a fixed number of bits in a register, there are decisions that need to be made as to how many bits will be allocated to the mantissa and how many bits allocated to the exponent.

Applications that deal with very large numbers, for example astronomy, are more interested in representing a big enough range of numbers. Whereas at the other end of the spectrum, scientists dealing with molecular data (where distances can be measured in 10^{-9} of a metre) place greater importance on precision.

1.2.10 Review questions

Q42: Convert the following integers to binary.

- 15
- 235
- 1623

Q43: What problems are there in representing negative integers using *sign and magnitude*?

Q44: Explain the functions that the *mantissa* and *exponent* have in floating point number representation.

Q45: Describe how negative numbers are represented in binary.

Q46: Show how the following integers are represented in Two's Complement.

- -5
- -62
- -114

Q47: Using the method shown in the Web animation, convert the decimal fraction 0.625 to its binary equivalent.

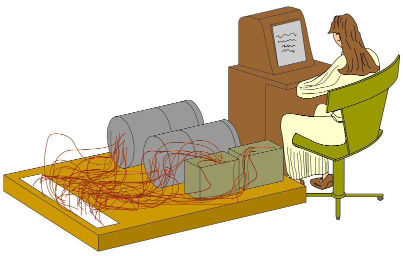
Q48: Use the same method to convert the decimal fraction 0.1875 to its binary equivalent.

Q49: What is the decimal fraction 0.1 expressed in binary to 10 binary places?

Q50: Convert the binary fraction 0.11 to decimal.

Q51: Convert the binary fraction 0.0101 to decimal.

1.3 Text



A teletypewriter.

Input, output, backing storage and data communications devices store and manipulate data in character code. During processing, text data stays as character code but numerical data is converted to one of the representations you have already met.

Programmers specify the type of data within their code and conversion is then carried out by hardware or software.

The set of characters that can be represented by the computer is known as the **character set**. Many computers have the flexibility of using several character sets, but we will restrict our discussions to **ASCII** and **Unicode**.

1.3.1 ASCII

American **S**tandard **C**ode for **I**nformation **I**nterchange (ASCII) was first developed for teletypewriters and is now an internationally agreed standard for storing information. It is a character set which forms the basis for almost all other character sets.

ASCII uses 7 bits per character, giving a possible 128 different characters. It has 96 displayable characters, enough to represent a letter or symbol of the English alphabet and numerical symbols. There are 32 special character codes known as **control characters**. A single code is used to represent each character which is stored in binary. When data is to be displayed or printed, the code is converted into an appropriate shape.

An extract of ASCII codes is shown in Table 1.4.

Table 1.4: Extract of ASCII codes

Symbol	ASCII	Symbol	ASCII	Symbol	ASCII	Symbol	ASCII
0	48	a	97	ACK	6	\	92
1	49	b	98	BEL	7	]	93
2	50	c	99	BS	8	^	94
3	51	d	100	HT	9	-	95
4	52	e	101	LF	10	'	96

When IBM PCs were designed, they moved and stored data in 8-bit units. The extra bit increased the character set from 128 to 256, producing an *extended* character set that could accommodate international dialects or alphabetic characters in foreign languages (French and German accented characters are examples).

This extended character set became the ISO 8859 standard, with variations of ISO *Latin-1*, *Latin-2* and *Greek*.

1.3.2 Unicode

An increase in worldwide communications brought a need to exchange information internationally. This presented a problem for Latin based character sets. Languages such as Japanese and Arabic, for example, do not use the Latin-based symbols of English, French or German and have entirely different symbol shapes. A Japanese symbol is shown in Figure 1.1.

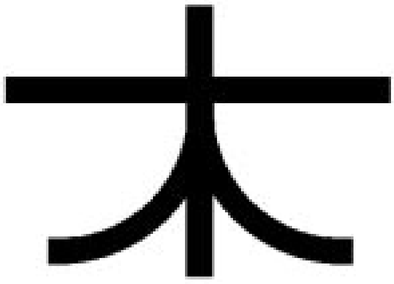


Figure 1.1: Japanese symbol

A solution to the problem of supporting multilingual text was to encode characters using 16 bits, thus providing 65,536 possible symbols. This is **Unicode** which is capable of including the characters from all known languages and alphabets in the world.

Applications such as Office97 use Unicode in document files, while Windows98 and MacOS allow you to set a language preference which is stored in the Windows registry on PCs or the system preferences folder on the Macs. Sun, IBM, Xerox, Lotus and Novell are also participating in the development of this standard.

Features of unicode

On the Web is an activity that asks you to identify the features of Unicode. You should now carry out this activity.

1.3.3 Review questions

Q52: The 7-bit ACSII code can represent:

a) 8 symbols
b) 320 symbols
c) 128 symbols

Q53: *Extended ASCII* means:

a) an additional symbol in Japanese can be represented
b) the Windows98 registry cannot be updated
c) an extra bit is used to provide encoding of 256 different character symbols

Q54: Unicode is:

a) a single code.
b) a way of encoding symbols using a 16-bit representation.
c) used only for Japanese symbols

Q55: Why is Unicode necessary?

Q56: Give one example of a language that would need a Unicode representation.

1.4 Graphics

In the early 1970s **raster graphics** displays, based upon the same technology as television, emerged and are now common to all desktop computer displays.

These displays store image data as a matrix of **pixels** (picture elements or pels) in an area of image memory known as a *refresh buffer*.

With earlier technologies image memory formed part of the CPU's main memory, but the advent of cheaper RAM modules has produced separate image stores called Video RAM (**VRAM**).

VRAM represents the entire screen area and the term **bit map** is used to describe the one-to-one mapping of pixels in VRAM to pixels on the screen.

Special display circuitry, called a **video controller**, is used to scan the VRAM, one raster line at a time, to build up the image on the screen. This is illustrated in Figure 1.2

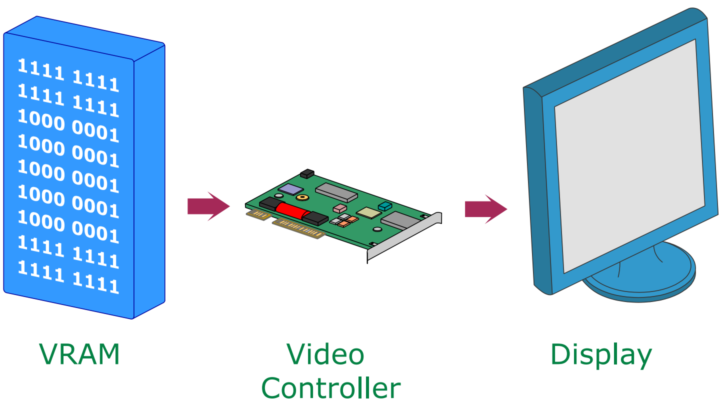


Figure 1.2: Raster scanning an image

How the video display controller works

On the Web is a simulation that shows you how an image is built up on the screen by the video display controller. You should now look at this animation.

Scanning occurs at a rate of 60Hz or more, depending upon the display controller technology.

Image scan inside a monitor

On the Web is an animation that shows you what occurs inside the monitor during an image scan. You should now look at this animation.

Although the image is stored in VRAM as a bit map, it can be represented at the application level in one of two forms: **vector graphic** or **bit mapped graphic**.

1.4.1 Bit mapped graphic representation

A bit mapped image is stored as a file of pixel data and is produced on screen at the same resolution at which it was created. **Resolution** refers to the total number of pixels in the width and height of the image. A bit map is a term that is applied to a 1-bit per pixel graphic system, i.e. where images can have only two possible colour values, 0 and 1, usually black and white.

Figure 1.3 shows a simple graphic produced in a painting package and its corresponding 9 x 9 x 1-bit representation.

0	0	0	0	0	0	0	0	0
0	0	1	1	1	1	1	0	0
0	1	0	0	1	0	0	1	0
0	1	0	0	1	0	0	1	0
0	1	1	1	1	1	1	1	0
0	1	0	1	1	1	0	1	0
0	1	1	0	0	0	1	1	0
0	0	1	1	1	1	1	0	0
0	0	0	0	0	0	0	0	0

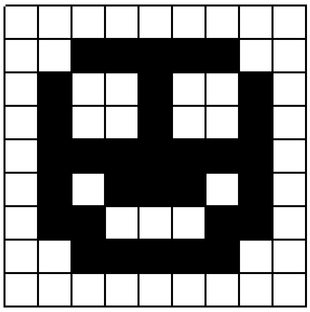


Figure 1.3: Bit map image representation

With multiple bits per pixel, colour and brightness is represented in a pixel map (*pixmap*). As the number of bits per pixel increases, the number of colours that can be represented increases. For instance, using 2 bits per pixel allows 4 colours to be represented, while 8 bits per pixel can represent 256 possible colours. The number of bits used to represent each pixel is called the **bit-depth**

Screen resolutions of 1280 x 1024 with 24-bits per pixel are available on today's personal computers. This requires 1280 x 1024 x 24 = 3.75 Mbytes of VRAM which is fairly low given current VRAM technology.

Scaling a bit map in a painting package is done through **pixel replication** where each pixel is replicated in the x and y direction by a scale factor to produce the enlarged image.

When a bit mapped image enlarges, it shows the same amount of detail spread over a larger area and thus appears coarser. This is frequently a problem in printing where the print resolution is greater than the image resolution.

Advantages of bit mapped graphic representation

- a bit mapped image can be manipulated at the pixel level. Thus a designer may apply particular colour values to a selected pixel area to produce shading or textured effects;

- it is possible to create a wider range of irregular shapes and patterns by simply deleting pixels anywhere on the image.

Disadvantages of bit mapped representation

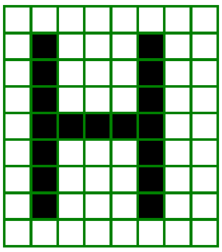
- requires large amounts of storage space;
- image becomes course (jagged) when scaled;
- does not take advantage of resolutions that are higher than the resolution of the image.

Identifying advantages of bit mapped representation

On the Web is an activity that asks you to identify the advantages of bit mapped representation. You should now carry out this activity.

1.4.2 Calculating memory requirements

The diagram shows a portion of the grid eight cells wide, so we could represent this in memory as a collection of bytes, with a bit value of zero to represent black, and a bit value of one to represent green. The top row of cells would thus have a byte value of 11111111_2 (255_{10} or FF_{16}), as would the final row. The second row would have a byte value of 10111011_2 (187_{10} or BB_{16}). The third, fourth, sixth, seventh and eighth rows would have the same value as the second row. The fifth row would have a byte value of 10000011_2 (131_{10} 83_{16}). Note that this example assumes that the byte boundaries line up neatly.



As can be seen from the diagram above, the display grid has some cells switched to give one colour, and the remainder are set to give the alternative.

These cells or dots on a display are usually referred to as pixels (picture elements, or *pels*). For the simple display approach shown above, one bit is all that is needed to determine whether a pixel is set or not set. Put another way, one byte is sufficient to control eight pixels. Knowing this, we can determine the amount of memory required to store an image on a monochrome bit mapped display of a given size by dividing the number of pixels by eight.

The number of pixels possessed by a device can be found by multiplying the number of pixels across the screen by the number of pixels down, so for a display of 80 pixels across by 36 pixels down, the number of pixels will be $36 * 80 = 2880$ pixels.

2880 pixels, with one byte controlling 8 pixels will thus require $2880/8 = 360$ bytes of memory.

Pixels across	Pixels down	Total	Bytes of memory needed
80	36	$80 * 36 = 2880$	$2880/8 = 360$ bytes

Calculating memory requirements - 1

How much memory will the following screens require? Use the headings in the table below to help you perform the calculation.

Q57: 640 * 200 pixels

Q58: 800 * 600 pixels

Q59: 1024 * 768 pixels

Pixels across	Pixels down	Total (pixels across * pixels down)	Bytes of memory needed (Total/8)

1.4.3 Arranging the bytes

There are many different ways of arranging the bytes that hold image information, but one way is to map them so that the first byte represents the top left pixels, the second byte represents the pixels to the right of the first pixel, until the end of the first row is reached, when the next byte holds the information for the left hand end of the second row. For a small (24 pixel by 4 rows) display the layout would look like this:

B	Y	T	E	-	-	-	1	B	Y	T	E	-	-	-	2	B	Y	T	E	-	-	-	3

One particular benefit of using this sort of bit mapped display is that we can easily save the image as a series of bytes to a storage device, such as memory or disk, and if we wish to review the image then it is a simple matter to transfer the image data back into the video memory as a direct copy.

As this image is also in a bit mapped format, we can still move it to and from other storage devices without any translation.

Number of colours	Number of bits
2	1
4	2
256	8
16 million	24

1.4.4 Greyscale

The next level of sophistication is to provide a rudimentary greyscale effect, providing a 'black', 'white' and two levels of 'grey'. As this comprises four different values we need two bits to represent each pixel (00 for black, 01 for darker grey, 10 for lighter grey and 11 for white). As each pixel now requires twice as many bits, we will require twice as much memory for a given screen size.

Continuing with the theme, we can provide more levels of grey by allocating more bits to each pixel. By the time we have eight bits (one byte) to one pixel we can represent 256 different intensities.

Monochrome displays are often clearer, especially for text than colour displays, but fashion, and the requirement to use colour for such items as colour pictures and user interface issues, dictates that colour displays are more likely to be purchased.

Calculating memory requirements - 2

How much memory will the following screens require to provide greyscale images? Use the headings in the table below to help you perform the calculation.

Q60: 640 * 200 pixels

Q61: 800 * 600 pixels

Q62: 1024 * 768 pixels

Total pixels (pixels across * pixels down)	Number of shades of grey	Bits per pixel	Total bits (total pixels * bits per pixel)	Bytes of memory (total bits/8)

1.4.5 Colour

At the simplest level, we can choose to represent one pixel with one byte, and use the 256 different values to represent 256 different colours.

This arrangement has the benefit of being fairly frugal with memory - one byte per pixel, but suffers from the limitation of only being able to show 256 different colours at any one time. Most users are believed to be capable of being able to distinguish several million different colours, and most users would not be content with 8 bit colour.

When viewing monitors, there are three primary colours from which all other colours can be obtained, these are Red, Green and Blue. These are known as the additive primary colours, and are obtained by adding coloured lights. (Note that they are different from the primary colours found in your traditional media paintbox as these the subtractive primaries.)

We saw earlier that we could use 8 bits to provide 256 different shades of grey. If we extend this idea so that we provide 8 bits to describe the amount of Red (0-255) where 0 is no red and 255 is full intensity red, 8 bits to describe the amount of Blue and 8 bits to describe the amount of Green, we can describe 256 $\times$ 256 $\times$ 256 different colours - somewhat over 16 million different colours.

Number of colours	Number of bits
2	1
4	2
256	8
16 million	24

The 24 bit codes made up of 8 bits for Red, 8 bits for Green and 8 bits from Blue are known as RGB codes and are usually expressed in hexadecimal. These codes can be used by a programmer to specify particular colours rather than selecting colour from a drop down menu.

Red	Green	Blue	Colour	Binary value
00	00	00	black	0000 0000 0000 0000 0000 0000
FF	FF	FF	white	1111 1111 1111 1111 1111 1111
00	FF	00	green	0000 0000 1111 1111 0000 0000
00	80	00	half green	0000 0000 1000 0000 0000 0000

Calculating memory requirements - 3

How much memory will the following screens require to provide 16 million different colours? Use the headings in the table below to help you perform the calculation.

Q63: 640×200 pixels

Q64: 800×600 pixels

Q65: 1024×768 pixels

Total pixels (pixels across × pixels down)	Number of colours	Bits per pixel	Total bits (total pixels × bits per pixel)	Bytes of memory (total bits/8)

1.4.6 Image Compression

An image using 24-bit colour graphics will be of an extremely high standard and will represent true colour . Each pixel (or dot) can represent over 16 million colours. On a high resolution display a single image may require several megabytes of memory for storage.

File compression techniques can be used to reduce storage requirements. Specialised techniques for image compression result in images that are indistinguishable from the original as far as the human eye is concerned, but much smaller file sizes than could be achieved using general purpose compression techniques.

1.4.7 Vector graphic representation

Vector graphic representation does not represent the image pixel by pixel. Instead, it stores a description of the objects that make up the image. For example, if an image contains a coloured circle and a pattern filled square then these object descriptions could take the form shown in Table 1.5

Table 1.5: Vector Graphic Representation

Square	Circle
lower left x	centre x
lower left y	centre y
upper right x	radius x
upper right y	fill colour
fill pattern	

A vector representation is **resolution independent**. Rasterisation is needed to convert the object descriptions to a bit map for subsequent display or printing.

Observing the differences between bit mapped and vector graphics

On the Web is an interaction that shows you the differences in image quality of bit mapped and vector graphics when viewed at different resolutions and as various zoom factors are applied. You should now look at this interaction.

Some printers are able to rasterise vector images of a certain format, such as EPS (encapsulated postscript format) while others need assistance from the comuter's CPU to carry this out. This can introduce additional processing requirements and many systems provide graphic co-processors to carry out these transformations, offloading the task from the CPU.

Advantages of vector graphic representation

Vector graphics representations have several advantages over bit mapped representations. These are:

- they do not lose their image quality on scaling. Scaling transformations are first applied to the object attributes before the image is rasterised to form a bit map;
- as values are not held for every pixel, this representation requires less storage space than a bit mapped image;
- they can be edited at the "object" level, thus allowing the user to reposition, scale and delete entire objects, or groups of objects, with ease;
- objects can be grouped to form larger objects that can then be manipulated as a single image;
- images are resolution independent meaning that the same quality of image will be rendered regardless of the resolution of the display or print device.

Disadvantages of vector graphic representation

See advantages of bit mapped graphics.

Identifying the advantages of vector representation

On the Web is an activity that asks you to identify the advantages of vector graphic representation. You should now carry out this activity.

5 min

Comparing the storage requirements of a bit mapped image and a vector graphics image

Working with a partner you should first create a full page image containing, say a filled circle with an overlapping filled square using a bit mapped package. Then create an identical full page image using a vector drawing package. Save both files.

10 min

- Use your operating system to list the file attributes and display the size of each file. Write down the difference in storage requirements between the bit mapped image and the vector graphics image.
- Now open up the original image documents. Try to move the overlapping square . Note down your observations of the following:
  - How easy it was to move the filled square.
  - The effect of moving the filled square.
- Finally, try to edit a small area of pixels in the filled square to change the colour. Note down your observations of the following:
  - How easy it was to edit at the pixel level.
  - The effect of editing at the pixel level.

Prepare to discuss your findings with your teacher.

1.4.8 Review questions

Q66: Which one of the following statements is true?

a) A bit mapped graphic takes up less storage space than a vector graphic
b) A vector graphic is a mathematical description of the objects that make up the image
c) A bit map graphic is resolution independent

Q67: Calculate the size in bytes of a bit map image of dimension 4 X 3 inch square using 72 dots per inch resolution and a pixel depth of 1-bit.

Q68: How many colours can be represented using a pixel depth of 16-bits?

Q69: An animation contains 1 minute of film. It is displayed in a frame size of 5 X 4 inches at a rate of 25 frames per second using a resolution of 300 dpi and colour depth of 24 bits. Assuming no compression, calculate the total storage requirements for this animation in Megabytes. Your answer should be given to 2 decimal places.

1.5 Summary

The following summary points are related to the learning objectives in the topic introduction:

- Data representation as positive binary numbers (using up to 32 bits) and negative binary numbers (using Two's Complement with up to 8 bits);
- Familiarisation and conversion between binary and decimal numbers;
- Representation of numbers using floating point technique and range depending on size of mantissa and exponent;
- Conversion of binary numbers to and from bit, byte, kilobyte, Megabyte, Gigabyte, Terabyte;
- Unicode and limitations of ASCII when representing character sets;
- Bit map methods of graphic representation including greyscale and colour;
- Relationship of bit depth and colour (up to 24 bits);
- Vector graphic methods of graphic representation (using objects);
- Relative advantages and disadvantages of bit mapped and vector graphics;
- Relationship between the bit depth and file size;
- The need for data compression with outline of methods.

1.6 End of Topic Test

An online assessment is provided to help you review this topic.

Topic 2

Computer Structure

Contents

Prerequisite knowledge

Before studying this topic you should be able to:

- *Describe the purpose of a processor and list its parts;*
- *Represent the data flow between the component devices of a computer system;*
- *Distinguish between main memory and backing storage;*
- *Describe the features and uses of Random Access Memory (RAM) and Read Only Memory (ROM).*

Learning Objectives

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

- *Describe the purpose and function of the ALU and the Control Unit within a processor;*
- *Describe the purpose and function of registers;*
- *Describe the purpose and function of the data bus and address bus;*
- *Identify control lines within a computer including reset and interrupt;*
- *Describe the purpose of read, write and timing functions of control lines;*
- *Outline the steps of the fetch-execute cycle;*
- *Describe and make distinctions between registers, cache memory, main memory and backing store in terms of function and speed of access.*

Revision

The following exercise tests the prerequisites for this topic. Ensure that you are happy with your responses before progressing.

Q1: What type of memory stores programs that must not be lost when the power to the system is removed?

Q2: A processor will frequently transfer data. Where is the data transferred to and from

a) Output devices
b) Input devices
c) Clock
d) Memory

2.1 Introduction

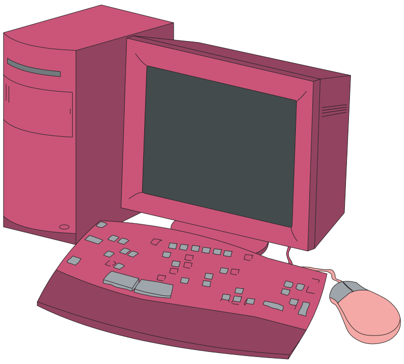
This unit on Computer Structure describes in detail the function of the component parts of a processor in the manipulation of data.

This is extended to the methods of transferring data within a processor and between a processor and memory.

The concept of a stored program is considered and the steps in the fetch-execute cycle to access and run programs. Memory types are considered, from registers to backing storage and how memory is defined and addressed.

2.2 Computer Organisation

Computers play a significant role in meeting our everyday requirements. You can now browse the Internet for a new home, order from a supermarket on-line and have the goods delivered to your door. You can import a new car from abroad at the touch of a button, order clothes from a catalogue company and communicate with friends overseas.



The ways in which you learn are also changing. You can complete a school or college assignment using a general purpose package, or solve programming problems at home on your own PC, emailing your results to your instructor for feedback. You can use computer based learning tools to assist you in understanding new concepts. You are using one such tool right now!

If computers have made such a significant impact, then it makes sense to find out a little more about how they work and how you can make use of them. For instance, how are they structured? How do they operate internally? How is data represented? Why are some computers faster and more powerful than others? What devices can you attach to them and how can you get them to communicate?

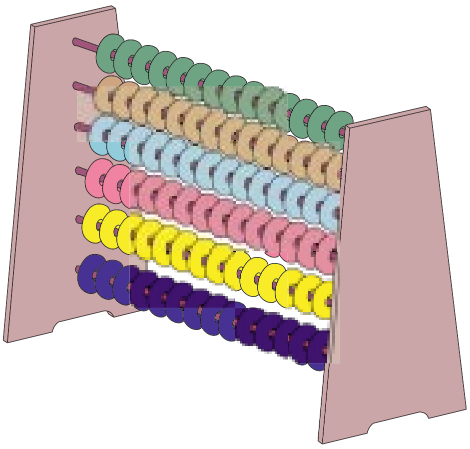
These are some of the questions we will be looking at in this topic. There are others. First let us take a closer look at the basic structure of a small computer system and at how it operates internally.

2.2.1 Calculating machines - from Babbage to integrated circuits

Learning Objective

At the end of this topic you will know:

- the contribution of Charles Babbage to modern computer design;
- major technological advances leading to the development of the Personal Computer.



Nowadays you use calculators to perform numerical calculations. This was not always the case, and for some societies, is still not. Since early times humans have tried to find ways to make calculations easier. For instance the abacus was developed by the Chinese around 1300 AD, although similar devices had been used by the Babylonians, since about 500 BC.

Logarithms were developed as a method of calculation by John Napier, who also invented a device known as Napier's Bones. The slide rule is a further example of an analogue device for multiplication that is based upon logarithms.

Simple, manually operated mechanical calculators were developed by many famous mathematicians, including Pascal and Leibniz.

The first steps towards automating the mechanical calculator were taken by Charles Babbage (1792-1871) with the design of the special purpose Difference Engine and the subsequent design of the more general purpose Analytical Engine.

The Analytical Engine was controlled by a set of instructions entered as punch holes on a set of metal cards. This idea was first introduced by Jacquard (1752-1843) who designed weaving looms. The pattern woven depended upon the position of holes punched in metal cards. Unfortunately the technology of the time was not advanced enough to allow Babbage to construct a working machine. However his ideas laid the foundations of modern computer design. These ideas included:

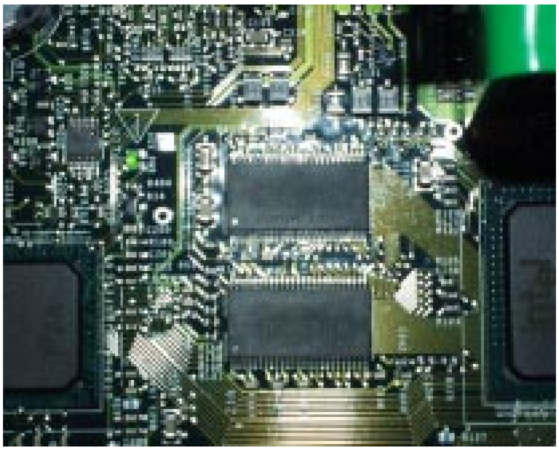
- a **memory** that could store 1000 numbers;

- a machine controlled by a **program** entered into it;
- entering a different program to perform a new task - **general purpose**;
- the program was a set of instructions that, if followed, would accomplish a task - an **algorithm**.

These ideas were not developed further until electro-mechanical relay technology produced computers in the 1930s. By today's standards these machines were large and slow.

Vacuum tube technology in the 1940s increased the speed of the computers and by the 1960s transistor technology reduced the size and power requirements.

From about 1965 to the present, the circuits for many operations have been incorporated into a single chip as **Integrated Circuits** (ICs). Chip fabrication techniques also improved, allowing further integration to the extent that we now have **Very Large Scale Integration** (VLSI) and a complete processor on a chip. This made possible the Personal Computer.



Review Questions

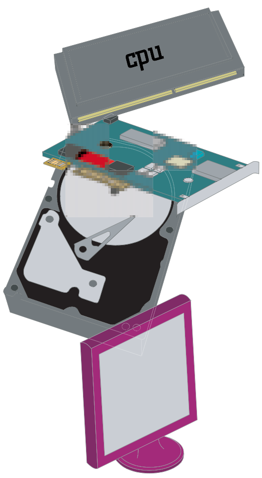
Q3: Describe the Analytical Engine designed by Charles Babbage.

Q4: How does the Analytical Engine relate to modern computers?

Q5: What technical development in the 1940s and 1950s helped reduce the size and power requirements of computers?

Q6: What affect did the development of the integrated circuit have on computers?

2.2.2 Computer organisation



Computers are digital machines that execute machine code programs and operate on data in binary form. By binary form we mean a representation of information as 0s and 1s.

We will now concentrate on how binary programs, also called *machine code programs*, are stored and executed. This requires some knowledge of the internal organisation of the computer.

2.2.2.1 The Organisation of a Simple Computer

Until now we have been looking at how we can use simple logic gates to produce devices such as adders, decoders and flip-flops. Although we have looked at simplified versions of these logic devices, it is devices such as this that are combined together to create a computer. We will now step back a level and look at the basic architecture of a simple computer.

A simple computer consists of the following components (see Figure 2.1):

- Processor;
- Memory;
- Input/output device;
- Communication channels (shown between the aforementioned components in Figure 2.1);

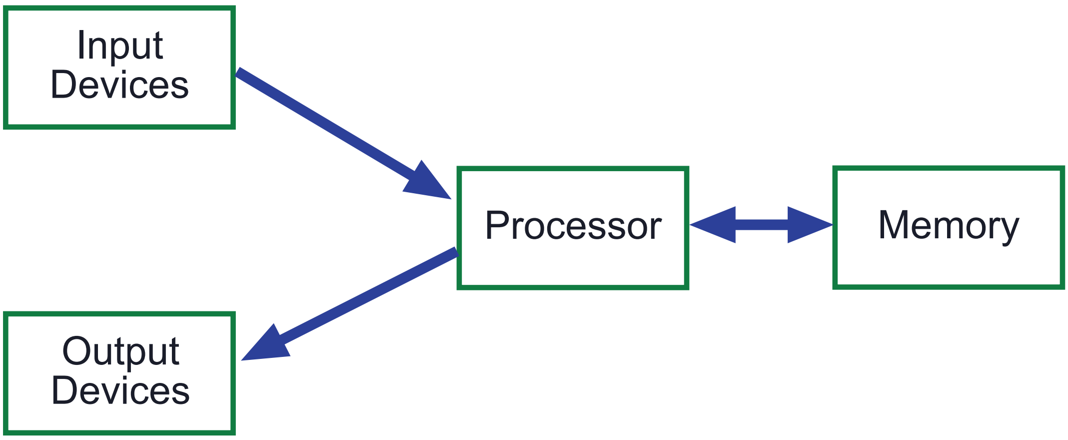


Figure 2.1: A Simple Computer

Input devices include the keyboard and mouse and can be used to supply input to the processor. Output devices include the screen and printers and these can be used to supply output from the processor. Input and output devices are often known as peripheral devices.

Some computers have more than one processor; however we will concentrate on single processor machines in this topic. Where there is only one processor it is known as the **Central Processing Unit**, or **CPU**. This is where instructions are processed and computations are carried out. This is the control centre of the computer.

The communication channels allow data and control signals to be communicated between the main components of the computer via the **system bus** or **external bus**. A bus is a collection of parallel wires each of which can carry a digital signal. Thus a 16-bit wide bus could transmit 16 bits simultaneously. The CPU has its own **internal bus** allowing the communication of data and control signals between its component parts.

It is worth noting here that the system bus contains lines to transmit data, lines to transmit memory addresses and various control lines. Frequently it is thought of as if it were separate buses: a data bus, an address bus and control lines. Sometimes the data and address lines are not separate at all and the same lines are used for different purposes at different times. For example, one moment sending an address to memory and the next transmitting data from the addressed memory location to the CPU.

A more detailed diagram of the main components of a simple computer are shown in Figure 2.2.

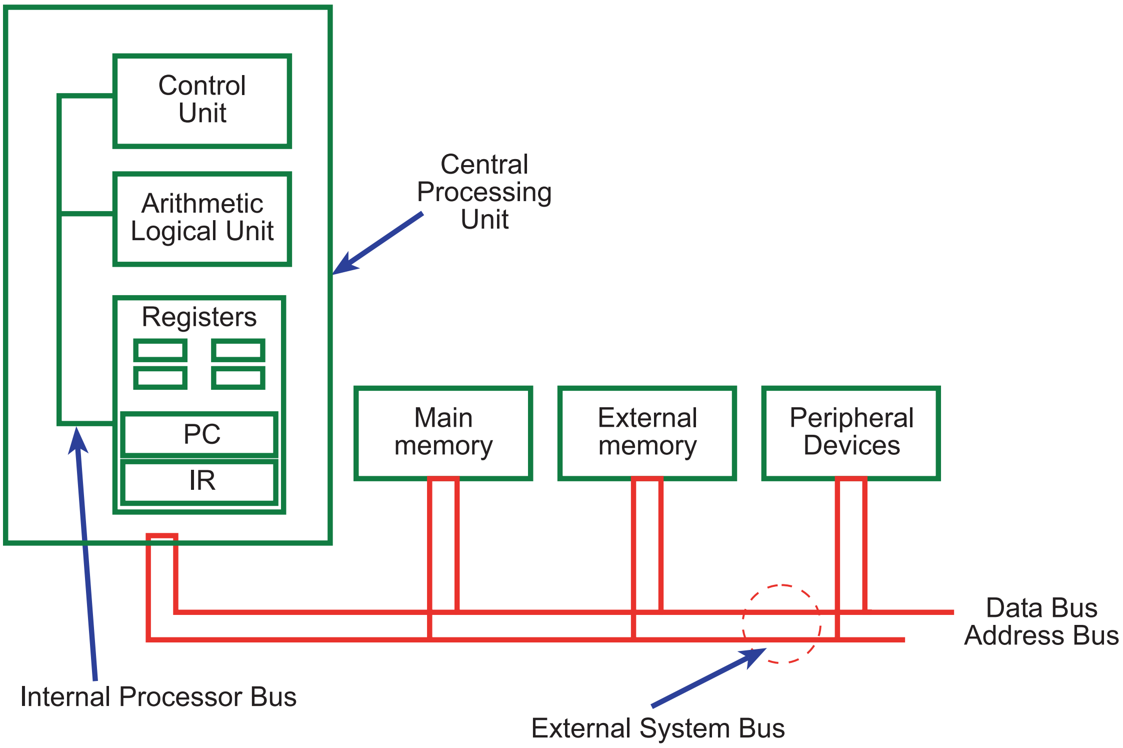


Figure 2.2: The Organisation of a Simple Computer

The computer illustrated in Figure 2.2 is a typical example of a **Von Neumann architecture**. Virtually all computers follow this architecture model that has its origins in the **stored program concept** proposed by John Von Neumann in 1945. The basic idea behind the Stored Program Concept is that the sequence of instructions (or program) to solve a problem should be stored in the same memory as the data. This ensured that the computer became a general-purpose, problem-solving tool, since to make it solve a different problem required only that a different program be placed in memory.

The component parts of the computer are:

- **Central Processing Unit (CPU).** Carries out computation and has overall control of the computer.
- **Main memory.** Stores programs and data while the computer is running. Has fast access, is directly accessible by the CPU, is limited in size and non-permanent.
- **External memory.** Holds substantial quantities of information too large for storage in main memory. Slower access than main memory, not accessible directly by the CPU but can be used to keep a permanent copy of programs and data.
- **Peripheral devices (input/output devices).** These allow the computer to communicate with the outside world.

- **External system bus.** This allows communication of information between the component parts of the computer.

Some possible transfers of information via the system bus are:

- data transmitted from main memory to the the CPU
- input data from an external device (e.g. the keyboard) travelling from the device to main memory
- information from external memory transmitted to main memory

The speed of the system bus is very important since, if it is too slow, the speed of the CPU is restricted by having to wait for data.

The CPU typically consists of

- A **Control Unit (CU)** which exerts overall control over the operation of the CPU;
- An **Arithmetic and Logic Unit (ALU)** which carries out computation;
- A set of registers which can hold intermediate results during a computation.

Two of these registers are of particular importance, namely,

- The **Program Counter (PC)** which holds the address in memory of the next instruction in the program;
- The **Instruction Register (IR)** which holds the instruction currently being executed.

These components are linked by an internal bus.

In practice, the architecture of a modern digital computer will be more complex than the description given here, with each component itself being an assembly of parts connected by various different buses. However, for the moment, this will suffice as a model for how the major parts of a digital computer are organised.

2.2.2.1.1 Review Questions

Q7: What are the main components of a computer?

Q8: What is the purpose of the system bus? What type of information is it likely to transmit?

Q9: With what important concept was John Von Neumann associated? What large advantage did this concept confer upon computers?

Q10: What is held in the main memory of the computer? Why is external memory also required?

Q11: How would an item of data that was entered at the keyboard finally find its way into a CPU register for processing?

Q12: All CPUs contain two particular registers. What are these registers and for what are they used?

2.2.3 The stored program concept

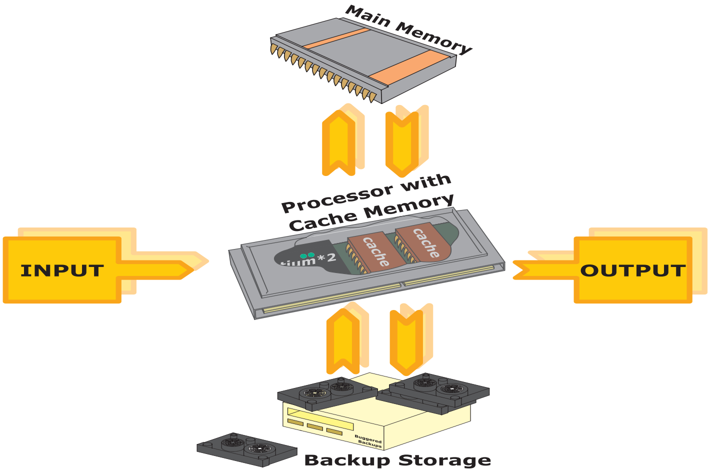
All computers are based upon the same basic design, known as the **Von Neumann Architecture**.

Computers carry out tasks by executing machine instructions. A series of these instructions is called a **machine code program**.

A machine code program is held in main memory as a **stored program**, a concept first proposed by John Von Neumann in 1945.

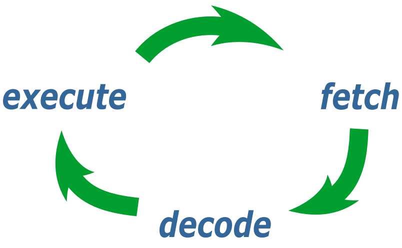
A unit, known as the Central Processing Unit (CPU) fetches, decodes and executes the machine instructions.

By altering the stored program it is possible to have the computer carry out a different task. As a user of a desktop computer you will already know this. You may have loaded a word processing program to enter and edit text. Using the same computer you may have opened a spreadsheet or drawing program to enter numerical values or create graphic images.



Being able to load and execute different programs allows the computer to become a **general purpose** problem solving machine.

2.2.4 Fetch-execute cycle



To execute a machine code program it must first be loaded, together with any data that it needs, into main memory (RAM). Once loaded, it is accessible to the CPU which fetches one instruction at a time, decodes and executes it at electronic speed.

Fetch, decode and execute are repeated until a program instruction to HALT is encountered. This is known as the **fetch-execute cycle**.

2.2.4.1 Fetch execute cycle in greater detail

Earlier we introduced the fetch-execute cycle and described the stored program concept where machine code instructions are repeatedly transferred from main memory to the CPU for execution.

We would now like to show you how the address bus, data bus, control bus and internal registers take part in reading a program instruction from main memory - essentially the *fetch phase* of the fetch-execute cycle. Figure 2.3 below illustrates in more detail the fetch-execute cycle.

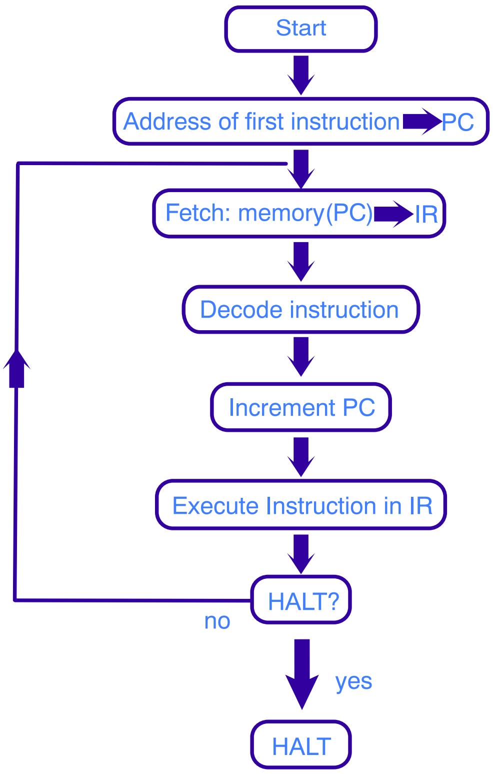
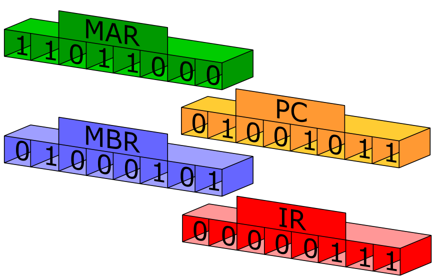


Figure 2.3: Fetch-execute cycle

Simulation of an instruction *fetch*

On the web is a simulation which shows you how the buses and the internal registers of the CPU take part in reading an instruction from main memory. You should now look at this simulation.

2.2.4.2 Registers used by the processor in the fetch-execute cycle



To accomplish the tasks, a processor has a collection of dedicated registers. These are used to hold information specific to this task. Note that these registers are in addition to the general purpose registers provided by the processor. Unlike the general purpose registers, these registers are not usually visible to the assembly level programmer.

Memory address register

The first register that we will discuss is the memory address register (MAR). This is used to hold a value representing the address in memory that the processor needs to access.

Usually the MAR will hold a bit pattern corresponding to the state (0 or 1) of the address bus. When the processor needs to access memory, the value of the location in memory is placed in the MAR and the processor circuitry will ensure that the address bus lines are set to the correct values.

Memory data register

The memory data register (MDR) is used to hold bit patterns that represent data values. For example, when reading from memory, the MAR will be used to set up the address lines to select a location. After a short delay, the memory device will set the lines on the data bus to appropriate values. When the values on the data bus have settled, the circuitry of the processor will set the value of the MDR to the value that appeared on the data bus.

Instruction register

The instruction register is a dedicated storage space used by the control unit when it is decoding instructions.

General purpose registers involved in the fetch-execute cycle

The program counter (PC) is the general purpose register most involved in the fetch-execute cycle. Remember that this register is used to keep track of where in the program execution has reached.

Other general purpose registers are only usually affected as part of the execution of the program and as such, are not fundamental to the operation of the cycle.

2.2.4.3 The *fetch* phase

The first of the two main phases of the fetch-execute cycle is the *fetch* phase, and consists of the following steps:

1. The contents of the PC are copied into the MAR;
2. The contents of memory at the location designated by the MAR are copied into the MDR;
3. The PC is incremented;
4. The contents of the MDR are copied into the IR.

Remember that the PC is used to keep track of where execution has reached. Thus the first step is concerned with establishing the location of the next instruction to execute. The second step is to get the value into the MDR.

The third step is to ensure that the PC points to the next instruction to be executed: if we did not increment the PC at some point, we would continually execute the same instruction over and over again!.

The fourth step ensures that there is a copy of the instruction in the IR ready for execution to begin.

Sequencing the steps in an instruction fetch

On the web is an assessment that requires you to place the steps of an instruction fetch in the correct order. You should now carry out this assessment.

2.2.4.4 The *execute* phase

The execute phase consists of the following steps:

1. Decode the instruction in the IR;
2. Execute the instruction in the IR.

Once the execute phase has completed, the fetch phase will be carried out again.

Animation of the *fetch-execute* cycle

On the web is an animation of the fetch, decode and execution of the instruction LOAD[16]. You should now look at this animation.

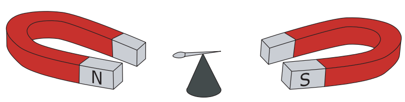
Pseudocode representation of the *fetch-execute* cycle

For convenience we can write this series of steps as a pseudocode representation:

```
loop forever
      PC -> MAR
      [MAR] -> MDR
      PC+1 -> PC
      MDR -> IR
      Decode IR
      Execute IR
end loop
```

Note that `->` means is copied to and that `[MAR]` means the contents of the location pointed to by `MAR`.

2.2.5 Two-state machine



The electronic components of a computer are designed to be in only one of two states. For example, a magnetic storage device records data magnetised in one direction or another, transistors conduct or do not conduct. The binary digits 0 and 1 are used to represent these two states and hence the computer is termed a **two-state machine**.

2.2.6 Review questions

Q13: Which of the following is true?

a) machine code is represented in binary

b) data is represented in decimal or hexadecimal
c) a stored program is executed from disk

Q14: What is meant by the term *stored program*?

Q15: The CPU:

a) was invented by John Von Neumann
b) holds a stored program
c) fetches, decodes and executes machine instructions

Q16: What is meant by the term *fetch-execute cycle*?

Q17: Which of the following is false?

a) magnetic storage devices are two-state devices
b) a two-state device can only be in one of two states
c) binary cannot be used to represent a two-state device

2.2.7 Structure of a small computer system

Learning Objective

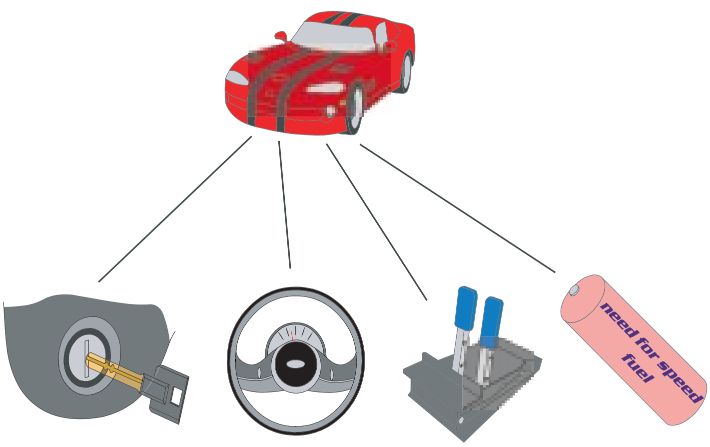
At the end of this sub-topic you will know:

- The main functions of the components of a small computer system;
- Organisation of Main Memory and types of main memory;.
- The purpose of the Address Bus, the Data Bus and the Control Bus;
    - Internal registers of the CPU and their purpose;
- The steps involved in a memory read operation.

We want to look at the internal organisation of the computer at a level of abstraction that describes the system as interacting, high-level *components*. For now, we are not that interested in the lower level detail of shunting bits around the machine.

To get a feel for this component level, imagine you are a car driver. When you step into the car, what is visible to you are interfaces to the components that allow you to drive it.

The ignition component allows you to start the car and to switch it off. You do not need to know how the ignition system works, this is hidden from sight. Accelerator pedals, brake pedals and a gear mechanism each interface to components which alter the speed of the car. To manouvre the car you operate a steering wheel which interfaces to a steering component.



From this description, you could produce a diagram of these interacting components. It would not be very detailed and certainly would not have enough information for a car mechanic to work with, but it would, at a higher level, describe how the driver interacts with the car. This is the component level we will introduce.

5 min

Identifying system components used in a task

You should find a partner to work with. Imagine that you have a PC or a Mac running a windows interface and that you have powered it up. Now think about the sequence of events that occur when you create a new word processing document. Try to identify the components of your PC that are involved in this task. For instance, there will need to be a transfer of the word processing application (program) from the hard disk component to the main memory component in order to run the program. What happens next and what components do you think are involved?

2.2.8 Computer components and their function

The components of the CPU and the connections to devices that are external to it are shown in Figure 2.4

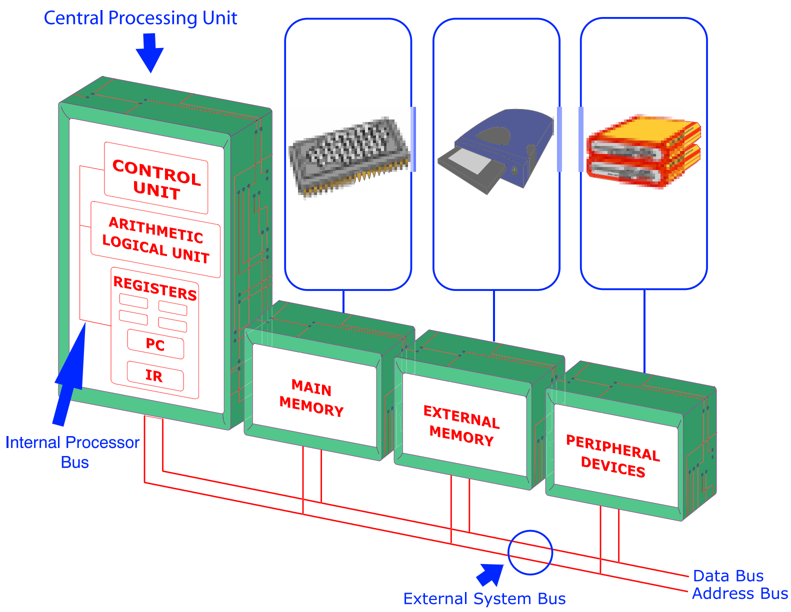
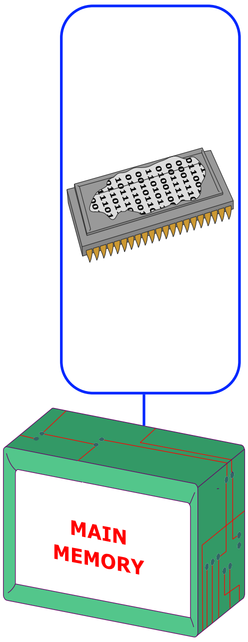


Figure 2.4: Components of a Small Computer System

2.3 Memory

Main memory (RAM and ROM) stores programs and data while the computer is operating.



Memory used to be soldered onto the system board of the processor (motherboard). The need to provide more readily upgradable computers led to the development of Single In-Line Memory Modules (SIMMs). These plug into a SIMM socket on the motherboard. Each SIMM contains a number of DRAM chips and varies depending on the type of computer and the amount of RAM required.

Later, we will take a closer look at the characteristics of RAM and ROM chips and the differences between DRAM and SRAM.

Main memory consists of a large sequence of bytes (typically 64 Mbytes in a PC) each of which may be directly accessed using its memory address. In a byte-wide memory, the first byte in memory has address 0 and subsequent bytes have addresses 1,2,3, etc. as shown in a simplified representation of RAM in Figure 2.5

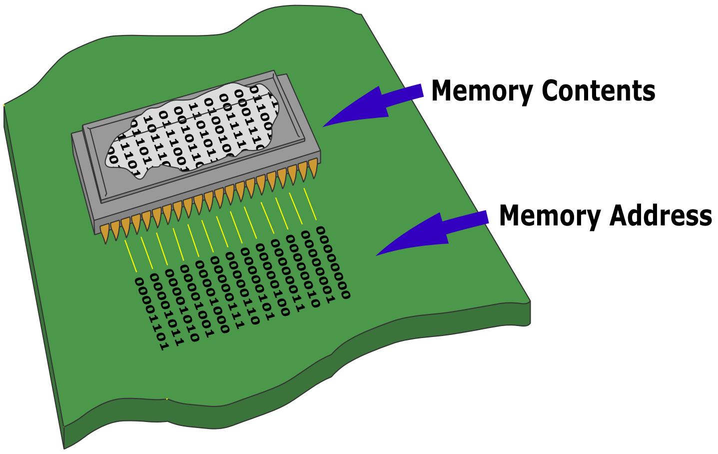


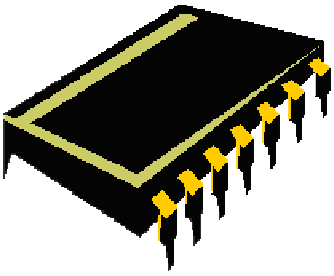
Figure 2.5: A simplified representation of RAM

Any location in memory can be read from or written to by referring to its address.

Memory can be organised as:

- 8- bit wide (PC-8088)
- 16-bit wide (XT-8086, AT-80286)
- 32-bit wide (386DX, 486SX, 486DX)
- 64-bit wide (Pentium)

2.3.1 Random access memory



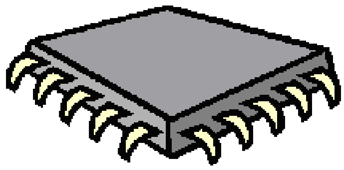
Random Access Memory (RAM) is a **volatile** memory. This means that the contents of RAM are lost when power is no longer supplied to the chip. RAM can be written to and read from. There are two types of RAM, namely *static* and *dynamic* (**SRAM** and **DRAM**)

SRAM chips are very fast but are not suited for very large amounts of memory. They are more suited to cache memory, where only small amounts are required. You will learn more about cache memory when we look at factors that affect system performance.

DRAM chips are more widely used. They are much cheaper to produce, can hold larger amounts of data in a smaller physical area and require less power. They are *dynamic*, requiring a continuous signal to refresh the contents of the chip.

2.3.2 Read only memory

Read Only Memory (ROM) is a **non-volatile** store which means that the contents are held permanently. The software and data stored on the ROM are fixed at the time of manufacture. Once programs and data have been entered into the ROM they cannot be subsequently altered.



ROMs are used to store programs and data that do not change during the operation of the system. These are known as *mask programmed* ROM.

Where *different* software and/or data is needed on a ROM chip, manufacturers produced a chip that allows existing data to be erased and new data written. These are known as *electrically programmable* read-only memory chips (**EPROM**s). Data is erased by shining ultraviolet light onto the chip.

EPROMs have the disadvantage that all the chip contents are removed during erasure. The entire chip has to be reprogrammed, even if only a single memory word needs to be changed.

Another type of ROM technology where the contents of the chip can be altered is the *electrically erasable programmable* read-only memory (**EEPROM**). By applying suitable electrical pulses, this chip can be *selectively* reprogrammed which means that the entire contents need not be erased.

2.3.3 Review questions

Q18: Which of the following is true?

a) RAM is volatile
b) ROM is volatile
c) Neither RAM nor ROM is volatile

Q19: Explain what is meant by the term *non-volatile*?

Q20: Which of the following memory chips can be selectively reprogrammed?

a) PROM
b) EEPROM
c) EPROM

Q21: Explain how EPROM chips can be reprogrammed and give **one** disadvantage of EPROM.

Q22: Which of the following statements is false?

a) RAM cannot be written to
b) ROM can only be read from and not written to
c) EEPROM is erased using electrical pulses

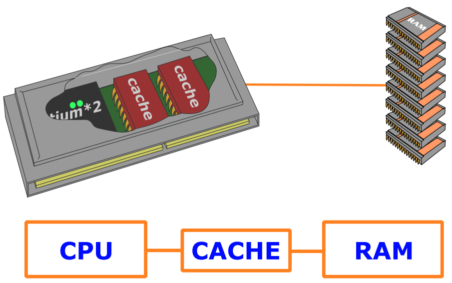
ROM technologies

On the web is an activity that asks you to match ROM technologies to their descriptions. You should now carry out this activity.

2.3.4 Cache memory

Program instructions are usually read sequentially. From one instruction it would be reasonable to assume that the next instruction required will be in the next memory location. This assumption is used to increase processor efficiency.

Although the movement of data within the processor is getting faster and faster, the system buses are not keeping up. This leads to wasted time while the processor waits for data to be fetched from memory.



To reduce this problem most machines now have a second, smaller, area of memory known as **cache 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.

When data or an instruction is read from memory, the memory locations following are copied into the cache memory. At the next read instruction the cache memory is read first. If the data is in cache the access time will be much lower than going to main memory. If the data is not in cache then main memory will be accessed, and although there is a slight loss of time from reading twice, the overall time saving in this method is quite significant.

Observing cache memory

On the web is a simplified simulation of the operations of cache memory. You should now look at this simulation.

The contents of cache are simultaneously held in RAM. When data is to be written back to memory it must be written to cache so that the cache is kept current. At some stage it will also have to be written back to RAM. Cache that has not been updated doesn't have to be copied back to memory it is just removed from cache when it is to be replaced by something the processor has a greater need for. There are 2 different ways to update cache memory.

Write through cache. When cache is updated memory is updated at the same time.

Write back cache. Cache is updated, but RAM is not updated until the content of cache is being cleared. Write back requires fewer write operations but there is an overhead in managing the selected updates. Write back cache is generally about 10% faster than write through cache.

2.3.5 Memory maps

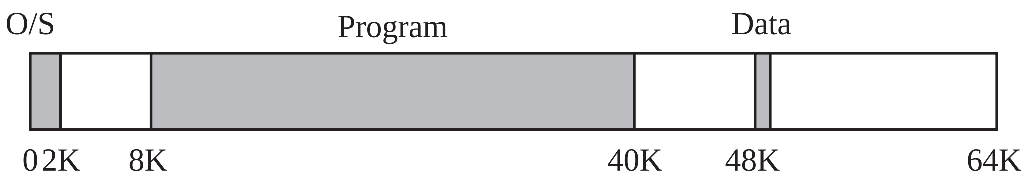
Main memory is made up of a matrix of cells. However when studying the content of memory it is helpful to use a logical view of a set of cells, arranged in numerical order in which the lowest memory position is location 0 and the highest is location M.

Different processors have different ways of organising information in RAM, the BIOS and part of the operating system will always be in the same protected areas. Certain areas are allocated for user applications.

Memory addresses are often written in hexadecimal as they would otherwise be awkwardly long binary strings.

Example : Memory addresses

A processor has a 16 line address bus. If a particular memory holds an operating system from position 0_{16} to 800_{16}, and a 32 Kb program starts at position 4000_{16}, is there enough free memory space for a 1 Kb block of data starting at position $C000_{16}$?



Step 1

There are 16 address lines. The maximum number of address locations = 2^{16} = 65536 = 64 K.

Mark 64K on the memory map.

Step 2

O/S from 0_{16} to 800_{16}.

2^{16}	2^{15}	2^{14}	2^{13}	2^{12}	2^{11}	2^{10}	2^9	2^8	2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
					1	0	0	0	0	0	0	0	0	0	0	0
			1	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

$400_{16} = 1$ Kb

$800_{16} = 2$ Kb

Operating system from 0 to 2 Kb

Mark this block on the memory map.

Step 3

Program from 4000_{16} for 32 Kb

$4000_{16} = 8$ Kb

so the program runs from 8 Kb to $(8 + 32) = 40$ Kb.

Mark this block on the memory map.

Step 4

The starting position for the data block is $C000_{16}$.

This is at 48 Kb and needs 1 Kb of space. Mark this on the memory map.

The data will fit.

Describing memory maps

Q23:

A processor has 16 lines. The memory has 4 KB s of BIOS data starting at position 0. There are device drivers positioned from 2000_{16} for 8 Kb A program runs from 32 K to the top of memory.

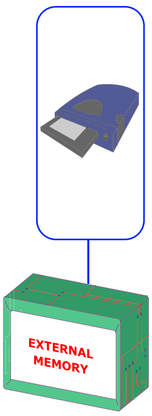
Identify where, in memory, there is free space and how much there is.

Q24:

A processor has 32 address lines. There is BIOS data from 0 to 8 Kb The DOS kernel lies between 3000_{16} and $6000\ _{16}$ There is a program from $F0000_{16}$ using 24 Kb There is a data block from F00 0000_{16} using 32 K.

Draw the memory map clearly identifying the start and end of each used memory area.

2.3.6 External memory



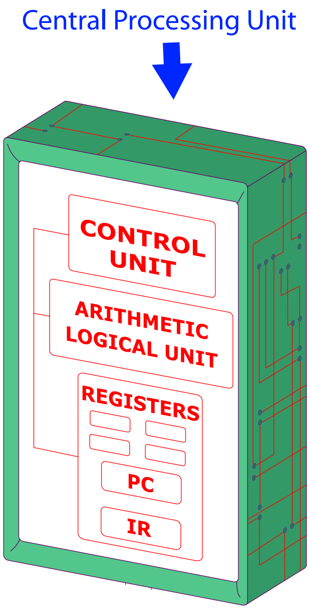
External memory, such as the hard disk, holds quantities of data too large to store in main memory. It is also used to keep a permanent copy of programs and data.

Examples of external memory devices are:

- hard disk;
- floppy disk;
- zip disk;
- CD-R;
- magnetic tape;
- flash drive.

2.4 Central processing unit

Central Processing Unit. The CPU coordinates and controls the activities of all other units in the computer system. It executes program instructions and manipulates data in accordance with the instructions.



The CPU is the most important component of a computer and it is essential to have a good knowledge of its internal organisation, i.e. its **architecture**. Just as the architecture of a building refers to a *structure* of rooms, facilities and access which links all parts of the building, processor architecture refers to its internal organisation of subsystems and how they interact.

CPUs are fairly complex and at this level of study we will concentrate on a simplified functional description of its structure using a **standard** architecture composed of the following three components:

- Arithmetic and logic unit (ALU);
- Control unit;
- Registers.

All three components work together to form the processor.

2.4.1 The architecture of the microprocessor

We will now study the internal architecture of the **microprocessor (CPU)** itself. Because of the **stored program concept**, any consideration of this architecture must consider the relationship between the CPU and memory. Figure 2.6 is a schematic diagram of a fairly typical microprocessor design, showing the internal structure of the CPU and its relationship to the memory of the computer.

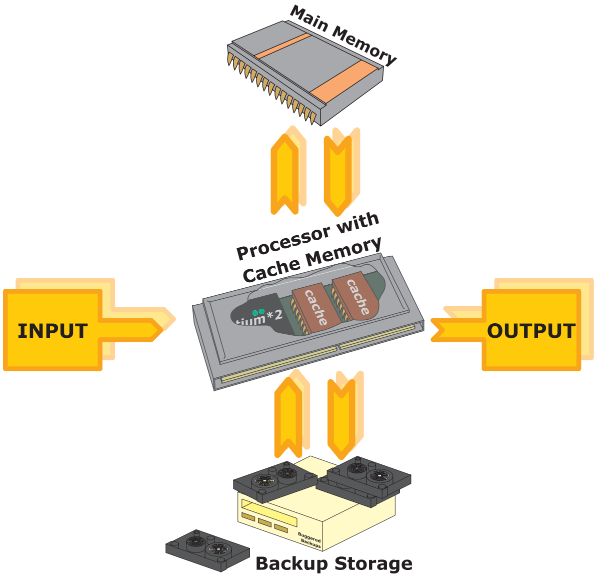


Figure 2.6: Typical Microprocessor Design

We will now look at the role that these components play in the operation of the processor.

2.4.2 Accessing memory

The **CPU** has to access memory both for instructions and to receive and transmit data from or to memory. For this purpose it typically has two internal registers, namely:

- **Memory Address Register (MAR)** - specifies the address in memory for the next read or write operation from or to memory;
- The **Memory Data Register (MDR)** or **Memory Buffer Register (MBR)** - contains the data to be written to memory or receives the data read from memory.

The MAR register is connected to the address portion of the system bus and the MDR register is connected to the data portion of the system bus.

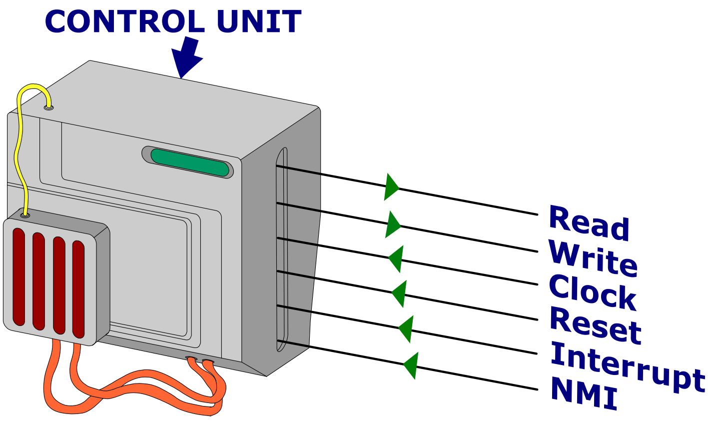
- **To read data from memory**, the CPU places the address of the required memory location into the MAR and activates the memory-read control line of the system bus. This will cause the required data to be transmitted from memory via the data bus to the MDR;
- **To write from the CPU to memory**, the CPU places the data to be written in the MDR; the address of the memory location where they are to be written is placed in the MAR; and the memory-write control line is activated.

The MAR and MDR registers have a large part to play in the fetch-execute cycle.

When fetching an instruction from memory during the fetch-execute cycle, the address contained in the **PC** will be copied to the MAR using the processor's internal bus. When the memory-read control line is activated, the instruction will be sent to the MDR using the data bus. From the MDR it will be copied to the IR using the processor's internal bus.

When executing an *add to accumulator* instruction, the address part of the instruction will be sent to the MAR so that the operand can be obtained from memory. The operand is then placed in the MDR from where it can be sent to the **ALU**, via the **CPU** internal bus, for adding to the contents of the accumulator.

2.4.3 Control unit



The machine code programs stored in main memory tell the computer what steps must be carried out to solve a problem. They also tell it the sequence in which it must carry out the steps.

The **control unit** includes *timing/ control logic* and the *instruction decoder*. It sends signals to other parts of the computer to direct the fetch and execution of machine instructions.

Using timing and control signals, it tells other parts of the system what to do and when to do it, i.e. it *synchronises* the whole system.

Signals are sent out and received on the control bus. For example, if data is to be read from main memory, the control unit will initiate a *read* signal on the control bus. Some signals, such as interrupts, originate from external devices, acting as inputs to the CPU.

The control bus is not really a bus at all. Unlike the address and data buses where bits are simultaneously transmitted along a set of parallel wires, the control bus is made up of *discrete* wires, each having a specific function. These functions are described in the table below:

Control Line	Function
clock	generates a constant pulse at a frequency measured in Hertz (Hz). Each pulse causes a machine operation to be carried out. Machine operations include reading data from main memory or adding numbers together.
reset	causes the processor to halt the execution of the stored program. All internal registers are cleared and the machine reboots.
interrupt	tells the processor that an external event has occurred, such as the transfer of data from an external device. The processor may ignore this type of interrupt.
NMI	a non-maskable interrupt that cannot be ignored by the processor. For example, low power failure.

When data is to be read from a memory location then the control unit will initiate a **read** signal on the control bus and when data is to be written to a memory location then the control unit will initiate a **write** signal. These operations are described later in greater detail when we take a closer look how an instruction is fetched from main memory.

Identifying functions of the control bus

On the web is an activity that asks you to correctly identify 4 functions of the control bus. You should now carry out this activity.

2.4.3.1 Getting the processor's attention

A computer receives signals from a number of different sources. Characters keyed in on the keyboard, the click of a mouse, data from a scanner. The arrival of this type of signal is not necessarily expected at any particular time and the computer has to have a way of detecting them.

There are two ways that this can happen, known as **polling** and **interrupts**.

2.4.3.2 Polling

The first is known as **polling**. This is what certain types of door to door salespeople do. They ensure that all their customers have a copy of the company catalogue and then they visit every house on a rota basis to ask if the customer would like to place an order. This means that the salesperson will visit every house, say, every month.

They may know that the average time between orders is about 3 months but they can't take the risk of leaving the customer unattended when they might want to place a big order. This means that the salesperson may be wasting a lot of time making unnecessary calls. From the customer's point of view this is not ideal either, because they might realise a week after the last visit that they have forgotten something important and there is no way to shorten the waiting time until next month.

For a computer system this would work quite satisfactorily if the processor was running a microwave oven because the processor would be dedicated to that one task and efficiency would be a meaningless concept.

The life of the door to door salesperson would be simpler and they would be able to handle far more customers, if the customer was given a phone number or e-mail address and asked to initiate contact when they wanted to order. This would provide the customer with a much better service and also allow the salesperson more time to do other things.

2.4.3.3 Interrupts

In computer terms, the signal from a peripheral device or program that the attention of the processor is needed is known as an **interrupt**. Every time a keyboard key is pressed an interrupt is generated. When the machine is designed, the handling of interrupts is planned for.

In an IBM type PC there is an allowance for 256 different types of interrupt. When one of these occurs the system is able to identify its type. With this information the processor then looks at an area in memory in which an address for each of the 256 interrupts is stored. At this address there is a program known as an Interrupt Service Routine. The address table is used to furnish addresses indirectly because this makes it possible for a programmer to control. Many of the ISRs are stored in ROM.

When an interrupt is received, the processor will:

- store the contents of its internal registers in an area of memory called the stack;
- find the address for the ISR;
- jump to the service routine and process it;
- reload the internal registers from the stack;
- continue processing from where it stopped.

There are different priority levels assigned to interrupts. If an interrupt arrives while the processor is already dealing with one, it can do one of two things. If the second interrupt is of a lower priority the processor will carry on until the current interrupt has been serviced then it will service the second. If the new interrupt is of a high priority the processor will store it's current state on the stack and start to process the newer interrupt. When that one is finished it will complete the processing of the first interrupt and then revert to the original process. This is known as nesting interrupts.

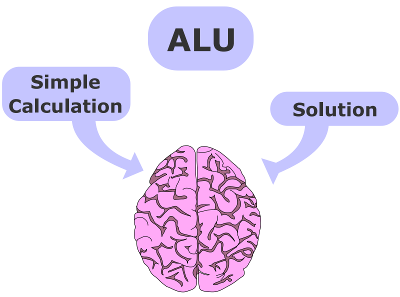
Interrupts can be generated by hardware or software.

Any signal coming from a peripheral device will prompt an interrupt. Each I/O connection has a physical link called an IRQ line. This serves to carry the interrupt signal and in turn to identify the source of the interrupt. Internal devices connected via the motherboard also use IRQ lines. There is a limited number of IRQ lines and this can be a limiting factor in adding hardware to a computer system. However the newer systems using USB (Universal Serial Bus) or Firewire can accept a number of hardware devices sharing IRQ lines.

A software interrupt is one generated from within a program. This includes routine activities like sending a character to the keyboard or the screen. A software fault like trying to divide by zero, or trying to write into a protected area of memory, will also call an interrupt.

Generally the processor can mask or delay the servicing of interrupts until it is ready. There is, however, a group of interrupts that cannot be ignored. These are known *non-maskable interrupts* (NMIs). Typically this interrupt would indicate a problem such as loss of power, requiring the computer to shut down immediately.

2.4.4 Arithmetic logic unit



The **arithmetic logic unit (ALU)** is where data is processed and manipulated and can be considered the "brain" of the computer.

Processing can involve either arithmetic operations and the ALU must contain circuitry to perform additions. Note that multiplication can be achieved through a series of additions, while division can be achieved through a series of subtractions.

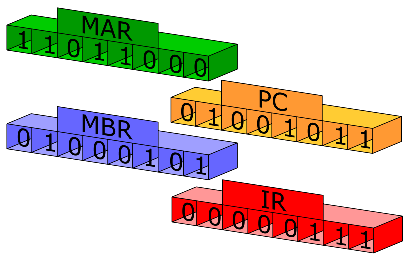
The ALU may also perform logical operations such as a logical OR operation. This requires in-built logic elements.

The ALU uses arithmetic registers which are storage locations used to hold data and temporary results of calculations. One special storage register used by the ALU is the **accumulator** which it uses to hold the results of additions. Typical operations performed by the ALU include:

- addition;
- subtraction;
- shift left;
- shift right;
- logical OR;
- logical AND;
- increment;
- decrement.

2.4.5 Registers

A register is a storage location used to hold instructions, memory addresses, data or the temporary results of calculations. Because registers are internal to the processor, they can be accessed at high speed.



The CPU has a set of *general* and *special purpose* registers and the number and type of registers in any one CPU will be different from those in another CPU. Special purpose registers typically found within a processor are listed below:

memory address register (MAR) is used to hold the address of a location in main memory.

memory buffer register (MBR) is used to hold data that has just been read from main memory or is to be written to main memory.

instruction register (IR) is used to hold the current instruction that is being executed.

program counter (PC) holds the address of the next instruction to be fetched from memory.

The processor also has a set of **general purpose registers**. They are called general purpose because their role is not defined at manufacture and can be used by programmers as appropriate.

2.4.6 Review questions

Q25: What is the function of the CPU?

Q26: Which of the following is NOT a component of the CPU?

a) ALU
b) RAM
c) Special purpose registers

Q27: How does the control unit synchronise operations within the computer?

Q28: Which of the following describes how the ALU performs multiplication?

a) using a logical OR operation
b) successive addition
c) using an increment operation

Q29: Why are *general purpose* registers provided within the CPU?

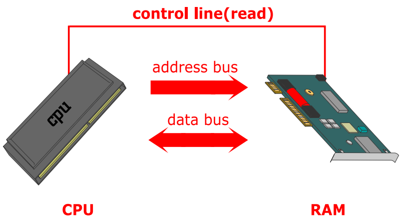
Matching CPU registers to their purpose

On the web is an activity that asks you to match some CPU registers to their purpose. You should now carry out this activity.

Matching CPU components to their descriptions

On the web is an activity that asks you to match CPU components to their descriptions. You should now carry out this activity.

2.5 Buses



The **system bus** is a group of parallel wires, each carrying a single bit of data.

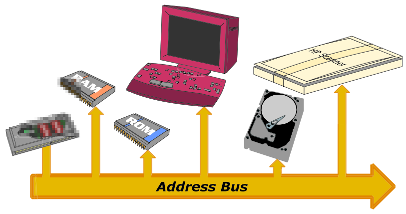
In a single bus system, input/output devices (I/O) and memory use the same communications channel.

A *two bus* system has a separate I/O channel and memory transfer channel. Larger systems make use of several I/O buses for more effective operation.

A single bus system is typical of small computer systems.

The system bus must provide components with the use of a *Data Bus*, *Address Bus* and *Control Bus*.

2.5.1 Address bus



The **address bus** is a **uni-directional** bus, transferring information in one direction only.

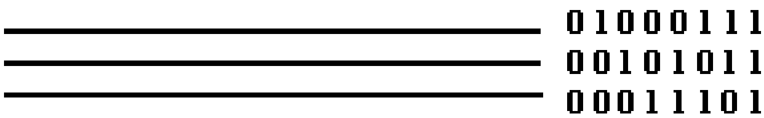
In a single bus system, input/output devices (I/O) and memory use the same communications channel. When the CPU needs to put data into memory or send it to a disk then it does so in the same way, only using different addresses. Devices are *memory mapped* and are treated by the system as if they *were* memory.

The address bus is made up of parallel wires, each capable of carrying 1 bit. The size of the address bus will determine how many memory locations can be *directly addressed*.

To understand this, consider first an address bus width of 1-bit. There are 2 distinct values that a single bit can represent (0 or 1). Thus a bus width of 1-bit can identify 2 unique addresses.

Now add another address line. There are now 2 lines, each of which can represent 2 distinct values. This results in 4 possible unique addresses shown below.

Now add another address line. There are now 3 lines, each of which can represent 2 distinct values, resulting in 8 unique addresses.



Can you think of a general formula to relate the number of directly addressable memory locations to the width of the address bus? If not, try answering the questions below and then think through the problem again.

Q30: How many memory locations can be directly addressed using 5 bits?

a) 10
b) 7
c) 32

Q31: How many memory locations can be directly addressed using 8 bits?

a) 8
b) 256
c) 16

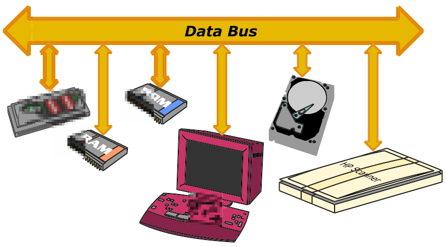
Generally then,

The number of memory locations = $2^{\text{width of address bus}}$

Thus, a 24-bit address bus will be able to distinguish between:

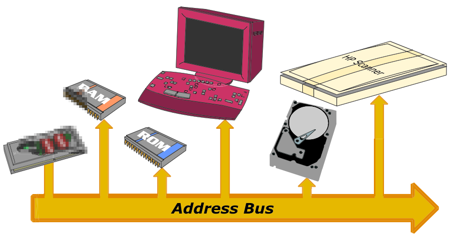
2^{24} = 16,777,216 memory locations.

2.5.1.1 Addressability



When the processor has to send or receive data or instructions from memory, cache or external devices, the medium along which the signal is carried is known as a bus. This is a set of wires or lines that can each transmit 1 bit at a time. Thus when an instruction is read from memory the bits are placed on the data bus and moved in parallel to the processor. Typically the data bus will be the same size as one memory location.

While the data bus moves data and instructions, a second bus, the address bus carries the address of the memory location to be accessed.



The size of the address bus affects the amount of main memory offered. If there were 8 address lines able to transmit 8 bits the maximum number of different addresses would be $2^8 = 256$

A 32 bit address bus addressing up to 4 Gb of memory is a more typical current size.

When an address is being specified there has to be a way to determine whether it refers to a main memory address or to one of the I/O interfaces that control communication with other peripheral devices. This can be done in two ways.

Memory mapped I/O

If the interface is **memory mapped**, a block of main memory addresses is mapped to the I/O interface. In this way the addressing process is exactly the same for main memory or I/O.

If there is no memory mapping the destination is defined by a signal on one of the control lines. This means that the same address values can be used for I/O and memory without confusion.

The address bus is effectively one way (uni-directional) while the data bus can transfer data in both directions (bi-directional), outwards for a write operation and inwards for a read operation.

A third bus, the control bus has the signal that identifies the type of operation, read or write. This can take 4 signals:

- read from memory;
- write to memory;
- read from I/O;
- write to I/O.

Therefore to fetch an instruction or data item from memory all 3 buses are used.

Address bus	address in memory or I/O device
Data bus	data to be written to or data being read from memory
Control bus	operation read / write memory or read/write I/O

This is illustrated in Figure 2.7

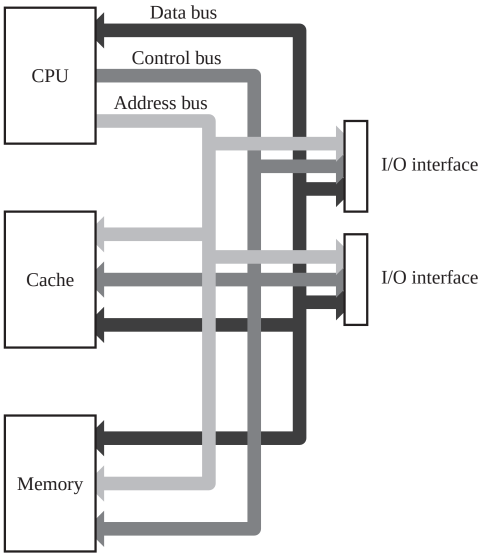
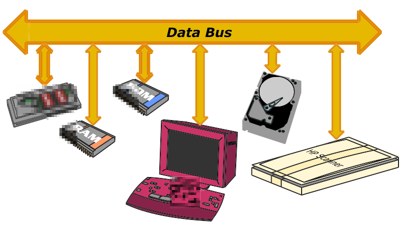


Figure 2.7:

2.5.2 Data bus



The **data bus** is used to transfer data to and from the CPU.

The data bus is a **bi-directional** bus which transfers data in both directions.

In a single bus system, the data bus is shared by main memory and external devices such as screens, printers and disk drives.

You can appreciate why this bus needs to be bi-directional if you consider some typical operations that are carried out. For instance, the CPU must fetch instructions from main memory which requires transfer in the direction from main memory to the CPU. If the stored program has instructions to calculate values and update variables, then the results of the calculations need to be stored in main memory. This requires a transfer of data from the CPU registers to main memory.

In the case of communicating with an external device, such as a hard disk, data must be loaded from the device and also saved to the device. This requires bi-directional data transfer.

2.5.3 Review questions

Q32: The purpose of the address bus is to:

a) initiate a read from memory operation
b) carry a memory address from which data can be read or to which data can be written
c) store results of calculations

Q33: Why is the address bus described as uni-directional?

Q34: The data bus is used:

a) to store the results of calculations
b) to signal a read event
c) to carry data/instructions from main memory to CPU or to carry data from CPU to main memory

2.6 Summary

The following summary points are related to the learning objectives in the topic introduction:

- Organisation of the component parts of a computer system;
- Description and structure of the components parts of a processor;
- The stored program concept and the fetch-execute cycle;
- The function of the processor components;
- Control lines functions and timings;
- The use of buses;
- The storage of data using registers, cache, memory and backing storage;
- Accessing memory.

2.7 End of topic test

An online assessment is provided to help you review this topic.

Topic 3

Computer Performance

Contents

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 toipc 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.*

Revision

The following exercise tests the prerequisites for this topic. Ensure that you are happy with your responses before progressing.

Q1: What supplies regular control signals to the processor that controls the timing of the data transfers and processes?

a) ALU
b) clock
c) control bus
d) control unit

Q2: A palmtop computer system is an example of a :

a) Embedded computer
b) Mainframe computer
c) Minicomputer
d) Microcomputer

Q3: Which of the following gives an indication of a computer's performance?

a) Modem speed
b) Modem bandwidth
c) Clock speed
d) VDU replenishing rate

3.1 Introduction

This unit on Computer Performance considers the factors that affect the performance of a computer system. Individual factors may well have an effect, but it is important to consider all factors collectively to allow for a meaningful analysis of a system's performance.

Different measurements of performance are considered along with current trends in improving computer specifications that aim for higher performance.

3.2 Measuring performance

When we talk about the performance of a computer we are usually interested in how quickly it can carry out instructions, or how many instructions it can execute per second. This is often talked about in terms of MIPS, or Millions of Instructions Per Second.

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.

It is important to note that improvements in technologies and design in each of these areas have a cost, and it is the developments with the best performance to cost ratio that succeed.

3.2.1 Clock speed

One of the prime factors affecting the performance of microprocessors is the clock speed at which they run. Every processor has an internal crystal-controlled clock which generates pulses at a regular rate. These pulses are used to synchronise the steps carried out by the microprocessor while carrying out the fetch-execute cycle. All processor activities will start on a clock pulse; for example, fetching an instruction, placing data in the Memory Data Register, transferring an operand from a general-purpose register to the ALU, etc.

The time between pulses is the **cycle time**. Early microprocessors had clock speeds measured in kHz (thousands of cycles per second) while modern processors such as the Pentium III can achieve speeds of about 1GHz (thousand million cycles per second), and the Pentium 4 above 2GHz. Technology such as VLSI allowed great improvements to be made in clock speeds.

Obviously clock speed is an important factor in determining the performance of a microprocessor. Thus a microprocessor running at 200MHz is likely to execute instructions faster than one which runs at 100MHz. However, when it comes to judging performance between competing processors, clock speed may not always be a reliable measure of relative performance. One of the reasons for this is that an instruction such as an *add* may take several cycles, the number of cycles required increasing with the complexity of the addressing method used for the operands. Table 3.1 tabulates the clock speed versus the performance of Intel processors as measured in **MIPS**. MIPS is now an outdated way to measure performance but it is the only measure applicable over the whole range.

Table 3.1: Clock Speed Versus Performance

Intel Processor	**MIPS**	**Clock Speed**	**Year**
4004 - first microprocessor on a chip	0.06	108kHz	1971
8008 - first 8-bit microprocessor	0.06	200 kHz	1972
8080 - first general purpose CPU on a chip	0.64	2MHz	1974
8086 - first 16-bit CPU on a chip	0.33-0.75	5-10 MHz	1978
8088	0.33-0.75	5-8 MHz	1979
80286 (286)	0.9-2.66	6-12 MHz	1982
80386DX - first 32-bit CPU	5-11.4	16-33 MHz	1985
80486DX	20-41	25-50 MHz	1989
Pentium - (*pentium* from Greek for five)	100	60 MHz	1993
Pentium Pro		150-200 MHz	1995
Pentium II		233-300 MHz	1997
Pentium III		450-600 MHz	1999
Pentium 4		1.4-1.8 GHz	2001

Table 3.1 shows that the performance as measured by MIPS has gone up at a higher rate than has the clock rate. For example, on these figures the average number of clock cycles required for an instruction in the Intel 8086 processor was about 10, while in the Pentium Pro, about two instructions per clock cycle are achieved. We shall look later at why this has happened.

3.2.2 MIPS

In an earlier topic you met the clock when we introduced the functions of the control unit. You learned that it was an electronic pulse, similar to a musical metronome, generated at a constant frequency and that on each pulse, machine operations were carried out.

Clock speed will clearly have an impact on performance. If more pulses can be generated per second, with machine operations carried out on each pulse, then it is safe to conclude that more machine operations are carried out as clock speed increases. This is readily observed in the slower performance of older PCs, running at clock speeds of 200 MHz, with most modern PCs, typically operating at around 1.2GHz.

The effect of clock speed on performance

On the web is a simulation that shows you the difference in performance of two computers operating at different clock speeds. You should now look at this simulation.

Care needs to be taken when simply comparing clock speeds. For instance if we compare one processor operating at say, 200MHz with a *different* processor operating at 200MHz can we say that they are *equivalent* in performance? No, we cannot.

Mip rate

A clock speed of 200 MHz does not mean that 200 million *instructions* are executed per second. It can take at least five clock pulses to execute an instruction. One to load the instruction, one to decode it, one to get any data that the instruction needs, one to execute the instruction and one to store the result. In this case, a 200 MHz processor would be capable of executing 40 million instructions per second. This is known as the **machine cycle time**, often expressed in *mips* (millions of instructions per second).

It may therefore be the case that two processors have the same clock speed but different mip rates.

3.2.3 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.

In the past 10 years processor clock speeds have increased at a phenomenal rate. The laws of Nature, however, limit how far clock speeds can be increased.

3.2.4 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.

A test to measure the efficient use of memory whilst running applications is the MemStone test. Basic system operations such as memory allocation, and reading and writing from various memory blocks are measured. Testing the speed for file access is done by randomly reading and writing data from a block of memory that is saved to a file. Also, a series of virtual memory allocations made to test performance for creating and freeing memory blocks.

3.3 Performance factors

3.3.1 Data bus width

The effect of data bus width on performance

On the web is an animation that shows you how the width of the data bus affects computer performance. You should now look at this animation.

Word size

A computer is described in terms of its *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. For instance, designers may build a 32-bit machine with a 16-bit data bus. This means that a 32 bit word must be fetched from main memory using 2 memory read operations; one to fetch the first 16 bits of the data and the second to get the remaining 16 bits. Clearly this is slower than a 32-bit machine designed to carry 32 bits on its data bus.

The overall performance is not necessarily *twice as slow* as there are other factors to consider. However we can say that a wider bus width will produce *increased* performance.

3.3.1.1 Trends in data bus width increases

64-bit computers appeared around 1993, however most of today's processors have a 32-bit word size. Increases in data bus widths and clock speeds of the Intel processor series, from 1st generation to the current 7th generation are shown in Table 3.2

Table 3.2: Intel processor series

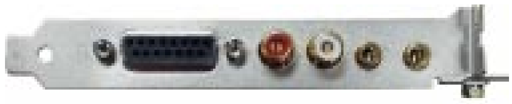
Type/Generation	Year	Data Bus Width
8088 - First	1979	8 Bit
80286 - Second	1982	16 Bit
80386SX - Third	1988	16 Bit
80486SX - Fourth	1989	32 Bit
Pentium - Fifth	1993	64 Bit
Pentium Pro - Sixth	1995	64 Bit
AMD Athlon - Seventh	1999	64 Bit

3.3.2 Data exchange with peripherals

Peripheral devices connect to the system bus, or I/O bus, via slots on the back of the computer. Such devices include printers, scanners, digital cameras, digital video recorders, mice, keyboards etc. They also include *mass storage* devices such as magnetic tape drives and disk drives.

Each device has an operational speed, uses its own language and deals with different amounts of data at a time.

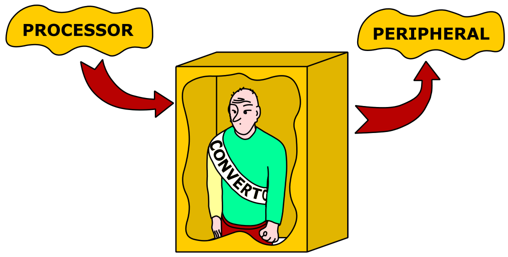
In order for these devices to communicate with the CPU they need to be interfaced. An **interface** is a unit that sits between the CPU and a peripheral device and compensates for the differences in speed, codes etc. to ensure compatibility.



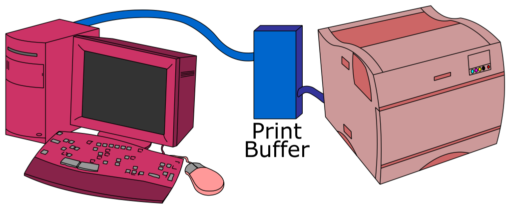
3.3.2.1 Standard functions of an interface

Every interface will need to carry out the following:

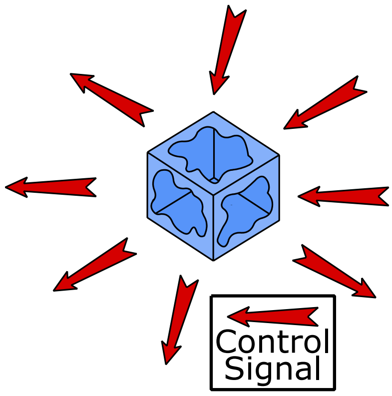
convert data from the format understood by the processor to the format understood by the peripheral



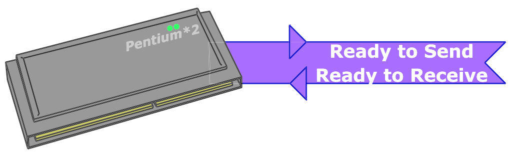
hold data in a *Buffer* as it is transferred from the processor to the peripheral and vice versa;



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.



Parallel to serial conversion

5 min

On the web is a simulation that illustrates parallel to serial conversion. You should now look at this animation.

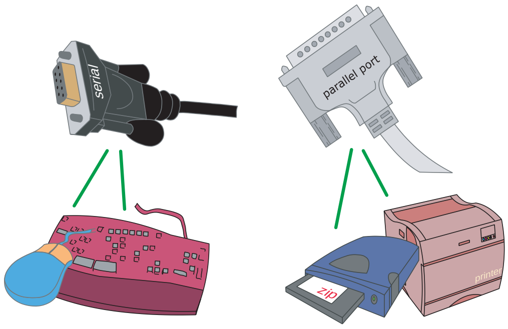
Identifying functions of an interface

On the web is an activity that asks you to identify the standard functions of an interface. You should now carry out this activity.

3.3.2.2 Data exchange with peripherals

Data can be transmitted in serial or parallel form. Most PCs have a combination of both with at least one parallel port.

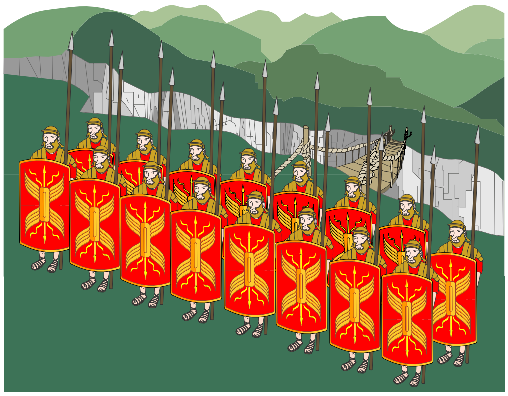
The keyboard and mouse connect through a serial interface, while printers, zip drives and CD-ROMs (requiring larger amounts and much faster transfer rates) connect to a parallel interface.



Serial and parallel communications operate differently.

3.3.2.2.1 Parallel transmission

With **parallel transmission**, each bit of an 8-bit byte is sent at the same time along a set of parallel wires. The intention being that all bits of the byte arrive at their destination at the same time.



Parallel transmission is clearly faster than sending out a single bit at a time, but is recommended where the distance between the transmitting device and the receiving device is fairly short, for example, connecting a printer to a PC.

Over longer distances there is a possibility of **skewing**, where the individual bits may arrive at their destination at different times. The data will lose its integrity. For longer distances, where speed is not essential, serial communication is more practical.

Parallel data transmission is illustrated in Figure 3.1

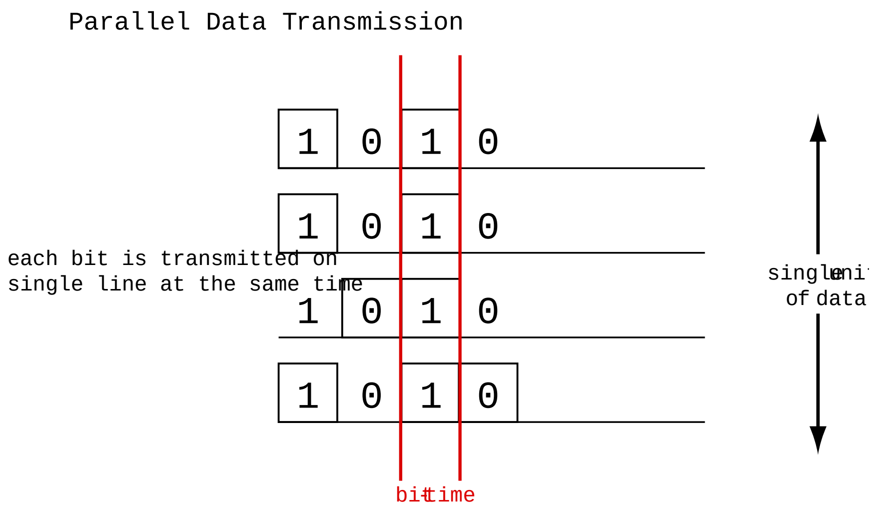


Figure 3.1: Parallel data communications

Example : Communicating with a Printer

A printer is connected to a computer using a standard, Centronics parallel interface. What follows is the sequence of events to transfer data to the printer.

1. The interface puts data on the parallel lines.
2. Signals the byte is ready to be transmitted.
3. If ready, the printer reads the byte transmitted.
4. Printer sends back an acknowledge signal to the CPU.
5. Interface prepares the next byte.

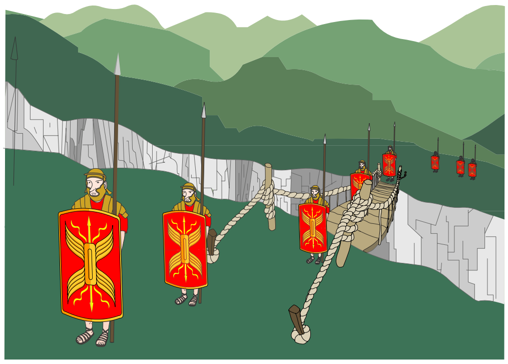
The interface also contains a set of status wires that can be used to signal events such as:

- *wait* because the printer is busy processing data already sent;
- *error*, such as a paper jam occurring or the printer is out of paper.

Identifying the characteristics of parallel communications

On the web is an activity that asks you to identify the characteristics of Parallel Communications. You should now carry out this activity.

3.3.2.2.2 Serial transmission



With **serial transmission**, each bit of the byte is sent out, one at a time, over the communications line. With **asynchronous transmission** the process of sending out the bits can be started as soon as the byte is available.

A **start bit** is used to signal to the receiving device that transmission is beginning, followed by each bit of the byte. There may or may not be a **parity bit** sent as part of the byte. Finally the transmitter waits a period of time, marking an output level that is known as a **stop bit**

Asynchronous serial communication of a single byte is illustrated in Figure 3.2

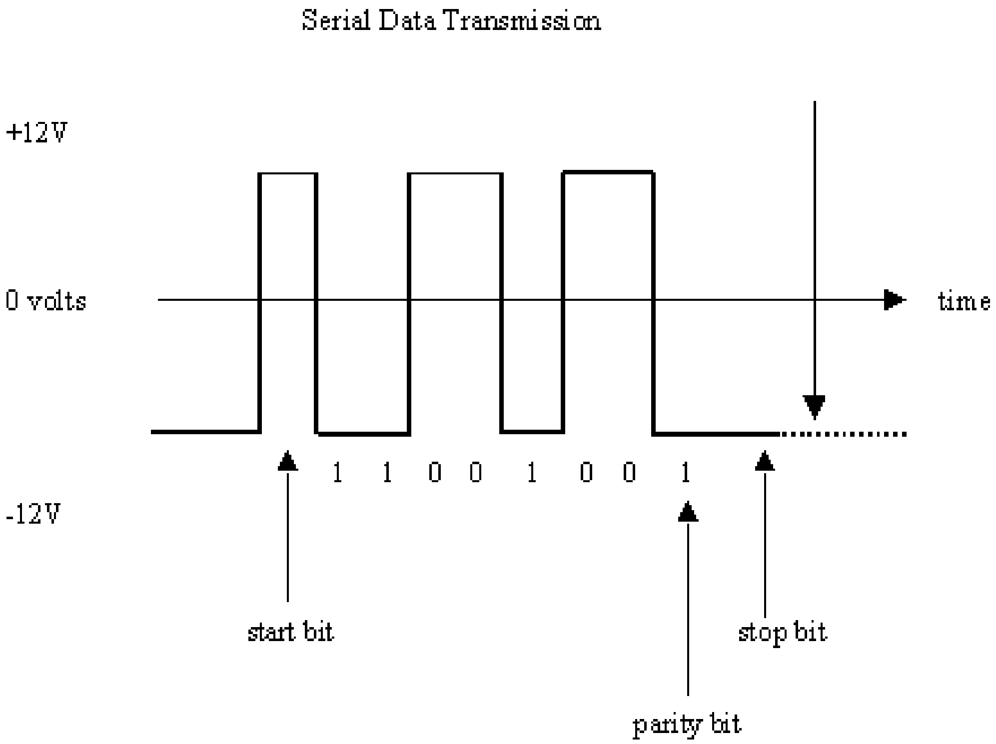


Figure 3.2: Serial data communications

With **synchronous transmission** data between two devices is timed to coincide with a clock pulse.

Example : Communicating with a mouse

As there is no pressing need for speed, a mouse is connected to a computer using a serial DB9 or DB25 connector. A universal serial bus (USB) may also be used. A USB interface is a typical port used to connect many devices.

1. A single data bit at a time sent to the CPU - typical rate of 1200 bits per second.

2. The interface converts this stream of bits to parallel to send on system bus.
3. Once the byte ready interface asserts an interrupt request.
4. CPU acknowledges request and interface places the byte on the system bus.

The CPU may choose to ignore the interrupt request which may lead to data loss. This is known as **data overrun**. In this case the interface must tell the CPU that data has been lost.

Identifying the characteristics of serial communications

On the web is an activity that asks you to correctly identify the characteristics of Serial Communications. You should now carry out this activity.

3.3.3 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.

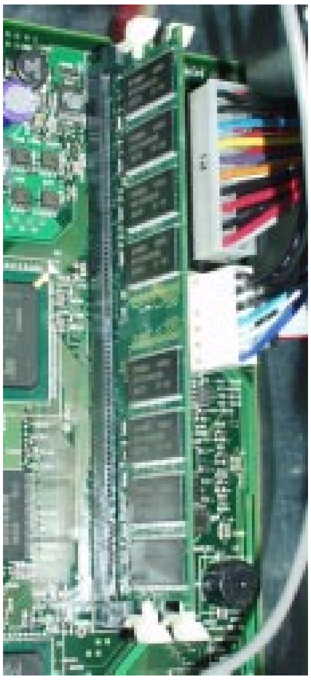
The first two aspects are dictated by the processor and logic board (motherboard). We will look in more detail at how the amount of memory and the use of cache memory affects system performance.

Empty RAM

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. This is known as a **memory upgrade**.



Full RAM

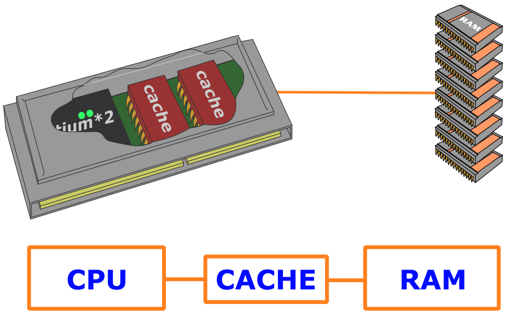
If your computer is struggling to run some software or you cannot load all the software you want at the one time, then adding extra memory will improve your system's performance. For example, if you are using an application that needs to manipulate large graphic files, video or sound then you should be thinking of upgrading RAM to the computer's maximum capability.

Machines typically come with 32 Mbytes or more RAM which can be upgraded to 256 Mbytes or even more.

When there is insufficient main memory then the hard disk can be used as an extension. This is known as **virtual memory** and results in slower performance since swapping data from main memory to hard disk and loading from the hard disk to main memory is much slower than directly accessing main memory.

Cache memory

Main memory bus speeds are not able to match the speed of the CPU and **cache memory** is used to speed up this transfer. This is a small amount of very fast SRAM that can reside *inside* the processor or sit *between* the processor and main memory.



When **writing** to main memory the CPU uses the cache to deposit data and then resumes its operations immediately. The data is transferred to main memory by the cache controller circuitry.

When **reading** from memory the CPU first checks whether the information is already available in the cache memory. If so then it can transfer this at high speed to the CPU.

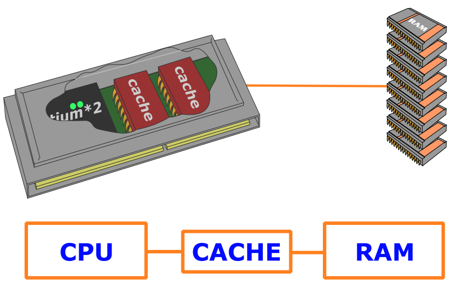
How cache memory is used by the processor

On the web is a simulation showing how the processor makes use of cache memory. You should now look at this simulation.

3.3.4 Cache memory

Program instructions are usually read sequentially. From one instruction it would be reasonable to assume that the next instruction required will be in the next memory location. This assumption is used to increase processor efficiency.

Although the movement of data within the processor is getting faster and faster, the system buses are not keeping up. This leads to wasted time while the processor waits for data to be fetched from memory.



To reduce this problem most machines now have a second, smaller, area of memory known as **cache 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.

When data or an instruction is read from memory, the memory locations following are copied into the cache memory. At the next read instruction the cache memory is read first. If the data is in cache the access time will be much lower than going to main memory. If the data is not in cache then main memory will be accessed, and although there is a slight loss of time from reading twice, the overall time saving in this method is quite significant.

Observing cache memory

On the web is a simplified simulation of the operations of cache memory. You should now look at this simulation.

The contents of cache are simultaneously held in RAM. When data is to be written back to memory it must be written to cache so that the cache is kept current. At some stage it will also have to be written back to RAM. Cache that has not been updated doesn't have to be copied back to memory it is just removed from cache when it is to be replaced by something the processor has a greater need for. There are 2 different ways to update cache memory.

Write through cache. When cache is updated memory is updated at the same time.

Write back cache. Cache is updated, but RAM is not updated until the content of cache is being cleared. Write back requires fewer write operations but there is an overhead in managing the selected updates. Write back cache is generally about 10% faster than write through cache.

3.3.5 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.

Prior to virtual memory, if data was too large to fit into the main memory the programmer had to break the data up into smaller sections, called overlays, each of which could fit within the main memory. The overlays would be stored on disk, and loaded into the main memory as required. However, the programmer had to manage the process of switching overlays and communicating between them. This approach was commonly used by applications forced to run on small and simplistic operating systems such as DOS.

Virtual memory not only automates the transfer of data from disk to main memory, but does it in such a way that the entire address space appears usable compared to overlays. Furthermore, virtual memory can be allocated on a per process basis, i.e. if the operating system is running multiple processes (multiprocessing), then each process appears to have the full address space entirely to itself. Multiprocessing operating systems are now common; typical examples include Windows XP, Linux, OSX etc.

Virtual memory and cache memory share some similarities, and it may be convenient to think of virtual memory as yet another layer on the bottom of the memory hierarchy. Unlike cache, virtual memory requires support of the operating system at the very least to access the hard disk. In addition, the algorithms used to manage virtual memory are typically far more advanced, since they must compensate for the slow speed of the hard disk and coordinate the sharing of memory between the processes that are running. It is therefore not uncommon to find sophisticated sharing, compression, and access prediction techniques being used to optimise virtual memory management.

3.3.6 Review questions

Q4: Computer A and Computer B each operate at 200 MHz. Why can it not be said that they are equivalent in performance?

Q5: What is meant by the term *Virtual Memory*?

a) memory that is not real
b) using the hard disk as an extension of main memory
c) adding more RAM modules

Q6: Explain why an increase in the width of the data bus can improve system performance.

Q7: A 24-bit address bus allows how many directly addressable locations?

a) 2 X 24
b) 26
c) 2^{24}

Q8: Explain why an increase in the width of the address bus have an affect on system performance?

Q9: A memory upgrade involves:

a) Redesign of the address bus
b) Insertion of additional memory modules
c) Replacement of the motherboard

Q10: How would you calculate the number of memory locations that could be directly addressed using a 16-bit address bus?

3.4 Summary

The following summary points are related to the learning objectives in the topic introduction:

- Indicators of computer performance include: clock speed, MIPS, FLOPS;
- Performance can be evaluated using these indicators;
- Benchmark testing is used to measure performance;
- Other factors affecting performance include: data bus width, cache memory and data transfer rates;
- Increasing clock speed, memory and storage capacity may improve performance.

3.5 End of topic test

An online assessment is provided to help you review this topic.

Topic 4

Peripherals

Contents

Prerequisite knowledge

Before studying this topic you should be able to:

- *Describe and compare typical input devices in terms of function, uses, and characteristics;*
- *Describe and compare typical output devices in terms of functions, uses, and characteristics;*
- *Describe and compare the features, functions and use of magnetic storage devices;*
- *Describe and compare the features, functions and use of optical storage devices;*
- *Describe the need for interfaces and their function.*

Learning Objectives

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

- *Describe the use of buffers and spoolers;*
- *Describe the hardware required to set up a LAN;*
- *Describe the hardware required to develop a website or a multimedia catalogue;*
- *Justify the selection of hardware in terms of its characteristics including resolution, capacity, speed, cost and compatability;*
- *Describe the features, advantages and uses of solid state devices;*
- *Describe developments in backing storage devices;*
- *Describe the functions of an interface;*
- *Make distinctions between serial and parallel interfaces;*
- *Describe interface speeds;*
- *Describe wireless communications between peripherals and CPU.*

Revision

The following exercise tests the prerequisites for this topic. Ensure that you are happy with your responses before progressing.

Q1: Which one of the following is a storage medium?

a) DVD
b) sound card
c) USB port
d) BIOS

Q2: What type of light do optical devices use when reading data?

a) ultra-violet
b) ultra-violet
c) infra red
d) LASER

Q3: A temporary store for data being transferred between two devices is called a :

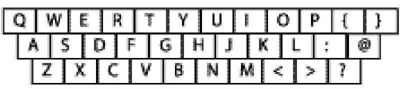
a) interface
b) register
c) buffer
d) cache

4.1 Introduction

This unit on peripherals considers a range of hardware devices required to carry out typical tasks using a computer system. For example, the peripherals required to set up a LAN, or to develop a multimedia product or a website. Hardware characteristics considered for such tasks include the speed, cost, resolution, capacity and compatibility. Developments and trends in the use of storage devices (including solid state devices) are also considered. Both parallel and serial interfaces are studied together with their various functions in the transmission of data between devices and the computer. This include wireless communications.

4.2 Input and output devices

4.2.1 Keyboards



The keyboard is one of the most common text-entry device in current use. Most keyboards have a standard layout and are identified by the first six letters of the alphabetic keys - QWERTY.

The layout of the keys in a QWERTY arrangement has its origins in the design of mechanical typewriters. In these designs, a character was formed by hitting a key, causing a hammer to drive a formed shape onto an ink ribbon which impacted onto paper. Depending upon the speed and sequence of keys pressed, it was possible for some hammers, due to their closeness in angle, to jam. To minimise this, the keyboard was designed so that letters commonly occurring in sequence were positioned at different ends of the keyboard. Thus the hammers could move onto the ribbon from different sides, preventing jamming.

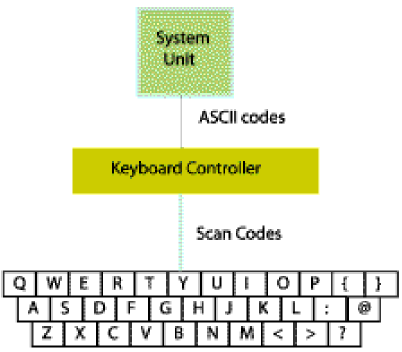
Although modern electric typewriters and computer keyboards do not have these mechanical problems, the original QWERTY design has remained. This simply saves typists, already skilled in using a QWERTY layout, from having to learn a new layout and undergo any necessary retraining.

There are typically more keys on a keyboard than those found on a typewriter. Additional keys include navigation keys (up, down, left, right), Home, Delete and Insert keys. Shift, Control and Alternate (Alt) keys are used to change the meanings of other keystrokes, such as capitalise, accent, etc.

Keyboard basics

When a key is pressed an electronic connection closes causing a **scan code** to be sent to the computer.

There are wireless keyboard variations that do not need a lead, an example of which is shown on the right.



On a PC the messages from a keyboard pass to the system unit through a special serial port. From this port they are sent to the **keyboard controller**, embedded on the VLSI motherboard or to a separate microprocessor with its own ROM and RAM.

As shift and shift-lock keys can alter the meaning of a keystroke, scan codes are interpreted by the controller in light of the current state of these key presses.

The keyboard controller generates a hardware interrupt to inform the processor that data is ready. An interrupt handler then translates the data into an ASCII or extended ASCII character code.

4.2.1.1 Modified keyboards

Modified keyboards are designed to suit the needs of the workplace in which they are used, or to improve user health and safety. For example, repetitive movements and awkward postures, traced to the use of computer keyboards, has increased cumulative trauma disorders (CTDs). One such disorder is **repetitive strain injury** (RSI) and some forms of wrist rest can help to alleviate this strain.

Some examples of modified keyboards are shown below. You will see examples of those customised to suit the needs of the workplace and others altered for ergonomic reasons.

Customised Keypads

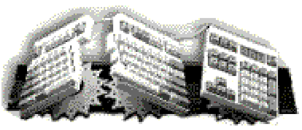
Companies provide **customised keypads** that can accomodate more or less keys than the conventional keyboard.

The keypads can be programmed allowing any key to be configured to send single or multiple scan codes.

Integrated keyboard and Touchpad

In this example the **integrated keyboard and touchpad** both keystroke and touchpad input, allowing more than one function in the same device.

Regarded as space efficient and used in situations where a mouse is impractical.



Adjustable-Split Keyboards

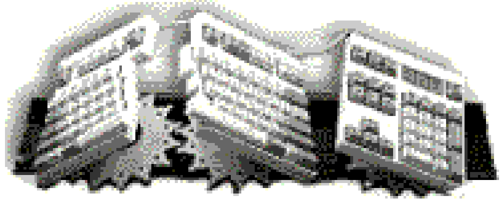
The **adjustable-split keyboard** is a three-piece folding keyboard, organised into three sections. Each section can be adjusted separately, allowing the user to type in as natural a posture as possible.

Below is a list of further keyboard alternatives and you will be exploring these later in your investigations.

- Dvorak Key Layout
- Adjustable-Split Keyboards
- Contoured Keyboards
- Vertical Keyboards
- Chording Keyboards

20 min

Investigating alternative keyboards



You have already met a few of the alternative keyboards that are currently available. Below are a further list that you should investigate using either the Internet or current magazines.

- Dvorak Key Layout
- Adjustable-Split Keyboards
- Contoured Keyboards
- Vertical Keyboards
- Chording Keyboards

As you carry out your investigations of each keyboard alternative you should try to answer the following questions:

- Why is it needed?
- What changes in design have been made?
- Who would buy this keyboard?
- What benefits does the new design bring?

4.2.2 Scanners

The **flat-bed** scanner can normally work with documents up to A4 size.

They are used in a similar way to a photocopier where a document is placed, face down, on a glass surface and scanned.

A light beam, the width of the page, is used to reflect light from the document. Photocells measure the level of light that is reflected.

This analogue data needs to be converted to digital, using an **analogue to digital converter** (**ADC**).

The resultant digital representation of the scanned image is a bit map.

Accuracy

Accuracy, in this context, is a measure of how close the computer image representation is to the original. This is influenced by the *resolution* and *bit-depth* capability of the scanner.

Resolution

Resolution is the number of dots per inch(dpi) that can be detected by the scanner hardware. A 600 dpi scanner requires 600 photocells per linear inch of the light beam.

Bit-depth

You have already met the concept of **bit-depth** when we introduced Graphical Data Representation. The same concept applies to scanner technology. For instance, representing 256 levels of light per image dot, typical of a grey-scale image, requires an 8-bit ADC. This provides the necessary 2^8 distinct levels of light.

Capacity

With this device, the feature of **capacity** is not considered, as image data is transferred directly to the computer for subsequent display on the monitor and/or storage to disk.

It is useful at this point to consider the backing storage requirements needed to record an image. You may wish to store the image in order to manipulate it within a graphic application, include it in a web document or simply examine it at a later date.

A simple calculation for an A4 image (typically 8.25 in x 11.75 in) at 600dpi with 256 levels of greyscale requires:

600 x 600 x 8.25 x 11.75 x 8 = 34897500 bytes = 33.28 Mbytes

Increase the colour definition to 24 bits and the file size increases 3-fold!!

You can appreciate the need to compress graphic data for delivery across the web where bandwidth can be fairly narrow. Two standards that will produce compressed image files are GIF (Graphic Interchange Format), better suited to drawings and cartoons that have only a few colours in them, and JPEG (Joint Photographic Expert Group) which can compress as much as 10 times more than GIF and is also more suitable for photo-realistic images. You will learn more about these graphic formats in studying the Multimedia Technology unit.

Investigating current scanner capabilities

20 min

(A) Use the Internet, or current technical magazines to find the names of TWO scanner manufacturers.

(B) For each manufacturer, make a list of the characteristics of their scanners under the following categories:

- **low to medium cost**
- **high cost**

You should produce a table of characteristics for each scanner using the following headings:

- **Manufacturer and Model**
- **Resolution**

- **Bit-Depth**
- **Cost**

(C) Find out the meaning of the term *interpolated resolution* and discuss why it would benefit a consumer to know what this means.

4.2.3 Sound

Computers have moved on a great deal from simply beeping when they boot up, or quacking at you when an error has occured. They can now playback sound quality to match that of an audio CD.

Sound as complex as a classical symphony can be captured, stored and reproduced. How is this achieved?

Sound is an analogue waveform that needs to be converted to a digital representation. It can be digitised from many sources e.g. radio, television broadcasts, CDs, long-playing records and cassette tape. It can also be captured directly using a microphone and sound recording software.

Digital representation is achieved by **sampling** the sound. This measures a value every *nth* fraction of a second and stores it digitally. The more often a sample is taken, the **sampling rate**, and the more data that is stored about each sample, the **sample size**, the better the quality of sound that is represented.

Resolution

There are three sampling resolutions most commonly used:

- 11.025 kHz (8-bit) - quality of voice mail
- 22.05 kHz (8-bit) - about the quality of AM radio
- 44.1 kHz (16-bit) - CD quality sound

The measurement of 44.1 kHz means that data is sampled 44,100 times per second.

Bit-Depth

Sample sizes can be 8-bit or 16-bit in depth and this determines the amplitude range. Thus 8 bits can represent a range of 256 amplitudes, while 16 bits can represent 65,536 amplitudes per sample.

Accuracy

How closely the digitised sound matches the original will clearly depend upon the sampling rate and sample size. The higher the sample frequency and the more bits used per sample, the closer the digital representation will be to the original analogue form.

Other factors that affect the quality of sound are **quantisation** and **clipping**. The value of each sample is quantised i.e rounded to the nearest integer which can introduce unwanted background noise. If the sound amplitude is greater than the sample range available then it needs to be clipped. This can distort the sound.

Capacity

Sound is stored directly to disk and so it does not really make sense to talk of the capacity of a sound card. As with all digitisers, you should be aware of the amount of backing storage that will be required to hold the data permanently.

A simple calculation with a sample frequency of 44.1 kHz (CD quality), a sample size of 16 bits, lasting for 2 minutes requires:

1. 44100 x 120 (secs) = 5292000 bps

2. 5292000 x 2 bytes = 10584000 bytes

= 10.09 Mb

Even high capacity media such as a hard disk will quickly run out of space if any reasonable length of sound at this level of quality is to be stored. There are several approaches to reducing file sizes. These include:

- downsampling - reducing the sample rate;
- reduce the sample size (e.g. 16-bit to 8-bit);
- use a compression technique to reduce the file size.

The technique chosen should be based on how much quality you can afford to lose in order to reduce the storage requirements. So for instance, could voice-level quality be used in parts of the application and CD quality reserved for, say, music recordings?

Investigating current sound card capabilities

Use the Internet or current technical magazines to carry out the following investigation:

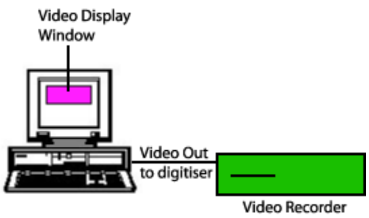
20 min

- **Find the names of two sound card manufacturers**
- **compare and contrast their products with respect to:**
  - **(a) sampling rate**
  - **(b) sample size**
  - **(c) compression techniques used**

4.2.4 Video

Current television video is based upon analogue technology. To display television video on a computer, the analogue signal must be converted to digital.

Video Digitising is performed by special video digitising circuitry installed on the motherboard of the computer.



Some digitisers simply allow the display of video images on a computer screen in a choice of window size. More sophisticated systems allow you to include special effects such as freezing and fading. Some digitisers can capture a single 1/30-second frame and save it as a still image.

File formats supported include **QuickTime**, **MPEG** (Motion Pictures Expert Group) and Audio Video Interleaved **(Avi)** format. AVI is a Windows based format that, like QuickTime, interleaves sound and video data.

With the advent of digital recorders there is no longer the need to convert from analogue to digital. You will be investigating these devices in a later activity.

20 min

Investigating digital video

Using the Internet, or current technical magazines, investigate the following:

- **What is Digital Video?**
- **Who were the pioneers of this technology and when did it first appear on the market?**
- **What advantages does it bring to the consumer?**
- **What is Firewire and how is this related to digital video?**

4.2.5 Digital cameras

A digital camera is one of the fastest ways to use pictures in a computer. There is no overhead in developing film and no need to scan images from a printed copy. Pictures are ready to use right away. Because they can be taken and downloaded to your computer in digital form they are ready to use in presentations or even send in email messages. It is now possible to print at home and achieve photo quality results.

The digital camera looks much like a conventional camera - only there is no film! Digital image data is stored in a memory card inside the camera.

Digital photographs are bit maps, made up of thousands or millions of pixels with values to represent image brightness and colour.

Digital cameras use hundreds of thousands, or millions, of photosensitive diodes, called charge coupled devices (**CCD**s), to record the intensity of light in an image. These analogue values are then converted to digital using an ADC.

Courtesy of Canon

Accuracy

As in scanners, accuracy refers to how well the computer representation of the image matches the original. This will depend on the number of pixels used to represent it i.e. the resolution. More pixels provide better detail and sharper edges. Please refer to your notes on scanners

Resolution

In cheaper range cameras the resolution is of VGA standard i.e. 640 x 480 pixels, capturing 307,200 pixels in each picture. Other cameras support "*megapixel*" representation, able to take high resolution shots at 1024 x 768 or 1280 x 960 pixel resolution.

Bit-Depth

You have already met the concept of **bit-depth** when we introduced Graphical Data Representation. The same concept applies to digital camera technology where the number of bits per pixel is proportional to the number of colours that can be represented. This will also affect the accuracy of the image.

Investigating digital cameras

20 min

(A) Use the Internet or current techical magazines to investigate the features supported in contemporary digital cameras. Try to compare features of an expensive, high-resolution camera and a cheaper option.

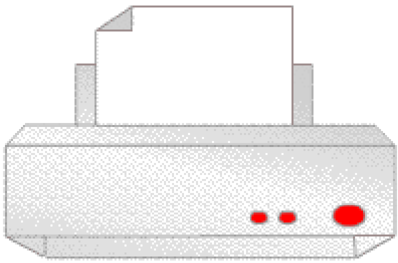
In particular, you should concentrate on

- **image capacity**
- **resolution**
- **bit-depth**

(B) Investigate how stored images are downloaded from the camera to a computer. Try to find out the image manipulation capability of any software that is used i.e. what can you do with the image after you have transferred it from the camera to a local hard disk?

4.2.6 Printers

Besides the monitor, printers are probably the next most commonly used output devices. They can be broadly classified as **character printers**, **matrix printers** or **page printers**. Character printers and dot matrix printers are pretty much outdated and have been replaced by affordable and more capable ink jet and laser printers.



Ink-jet Printers

Ink-jet printers are based on one of three different types of technology: *continuous flow ink-jet*, *liquid ink-jet* or *phase-change ink-jet*. We will look at how a liquid ink-jet printer works.

Liquid ink-jet

Also known as **bubble-jet**, this device operates by squirting tiny droplets of ink onto the page. The ink is first heated by passing an electric current through a coil. In milliseconds a bubble of vapour appears, forcing a tiny drop of ink from the nozzle onto the paper.

How a bubble-jet printer works

5 min

On the web is an animation that shows you how a bubble jet printer works. You should now look at this animation.

Characteristics

Resolution is typically 300 to 600 dots per inch. They support the printing of text and graphics, colour and a range of shades. Speed is pretty slow with a range of 4 pages per minute to 8 pages per minute, depending upon the model.

Laser Printers

This type of printer uses lasers to "write" a page image onto a special drum as an electrostatic charge. The charged drum attracts toner particles which are transferred to the page and heated to set the image.

Characteristics

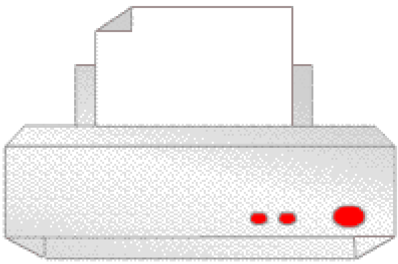
Resolution is typically 300 to 600 dpi, although higher resolutions are available if you are prepared to pay the price. They print a complete page at a time to a predefined maximum page length and width. Colour and a range of shades is supported. Speed ranges between 4 pages per minute and 40 pages per minute.

Investigating printer capabilities

20 min

For each printer type, you should construct a short report that describes what they are capable of in terms of:

- resolution
- buffer capacity - where appropriate
- colour
- speed
- cost



You will need to use the Internet, or current technical magazines to find out about the following two types of printer capabilities:

- Laser
- Bubble Jet

4.2.7 Multiscan monitors

The monitors for desktop computers are based on similar technology to that of a television screen, the Cathode Ray Tube (CRT). However, they are of a much higher quality. They operate essentially as follows:

- **Find the names of two sound card manufacturers**
- **compare and contrast their products with respect to:**

- **(a) sampling rate**
- **(b) sample size**
- **(c) compression techniques used**

Colour is rendered using *three* electron guns: one for Red, one for Blue and one for Green. These guns illuminate the red, blue or green phosphors that are arranged on the screen in groups of three as a triad.

5 min

How the monitor works

On the web is a simulation that shows you how a monitor works. You should now look at this animation.

The number of colours that can be represented will depend upon the intensity levels that each gun is capable of generating. For example, using 8-bits per RGB gun provides a total of 24 bits for colour. This allows for 2^{24} = 16,777, 216 possible colours. This level of colour is enough for most professional film and video graphics products, referred to as *photo-realistic*.

How many times the scan from the top of the screen to the bottom takes place per second is known as the *scan rate*. In order to prevent a flickering image, this rate needs to occur at no less than 50 times per second, which is measured in Hertz as 50 Hz.

The scan rate is set by a *video adapter* and some manufacturers provide adapters with different screen proportions and resolutions. A resolution of 640 x 480 pixels (VGA) will require a different scanning rate than a resolution of 800 x 600 pixels (SVGA). **Multiscan monitors** are able to deal with these different scanning frequencies. First appearing in 1985, multiscan monitors are now prevalent on medium to high-resolution displays.

20 min

Investigating multiscan monitors

Use the Internet, or current magazines to investigate the capability of multiscan monitors. You should compare and contrast the products of at least two manufacturers, paying particular attention to:

- screen resolution
- scanning frequencies

4.2.8 TFT Monitors

TFT (thin-film transistor) technology is a liquid crystal display (LCD), common in notebook, laptop and other flat computer screens. Each pixel has a transistor that controls the illumination of the display. The current required by each transistor for pixel illumination is very small and therefore can be switched on and off more quickly.

4.3 Selecting hardware to match operational requirements

This section will take you through the process of matching hardware to the operational requirements of a given situation. An example of setting up a library system is used to outline the process. This is followed by another scenario and an activity fin which you will consider what hardware and peripherals are required. You should draw from your knowledge gained earlier in this topic. (Later on in your course you might wish to return to this section and tackle the activities from a wider perspective).

4.3.1 Analysing the problem

Imagine that you are a computer consultant specialising in the recommendation of hardware and software to suit clients' needs. Clients are customers with whom you sit down and try to figure out their operational requirements. You will know all about the processes of eliciting client requirements, having studied the Software Development unit.

Imagine that one of your clients is a librarian who wants to look at using computers to manage book cataloguing and book issuing. She describes her operational requirements to you as follows:

Dear Student,

I work in a library and it is my job to catalogue and issue books that the library owns. At present the library has 50,000 books and 15,000 members. Each book is labelled with a unique book code, up to 10 characters in length, together with the library name and address. A catalog of library books is maintained with the book title, author and replacement cost. Members have a library card that they must present in order to borrow a book. They are allowed to borrow up to 3 books at a time. Each book can be on loan for a maximum period of 4 weeks, thereafter a letter of request to return the book is sent out and an overdue fine of 50 pence per day is applied. We would like to look at using a computer to help us with the recording and issuing of our books. We have a budget of 3000 pounds to spend. Can you help?

The first stage is to do some *close reading* of these operational requirements. Try to make a list of all *objects* and *operations* that you can identify. One approach is to think of the *verbs* as operations and the *nouns* as objects.

You could end up with the list in Table 4.1 below.

Table 4.1: Objects and Operations of The Library System

Objects	Operations
book	enter into catalog
	issue to member
	return to stock
member	add a new member
	change member details
book label	create new label
	add label to book
reminder letter	create reminder letter
	send to member
penalty	add penalty to member

4.3.2 Hardware requirements

The hardware that you recommend will need to match the operations, support all identified software and fall within the specified budget. Remember, you are not simply recommending your favourite PC. You are setting out to provide an environment that meets all operational requirements and is within budget.

RAM requirements

The amount of memory you specify will need to support the application software and the data. Software developers usually provide details of how much RAM is required and whether it will operate using a virtual memory environment. Add to this figure an estimate of the data size by calculating the total number of records in the database, together with the size of each record. Make this future-proof by including a reference to how large the database is likely to grow.

If you wish the operations to support multitasking then you will need to include the RAM requirements for each different package type.

Backing Storage

The specification of backing storage should consider how frequently the data is updated, the volume of data and any backup procedures you feel are required. The librarian will wish to have direct access to the data on an ad-hoc basis and so you will be looking at some form of disk based storage. If using a hard disk you may consider partitioning the disk and using one of the partitions as a backup area. This, however, is unwise as disk failure will render the entire system unusable. You should perhaps consider a removable disk medium as backup.

Clearly the backing storage medium should also be able to hold the operating system, the applications software and the user data.

Processor performance

This is not a particularly high performance area and does not require any specialised co-processor hardware. A modern PC operating at 1GHz is more than adequate to meet the librarian's needs. You should, however detail the processor specification and show that it meets the necessary performance levels.

Peripherals

The librarian will clearly need printing capability. You need to decide whether this should take the form of high quality laser printing or a lesser quality inkjet printer. Will colour be required? Does the print have to be rendered on special media e.g. acetate, glossy paper? How often is printing performed? Is print speed significant in this application?

Communications

Is it necessary and affordable in this application to have more than one computer system that is connected to a library network? It cannot be fully justified given the current budget, but you should consider future possibilities. Perhaps at a later date the library would like to provide members with access to the Internet, or allow them to browse through the library catalogue on-line. Your computer specification should then consider network interfacing. Does it have a network interface card and network operating system?

4.3.3 Justification

This is the most important section of your report. It gives you the opportunity to demonstrate your reasoning and shows:

- how well you have analysed the operational requirements;
- your ability to select appropriate hardware and software to meet user needs;
- your understanding of issues of hardware and software compatibility.

For each hardware and software component you should describe which part of the operational requirements are being met. The justification is in terms of:

- **Functionality** - has the operational requirement been met. So for instance if you have selected a particular output device, then does its characteristics (speed, capacity etc.) meet user needs.
- **Cost** - is the total amount of hardware and software within the specified budget.
- **Compatibility** - have you ensured that the software versions are supported by the selected operating system? Will the software run on the hardware? Can the peripherals selected be supported by the system?, etc.

Simulation of the specification of a system

On the web is a demonstration that will show you how to select a computer system to match a set of operational requirements. You should now look at this demonstration.

15 min

4.3.4 Setting up a LAN

4.3.4.1 A small secondary school using RM Community Connect

This scenario outlines the hardware (and other) requirements when setting up a LAN.

Research Machines (RM) provide Community Connect as a networking solution for schools who cannot afford to commit a large number of technical support staff to running their networks. The Community Connect software is based on Windows NT4 as the server operating system, Internet Information Server running the Intranet and Windows 98 on network stations. Schools are unusual network environments in that they require high bandwidth connections, have peak periods when users log on and off, and need a high level of security. They are also unusual in that users do not normally always work from a specific network station, but may log on to several different stations in the course of a day. RM provide a comprehensive suite of network management tools specifically designed for schools.

Hardware: Two servers, one acting as a file and applications server the other acting as a communications server. 150 network stations with a variety of different specifications. Both servers and a router are connected to a 24 port 100Mbps switch. Cabling throughout the school is category 5a UTP cabling connected using 100Mbps hubs. The *router* is a Cisco 2600 series connected to two ISDN lines giving access to four 64Kbps data channels.

Software: Both servers are running Microsoft Windows NT4. The network stations run Microsoft Windows 98. For Internet access the communications server runs Microsoft Proxy server. RM provide a suite of network management applications for adding users and groups to the network and for installing software on stations remotely.

Communications: The network has a Gigabit per second (Gbps) backbone between the servers and the switch and the rest of the cabling runs at 100Mbps. The protocols used are **NetBEUI** and TCP/IP.

Functionality: Printers can be shared throughout the school using dedicated print servers. Internet access and email are available from all network stations. The Community Connect software allows users to access files on their network drive via an Internet log-in using the Point To Point Tunneling protocol. The RM Connect system allows users to access shared areas of the hard disk which the teacher has set up containing class materials and work-sheets. A school Intranet is available containing subject specific web links and pupil projects with a staff area for the school handbook and other administrative documents.

Security: Network security is imposed using NT identities and passwords. Users are made members of NT groups whose permissions determine their access to areas of the Intranet, levels of access to software and access to shared resources.

Station security is imposed using proprietary software which backs up all crucial system files on to a hidden directory on drive C and sets them to read-only status. These files are not visible to users when they are logged on to the network. The software restores these files when a station is re-booted. This means that any configuration changes made to a station are removed by switching it off and on again. In the event of the system files of a station being permanently corrupted, the operating system and other software can be installed remotely and the station "rebuilt" to the same state it was in prior to the problem.

Developing a website or a multimedia catalogue

Make a list of the hardware you would need to build a website or a multimedia catalogue stored on a CD. For each item of hardware, identify the operational requirements. For example, the resolution of the screen (and the camera), and the capacity of the media. Record your findings in a table format and discuss with your teacher.

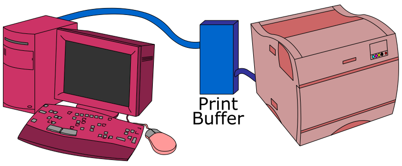
If you have time you should consider the doing both tasks

4.4 Buffers and spoolers

4.4.1 Buffering

A **buffer** is an area of computer memory that is allocated to transferring data between the computer and a peripheral. A buffer will be used when a fast acting part of the system is exchanging data with a slow acting device. The buffer stores data until it can be dealt with.

Peripheral Buffers



A printer operates at a much slower speed than the computer. A program can continue operating without waiting for each character to be printed if the data is sent to a buffer.

The buffer is normally managed by the operating system which sends the data to the printer when it is ready to receive data.

On page printers, such as laserwriters these buffers are memory stores within the device. Some typically hold up to 8Mbytes of data, with more expensive printers holding 20 or more Mbytes.

Disks and Tapes

Mass storage devices require data to be read or written in such a way that a *block* of data is moved in a single operation. The computer and the peripheral must be capable of sending and receiving a whole block of data at high speed. Block data transfer is managed by buffering.

An area of memory is allocated as the buffer and when information is to be transferred then it is stored in the buffer until the entire block is complete. From there it can be processed.

Interface Buffer

The component that handles conversion of serial data to parallel and vice versa is called a Universal Asynchronous Receiver/Transmitter (**UART**). With the present high rates of transfer in modern PCs, one method of relieving the CPU of having to continuously stop what it is doing, initiate a transfer of a single byte of data and then resume its operations, is to use a buffer in the UART. This is simply a **F**irst **I**n **F**irst **O**ut, FIFO queue. (You will learn more about this data structure if you study the Programming Unit).

The FIFO queue in the 16550 UART uses two 16-byte buffers. In this way the CPU can place up to 16 bytes of data in the buffer, ask for it to be sent out and resume its current task.

When receiving data, the UART can build up to 16-bytes of data before interrupting the CPU to transfer them. This means that the CPU does not need to be interrupted as often and reduces the possibility of data overrun.

4.4.2 Spooling

When large amounts of data are to be sent to a peripheral device, or when the peripheral is shared across a network then spooling is a preferred method of compensating for the difference in speeds of the processor and the peripheral.

Spooling involves the input or output of data to a tape or a disk.

This, for example, allows output to be queued from many different programs and sent to a printer by a **print spooler** (special operating system software).

The print spooler stores the data in files and sends it to the printer when it is ready, using a **print queue**.

Spooling to a printer

5 min

On the web is a simulation that shows how spooling is handled by the operating system. You should now look at this animation.

4.5 Storage devices

4.5.1 Magnetic

Magnetic storage devices include *hard disks*, *floppy disks*, *Zip disks* and *magnetic tape*. They are called magnetic storage devices because their recording surfaces are coated with a material that responds to magnetic fields to enable data to be stored.

These storage devices can be fixed or removable. **Removable storage** devices allow the user to disconnect the device and physically transport data from one computer to another. Varieties of removable devices include the Iomega and Syquest hard disks and Jaz cartridges.

4.5.1.1 Hard disk

All the **sectors** around the disk, equidistant from the centre, form a **track**. With multiple platters, the collection of tracks on each platter, equidistant from the spindle is called a

cylinder.

How the hard disk is configured

5 min

On the web is a simulation that shows you how the tracks, sectors and cylinders are configured on a hard disk. You should now look at this animation.

When data is to be read or written, the read and write heads are moved to the appropriate track, where they wait until the relevant sector spins past.

Speed

The **rotational speed** of hard disks has improved, from 3000 revolutions per minute (rpm) of very early disks, to current rotational speeds of 5,400 and even 7,200 rpm.

Performance is also measured in terms of the rate of data transfer from the disk. This will depend on the type of interface being used. For example, **SCSI** (Small Computer Systems Interface) can transfer at a rate of 5Mbytes per second, while Ultra Fast SCSI III can transfer at rates of 40 Mbytes per second.

Capacity

Hard disks have improved tremendously in their capacity to store data in the last 10 years. From the modest 10Mb disks of the early 80s to current 8Gbyte disks on many of today's PCs.

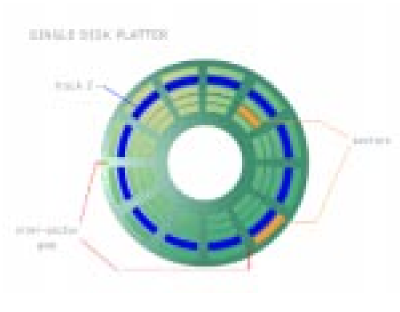
Access

The hard disk is a **direct access** device, meaning that data can be directly read or written to any portion of the disk

Investigating hard disk capability

20 min

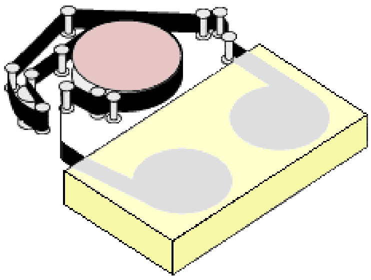
You should use the Internet or current technical magazines to investigate the following characteristics of hard disks:



- **Capacity**
- **Speed**
- **Cost**

Using textbooks or the Internet, find out how accuracy of stored data is achieved. Keywords you might like to explore are :*cyclic redundancy checks* and *error correction codes*.

4.5.1.2 Tape storage



Storing data on tapes used to be the only solution to backing up hard disks of large capacity. Now, with the advent of large, *removable magnetic disks* and optical CR-RW technology, this is no longer the case. However, removable storage media is comparatively expensive, with overall costs up to ten times that of tape. Tape, therefore, still has the edge in this market.

Tape is read and written on a *tape drive*. This drive winds the tape from one reel to the other causing it to move past a read/write head. Data is written to tape in blocks with **inter-block gaps** between them. The tape runs continuously and a single operation writes each block

Data is stored on magnetic tape as magnetised regions on the surface of the tape induced by the magnetic recording head. To read data, the tape passes under the read/write head and the stored magnetised regions produce very small voltages in the head, leading to a current in the head coil. This current can be analysed to give a representation of the stored binary data.

Capacity

Magnetic tapes have large capacities, reaching up to several gigabytes and come in a variety of sizes and formats. These are shown in Table 4.2

Table 4.2:

Type	Capacity	Description
Half-inch	60MB to 400MB	Half-inch tapes come both as 9-track reels and as cartridges. The tapes themselves are relatively cheap but they require expensive tape drives.
Quarter-inch	800MB to 20GB	Quarter-inch cartridges (QIC tapes) are fairly cheap and support fast data transfer rates. The drives are made to work with different lengths of tape and there are two popular cartridge sizes: full (data-cartridge) and minicartridge.
Helical-scan	5GB to 40GB	8-mm helical-scan cartridges are based on VCR tape technology. They have the largest capacity but need fairly expensive tape drives. Quite slow data transfer rates.
4mm DAT	2GB to 24GB	DAT (Digital Audio Tape), using the same technology as helical scan, has the greatest capacity but also requires a fairly expensive tape drive and has a slow data transfer rate. However, the DDS format can support 20 and 40GB capacity with a transfer rate of 2.4 and 4.8MBps

Since their introduction, tape drives have passed through many stages of improvement with extremely reliable Digital Audio Tape (44.1 kHz, 16-bit record and playback DAT) drives representing the current state of the art. A 4mm DAT tape can now store up to 24 Gbytes of data!

Access

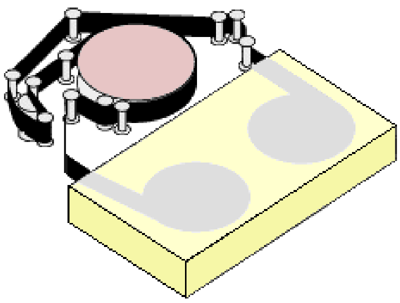
Tapes are *sequential access* devices which means that to get to a particular block of data on the tape, it must go through all the preceding blocks of data. Accessing data on tapes is therefore much slower than accessing data on disks.

They are not suitable as storage media for applications where data needs be used regularly - where a disk is a more appropriate medium. Because tapes are so slow, they are generally used only for long-term storage and backup.

Investigating Magnetic Tape Technologies

20 min

Using the Internet or current technical magazines, investigate the cost, capacity and access times of the following tape media



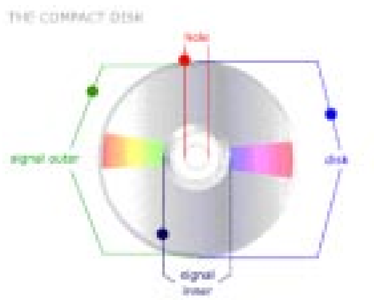
- **half inch**
- **QIC**
- **Helical scan**
- **DAT**

You may wish to look at the following manufacturers in your search:

- *Athana, Gigatek, Hewlett-Packard, Imation, Verbatim.*

4.5.2 Optical

The most common optical storage device is the CD-ROM (Compact Disk Read-Only Memory). It is a read-only medium whose contents cannot be altered once data is written to it.



During the mastering of a CD, a reasonably high powered laser is used to burn **pits**, typically 0.5 microns wide, 0.83 to 3 microns long and 0.15 microns deep, in a spiral, outwards from the centre of the disk. The disk is then coated with a protective layer of aluminium.

If the spiral was straightened out then the data would stretch for four miles!!

The CD player uses a lower powered laser to track the spiral as the disk turns. A spot of infrared light is shone onto the spiral to detect the data stored there. Pits on the disk surface encode digital data and signal the path of the spiral.

How data is read from a CD-ROM

10 min

On the web is an animation that shows you how data is read from a CD-ROM. You should now look at this animation.

Recordable CDs (CD-Rs)

The read-only limitation of CD-ROMs has been overcome by the creation of writeable CDs. The CD-R drive operates a laser light at one of 3 different levels. At low levels it detects the presence or absence of pits, while at the highest level it can burn data onto the surface.

Recordable/Writeable CDs (CD-RW)

The next stage in CD development was the introduction of CDs that erase current data and store new data. These operate using the highest power of laser to melt a small region of the recording layer. After freezing, the middle power laser is used to warm the surface to a temperature that is less than melting point but high enough to create the highly reflective crystalline form. The highest power is then used to write the data.

Capacity

CD-ROMs are typically 650 Mbytes in capacity. However, the future use of blue lasers, instead of red laser light, will deliver increased CD-ROM storage capacity. This is due to the shorter wavelength of the blue light compared to red light used in CD-ROM drives. Using a shorter wavelength results in smaller pits, allowing higher pit densities and more data storage. In later generations of CD-ROM we could see the capacity increase four-fold, from a mere 650 Mbytes to around 3 Gbytes.

Speed

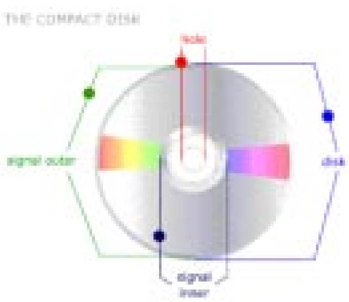
In a single speed CD reader, data can be read at a rate of 150Kbytes/sec. This is sufficient for audio but is very slow for motion video or large image files. Multiple speed CD readers, such as 8xCD-ROM or 24xCD-ROM can read data at a rate of 1.2 Mbytes/sec and 3.6 Mbytes/sec respectively. Notice that these are multiples of the single-speed rate. Although 48xCD-ROMs are available, at data rates of 7.2 Mbytes/sec, they pale in comparison to SCSI disk drive rates.

Access

CD-ROM access is direct.

Investigating CD-ROM capabilities

20 min



Use the Internet or current technical magazines to investigate the current capabilities of CD technology in terms of **cost**, **capacity** and **speed**.

During your investigation, try to trace the technological advances that have led to current erasable and writeable CD media.

4.5.3 Magneto-optical

Magneto-optical storage devices combine magnetic and optical technologies to read and record data. With a magneto-optical disk, a laser beam and a magnetic field is used to write the data. Only the laser is used to read the data.

The device works using the property of **coercivity** and the effect of heat on this property.

Coercivity describes how easily magnetisation can be switched from one orientation to another. If coercivity is low then magnetisation can be achieved using the magnetic field strength of the write head. However, if the coercivity is high then the magnetic field that must be generated needs to be a lot stronger.

Coercivity falls as temperature rises. At a critical temperature, the **Curie Temperature**, the material stops being a permanent magnet. After cooling, magnetic property returns. If the disk surface is heated it becomes easier to change the magnetic orientation using the magnetic field strength of the tiny write head.

This is how laser and magnetic fields combine to store data on a magneto-optical device.

4.5.4 Solid state storage devices (SSSD)

Solid-state storage devices are made up entirely from electronic components i.e. they have no moving parts. They are also called RAM disks, as they take the place of a magnetic disk as a mass storage device. They can be in the form of a plug-in card or cartridge containing memory chips. The chips of a SSSD are typically static RAM or Electrically Erasable Programmable ROM (EEPROM or Flash EPROM).

SSSD are used with devices where space is at a premium e.g. in a camera, or when portability is desirable e.g a USB flash drive (Figure 4.1).

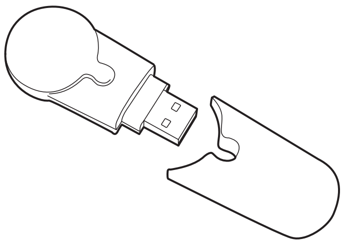


Figure 4.1: USB flash drive

4.6 Interfacing

Interfacing hardware means making the appropriate connections so that two pieces of equipment can communicate or work together effectively. This section considers the factors that have to be resolved between the devices before communication can happen.

4.6.1 Data format

Data to be communicated has to be arranged in such a way to be understood by the hardware or software doing the sending or receiving. The data format may have to be converted from serial to parallel and vice versa, or from analogue to digital and vice versa.

4.6.2 Parallel/Serial interfacing

In data transmission, the techniques of time division and space division are used, where time separates the transmission of individual bits of information sent serially and space (in multiple lines or paths) can be used to have multiple bits sent in parallel.

Serial connection, operation, and media usually indicate a simpler, slower operation (e.g. a serial mouse attachment).

Parallel connection and operation (e.g. multiple characters being sent to a printer) indicates faster operation.

This indication doesn't always hold since a serial medium (for example, fibre optic cable) can be much faster than a slower medium that carries multiple signals in parallel.

4.6.3 Digital/Analogue interfacing

Digital-to-analogue conversion is a process often carried out by a device in which signals with defined levels (usually two) are converted into signals having a theoretically infinite number of levels. For example, a modem will convert computer data into audio-frequency (AF) tones that can be transmitted over a telephone line. The circuit in the modem that performs this function is a digital-to-analogue converter (DAC).

Analogue-to-digital conversion is the opposite. The modem will convert an AF signal from the telephone line to a digital signal that the computer will understand.

4.6.4 Voltage

Voltage is the electrical force that allows devices to work. It is also required to send data signals between devices. As devices operate and send signals at differing voltages then these quantities must be interfaced to allow successful communication.

4.6.5 Protocols

A protocol is a set of rules that govern the transmission of data. Certain standards are set to allow for successful communication. The standard (or protocol) will dictate things like data format, timing, voltage levels etc required by devices to allow for data exchange. Protocols exist within devices as well as between devices to allow for data to be passed through the systems layers. Protocols are also used extensively when transferring data across the Internet.

4.6.6 Status signals

Status signals from a device indicate what the device is doing at any given moment. This data can be used by other devices when preparing to transmit data. For example, if a device is unable to receive data, then a transmitting device can delay transmission and retry later.

4.6.7 Speed

Different devices send and receive data at different rates. If a device sends data at too high a rate, then the receiving device must compensate for that, possibly by buffering the data. Alternatively, the devices can agree a rate prior to transmission by utilising a protocol.

4.6.8 Wireless communication

Wireless communications can be achieved using WAP (Wireless Application Protocol). This, like any protocol, is a specification for a set of communication rules to standardise the way that wireless devices, such as cellular telephones and radio transceivers, can be used for Internet access, including e-mail, the World Wide web, newsgroups, and Internet Relay Chat (IRC). Interconnecting devices centred around an individual person is called a wireless personal area network (WPAN).

Typically, a WPAN uses technology that permits communication between devices in a short radius of about 10 metres. One such technology is Bluetooth.

4.7 Summary

The following summary points are related to the learning objectives in the topic introduction:

- Characteristics of peripherals in a system design, including speed, capacity, cost, and compatibility;
- Differences between devices resolved through interfacing;
- Interfacing techniques including buffering, spooling, data format conversion, voltage conversion, handling status signals;
- Advantages in buffering and spooling;
- Data format conversion for parallel and serial transmissions;
- The use of solid-state technology in mass storage;
- The use of wireless technology in data transmission;
- Selection techniques to support typical tasks (LANs, web sites, multimedia).

4.8 End of topic test

An online assessment is provided to help you review this topic.

Topic 5

Networking

Contents

Prerequisite knowledge

Before studying this topic you should be able to:

- *Describe the features of LANs, WANs and the Internet;*
- *Describe the function of a client server on a network;*
- *Describe the benefits of networks;*
- *Describe the functions and uses of e-mail.*

Learning Objectives

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

- *Make comparisons between LANs and WANs;*

- *Make comparisons between intranets, internets and the Internet;*
- *Distinguish between a mainframe and a network;*
- *Make comparisons of peer-to-peer and client server networks;*
- *Describe the functions of file, print and web servers;*
- *Describe network topologies and the effects of failures.*

Revision

The following exercise tests the prerequisites for this topic. Ensure that you are happy with your responses before progressing.

Q1: What category of networks would be used to connect the computers in an office block?

a) LHN
b) WAN
c) MAN
d) LAN

Q2: What is the main factor that determines the category of a network?

a) distance
b) speed
c) price
d) number of clients

Q3: Using e-mail for unsolicited adverts is called:

a) e-commerce
b) spamming
c) airmail
d) snail mail

5.1 Introduction

This unit on Networking considers the basic system topologies and functions that allow computers to operate on a network. The various types of networks are considered ranging from small peer-to-peer networks to LANs, WANs and the Internet. Comparisons are made between the various networks in terms of type and scale and the effects of network failures.

5.2 Networks

If you purchase a computer nowadays, it will almost certainly be equipped with the means to connect it to another computer - either via a telephone line or a network cable, or wireless. We now take it for granted that we can use computers to communicate with other people and retrieve information from other computers anywhere in the world. The ability to connect computers together so that we can use them as communications devices as well as processing devices has revolutionised the way we work and the way we spend our leisure time. At its simplest, connecting two or more computers together turns them into a computer **Network**. Connecting two computer *networks* together creates an **Internetwork**. At its most complex, connecting millions of computers and computer networks together forms a huge internetwork or what we now refer to as the **Internet**.

There are many economic and technical reasons for the development of computer networks, but underlying all of these is the natural human desire to communicate and to share information. We are going to look at some of the reasons for the enormous growth in computer networks over the last few years.

As with any new development, many of the things we take for granted now such as **Email**, the Internet, **Video Conferencing**, file transfer and many other applications, were not obvious or predictable when the idea of connecting computers together first appeared. We are going to study the main applications of computer networks both within a single organisation and on a world wide level.

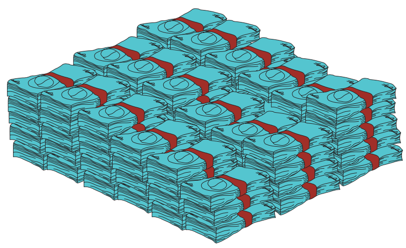
Categorising computer networks

Computer networks can be categorised by the rate at which data can be transferred between machines (Bandwidth), and the physical distance between them. Usually there is a relationship between these factors - the greater the distance, the lower the bandwidth, although of course if you have the resources, it is perfectly possible to create a high bandwidth long-distance connection or if you do not have sufficient resources, a low bandwidth short-distance connection. Using these criteria, networks tend to fall into two groups: **Local Area Network**s (LANs) and **Wide Area Network**s (WANs) and we will study the applications and benefits of both types.

There are many applications which apply to both sorts of network because they are applications which improve the way in which we communicate, and this is to our advantage whether we are communicating with the person in the room next to us or someone on the other side of the world. It could be said that computer networks have made the physical distance between us an irrelevance when we communicate with each other.

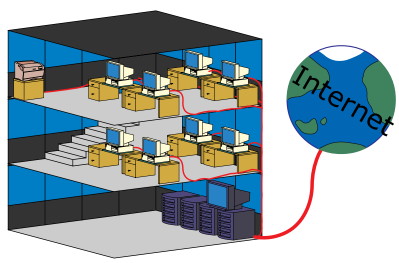
There are both benefits and pitfalls which result from the increased communication facilities which computer networks provide us with, and the new applications available have resulted in a number of social, ethical and legal issues which we need to consider if we are going to make the best use of this technology.

There are many economic factors which have led to the recent huge growth of computer networks. In general, anything which has either made a way of doing something cheaper, or has made it easier for large numbers of people to want to do it has encouraged this growth. We are going to look at the economic factors which have led to the growth of both local area networks and wide area networks.



5.2.1 LANs

Local area networks have become more and more popular over recent years, and the technology has become so cheap that that even home users with more than one computer are likely to have them connected together to form a mini-network.



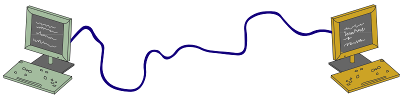
The economic factors which have encouraged the growth of local area networks are similar to those which have influenced many other technological developments. As the economic benefits of connecting computers together in a network such as the ability to share resources and share information have become more obvious, there has been an increased demand for the hardware and software to make it possible. The demand for this hardware has increased competition and manufacturers have been forced to make the technology *easier* to use and *cheaper* to purchase.

Sharing resources

A local area network usually belongs to a single organisation. This means that there is a higher degree of trust between its users than is possible over a Wide Area Network like the Internet. It also means that the network is physically located on a single site.

Both these things mean that users on a LAN can share expensive physical resources such as printers, scanners and hard disk space. It also means that they can share software resources such as an Internet connection and installed applications. Although these benefits are compelling in themselves, organisations which install local area networks also use them as a means of improving communications within that organisation.

Sharing information

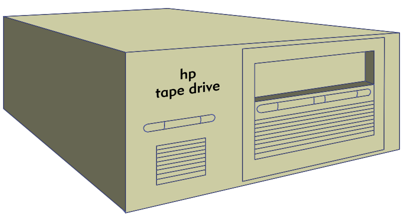


Once users are connected together using a network they can share files and data, they can use internal email and appointment management software, and they can create a private **Intranet** which allows them to distribute information electronically within the organisation.

Distributing information electronically means that as well as saving on copying costs, the information can be centrally organised and kept up to date.

Sharing services

Users on a LAN will also have access to a number of services provided by their network manager which reduce costs and improve efficiency. These services include centralised backup, technical support, virus protection, software installation and software updates. Providing these services centrally is usually cheaper and easier than expecting users of a LAN to do it themselves.



Reduced hardware costs

The use of local area networks (LANs) has also increased dramatically as a result of the reduction in cost of networking hardware such as the **Network Interface Card**, and cabling technology such as the **Hub** and the **Switch**. The availability of operating systems which include networking software has also contributed to the enormous growth in this area.

We will look at these items of LAN hardware in more detail in the section on the Characteristics of Computer Networks (Outcome 2).

LAN investigation

Spend some time investigating the services which your school or college local area network provides.

What hardware resources are shared?

What software resources are shared?

What information is available on your LAN Intranet?

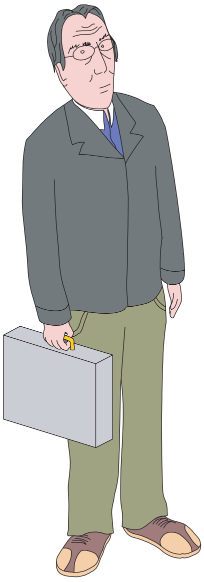
What are the economic benefits of the services you have listed?

5.2.1.1 Sharing resources on a LAN

We have already mentioned that one of the main benefits of connecting computers together in a LAN is that users can share hardware and software resources across the network. Not only does it cost less if expensive peripherals like printers, plotters and hard disks can be shared, but it makes installation and maintenance easier as well because everyone is using the same equipment.

If users can have shared access to hard disk space, then they can use this to give other users access to files, set up an **Intranet**, use internal email and **Newsgroups** (A newsgroup is a bulletin board system where users can post messages which can be seen by all other users) and all the other communications facilities which users of a network enjoy. Most networks also give their users access to some form of search facility which allows them to search for information within their Intranet.

5.2.1.2 Sharing services on a LAN



It is not only physical resources which can be shared on a LAN. There are many other benefits of connecting computers together to create a single entity. Any network made up of more than two or three machines needs someone to organise it and to control access to the resources which are being shared. This person is often called the **Network Manager**. Because all the machines on a network can be accessed from any one machine, the network manager can control software installation and upgrades centrally. The network manager will control shared Internet access and **Firewall** protection, provide automatic virus checking, do daily backups, provide technical support, etc. On very large networks there may be several Network Managers providing a number of these services.

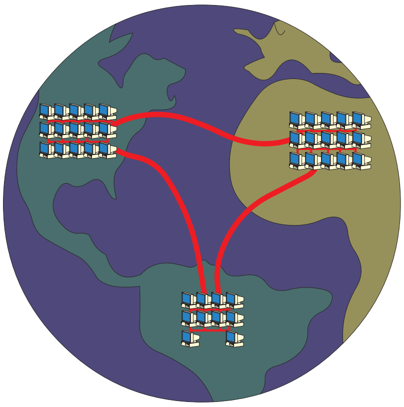
Network Manager investigation

Find out from your network manager what kind of work they have to do.

What are the routine tasks they have to perform?

What kind of event requires them to act immediately to solve a problem?

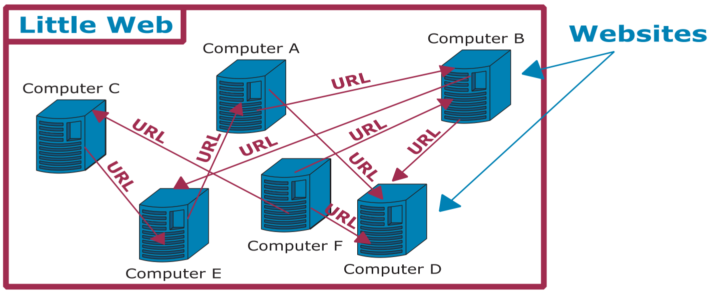
5.2.2 WANs



A WAN allows an organisation to maintain its management structure despite being geographically distributed around the world.

Wide Area Networks (WANs), and in particular the Internet, have become increasingly popular due to the decline in the cost of communication and telecommunication technologies. This has meant that as more and more people are connected, new markets have opened up for the sale of goods and services.

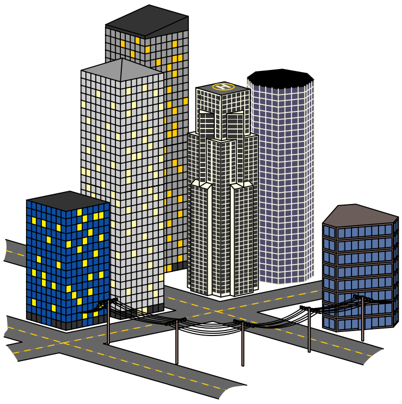
The increased availability of the technology and the improvement in communications facilities has also meant that it becomes easier and cheaper to conduct everyday business and even to live in one country and to work for an employer in another without ever actually having to physically travel to that country.



Other than sharing physical resources like printers and scanners, which does not really make much sense in the context of geographically separate computers which are connected together, (after all there is no point in sending your homework essay to a printer on another continent) most of the benefits of using a LAN also apply to using a WAN, with the added advantage of reducing the need to travel long distances.

5.2.2.1 Metropolitan area networks (MANs)

A MAN is a wide area network which has a city or metropolitan area as its geographical limit. MANs are usually under the control of a single local authority and will typically consist of a WAN which is administered centrally, providing networked services to local government offices, schools, libraries, community centres and other organisations under the control of the authority. The advantages of a MAN are the economies of scale and the improved communications which can be achieved by connecting local government offices together.



If all local authority departments are connected together using the same network, then an authority Intranet can be set up, making distribution of policy documents and access to information much easier. Employees of the local authority can use an internal email system to communicate with each other, and can benefit from access to centralised technical support data backup and virus protection. When a large organisation like a local authority is in control of purchasing hardware and software, then it is usually possible to force more advantageous prices from suppliers than would be possible if a large number of small organisations were making the purchase.

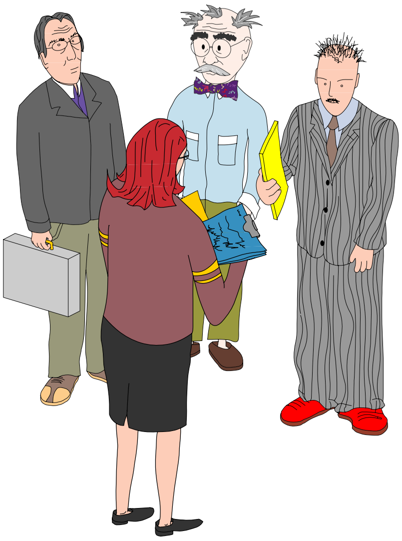
MANs depend on a high bandwidth connection between offices and organisations within the group, and so are more common in cities or densely populated areas where cable connections are available.

Identifying characteristics of local and wide area networks

On the web is an activity which requires you to identify correctly the characteristics of local and wide area networks. You should now complete this activity.

5.2.2.2 Improved communication on a WAN

Email allows users to communicate with each other by sending text and graphic messages from one machine to another on a wide area network. Users can also use email to transfer files from one machine to another by attaching these files to an email message.



Newsgroups are electronic bulletin boards for text based discussion on any subject. They allow people with similar interests to create specialised discussion groups and allow others to ask questions of the experts. Newsgroups are organised by threads so that you can watch a particular discussion develop while other discussions also continue in the same newsgroup. Anyone who can connect to a **News Server** can read or post messages on a newsgroup. Newsgroups exist on a wide variety of subjects and users can create new newsgroups if enough people agree that a need for a new topic exists.

Newsgroups can be very useful for obtaining technical information or advice on problems with software, or for discussing hobbies or politics with other people who have similar interests. To contribute to a newsgroup, a user e-mails a message to the appropriate newsgroup and the message is distributed around the world and stored on a local news server for people anywhere in the world to read.

Electronic Forums are similar to newsgroups, but are often controlled by a particular organisation and users need to join them before they can read or post messages. Forums are stored on a single server and are accessed via a web browser. One example of an electronic forum is the Scholar forum.

List-Servers are like electronic forums, but use email to distribute messages rather than users having to connect to a particular web site to view them. If you subscribe to a list server, then any message you send to the list server will be automatically emailed to everyone on that list. This enables discussion groups or electronic conferences to be set up for people to communicate on a topic who would otherwise find it difficult to meet physically.

File transfer facilities

Being able to transfer files over a WAN means that if the product you sell can be stored or transmitted electronically, then you can sell it to anyone who is connected to the Internet without having to transfer a physical object. Items on sale over the Internet which can be electronically distributed are things like software, music, photographic images, video and information in the form of data files. Some companies are now offering services such as remote backup and data storage over the Internet.

Distributed processing

Once computers are connected together into a network it becomes possible to create systems where different parts of a suite of programs can run on different machines across the network. It also allows software distributors to "rent" software rather than sell it, so that customers can be charged for the time that they use a piece of software as it is located on a remote server rather than installed on the machine they are using it on. Another interesting application of distributed processing is the Search for Extraterrestrial Intelligence (SETI) where astronomical data is distributed to users and whose computers process it whenever their machine is idle.

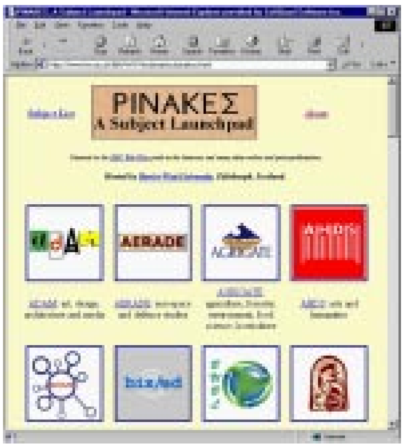
On the web you can find out more about SETI using this link: http://setiathome.ssl.berkeley.edu/

Accessing the Scholar forum

Log into the Scholar forum and read some of the messages for this and the other computing topics. Post a question of your own if you have one or answer one of the queries if you think you could help. Resist the temptation to be "funny" - humour is very difficult to communicate electronically.

5.2.2.3 Access to information on a WAN

Information services

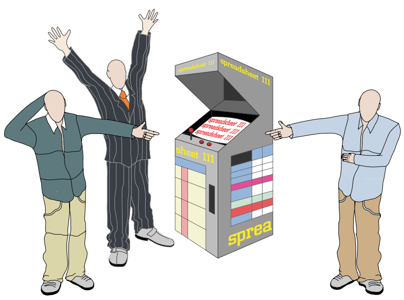


Access to the Internet means that you have instant access to news, technical information and other data at any time. Although all of this information would be available through normal channels, it is the fact that it can be accessed at any time of day or night, and does not require the user to physically travel anywhere to acquire it that makes it so useful. Just as LANs give users access to a search facility within their Intranet, users of the Internet can use a **Search Engine** to search for information on the web. With the huge amount of information available and the increasing complexity of the Internet, search engines have become an indispensable means of tracking down resources.

On the web you can try searching for information using this link:

- http://www.google.com

Entertainment services



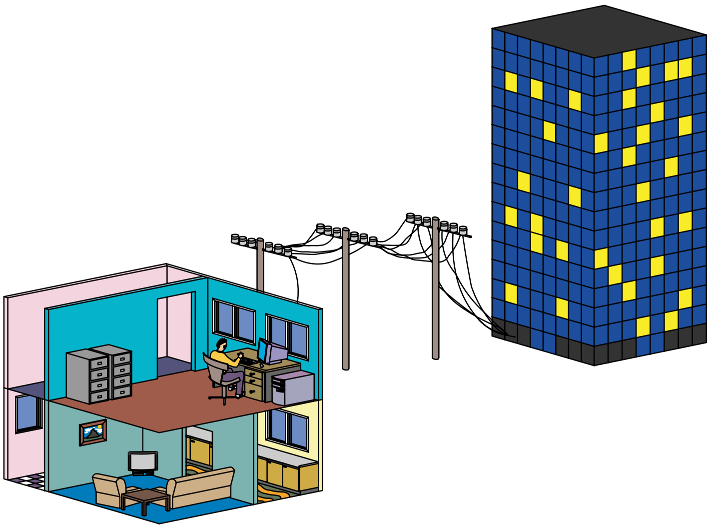
Access to a WAN like the Internet gives users access to entertainment services such as Internet radio, live video, networked games, etc. Internet radio and video are possible over a relatively low bandwidth connection because the information is being broadcast once, rather than being individually requested and transmitted separately for every user.

Networked games do not require a lot of information to be transmitted, only that the information itself is transmitted quickly to give a fast response time for the player. The game itself and the complex graphics which are part of it are running locally on the user's machine. The information which needs to be transmitted over the network is just the position of other players, and the changes in the game environment which they make. Networked games allow users to participate in tournaments or other competitions even when separated by large distances, although the speed of your Internet connection will affect your performance in the game if success depends on reaction time rather than strategy.

5.2.2.4 Tele-working

Tele-working has revolutionised the way in which companies utilise their employees.

Tele-working can mean working from home instead of travelling to an office every day, using communications technology to keep in touch with your employer. The work you do can also be transferred in this way as it will be stored in electronic form on the machine you work on at home and can be transferred electronically to your employer's network.



Tele-working can also mean working in a different country from your employer or even not ever meeting your co-workers or collaborators because they are all living in different parts of the world. Any job where the result of your labour can be transmitted electronically to your employer can become a tele-working job.

Advantages for the employee

- because you are working from home you can save on travel time and transport costs;
- as long as you produce the required results, it does not matter where or when you work. This means that you can save on child-minding or other expenses such as high housing costs;
- if you are self employed, you can deliver your work to your customers electronically.

Disadvantages for the employee

- because you are not in touch with your fellow employees, you can get a feeling of isolation, missing out on office gossip or possible promotion prospects;
- you may actually work harder and put more hours in than you would if you were

working in an office because you worry that your employers may think you are having an easy time of it at home;

- you will need space in your home to work, and you may have to insure your employer's computer equipment.

Advantages for the employer

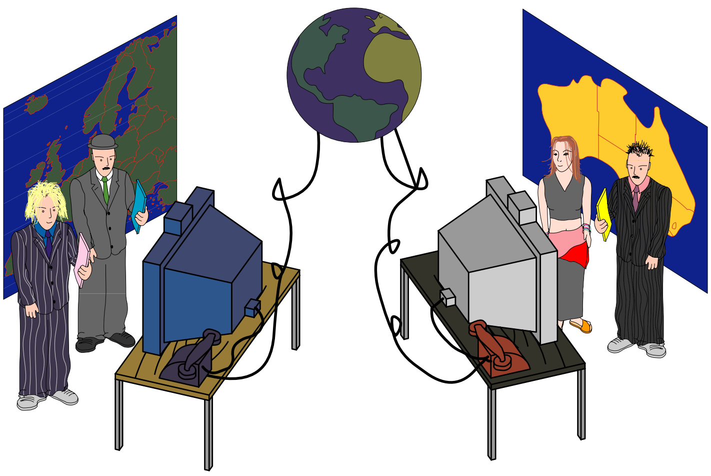
- you can save money on office space because your workers are using their own homes as their work-space;
- tele-workers often work harder than they would if they were working in an office;
- you can save money on wages as you can employ people in parts of the world where wages are low. For instance, many Western software companies are now looking to the Far East to employ programmers and data processing workers.

Disadvantages for the employer

- you will need to train your tele-workers and pay for communications costs and computer equipment;
- there may be security issues you will need to consider if you want your employees to be able to log into your company network from home.

Video Conferencing

Video Conferencing allows a number of people to communicate with each other using sound, video and to share data such as text and graphics even though they are thousands of miles apart. Video conferencing requires a high bandwidth connection and is expensive, but can be dramatically cheaper than flying several people from one continent to another.



At the moment video conferencing requires specialised equipment and a dedicated communications channel. As the technology becomes cheaper and more bandwidth becomes available, it is likely that more people will start to use it as an alternative to travelling long distances for meetings.

Identifying advantages and disadvantages of tele-working for employers

On the web is an activity which requires you to identify correctly the advantages and disadvantages of tele-working for employers. You should now complete this activity.

5.2.3 Intranets

It is often difficult to tell whether you are using a wide area network or a local area network.

The distinction between LANs and WANs is becoming increasingly blurred, particularly form the user's point of view. Many organisations now use an **Intranet** to distribute information, and as the user interface to an Intranet is usually a web browser of some sort, it is often difficult for a user to tell whether they are accessing information or services *within* an organisation or *external* to it.

Many companies are international and have branches throughout the world. They may wish their employees to share access to an Intranet which is private and internal, yet uses a WAN to connect users. It makes sense that email should be just as easy to use whether you are sending a message to the person at the next desk or someone on the other side of the world. web browsing should be a similar experience whether you are accessing private information within your organisation or public information on a server elsewhere. Obviously you need to take some account of the geographical realities - it would be nonsensical to set up a video conference with someone in the next room to you, but most of the applications which make WANs useful also make LANs useful and vice-versa.

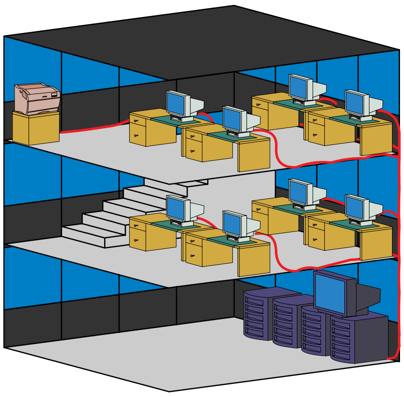
5.2.4 Mainframes and networks

5.2.4.1 Mainframe network with "Dumb" terminals

The original method of communicating using computers was to have everyone using the same computer. The organisation concerned would install a single mainframe computer and everyone who needed access to computing resources would have a "dumb" terminal with no processing power of its own which allowed them to run programs on the mainframe machine, sharing the processor of the mainframe processor with all the other users connected at the time.

Many mainframe systems are still in existence, but are often combined with conventional client-server networks. Users wishing to access the mainframe resources do so using terminal emulation software on their PCs.

5.2.4.2 Thin client networks

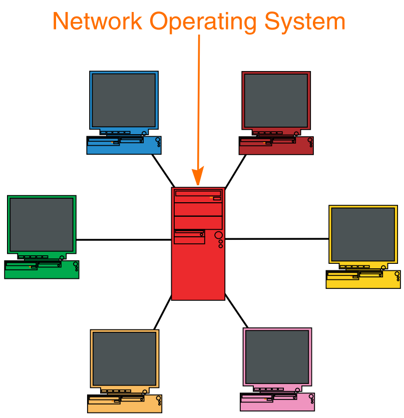


Thin Client networks are more like the old fashioned Mainframe - Terminal arrangement. Thin clients are workstations which can be relatively low specification since they run virtually all their applications on the server. The thin client loads the operating system from the server and runs a copy of any application on the server as well. This means that the server must have a very high specification and have enough memory to run multiple copies of the same application. Thin client systems are very low maintenance because all upgrades are done on the server, not the stations. They are best for situations where users are restricted to a small number of applications, or even just an Internet browser.

On the web you can find out more about thin client networks from:

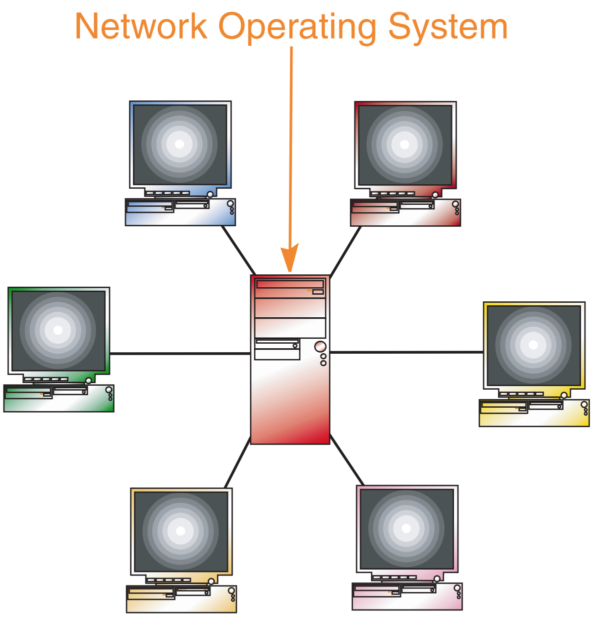
- http://www.citrix.com/demoroom/login.htm

5.3 Network operating systems



Improved networking operating systems have made it much easier to connect computers together.

In the past, network operating systems were expensive, complicated and very difficult indeed to configure. Modern operating systems have networking built in to them and make setting up a network much easier than it used to be. There are many free network utilities such as **web Server**s, **Email Clients** and other collaborative software available. In addition the adoption of common software standards such as TCP/IP for connecting networks has made connecting computers together much easier and cheaper.



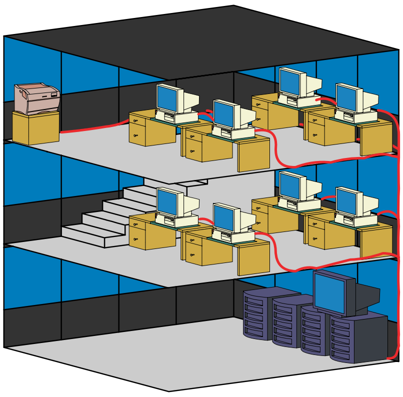
The main function of a **network operating system** is to allow users to treat resources on other computers as if they were local resources. Resources could be physical devices, software applications or files.

In more detail, functions provided include:

- mapping directories or drives of remote machines;
- copying remote files to local drives;
- managing different operating systems on the same network;
- ensuring data integrity;
- providing privacy and security of network resources;

There are two models of network operating systems. Namely, the **Client/Server** model and the **Peer-to-Peer** model.

5.3.1 Client-server



Larger local area networks use one or more **File servers** to provide resources to other machines on the network. Typically a file server will be one of the more powerful machines on the network, though as file servers tend to be replaced less frequently than normal network stations, this depends on the age of the network. Since file servers need to provide fast access to shared resources like hard disk space they are likely to be fitted with a lot of memory and high speed, high capacity hard disks.

They are also likely to be fitted with backup facilities like tape drives and **RAID** hard disk systems and may be running a different operating system from the network stations. File servers will also have fail-safe systems fitted such as extra power supplies and Un-interruptable Power Supplies (UPS) to cope with power failures. The advantages of a client-server system are that the file server can provide access to shared resources without affecting the performance of any one machine. On a large network there may be several machines acting as file servers, each one providing different services to the users. This means that the network manager can share the load between them, so that not all network services depend on one machine. File servers on a local area network may act as mail servers, **Intranet** or proxy servers, networked storage, applications servers or news servers.

Services provided by a client-server local area network

- **Security:** The network software on a client-server network will only allow those users with the appropriate password to access the files or resources that they have permissions for;
- **Networked storage:** If your files are stored on a network, then you can access them from any machine on that network;
- **Communications:** Your local area network will normally provide internal and external electronic mail, a local news server, access to an Internet connection, and the ability to "chat" live with other users of the network. It will also provide you with the facility to give users access to shared files. Your local area network can provide access to an Intranet - a shared area which can be accessed via an Internet browser, but which is private to the organisation which owns the network. An Intranet can be used to give users access to information, ensure that administrative documents are always the latest version, etc. Your Internet connection will probably be protected and filtered using a firewall. A firewall is a system which can block access to the network from outside, and can block access to certain external sites from inside the network;
- **Applications:** If applications are installed centrally on the file server of your local area network, then you do not need to worry about keeping installations up to date or about whether the application you need is installed on the machine nearest to you;
- **Support services:** Your network manager can provide services like virus checking, remote control and remote monitoring. You can rely on the network backup software to back your files up regularly.

5.3.1.1 Advantages and disadvantages of client-server networks

Advantages of client-server networks

- client-server networks allow the centralised control of software installation, control over software versions and the ability to configure stations on the network to a common format;
- resources such as hard disk space, expensive printers, modems and scanners can be shared easily among users;
- security can be easily implemented using the network operating system to control access to files and other resources;
- backup routines can be easily implemented, particularly if users files are stored on a shared resource like a server hard disk;

- most LANs are easily extended by adding cables and stations.

Disadvantages of client-server networks

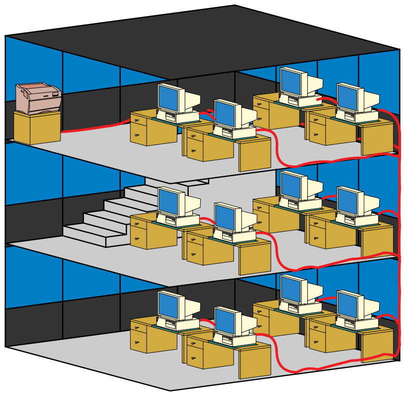
- cable faults in a bus topology can bring down a whole segment of the network;
- server failure can bring down the entire network;
- some LANs use proprietary software or communications protocols which require expensive equipment to link them to other LANs;
- large LANs require network management and control. The individuals who perform this function need to be trained and their salaries need to be paid.

Identifying advantages and disadvantages of client-server networks

On the web is a an activity which requires you to identify the advantages and disadvantages of client-server networks. You should now complete this activity.

5.3.2 Peer-to-peer

Small local area networks (LANs), usually up to 5 machines, can be easily set up as "peer to peer" networks where the expense of dedicating a machine solely to act as a network server cannot be justified. (The term peer means "equal" so all machines on the network have equal status, with no one machine having any controlling role.) Connecting machines together in this way means that machines can share resources such as printers, folders on hard disks and an Internet connection.



This kind of system works well in a home, small office or work-group, but means that security is difficult to implement, and can lead to bottlenecks if several users want to access the same resource at the same time. For instance if several people want to access files on the same hard disk on a particular machine simultaneously, this will adversely affect the performance of that machine for the person who is using it. For this reason, once the number of computers connected together goes above 5 or so, it becomes more practical to use a dedicated machine to control and provide access to the network resources.

Advantages of peer to peer networks

- peer to peer networks are cheap and easy to implement as modern operating systems like Windows 98 and Apple OSX have this sort of networking built in, and you do not have the expense of providing a dedicated machine to function as a server;
- if security is not a problem, then it is very convenient to be able to share files, access peripherals and have several people using the same Internet connection without having to set up a more complex network structure.

Disadvantages of peer to peer networks

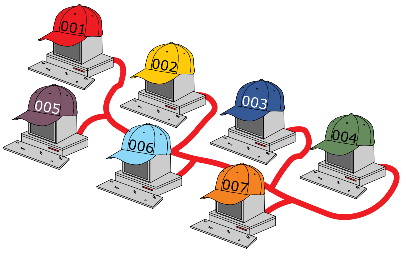
- organisation and management become a problem once several machines are connected together and sharing files. How do you keep track of where files are stored or which one is the latest version? Unless users are very disciplined, chaos can develop very quickly;
- versions of installed software are difficult to control since they will all be installed on individual machines;
- backup becomes difficult when files are distributed across a number of separate hard disks;
- proper security is very difficult to implement on a peer to peer network. Although shared resources can have a password which is required to access them, there is no easy way of implementing proper security or restricting access to files or resources which are shared in a peer to peer network;
- it is difficult to set up individual personal email accounts on a peer to peer network so most organisations using a system like this have to use a single shared email account.

5.3.3 Stations and servers

A **server** usually refers to a program and the computer it is running on. Servers provide access to **resources** on a network. On a network, one machine may provide access to a number of different resources, or there may be a number of different machines acting as servers, each one providing access to different resources.

Machine identification

All nodes on a network, whether they are servers or network stations, will have a unique identity which identifies that particular machine. The type of identity a node has will be determined by the protocols running on the network.



For instance the Ethernet standard requires that every network interface card has a Media Access Control (MAC) address in the form of a 6 byte number. The TCP/IP protocol used for access to Internet services requires that every node has its own IP address of the form of a 4 byte number. Nodes on a network may also have a name which identifies them to users on the network. The network operating system is responsible for making sure that all of these different ways of identifying a machine are mapped to the same unique entity.

Computer Name	**Room11PC**
MAC Address	00-A0-C9-AB-12-59
IP Address	192.168.0.45

Network station

Network stations and network servers are both nodes on the network. The difference between a station and a server is often one of function rather than specification as it is always easier to replace a network station than to re-configure a network server.

Network servers

A **File server** controls logins to a network and gives users access to file areas. A file server is likely to be a powerful machine with one or more high capacity hard disks.

A **Print server** allows a user to access a printer attached to it. Print servers are often small dedicated devices attached to the network wherever a printer is required.

An **Applications server** distributes applications software to stations. It is effectively performing the same function as the file server, but only allowing users read access to files stored on it.

An **Email server** stores users' email and details of their email accounts, giving them access to their mailbox when they request it. An email server is likely to regularly log into another email server on the Internet, receiving and transmitting any external mail which needs to be sent or received outside the LAN.

A **News server** stores and forwards messages posted by users on a bulletin board system. Although more common on WANs some organisations use a news server on their LAN to keep users up to date with company information.

A **web server** transmits and stores web pages. This would normally be used to provide access to an Intranet. As a web server and a web browser (client) can be running on the same machine, web servers can be used to test out a web site or Internet application without having to publish it to the Internet first.

A **Proxy server** connects a LAN to the Internet by forwarding requests for web pages or other Internet activity. A proxy server allows a number of users on a network to access the Internet through a single connection by making the request for pages for them. As far as the outside world is concerned, there is only one machine connected to the Internet. As far as the machines on the network are concerned they are all connected to the Internet, but they are connected via the proxy server.

Some or all of these functions may be performed by a single machine on a small network, though it is wise to try and spread the load between a number of different machines in order to avoid the network slowing down when a particular service is requested by a number of users simultaneously.

Un-interruptable power supplies

An Un-interruptable Power Supply (UPS) provides a backup power supply in the event of a power cut and can also smooth out voltage variations and power surges. Although is useful for any computer system, it is particularly important for network servers as data can be lost and open files corrupted if a network server suffers a power failure.

Matching servers to their correct descriptions

On the web is an activity which requires you to match correctly the following servers and descriptions. You should now complete this activity.

5.4 Network topology

As the number of computers in the workplace grew, new ways of utilising their potential developed. Smaller, more powerful computers could be linked together in a network to share storage space, information and applications software.

People could now work remotely, logging in to their workplace computers to transfer information. Companies established *intranetworks*, allowing company-wide distribution of information and data.

Computers could be connected to form a privately-owned, **Local Area Network** (LAN), contained within a building, or a University campus and extending to a maximum of a few kilometers in size. **Metropolitan Area Networks** (MANs), interconnecting buildings or other facilities and **Wide Area Networks** (WANs) such as the Internet, could link machines into a network on a countrywide or worldwide basis.

In this unit, we will concentrate on how computers are arranged to form a Local Area Network. Local area networks are characterised by their *size*, *transmission technology* and *topology*. A **network topology** describes the arrangement of computers to form a network. The actual physical layout is determined by the buildings or other locations that house the parts of the network.

We will look at four local area network topologies: **Bus, Star, Ring**and **mesh**.

Network topologies

On the web is an animation that shows you three network topologies: Ring, Star and Bus. You should now look at this animation.

5.4.1 Bus

A **bus** network has a single communications cable, running the length of the network onto which nodes are connected. One or more of the nodes on the network may act as **file servers**.

A typical bus system is **Ethernet,** a network setup with coaxial cable that can transmit at a rate of 10 or 100 megabits per second (Mbs).

A bus topology is shown in Figure 5.1.

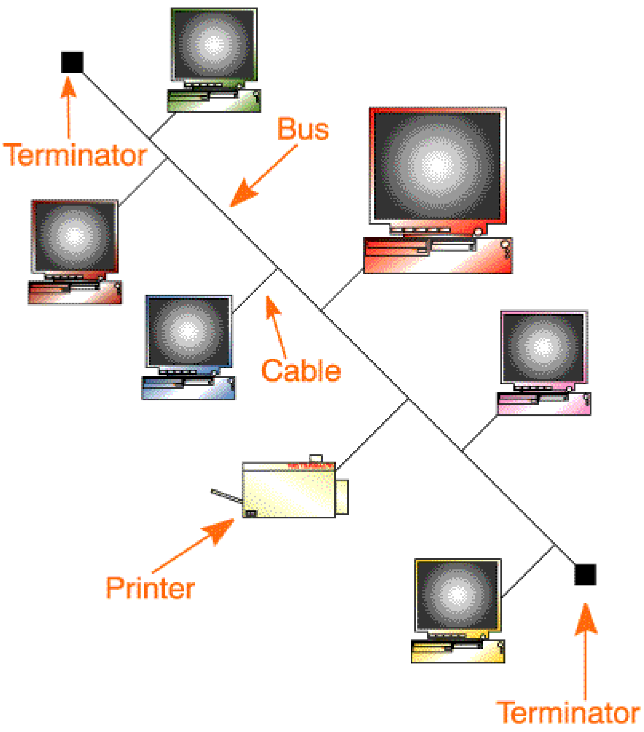


Figure 5.1: Bus Network Topology

Devices may be connected to the bus cable using a spur, or the connections may pass through the network interface in each device. In the latter, there must be some provision made for the network to continue to work if one of the nodes is switched off, or fails to operate correctly.

A bus network implies a very high speed circuit and a limited distance between the computers, such as all nodes being in one building. Distances can be extended using repeaters. This can be achieved by replacing one of the terminators with a repeater and running a length of cable from the repeater to another terminator.

Data Security

Data can be disguised using a method called **data encryption** to encode the data.

Bandwidth

A bus network can have slow response times when many computers are simultaneously accessing the communications link. This is because the available bandwidth is being shared. However **data compression** techniques can be used to compact data to produce more acceptable response times.

Reliability

The loss of a single node on a bus does not hinder the operations of the network unless the cable itself is cut.

Cost

This is a popular topology because hundreds of computers can be connected to a single bus fairly cheaply.

Identifying the advantages of a bus network

On the web is an activity that asks you to identify the advantages of a Bus Network. You should now carry out this activity.

Identifying the characteristics of a bus network

On the web is an activity that asks you to identify the characteristics of a bus network. You should now carry out this activity.

5.4.2 Star

A **star network**, sometimes called a hub topology, connects all nodes to one central node that routes traffic to the appropriate place. Each computer is linked by a separate circuit through a central connection point or controller. A high speed computer generally acts as the central controller which allows for very fast communications with all computers on the network. A star network arrangement is shown in Figure 5.2

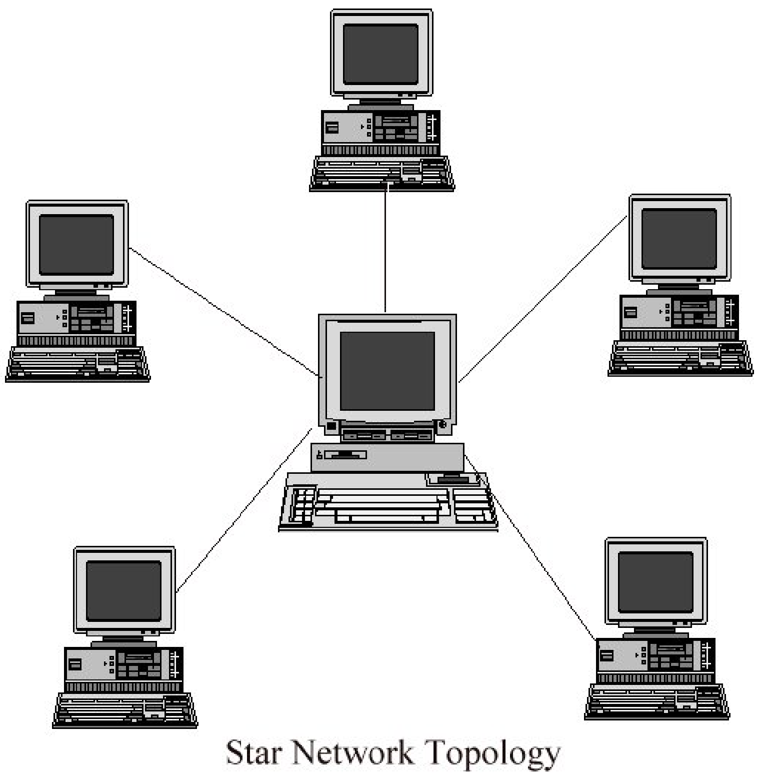


Figure 5.2: Star Network Topology

Data Security

A star network provides higher security in data transmission as data is routed only to the computer that is to receive the data and so there are no data collision problems.

Reliability

If a link fails then only the node connected to that cable will be affected and the system will still be operational. Failure to the central controller will cause the entire network to fail. This network also performs well under heavy usage i.e. the system does not slow down when there are lots of users.

Cost

This can be an expensive option due to the amount of cabling required if there are many computers to connect to the network.

Identifying the advantages of a star network

On the web is an activity that asks you to identify the advantages of a Star Network. You should now carry out this activity.

Identifying the characteristics of a star network

On the web is an activity that asks you to identify the characteristics of a Star Network. You should now carry out this activity.

5.4.3 Ring

The **ring network** shares many of the characteristics of a bus network. For instance, data sent around the network can potentially be accessed by any device hooked onto it. This topology therefore has the same security problems as a bus network. Available bandwidth is shared amongst the computers on the network. This arrangement is shown in Figure 5.3

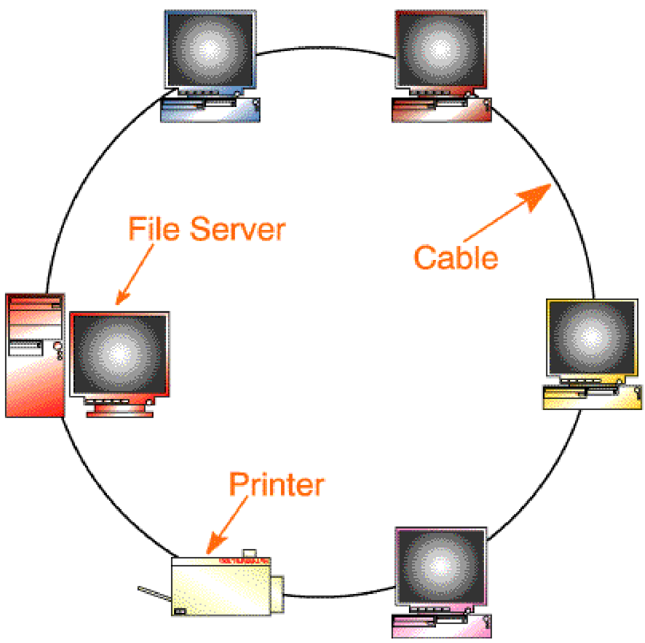


Figure 5.3: Ring Network Topology

As in the bus network, devices may be connected to the ring cable using a spur, or the connections may pass through the network interface in each device. In the latter, there must be some provision made for the network to continue to work if one of the devices is switched off, or fails to operate correctly.

One method of avoiding data packets colliding on the network is to use a **token**. In a token ring network the token circulates around the ring, picking up packets from a source computer, carrying and dropping them off at their destination. A computer may grab hold of this token and thus gain control to send data.

Identifying the advantages of a ring network

On the web is an activity that asks you to identify the advantages of a ring network. You should now carry out this activity.

Identifying the characteristics of a ring network

On the web is an activity that asks you to identify the characteristics of a ring network. You should now carry out this activity.

5.4.4 Mesh

A **mesh** topology is where every computer is connected to at least two other computers. The original Internet as envisaged by the US Department of Defence was devised as a *mesh*, the idea being that in the event of a nuclear strike, there would still be some connections intact and therefore the network would still be able to function. In reality the topology of the Internet is a partial *mesh*, with some sections connected together in a *tree* topology.

How a mesh topology tolerates damage

On the web is an animation that illustrates the way a mesh topology can tolerate damage and still function. You should now look at this animation.

5.5 Summary

The following summary points are related to the learning objectives in the topic introduction:

- Scale of network - peer-to-peer, LAN, WAN & Internet;
- Comparisons in terms of transmission media, bandwidth, geography and functions;
- Differences between peer-to-peer and client server operating systems;
- Functions of various servers on a network;
- Network topologies - bus, ring, star and mesh;
- Differences between mainframe systems and networks.

5.6 End of topic test

An online assessment is provided to help you review this topic.

Topic 6

Using Networks

Contents

Prerequisite knowledge

Before studying this topic you should be able to:

- *Describe the features and functions of the World Wide Web;*
- *Describe the economic factors which have led to the development of networks;*

- *Describe the main legal requirements of network related laws.*

Learning Objectives

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

- *Describe the functions and use of network hardware including hubs, switches, routers and NICs;*
- *Describe network bandwidth and wireless communications;*
- *Describe developments in hardware and software used in networks;*
- *Describe the misuse and illegal use of networks.*

Revision

The following exercise tests the prerequisites for this topic. Ensure that you are happy with your responses before progressing.

Q1: Buying and selling products and services over the Internet is called:

a) e-mail
b) HOBS
c) e-commerce
d) EFTPOS

Q2: Working from home with work being transmitted electronically is called:

a) commuting
b) firmware
c) point-of-sale
d) teleworking

Q3: Illegal downloading of commercial music from the Internet as MP3 files may be contravening the:

a) Copyright, Design and Patents Act
b) Computer Misuse Act
c) Data Protection Act
d) Health and Safety at Work Act

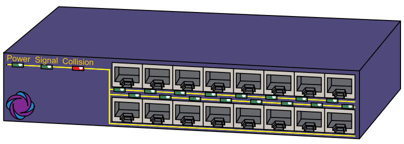
6.1 Introduction

This unit considers the hardware and software required to use networks and the associated communications issues. The misuse of networks is also considered and the implications discussed.

6.2 Network hardware

6.2.1 Hub

A **hub** in an Ethernet network is a multi-port *repeater*. *Hubs* are most often used on networks using UTP cabling as the limit for this type of cable is 100 metres.



A *hub* will also effectively divide the network up into a series of different segments, thus reducing the likelihood that a cable fault will bring the whole network down. *Hubs* are commonly supplied with 12 or 24 ports, and are suitable for distributing a UTP cable to a room or office full of network stations.

A *hub* also functions at the lowest level of the OSI networking model, the **Physical Layer**.

6.2.2 Switch

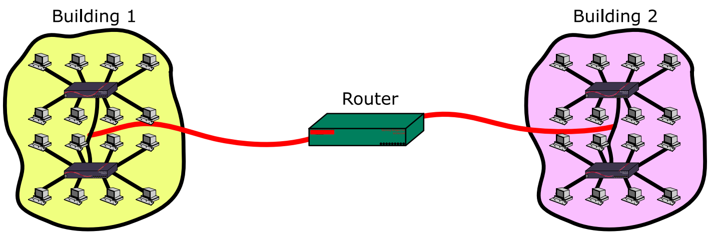
A **switch** (sometimes called a switched hub) divides the network up into *collision domains*. A **Collision** occurs when a station begins transmission and then receives the beginning of a frame from another station. The station will immediately stop transmission and issue a JAM signal onto the segment. This will indicate to the other transmitting station that a collision has occurred and both stations will back off for a random amount of time and try to re-transmit. This back-off time is dependent on the number of consecutive collisions that were issued before a successful transmission. The more collisions, the longer the maximum back-off time. This mechanism requires that stations be close enough together for each station to see any possible attempted transmission before the first 64 bytes of its frame have been transmitted. This is because 64 bytes is the minimum frame size for an Ethernet network.

The difference between a hub and a switch

On the web is an animation that shows the difference between a hub and a switch. You should now look at this animation.

6.2.3 Router

A **router** works much like a bridge, but works on inter-networks - typically a *router* uses an IP address to determine where data is to be delivered rather than the Ethernet MAC address.



This is why routers are used to link Ethernet networks to the Internet via fibre or ISDN line. Routers can often be programmed to restrict access to certain IP addresses in either direction between the networks it connects. *Routers* are usually dedicated hardware devices which are programmed remotely over the network, though their functions can also be emulated in software on a normal network station or node fitted with two Network Interface Cards. This arrangement is sometimes called a Gateway.

6.2.4 Network interface card (NIC)

A **network interface card** (NIC) is a circuit board or card that is installed in a computer so that it can be connected to a network. Personal computers and workstations on a local area network (LAN) will have installed a network interface card specifically designed for that LAN e.g Ethernet.

Network interface cards provide a dedicated, full-time connection to a network.

6.3 Hardware and software factors

6.3.1 Processors

The continuing bandwidth explosion of the past few years has had a dramatic impact on what can be expected of networks. Coupled with the falling costs more of us are increasingly taking full advantage of the connectivity networks (and particularly the Internet) provides.

The increase in bandwidth available has been paralleled with the development of bandwidth-hungry and computationally intensive applications. For example streaming audio and video, Peer-to-Peer (P2P) applications, Virtual Private Networks (VPNs) are now demanded.

Networks are demanding equipment with very high throughput and need the flexibility to support these applications. For networks to effectively handle these new applications, they need to support new network management functions such as protocols that differentiated services and security functions. In addition, network equipment associated with these application also have to be supported.

Features to consider in network processors include:

- Performance - the ability to execute key computational kernels and perform many applications at high speed;
- Flexibility - the ability to adapt to changing standards in operating systems and applications software;
- Power - power consumption is an important factor when considering where processors are situated e.g. in energy-sensitive areas and the cost implications of housing and packaging;
- Software support - consideration must be given to what compilers, operating systems and libraries are available with, and supported by the processor.

6.3.2 Memory

Factors defining the memory that supports a network processor can be categorised as follows:

- shared, distributed or a combination of memory across the network;
- size and type memory used;
- caches that are included on or off the chip.

In addition there are many special features relating to network memory. These include: memory / buffer controllers, packet management units and address generation units to name a few.

A method of providing clients with a large amount of RAM on a network can be achieved by using network memory server (NMS). This is a device that provides of RAM via a fast network memory paging service, in the same the way that a network file server will provides clients with access to a large amount of disk storage.

The speed of data transfer via a NMS can be much faster than paging to and from local disks and can be in excess of 1Gb/sec.

6.3.3 Backing storage

The speed of CPU and semiconductor memory has increased rapidly over the years. However, hard disks have been limited by the physical constraints of their moving parts. This has led to paging overheads when considering the performance of stand alone system and more so on networks. In the old days of mainframes, data was stored separate from actual processing units. When PC based servers came along, storage devices became integral parts of the system. Neither of these approaches were feasible as network systems developed. The need to store and access increasing volumes of data over networks required a different solution.

Network storage is simply about storing data so that it can be accessed by network clients. The simplest form of storage on a network is the hard drive of the the network server and is very common on small networks. For larger networks, the storage requirements are larger and more complex to implement. Devices can connected directly to the network media. These devices are assigned a network address and can then be accessed by clients via a server that acts as a gateway to the data. For a very large network, such devices can themselves, form a storage area network and be accessed via a server.

6.3.4 Browsers

A browser is the application program that enables a computer user, to look at, and interact with World Wide web. It makes requests to web servers throughout the Internet. Netscape Navigator and Microsoft Internet Explorer are the only two browsers that the vast majority of Internet users are likely to use.

The notion of a browserless web describes communication over the World Wide web between programs rather than between people using browsers and web servers and does not involve an interactive user.

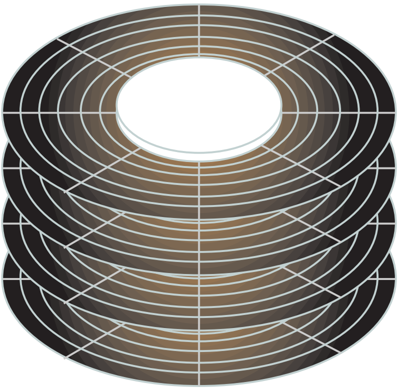
6.3.5 Network operating systems

Network operating systems will usually run on the server or servers installed in a LAN. Common operating systems are Novell, **Windows NT**, **Windows 2000 Server** and various flavours of **UNIX** including the open source operating system **LINUX**

Workstation operating systems need not be the same as the operating system running on the server. Typically an NT or LINUX server will have client machines running NT Workstation, Windows 98 or Mac OS connected to it. There may be several client operating systems on a network. The version of network operating system running on the server should not make any obvious difference to the user.

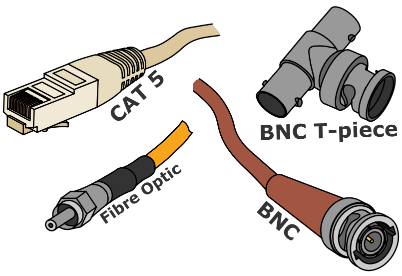
Most operating systems use a command system which treats all devices connected to the machine in a similar way. A workstation on a network needs to be able to treat devices on the network such as disk drives or printers in exactly the same way as if they were directly connected to that machine. Users should not need a different command to print to a computer connected to their machine than to print to a printer on the network. Saving a file should be done the same way whether you are saving the file on a local disk or saving it on a network drive This kind of **transparency** is achieved by using a network driver called a **Redirector**. A redirector takes a request for access to a network resource and sends that request to the appropriate protocol running on the network.

6.4 Technical factors affecting communications



The technical factors which have led to the growth of computer networks have emerged in parallel with the economic factors which have driven the research into networking technology. As the economic demand for networking technology has grown, the trend has been for equipment prices to fall and performance to increase. Although still in its infancy, the development of wireless networking is likely to follow the same pattern.

6.4.1 Cabling



Advances in cabling technology for local area networks has resulted in the equipment becoming easier to install, cheaper to purchase and has provided ever increasing **bandwidth**. As usual, economies of scale have meant that as more people purchase the equipment, it becomes more economical to manufacture it in bulk and as a result it becomes cheaper.

The ensuing competition between manufacturers has encouraged research into the technology which has produced faster and more efficient networking infrastructures.

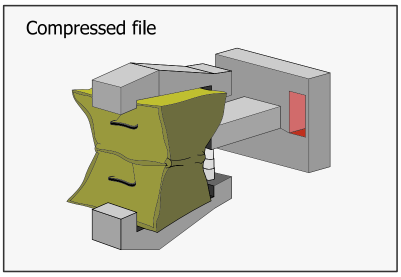
Most LANs currently in use are based on the **Ethernet** technology developed by the Xerox corporation, enabling computers to be connected together using **UTP** cabling running at 100Mb/sec. This kind of bandwidth allows users to run multimedia applications over their network, and to share high speed Internet access. We will look at network cabling technology in more detail in The Characteristics of Computer Networks (Outcome 2).

6.4.2 Bandwidth

The technical factors affecting the development of wide area networks have been similar to those affecting local area networks, particularly the expectations of users to have the same facilities over a WAN as they were used to over a LAN.

The most obvious technical advance has been the increased **Bandwidth** available to users of networks. The term bandwidth describes the rate at which data can be transmitted over a segment of a network. Unfortunately the demand for bandwidth, particularly where WANs are concerned has always outstripped the ability of networks to provide it. Many service providers charge their users by the amount of data they receive rather than by the time they are connected. This makes sense if the commodity in short supply is bandwidth rather than an Internet connection. As users share networks, the bandwidth available to any individual depends on the number of people they are sharing that part of the network with as well as the rate at which data can be transmitted. From an economic point of view there are two possibilities - increase the bandwidth available or reduce the demand.

Reducing the demand



Data Compression - if data is compressed before it is transmitted then it requires less bandwidth. There are a variety of methods for compressing data - most rely on removing redundant information by detecting repeating patterns in the data transmitted and then they only need to transmit the pattern once together with details of how often it is repeated.

Caching - if data can be stored locally, then the data need only be re-transmitted over a short distance when users request it. This system is used by your Internet browser to reduce download speeds by storing files from Internet pages you have viewed on your hard disk. When you ask to see the page again it reloads them from the hard disk rather than downloading them again. This technique is also used by Internet Service Providers (ISPs) to reduce network traffic.

Broadcasting - if data such as video or audio is transmitted live, then although the user only has one chance of viewing it, the bandwidth requirements will be a lot less than if users are continually downloading their own copy of the data.

Increasing the available capacity

Improved communications technology has increased user expectations of what can be transmitted over wide area networks.

Just as building a motorway between two points does not always reduce congestion, networks have to be carefully planned to make sure that bandwidth is used efficiently. Improved communications technology including fibre optic cable and satellite links and have made the Internet a much faster service to use. From the individual point of view, operating systems with built in networking software have made connecting to the Internet much easier. Faster modems and services like **ISDN** and **ADSL** have made data transmission much faster so that it has become feasible to receive multimedia services like graphics, sound and video over an Internet link

On the web you can find out more about ADSL from:

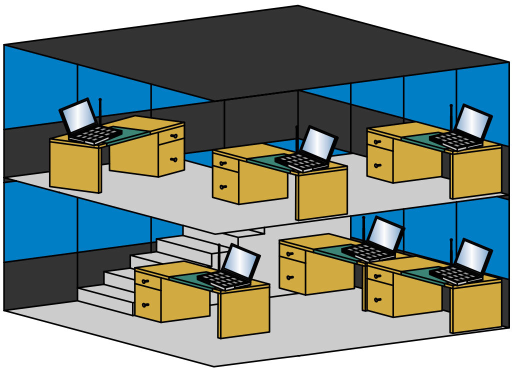
- http://www.adsl.com/adsl_tutorial.html

On the web you can find out more about ISDN from:

- http://www.alumni.caltech.edu/~dank/isdn/isdn_ai.html
- http://www.ralphb.net/ISDN/index.html

6.4.3 Wireless

Wireless networking in LANs is still slower than conventional cabling, but the advantages of being able to move around a building and still be connected to the network often makes wireless networking worthwhile.

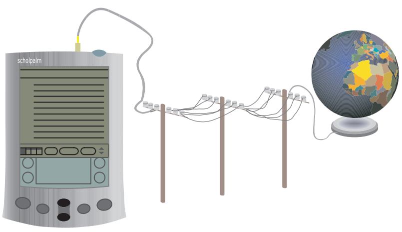


Wireless networking can be advantageous for organisations who need to create a network quickly within a building without suffering the disruption of installing cabling, or who wish their users to be able to move around while using laptop computers. A wireless network requires each computer to be fitted with a wireless network card and to be within range of a wireless hub. The available bandwidth is often substantially less than is available using a cabled network, but is still adequate for many applications.

If you are interested in wireless networking technology you may wish to study the topic in more detail, though it is not necessary for this course. Still in its infancy, a new wireless technology known as **Bluetooth** allows devices like printers, mobile phones and laptop computers to form ad-hoc mini-networks when they come close enough to each other.

On the web you can find out more about Bluetooth from:

- http://www.palowireless.com/infotooth/whatis.asp
- http://www.nokia.com/bluetooth/



Mobile communications technology is likely to be very important in the development of the Internet in the next few years. There is considerable interest in the capacity for mobile phones to access the Internet via a software standard called **Wireless Application Protocol** (WAP).

These services are currently very slow and have not yet been taken seriously by many providers. As the available bandwidth increases however the situation is likely to improve.

On the web you can find out more about WAP from:

- http://www.nokia.com/wap/

6.5 Misuse of networks

6.5.1 Social issues

Personal privacy

The monitoring of individuals' Internet and email use is a contentious issue. Many governments would like to have access to this sort of information. Many individuals would not like them to have this kind of access. With modern communications networks it is theoretically possible to trace the movements of any individual around the world - they use cash machines, pay with credit cards, use the Internet, use mobile phones, appear on surveillance cameras, and may use electronic road tolling systems.

All of these systems could be combined to track an individual's movements. Even organisations who manage LANs are becoming concerned about the security and employment issues surrounding email, and are starting to monitor all email activity on their networks to make sure that their employees are not divulging company secrets or using the network facilities for their own personal use.

Encryption

Encrypting data is a method of coding it, in order to make it difficult or even impossible for someone to read it unless they have authorisation from you. Some governments make encrypting data illegal, others regard it as an individual right. As with all security systems, there is a trade-off between security and convenience. The more secure you make a communication system, the more inconvenient it is to use.

Encryption is classified according to the number of bits needed for the key used to encode the data. The more bits that are used, the longer it takes for the code to be cracked. As computers increase in power, the time and resources needed to crack such codes decreases, so the minimum encryption key size needs to increase as computer technology improves.

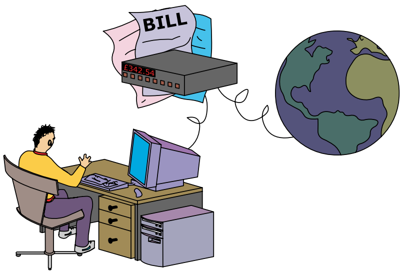
At the moment email on most networks is un-encrypted and insecure, as messages are sent as plain ASCII text. One popular and free encryption system currently available is Pretty Good Privacy (PGP), although until recently PGP was regarded as "munitions" by the USA government and its inventor was prosecuted for allowing details of how his system worked to be posted on the Internet.

On the web you can find out more about PGP from:

- http://www.gildea.com/pgp/

Disconnection

Many people feel that the Internet has accentuated the difference between the rich and the poor. People in western countries are described as "Information rich" whereas most people in the Third World are "Information poor", since connecting to to the Internet requires expensive equipment and access to a networking infrastructure which does not exist in many poor countries.



Social isolation

There is a fear that the increased use of electronic communication will mean that many people will become physically more isolated from each other. The Internet encourages the creation of global communities, but may result in neighbours not seeing each other for weeks on end.

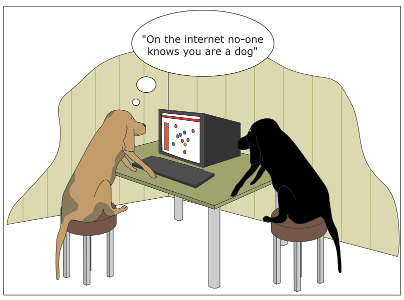
6.5.2 Ethical issues

Netiquette

There are several conventions and a whole new vocabulary which has developed with regard to how you should behave on Internet newsgroups, using email, etc. Simple rules such as not sending "**spam**" (unsolicited commercial email), not shouting (writing in capital letters) and respecting people's privacy may seem obvious, but it is easy to make mistakes without realising it. The best policy when subscribing to a newsgroup or joining a list server is to "listen" quietly for a while to see how others behave before you submit any contributions yourself.

Misrepresentation

The ability to post messages on newsgroups or discussion lists anonymously means that offensive statements can be made without the person making them being held to account.



Surveillance and monitoring

The email you send and receive, the pages you view and the files you download via the Internet can all be logged by a number of organisations. These organisations may not always have your best interests at heart.

Chatrooms

Chatrooms are popular with young people who use the Internet as they enable users to communicate in real time with others from all over the world. A chatroom on the Internet is an area where users can type messages which can be read immediately by anyone else connected to the same server. Communication is slow because of the need to type the messages, but this disadvantage is outweighed by the immediacy of the system. One problem with chatrooms is that there is no way of verifying that the person you are communicating with is who they say they are. They may have lied about their age, their sex, or why they are interested in chatting with you.

6.5.3 Network etiquette

While there are no international standards of behaviour on networks, there is an accepted code of conduct often referred to as netiquette. Netiquette is usually just common sense. Network administrators are the only users of a network who have access to all the files on a network - it is considered unprofessional for a network administrator to access any personal file unless they are required to do so to perform their job.

Some network rules of etiquette are:

- do not use other peoples identities and passwords;
- do not attempt to access files which are not yours;
- avoid wasting disk storage space or bandwidth unnecessarily. (Keeping your entire music collection in MP3 format on your network drive or downloading large video files at peak times of Internet use are activities unlikely to endear you to your network administrator);
- avoid causing other people unnecessary work. It may be entertaining to install the latest screen-saver on your network station, but the next user is unlikely to appreciate having to re-boot the machine because they do not know the password, and the network administrator is unlikely to enjoy un-installing it.

Just as the resources on a LAN are limited, a WAN such as the Internet also has limited resources, specifically bandwidth. Some sensible precautions:

- Avoid wasting bandwidth by sending unnecessary emails, attaching uncompressed files or creating web pages with large or unnecessary graphics on them;
- Be aware that the Internet is used by people with a large variety of different machines running different operating systems and different browsers. If you want other users to access information, try to store the files in a format which is going to be easy to read for everyone. ASCII code is the simplest format for text. CSV is

the simplest format for a database;

- Try to create web pages which are readable no matter what resolution your monitor is set to use.

Email etiquette

- email messages should be concise and to the point;
- remember to make the subject line relevant to the message you are sending. Keeping to one subject per email makes it easier for the recipient to file the message if they want to keep it and makes it easier for them to see what the message is about without reading it all first;
- do not repeat email messages. Give the recipient time to read your email and to respond to it;
- be professional and careful what you say about others. Email is easily forwarded and often archived or stored, so whatever you say may come back to haunt you;
- email is not a secure means of communication as it is transmitted without encryption. If you are concerned about the security of your email use a system like Pretty Good Privacy (PGP) to encrypt your messages;
- when being humorous, use emoticons to express humor. (Tilt your head to the left to see the emoticon smile) :-) means happy face;
- typing mail messages all in upper case is considered SHOUTING! and rude;
- do not send people chain letters or hoax virus warnings. Such email just reduces the bandwidth available for everyone;
- avoid sending anyone **spam**;
- do not attach large files to an email message without asking the recipient's permission first. They may be using a slow modem connection and will resent paying for an hour or so of connection charges just to view that amusing video you made of yourself!

Usenet etiquette

All of the email etiquette rules apply to Usenet. In addition, since posting to a newsgroup is in effect sending a message to a large group of people who you do not know, remember to be courteous and respect others:

- when quoting someone else, remove what is not directly applicable to your reply. Do not automatically quote the entire body of messages you are replying to when it is not necessary. Leave only the minimum necessary to provide context for your reply. Nobody likes reading a long message in quotes for the third or fourth time.
- it is extremely rude to forward personal email to mailing lists or a newsgroup without the original author's permission.

- only messages meant to be read by the entire group should go to the list. Send a personal mail message aimed at one person to that person.
- posting an advertisement in news groups, unless it is specially chartered for that purpose like the for-sale newsgroup, or sending unsolicited advertisements with email is considered rude.

Listserver etiquette

What to type	Meaning	What will be displayed
:D	big grin	
:cool:	cool	
:eek:	eek!	
:o	embarrased	
:(	frown	
:mad:	mad	
:rolleyes:	sarcasm	
:)	smile	
:p	stick out tongue	
;)	wink	
:confused:	confused	

All of the Usenet etiquette rules apply to listservers, though as your messages will arrive as emails to the members of the list, it is best to be even more careful than when posting to a newsgroup. Some list servers generate large amounts of email every day. It is often wiser to receive the messages in digest form as one single email containing all the messages posted that day. Here are a few guidelines:

- when signing up for a group, save your subscription confirmation email for reference, so that if you go on holiday or no longer wish to receive the list messages, you will have the subscription address for suspending mail;
- if you lose interest in a list-server, un-subscribe from it so that the messages being sent to you do not clog up the network. Use the instructions in the subscription confirmation to unsubscribe. Sending a subscription or un-subscription notice directly to the list instead of to the listserver is annoying to others.

Some Netiquette Vocabulary

Acronym: Code used in emails and Usenet postings made up of the first letter of each word in a commonly used phrase. Acronyms are quicker to type than the phrase itself. Examples are FWIW (for what it's worth) IMHO (in my humble opinion) ROTFL (rolling on the floor laughing) FYI (For your information) FAQ (frequently asked questions).

Archive: An indexed database of previous postings to a list or newsgroup. It is best to consult the archive for a list if possible before asking questions which may have been asked (and answered) several times before.

Cross-Posting: To send the same message to more than one newsgroup.

Emoticon: A symbol created from ASCII symbols used to communicate emotions or body language equivalents. Examples of emoticons are: :-) Smile ;-) Wink :-/ Perplexed

FAQ: A list of frequently asked questions and their answers provided to help users understand how to use a system. Often netiquette rules for the list or newsgroup are included in the FAQ .

Flame: An aggressive or emotional outburst criticising another member of the list or another contributor. Flames are always best avoided.

Lurker: Someone who joins the group and reads all the messages but does not contribute to it. It is advisable to lurk when first joining a list.

Moderator: A person who controls or administers a newsgroup or list server. Not all lists or newsgroups are moderated.

Newbie: A newcomer on a list or newsgroup.

Post: To send a message to a newsgroup or list server.

Spam: Unsolicited email containing advertising material. Usually sent out anonymously. Spam is almost universally loathed by everyone except those who send it out.

Thread: A theme for discussion on a newsgroup or list server.

Troll: Someone who posts controversial or unacceptable opinions in the hope of starting an argument.

On the web you can find out more about netiquette from:

- http://pixel.cs.vt.edu/class1/spinners/InternetSpeak/nettiquette.html

- http://www.rangenet.com/mainhtml/nettiquette.html

Finding out about netiquette

Use the web to find out as much as you can about netiquette.

Make up your own list of "10 Commandments" for appropriate behaviour on the Internet.

6.5.4 Hacking

Controlling access to the network

The network operating system is responsible for security on the network. The most obvious example of this is when a user logs on. The user must supply an identity and a password. The operating system compares the data entered with the identities and passwords in its database and if the two do not match up then it will not allow that user any access to the resources on the network.

If the identity and password do match, then the resources which the user has access to will depend on the level of access that user has been given by the network manager. The access a user has to resources depends on that user's level of **permissions**.

Controlling access to network resources

The easiest way to control different levels of access is to divide users into groups. Each group can have different levels of access to the shared network resources. For instance one group may need access to a particular printer or need to be able to alter files in a particular section of the organisation's Intranet. Once groups have been set up with different permission levels, it is easier to give one or more users access to a resource by making them a member of a group than to individually change the permissions for each user. When a user tries to access a resource, whether it is a printer, a file or an area of disk storage, the operating system checks against the permissions for that resource. If the user or the group they belong to does not have permissions for that resource then their access is denied. The access to files and disk storage given to users may vary according to how these resources are to be used. For instance, users will probably be given **full** permissions to their own file area, but only **read** access to parts of the Intranet, **read/execute** access to applications areas, and no access at all to network management areas or other user's files.

Security on the Internet is of particular concern to anyone purchasing goods or services using e-commerce. Many people are concerned that their credit card details are being transmitted over the Internet when they use e-commerce systems. Anyone using an on-line bank account also needs to be reassured that the system they are using to access their financial details is secure. A secure web browsing protocol named Secure Hyper Text Transmission Protocol (SHTTP) has been developed for this kind of transaction. **SHTTP** ensures that data transmitted using this protocol encrypts the data before it is transmitted using SSL (Secure Sockets Layer) originally developed by Netscape.

Other security systems on the Internet depend on the issuing of digital certificates which guarantee that you are who you say you are.

On the web you can find out more about digital certificates on the Internet from:

- http://www.verisign.com/

6.5.5 Viruses

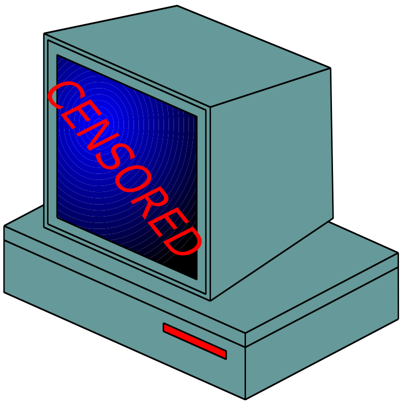
A virus is a piece of programming code that causes some unexpected and usually undesirable event in a computer system. They are often designed so that they automatically spread to other computer users on a network. Viruses can be transmitted as attachments to an e-mail, as a download, or be present on a disk being used for something else. Some viruses take effect as soon as their code takes residence in a system whilst others lie dormant until something triggers their code to be executed by the computer. Viruses can be extremely harmful and may erase data or require the reformatting of a hard disk once they have been removed.

Email and viruses

On the web is an animation which illustrates how an email attachment that is infected with a virus can be spread to other machines. You should now look at this animation.

6.5.6 Legal protection

Censorship and pornography



What is legal in one country may not be legal in another. Not all governments encourage the free debate of political subjects and many governments try to block access to opinions or political debate which is critical of them. Many people believe that children should be protected from violent, pornographic or extreme political material, though of course what is considered to be extreme in one society is not always considered to be extreme in every society.

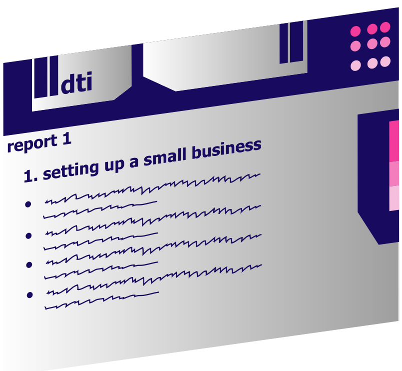
From the point of view of those wishing to impose censorship, the world-wide nature of the Internet makes it very difficult if not impossible to control. An international standard of "rating" web pages for violent or sexual content has been developed which makes controlling access easier.

Many educational institutions and some Internet Service Providers install filtering software on their **Proxy Server** which can be set to deny access to web pages or news groups which contain offensive, violent or pornographic material. Commercial organisations may deny their users access to entertainment or other services because they believe that their employees should only access material which is relevant to their work. Inevitably what should or should not be banned is a matter of considerable debate, and is difficult to get right. Too strict a policy on banning access to parts of the Internet can deny access to legitimate material. Too lax a policy can allow access to unsuitable material.

On the web you can find out more about censorship from:

- http://personal.dis.strath.ac.uk/control/

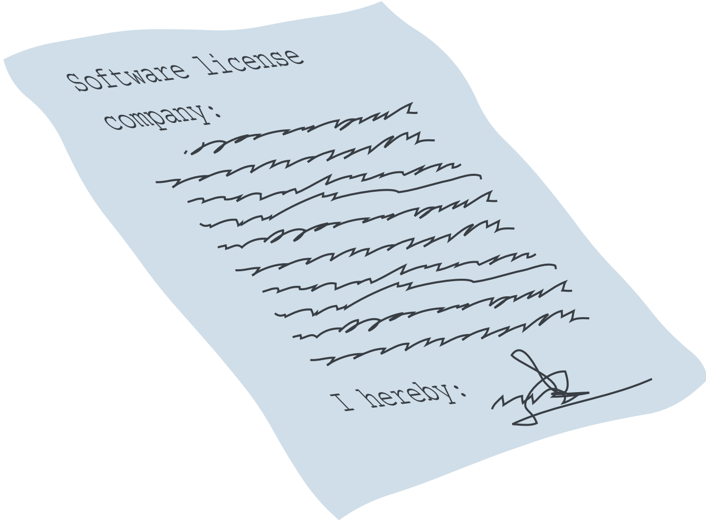
Taxation



Buying and selling on the Internet makes it very difficult indeed to impose taxes, import duties or other fees, particularly if the item purchased such as software, music, or access to information can be transmitted electronically and does not have to be physically delivered to the purchaser. If you work in one country but are employed in another (tele-commuting) where should you pay tax? Who should pay for your National Insurance? Which country's employment laws should apply?

Copyright

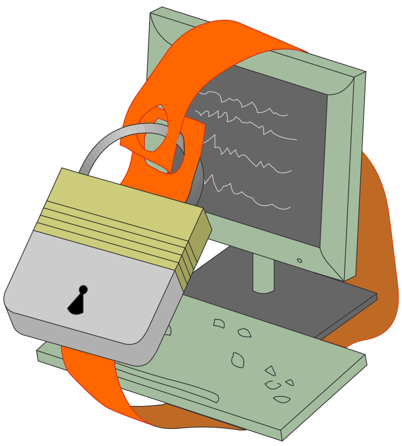
The Internet has made the distribution of software cheap and easy, but of course this also means that the distribution of illegal copies of software is just as easy. Software piracy is a major problem in many parts of the world. Many people argue that software piracy increases the cost of software because software distributors need to charge more to recover the cost of research and development if only a percentage of those using the software are actually paying for it.



In addition to conventional software licencing, there are a number of alternative software distribution models in existence, including **Shareware**, **Adware** and even **Freeware**. The Internet has made many of these distribution models viable because it provides a large enough market and removes the cost of distribution from the developer.

There is also concern over copyright concerning images and other data available on the Internet. The fact that this data is on a web page and anyone can access it does not make it freely available for anyone to take for themselves and publish it under their own name.

Computer Misuse Act



In the United Kingdom, the Computer Misuse Act (1990) covers using computers to damage or steal data. The Computer Misuse Act covers crimes such as breaking into computer systems or networks to destroy or steal data and propagating viruses which destroy or damage information or computer systems.

Data Protection Act

In the United Kingdom, the Data Protection Act (1998) describes the duties and responsibilities of those holding data on individuals. It also describes the right of these individuals. In general, it is the duty of those holding data on individuals to register with the Data Protection Registrar, to keep the information secure, make sure it is accurate, and to divulge it only to those persons who are authorised to view it. It is the right of an individual who has data stored concerning them to view that information and to have it changed if it is inaccurate. There are a number of organisations which may be given exemption from this act - namely the Police, Customs, National Security and Health Authorities.

6.5.6.1 Copyright Designs and Patents Act

Copyright is the ownership of intellectual property outlined by a particular nation's or international law. In the UK, the Designs and Patents Act of 1988, and legislation in other countries who signed the Berne Convention, provide protection.

6.6 Summary

The following summary points are related to the learning objectives in the topic introduction:

- Function of network hardware including hubs, switches, routers and NICs;
- Demand for bandwidth;
- Use of wireless communications;
- Developments in processors, memory, storage and data transfer across networks;
- Developments in network related software;
- Misuse of networks;
- Legal protection related to the misuse of networks.

Accessing a LAN

This activity will provide the evidence you require to show that you have accessed and used a Local Area Network

a) Log on to your school network in the usual way

b) Load a file (or run a program) from the network file server

c) Print out a page of your file (or a screen shot of the program) using a shared printer

d) Write up a short report outlining how you used the LAN. This should include your name, the date and a list of the steps you took with the commands used.

Accessing the Internet

You are also required to produce evidence that you have accessed and used the Internet for WWW. Email and file transfer.

WWW

Since these learning materials are web based the first part should be straightforward:

a) Log on to your Internet server and access the SCHOLAR site (or any other).

b) Print out a screen shot from the site showing the URL

c) Write up a short report listing the steps you took.

Email

To complete this task you will need an email account that has been set up for your use.

a) open your mail reader (email client) software

b) create a new message to send to someone you know

c) before sending it request a delivery receipt (in case it isn't read or you don't get a reply)

d) save and print a copy of your message and the receipt

e) write up a short report listing the step you took

File transfer

Downloading and uploading files on the Internet is done using file transfer protocol (ftp). Your browser will have limited ftp capability to allow at least the facility to download files (e.g. every time you access a website you are downloading a file).

You may also be able to download a file from a website, e.g. a piece of software to run on your computer. This might depend on the configuration and restrictions on your computer.

If you've been involved in creating websites, you may well have uploaded files to an ftp server.

So this task really depends on what facilities you have at your computer and you (or your teacher) will have to choose how you can demonstrate that you have used file transfer.

Whatever you choose to do, follow the same advice as in the previous tasks by printing out any screens you use and listing the steps you took

6.7 End of topic test

An online assessment is provided to help you review this topic.

Topic 7

Computer Software

Contents

Prerequisite knowledge

Before studying this topic you should be able to:

- *Distinguish between an operating systems and an application package;*
- *Explain the need for standard file formats;*
- *Identify data objects and operations in the context of general purpose packages.*

Learning Objectives

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

- *Describe the function of a bootstrap loader;*
- *Describe the main functions of a single user operating system;*
- *Describe the function of typical utility programs;*
- *Describe standard file formats;*
- *Describe the software required to support typical tasks including multimedia catalogues, setting up a LAN and developing a web-site.*

Revision

The following exercise tests the prerequisites for this topic. Ensure that you are happy with your responses before progressing.

Q1: Which one of the following is a text file format?

a) GIF
b) TIFF
c) JPEG
d) ASCII

Q2: Windows XP and Mac OS X are examples of:

a) operating systems
b) viruses
c) application packages
d) utility programs

Q3: An operating system is best describe as:

a) software to monitor users' actions
b) software to carry out specific tasks
c) software to control a computer's resources
d) software to backup data files

7.1 Introduction

This unit on computer software considers the software that is required to support the use of computer systems. The emphasis is on systems software, in terms of purpose and characteristics, but factors affecting the use of applications software is also considered with particular reference to graphics. The selection of software and file formats is considered further in particular situations including a LAN, multimedia applications and web sites.

7.2 Systems software

System software is designed to enable you to run a computer without having to know exactly what's going on inside. System software controls the actual operation of the computer system. You can enter instructions to the computer by typing at the keyboard or clicking on a mouse and the system software will convert these instructions into the thousands of low-level operations needed for the computer to carry them out appropriately. The operating system is part of the system software.

The purpose of an operating system is to provide the user with a means of operating the computer. As a long established category of software, there is a wide range of different operating systems, each with the task of providing a user with an element of control. The operating system can be viewed as providing a layer of software between the user, applications and the underlying hardware of the machine. This operates both to the benefit of the user who is insulated from needing to know the detail of how the hardware works and to the benefit of the system which is protected from inappropriate actions by the user.

Each operating system provides a different amount of control over a machine, with some, such as the operating system on 1980s Home Computers providing a fairly rudimentary level of service, while others, such as Microsoft Windows XP or recent Linux distributions provide an enormous range of facilities.

Opinions vary as to the best way of writing the software that comprises an operating system, with some preferring to write much of the operating system as one enormous program, while others suggest that an operating system should be divided into several different programs that cooperate to provide services. In either approach there will usually be a categorisation of functions into different areas.

One common approach is that an operating system can be described as having five areas of functionality. These five areas are:

1. user interface;
2. file management;
3. input and output;
4. memory management;
5. kernel.

An operating system is meant to provide these features, but it is not only these alone by which a system is judged. Other criteria include:

- cost;
- ease of installation;
- ease of management (how easy it is to add new devices, how much time you have to spend working at the operating system rather than on the applications);
- reliability (how often it hangs);
- performance (how fast it works, how much throughput it can handle);
- range of platforms (how many different kinds of computer it can run on);
- range of software (what applications are available for it);
- customisation (how easily and how much the system can be adapted to particular uses and users).

What is a process?

The term *process* is basic to the understanding of the workings of an operating system.

A programmer writes programs. The source code of these programs will be in assembly language or in some high level language like Java. Before a program can run on a computer, it has to be converted into machine code. When a program is actually running, it contains run-time data and a stack; other information necessary to keep the program running include a program counter and a stack pointer. All of these things together make up a process.

A **process** can be defined as a program in execution.

When you run a program, the operating system has to convert it into a process. It is the process, rather than the simple program, that the processor executes. When the process terminates, the program itself remains (on backing store), fit to be converted into a process again if needed.

Another name for process is task.

7.2.1 Starting up

The most essential and frequently used instructions of the operating system are held in main memory. This part of the operating system is known as the *supervisor*, *executive* or *kernel*. The remaining part of the operating system can be loaded into memory whenever it is needed.

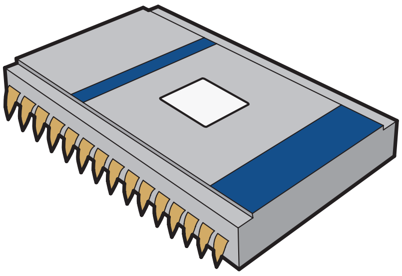
When a computer is powered up it carries out a series of tasks to check that the basic hardware is operating correctly. It then tries to load the operating system program into the main memory. Checking hardware, locating and loading the operating system is carried out by a small program called a **boot program** that is held in ROM. Even if the operating system is disk based, a small part of it, the boot program still needs to be held in ROM.

The term *booting* comes from the expression "*to pull yourself up by your own bootstraps"*. This means that the operating system builds its capability using what it already has, the ROM boot program, to load itself from disk into RAM.

Once the operating system is loaded into main memory, it runs until the computer is switched off.

While this process is not identical to that used on large computers, the functions performed are similar.

7.2.1.1 ROM based operating systems



In **ROM based operating systems**, the operating system is held on a ROM chip. This has the advantages of releasing main memory for applications and preventing the operating system from being corrupted. It also has the advantage of taking very little time to load and be up and running.

Changes to the operating system require an alteration to the ROM program and subsequent chip replacement. This is a procedure that most users would rather not undertake themselves, preferring to pay professionals to do it.

7.2.1.2 Disk based operating systems

With a **disk based operating system** the system software is held on backing storage and loaded into main memory when the machine is switched on. Some computers have the operating system and utility software pre-installed on the hard disk. With others, the user must install and configure the software.

In older computer models, systems software was distributed on floppy disks which also served as a backup. These days, desktop operating systems are so large that they are distributed on CD-ROM, or upgrades can be downloaded directly from manufacturer's web sites.

7.2.2 User interfaces

There a many different types of user interfaces currently in use. Each different type of interface can be characterized by the style of interaction it supports. Style of interaction refers or manner in which you interact. The most common styles include:

- form filling;
- menu driven;
- command and graphical user interfaces;
- natural language.

Styles are chosen by interface designers to support particular tasks. Many systems have a mixture of different styles of interaction to support different tasks and subtasks. If you are using a Graphical User Interface such as Windows you will also find a form filling style of interaction if you are asked to register the software. It is important to note that the different styles of interaction exert different levels of control over the user. Command line interfaces exert little control, i.e. the user can type what s/he wants. On the other hand a form filling interface restricts the kind of interaction that the user can peform.

7.2.2.1 Form filling

Form Filling menus are commonly seen in software such as database systems or as subsystems of other styles of interaction such as GUIs. The forms are used to control the type of information the user can input and many form filling interfaces can be quite smart in detecting that the wrong type of data has been input. This style of interaction is very common for registration or purchase where detailed information in various categories is needed.

7.2.2.2 Menu driven

Menu based interfaces provide the users with a set of options from which they choose. The way that menus are presented on screen varies but at the heart of menu based systems lies the knowledge that users are much better at recognising things as compared to remembering. The menus provides clues and this form of interaction is very good for non-computer oriented users and shortens the amount of time required to learn how to use an interface.

7.2.2.3 Command and graphical user interfaces

The user interface is provided to allow the user to issue instructions to the operating system and receive the results of the instructions that have been issued in the past. Current systems provide two different types of user interface:

1. the command line interface (CLI);
2. the graphical user interface (GUI)

The command line interface (CLI)

The command line interface is a mechanism where the user is required to issue commands to the system. These commands are usually words (or contractions of words) chosen to represent the nature of the activity. For example, in MS-DOS the command to perform a copy command is started with the word `COPY`. In UNIX type systems, the copy command is usually started with the characters `cp`. As these commands are usually entered from a keyboard, the designers of the system try to limit the amount of typing that the user has to perform.

The system responds to the user through an output device. In the past this could have been a device such as a teletype, which printed output onto paper. Current systems are much more likely to display onto a computer screen, still sometimes referred to as a

'Glass teletype'. Used in this way new text will appear at the bottom row of the monitor and scroll up, 'old' text disappearing off the top of the screen. CLI based systems tend to be fairly small pieces of software that consume limited amounts of system resources, making them more suitable for low powered hardware.

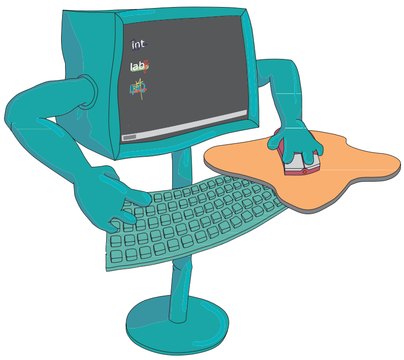
Inexperienced users tend to be uncomfortable when using CLI systems, as the commands are not always easy to remember. The error reporting from CLI systems tends to assume that the user has a certain amount of expertise and can understand the messages that some systems produce when presented with unexpected or erroneous input. For more experienced users CLI based systems can allow complicated and sophisticated operations and when combined with scripts (small programs written using the command language of the Command Line Interpreter) provide a convenient way of automating some of the more time-consuming tasks of administering a computer systems.

Note that some GUI based operating systems require that a CLI based system is present in order to perform the installation. In the event of a system failure it is quite likely that command line intervention will be required to repair the system.

Identifying the advantages of a command based interface

On the web is an activity that asks you to identify the advantages of a command based interface. You should now carry out this activity.

The graphical user interface (GUI)



The graphical user interface is a way where the user is presented with a series of graphical images and controls the operation of the computer by interacting in some way with those graphics. One common form of the GUI is the desktop metaphor where elements such as documents are represented by on screen pictures of the document. To interact with a document, the user will use a pointing device such as a mouse.

GUI systems tend to be much more reassuring for the novice user, as real world entities are represented by recognisable pictures. Operations such as moving files can be performed by 'clicking and dragging'. This means that the user does not need to remember a full set of commands.

Since a GUI uses graphics extensively, a more powerful processor and higher specification monitor is needed than if a CLI is used.

Some complicated procedures may require the users to perform a large number of simple procedures, a time consuming and error prone process. Under these circumstances experienced users may choose to use the facilities of a CLI (if one is provided).

7.2.2.4 Natural language

Natural language interfaces are quite specialised but we are all familiar with what such styles of interaction could do from science fiction films. There are two main types of interfaces, speech input and output and written language. There are many technical challenges in designing and implementing natural language user interfaces. The most common usage is in speech input and output. Speech recognition systems are now widely available as time saving devices for various kinds of office work. Also, increasingly common are natural language interfaces which use writing with a special pen. These are commonly seen in hand-held or palm-top computers.

The idea of being able to have a conversation with a computer is a big research challenge but there is evidence that while such an interface might be natural as in the words natural language interface, users do not want such interfaces for the usual work done on a computer. Other forms of interaction such as GUIs are simply more efficient.

7.2.3 File management

ID	Name	Size	Read/ Write
1	My Files	0kb	rw
2	My File	200kb	rw
3	Program Files	0kb	r

The file system is the portion of the operating system that manages how files are stored. Current systems typically provide storage in the form of magnetic media, such as floppy disks, hard disks and various forms of optical media, such as CD-ROM and CD-RW. For each of the media types a different way of physically storing data is employed, and a large variety of ways of laying out files on a system are used. Some systems will have several different ways of storing data on the devices attached to them.

Magnetic disks provide storage in the form of a series of concentric circles (usually called tracks) on the surface of a platter. Each track is sub-divided into a number of sectors. Many hard disks provide more than one platter in a unit, and the corresponding tracks on different platters are sometimes called cylinders.

In order to determine where on a disk to place a file for later retrieval the file system must be able to distinguish which sectors are free, and which sectors are already in use.

The simplest way of finding free space would be to search the disk until a suitable space can be found. This would be too slow for practical purposes, and a better approach is to use an index file to keep track of the files and space on a disk. The use of an index means that some of the space on the disk is used to keep the index file, reducing the amount of space available for user files.

One such approach is to use a File Allocation Table (FAT). Systems that employ this approach set aside a fixed amount of space on the disk for the index, the amount of space is related to the number of files that the system can manage. For example, a FAT12 system is capable of dealing with 2^{12} entries. FAT12 is thus limited to 4096 entries, but consumes little space on the disk. FAT12 is used for MS-DOS floppy disks and very small hard disks.

FAT16 systems provide for 2^{16} entries in a table and as such provide for much larger hard disks, but at the cost of taking more space. The vast increase in hard disk sizes triggered the development of FAT32, which provides 2^{32} entries in a table which is larger still than a FAT16 table.

Rather than allocate files to sectors, it is possible to allocate files to groups of sectors, called clusters. This has the result of reducing the number of entries required in the File Allocation Table. Each of these entries in the FAT is called an allocation unit. It may be that a file may not completely fit into a single allocation unit, in which case more than one allocation unit will be needed. The linkage between sections of the file must be maintained, otherwise a file might be damaged.

Note that only one file may occupy a given allocation unit, so any space left by a file incompletely filling an allocation unit is not available for other files, giving rise to wasted space. As allocation units increase in size, the amount of space wasted in incompletely filled allocation units increases.

In addition to the actual data held in a file system, there is therefore an amount of metadata (metadata is data about data). This metadata includes the locations where the file is stored, the size of the file and other properties of the file, such as Read Only attributes. Multi-user systems, where multiple users share a machine, will typically store information about who is allowed to access given files as metadata within this system.

Certain file systems tend to be associated with different operating systems, some of the better known are:

File System	**Operating System**
ext2	Linux
FAT	MS-DOS / Windows 3.X Windows 9X
NTFS	Windows NT/2000/XP
HPFS	OS/2

An introduction to file systems can be accessed at:

http://physinfo.ulb.ac.be/cit_courseware/opsys/os05.htm

Detailed information about various file systems can be found at:

http://dmoz.org/Computers/Software/Operating_Systems/File_Systems/

Hierarchical filing systems

A hierarchical filing system is shown in Figure 7.1. The folder icons represent directories, while the rectangle icons represent files. This representation allows users to group related data into directories and helps them to model their own filing methods.

It is possible with this structure to have many files of the same name existing in different directories. The file management system holds *pathnames*, providing unique identifiers for every file and directory that is created.

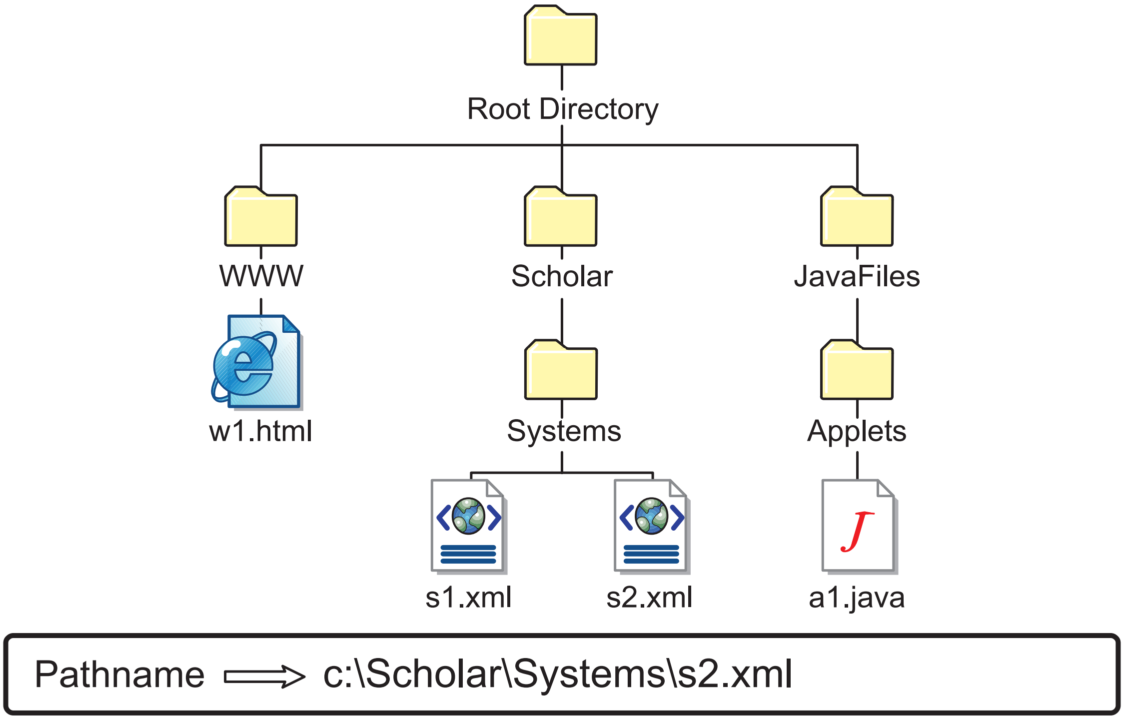
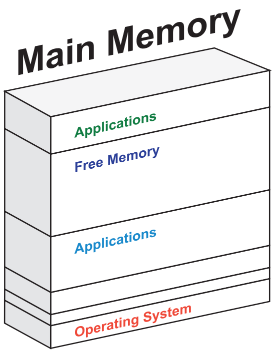
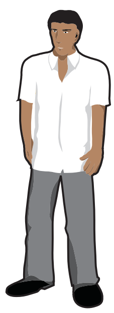


Figure 7.1: Hierarchical Filing System

7.2.4 Memory management





The memory manager is the part of the operating system that regulates the way that memory is used. Memory can be viewed as a scarce resource, with the different processes in the system competing for the use of memory.

Consider the following memory map:

OS	A	B	C	D	E	F

- the lowest area of memory is used by the operating system (OS);
- the second block of memory (area A) is currently unused;
- area B holds the program and data of Process 1;
- area C holds the program and data of Process 2;
- area D is currently unused;
- area E holds the program and data of Process 3;
- area F is currently unused.

In an unregulated system it is possible that Process 1 (currently residing in area B) could require more memory, and expand upwards into area C. Unfortunately area C is currently occupied by Process 2, and the expansion of Process 1 would result in data loss/ corruption in Process B.

The role of a memory manager is to prevent this sort of occurrence, by managing the memory resource. In order to access more memory, a process must request the extra memory from the system. Assuming that the system actually has enough free memory to grant the request the process will be able to access the extra memory. The memory manager may well also provide a logical to physical mapping where the process is not actually aware that the new block of memory that it has been allocated is not immediately after the block it already had.

A second facet of the role of the memory manager is that of security. Using the memory map above, we can assume that Process 1 is owned by User 1, and that Process 2 is owned by User 2. Basic security provision requires that User 2 should not be able to read any of the data owned by User 1, unless with explicit permission from User 1.

An introduction to memory management can be found by following the links at:

http://www.howstuffworks.com/operating-system.htm

A more technical discussion on memory management can be found at:

http://www.memorymanagement.org/articles/begin.html

7.2.4.1 Multi-tasking

The simplest method of memory management is to have just one process in memory at a time, and to allow it to use all the memory there is. One obvious problem with this approach is that a program needs to be entirely self-reliant and contain, for example, drivers for all I/O devices it needs to use. This method is no longer used.

The technique that evolved for simple PCs was to have the operating system and one user program at a time in memory, as well as a Basic Input Output System (BIOS) in ROM. The BIOS contained the device drivers.

This technique can still be called monoprogramming because there is only one user program in memory at a time. DOS basically works in this way.

A sketch of what memory might look like is shown in Figure 7.2

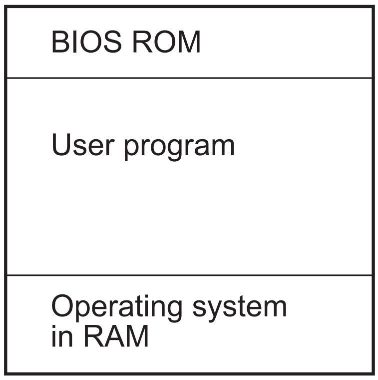


Figure 7.2: Memory in a monoprogramming system

However, Windows does offer **multi-tasking**. In multi-tasking, several applications or tasks are (apparently) available at the same time, and the user can switch easily between one application and another.

Similarly, Windows offers foreground/background processing. The foreground task can claim sole use of the processor but other background tasks can make use of the processor whenever the foreground task doesn't need it.

Multitasking animation

On the web is an animation that demonstrates how the processor is switched between tasks resident in memory. You should now observe this animation.

10 min

7.2.5 Input/output

Computers are, essentially, input/output machines. They take in data, process them, and put out the results. Many users of PCs aren't much bothered about explicit processing, they don't want the thing to solve differential equations for them; they want to be able to write things like reports, and they want to use e-mail and the Internet; they want to type things in, see things on the screen, and print things out. It follows that how the operating system handles input and output is now of the first importance. Users expect input and output to be easy and straightforward. Input/output is often abbreviated to I/O.

The I/O system handles the interactions between the processor and all the devices that are part of the computer system as a whole. Among devices that are regarded as part of the computer could be listed the keyboard, the mouse, disk drives, modems and screen (often called a monitor). The I/O system also has to deal with obviously external devices such as printers and scanners. Devices are sometimes called peripherals (because they are peripheral to the CPU and main memory which are the central devices of any computer system).

These devices are, in general, different from one another: they work in different ways and at different speeds. All these speeds, however, are slower than the computer's processor and memory. One of the I/O system's objectives is efficiency: making the most of the attached devices and getting them to work at their highest useful rates.

In the I/O system, the software can be thought of as organised in layers: the lower one deals directly with the hardware and the upper one presents an interface to the user or application. (There are so many different kinds of device that no operating system could include the code to handle them all.) The upper level is the I/O control system, which takes commands from the user (either directly or through an application), checks them, and sends them to the appropriate part of the lower level. The lower level consists of device drivers.

A device driver is a piece of software that handles one device or type of device. For example, every printer on a computer system will need a device driver, but that device driver might work with more than one model of printer. The device driver converts commands originating from the user into instructions for the device to carry out.

The device driver communicates with the device controller, which is part of the hardware, plugged into the computer's I/O bus. The controller, in turn, communicates with the device itself, converting the electronic instructions received into physical action.

Figure 7.3 shows how these levels are related. An application starts things off by making a system call that demands action from the I/O system.

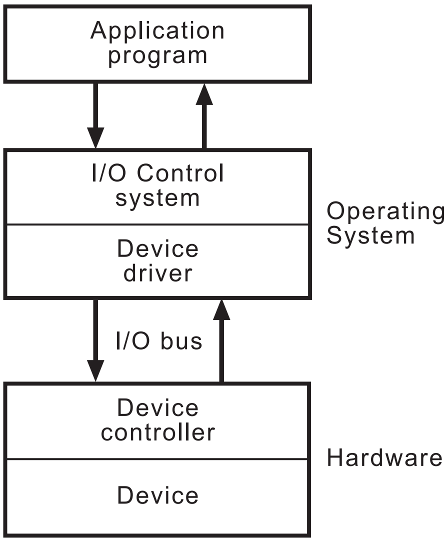


Figure 7.3: Levels of an operating system

Block and character devices

Roughly speaking, there are two categories of device: block devices and character devices.

Character devices accept stream of characters without any regard for a block structure. The screen can be regarded as a character device.

Block devices store information in blocks of fixed size. Discs are block devices.

7.2.5.1 Direct memory access

As a rule, a device controller will have its own memory on board. This enables it to use Direct Memory Access (DMA) to get data to and from the computer's memory. In DMA, the processor only sets up the transfer. Thereafter, some clock cycles are used by the I/O system for the data transfer. During these cycles, the processor is doing nothing but, during the other cycles, the processor can be getting on with its work: it doesn't have to suspend the current process or spend time on context switches. On the whole, DMA saves time by saving the processor for other things than supervising data transfer.

7.2.5.2 Virtual devices

A virtual device seems, to the user, like an actual device but is, in fact, a simulation maintained by the I/O system.

A print spooler is a virtual device, that takes data sent to the printer, stores them temporarily on disk, and prints them when the system is ready to do so. This means that several print jobs can be sent to the printer at the same time and will be dealt with in an orderly manner. The jobs are placed in a print queue and serviced one by one.

It also means that a user does not have to hang around while the actual printing takes place but, having sent a print job to the spooler, can get on with other work on the computer.

7.2.5.3 Buffering

All I/O is relatively slow. For most of us, input by typing is painfully so. Even very fast typists are, from the CPU's point of view, so slow that they're hardly doing anything at all. Screens and printers work a fair bit faster (as a rule) but all the same, the I/O system needs to use the devices efficiently so as not to waste the CPU.

A buffer is an area of memory set aside to help in the transfer of data between the computer and a device. A buffer provides a sort of barrier between parts of the computer system that work at different speeds.

Buffering is used in sending blocks of data. Data is transferred into the buffer until it is full. Then the entire block is sent at once. This is more efficient than having the data trickle through the system.

With single buffering, the receiver has to wait for a block before it can do anything with it, and the transmitter of the data has to wait for the receiver to have processed the block before it can send another.

With double buffering, two buffers are used and, as one is emptied, the other is being filled up. This makes more efficient use of both transmitter and receiver of the data.

7.2.5.4 Device drivers

Device drivers are specialist software components that facilitate the operation of physical devices attached to the system. Without an appropriate device driver the system cannot communicate with a device, rendering the device useless.

Each variety of device will have different hardware. This means that for optimum performance, each device should be used with a device driver specifically written for use with that device. Some devices may be used with device drivers built for slightly different devices, but at the cost of reduced performance or functionality. Some devices such as CD-ROM drives are currently so similar that it is possible to use a 'generic' driver in conjunction with just about any CD-ROM.

Modern operating systems offer the facility of adding device drivers to a system, at the same time that the device itself is added.

This is in contrast with earlier approaches where all of the device drivers had to be included in the operating system when it was built. This had the effect of increasing the size of the operating systems, and ensuring that new devices could not easily be added. On a practical note it is recommended that particular care is taken with media containing device drivers, as they can be difficult to replace.

Device drivers relieve the writers of an operating system from having to know the exact details of a device. Instead the operating system can act as though it is dealing with a generalisation of that type of device, and leave the device driver to sort out the details.

Device drivers can be found at many sites, including:

http://www.driverguide.com/

and:

http://www.driverzone.com/

Many vendors also place device drivers for their products on their web sites.

7.2.5.5 Plug and play

Adding a new device to a system has in the past been rather a fiddly business. Often a new control card has to be plugged into the computer. Very often, switches on the device would have to be set by hand. Then the computer system would need a lot of tweaking before the operating system and the device could communicate in a satisfactory way. The concept of plug and play tries to make the business more automatic.

Plug and play can be regarded as a sort of convention or standard agreed between the companies responsible for the different parts of a computer system: computer manufacturers, chip manufacturers, device manufacturers, and so on. Things are designed to suit the plug and play standard; devices are designed so that they can be set by software (rather than having switches set by hand).

Windows automatically detects all plug and play devices on the system. The user has to supply very little information during installation and nothing at all from then on.

All devices need to use the computer's resources: the processor, memory, interrupts, DMA channels, I/O ports. Windows handles these needs automatically for all plug and play devices on the system.

Windows looks after the system configuration (so that you don't have to do any of that yourself by hand) and attempts to resolve conflicts between devices for the computer's resources.

Windows can dynamically load and unload device drivers (as compared, say, with having to edit various batch files by hand and re-boot the computer). Furthermore, if one of the devices fails, Windows will detect that it is not on line and display a dialog box to inform the user of what's happened. (That, at least, is the idea - on a machine the writer has at home, the CD-ROM drive doesn't work, but Windows, when asked about it, blandly displays assurances that it's working fine.)

Identifying operating system layers

On the web is an activity that asks you to identify the layers of an operating system. You should now carry out this activity.

7.2.6 Kernel

The kernel is the central component of the system, upon which all of the other components depend. In effect the kernel co-ordinates the activities of the system. Again there is wide variety in exactly how systems are implemented, but typical operations of the kernel include:

- if a key is pressed, the device driver will notify the kernel which will pass the request to the user interface.
- If the user interface requests that a file be moved, the kernel will pass the request to the file system. Once the file system has finished, any messages from the file system will be sent to the user interface;
- if the user interface wishes to send a message to the user, the kernel will notify the device drivers for the output device (such as monitor or printer).

In single user/single tasking systems, the kernel may not be required to do much more than this. In multi-tasking / multi-user systems the kernel will usually also be responsible for managing the processes as well. There are two types of multi-tasking system to consider here. The first type uses a model known as *co-operative multi-tasking*.

In this model a process is allowed to run until it voluntarily gives up the processor. The intention is that each program should contain program instructions that indicate to the OS kernel that the program has finished its turn, and is ready to allow another process to use the processor. This approach can work well as long as each process is well-behaved, and relinquishes the processor as agreed. However, some programs are not as well behaved, and can consume all of the computer's resources without any of the other processes being allowed a turn. This is particularly serious when a program 'crashes' and locks up the system completely. The second model is usually referred to as pre-emptive scheduling. In this model, the kernel takes the additional responsibility for ensuring that a process is given an amount of time to run, and then stopped to allow the next task a turn. Each time the kernel stops a process, information about that process must be preserved, so that the process can be restarted at exactly the same point next time it is given a chance to run on the processor.

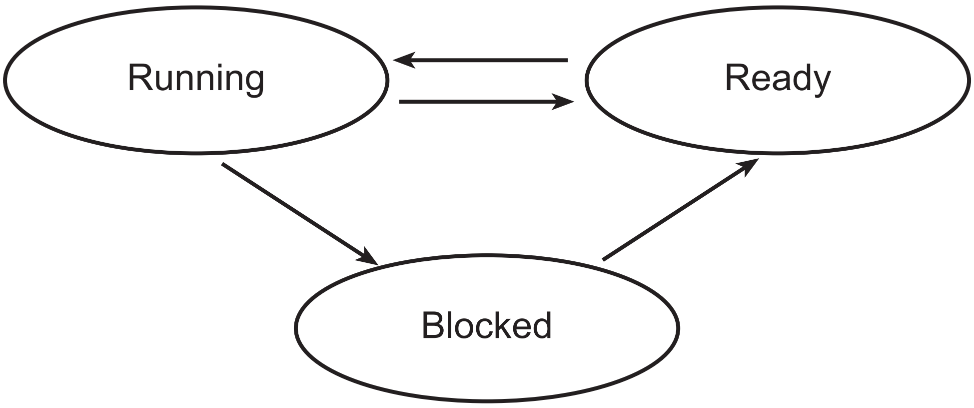
The operation of stopping one process, saving its status, loading the status of the next process and starting to run the next process is known by the term 'context switch'. A context switch may take a significant amount of time to complete.

In order that important processes (such as disk accesses) are not held up by slower processes (such as printing) most systems implement a notion of priority. The kernel itself will have the highest priority (otherwise some other process could lock the kernel out and gain total control of the machine!), and will determine the priority of all of the other processes in the system. As the kernel will require control over all of the other components of the system it will also have the highest set of privileges in the system.

A common way of distributing processor time amongst a set of competing processes is known as 'round robin' - each process will have a turn, and then be sent to the back of the queue to await its next turn.

There will be occasions when a process is not capable of taking advantage of its turn at the processor. This is usually when the process is waiting for input or waiting to do output. Even if the process were given access to the processor it can do no useful work. Granting access to the processor to a process that was waiting for I/O would slow the rest of the system down.

In order to quickly determine which processes are capable of being run, and those which are currently waiting for I/O each process can be described as 'ready' or 'blocked'. There is a third state in which a process can exist, 'running', used to describe the process actually in the processor at the current time. Note that a single processor system will only have one process running at any one time.



The diagram above shows the different ways in which a process can change state.

1. a running process can complete its allotted time, and as it is not waiting for I/O can return to the ready state.
2. a running program may request I/O, and as such its state will be changed to blocked. If this happens it will immediately be removed from the processor.
3. the I/O transaction that a process was waiting for may occur. In this case the process will move from the blocked to ready state.
4. the process that is next in the queue can be scheduled to run, changing state from ready to running.

Note that other transitions such as blocked to running are not supported, as it makes no sense to try and run a process that is blocked.

Using an operating system

This activity will give you evidence that you have used the functions of the operating system of your computer. In fact every time you switch on your computer you start using the operating system. For each of the points below you should record and file your responses.

Describe what happens when you switch on your computer. Make reference to what the operating system does and what files it uses.

Where is the operating system loaded from? Has your operating system been customised in any way? That is, are there specific files, settings or programs loaded or run on start up?

If so, list them.

Once your computer is ready for use, click on an icon to load one of your files. Make a list of the actions that the operating system carries out in order to load your file. You should make reference to the layers of the operating systems and there respective tasks.

Now close down your system, but as it does list the tasks that the operating system is performing to ensure a successful shut down.

7.3 Utility programs

Utility programs are those programs that are used to enhance the operating system, or in some other way improve the usefulness of the system. The line between utility programs and application programs can be a little blurred at times, but in general a utility program will be used to maintain a system, while an application will be used for a purpose external to the use of the system. For example, a word processing package can be used to write correspondence, while an anti-virus package can be used to reduce the risk of a virus infection on a system.

Most operating systems will ship with a collection of utility programs. Examples of these include:

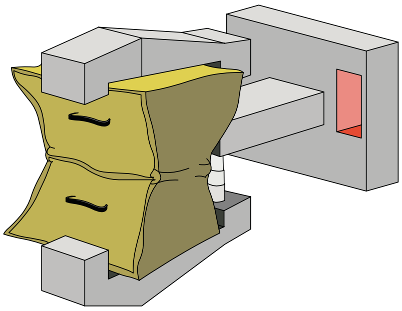
1. disk partitioning tools;
2. disk formatting programs;
3. file compression utilities;
4. disk defragmentation tools.
5. virus checking tools

7.3.1 Disk partitioning tools

Disk partitioning tools are used to divide one physical device into more than one logical device. Some operating systems require a disk to be partitioned before installation can take place, other systems provide the user with the facility to partition a disk for the purpose of convenience. (It may be convenient to provide a separate partition for user data). The disk partitioning tools provided with many operating systems can be regarded as a little primitive and unfriendly, but are usually sufficient for all but the most exotic of requirements.

7.3.2 Disk formatting programs

Disk formatting programs are used to prepare the surface of writeable media for use. The process often involves laying down 'markers' for future use. Formatting a disk usually deletes any existing data on that disk, so caution should be exercised.



File compression utilities

File compression utilities are used to reduce the amount of space that a file takes up on disk. It will take a small amount of time to decompress a file, and also some time to compress a file before storage.

7.3.3 Disk defragmentation tools

Disk defragmentation tools are used to combat the problem of file fragmentation. Ideally, all of a file should be located in the same area of a disk, as this speeds up file accesses. However it is possible for files to be broken up, with the various portions of the file to be dispersed across the disk. This usually happens when a number of small files are deleted from a disk, and the file system chooses to fill in the gaps left by the deleted files when the next file is saved. A defragmentation utility reorganises the way the files are laid out on the disk so that the components of a file are near to each other.

Some operating systems are more prone to fragmentation than others, so not all systems will ship with a utility of this type.

7.3.4 Third party utilities

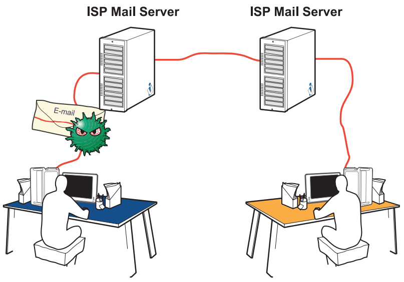
Third party utilities are tools provided by others than the supplier of the operating system. These programs can either be free (freeware) or purchased on either a shareware or commercial basis. Third party software often offers enhanced functionality over utilities supplied with a system, or may provide functionality not envisaged by the system vendor. Such tools include:

1. disk partitioning software;
2. disk formatting software;
3. anti-virus software.

Disk partitioning software performs much the same function as the utility supplied with the system, but usually has a unique selling point, such as ease of use or applicability over more than one system. One reason that hard disks are partitioned is to allow the installation of more than one operating system on a single machine, with the user able to choose which system is run when the system is started. These systems often referred to a dual-boot (for two OSes) or multi-boot systems (for more than two OSes). Some of the third party partitioning utilities are particularly useful when creating a dual boot system. At least one commercial utility allows for a disk to be dynamically re-partitioned without reformatting the affected partitions.

Disk formatting performs much the same function as the utility supplied with the system software but usually offers a wider range of facilities, and may be able to format a disk to higher capacity than the 'official' software. Such formats may not meet with the approval of the system vendor.

7.3.5 Anti-virus software



Anti-virus software is used to prevent the spread of small, usually malignant programs that spread amongst machines using a similar approach to that used by biological viruses. As programs, they need to be executed before they can be effective, so viruses often attach themselves to program files. Another place that computer viruses can be found is in the boot sector of disks.

As this is where the disk stores the program that is used to start the computer, any virus in this are can be activated automatically each time the machine is started. Viruses are also capable of hiding in other places, such as the partition table. The damage that a virus can inflict is called its payload, and this can vary from no effect (apart from taking up space and slowing the machine down slightly) to viruses that can overwrite a part of the operating system that resides in certain types of hardware. Such an action can result in the hardware becoming useless.

Anti-virus software can operate in several different ways, including:

1. monitoring the size of files (if a program suddenly changes size this may indicate that a virus has attached to it);
2. monitoring the system for 'suspicious' activities, such as trying to modify important files;
3. checking to see if a characteristic series of program instructions known as a signature appears in a file. The appearance of a recognised signature is an indication that a known virus is infecting a system.

The virus writers are aware of the efforts being made to prevent the spread of viruses, and are developing different techniques. One of these is the polymorphic virus. This attempts to defeat detection by encoding parts of the virus into many different forms. This generates a huge number of different signatures for the anti-virus software to try and detect.

The large vendors of anti-virus systems provide daily updates for their products, and given the speed with which viruses can propagate through the system, it is a good policy to update virus protection products at least once a day.

There are a large number of web sites that contain information about computer viruses and how to remove them, not least from the vendors of software to deal with the problems of infection. These include:

http://vil.nai.com/vil/

Use of utility software

This activity requires you to use a utility program on your system. Utility programs range from the complex and time consuming e.g. disk defragmenter to relatively straight forward and simple e.g. a search. Here are a few examples of utilities that you might choose from:

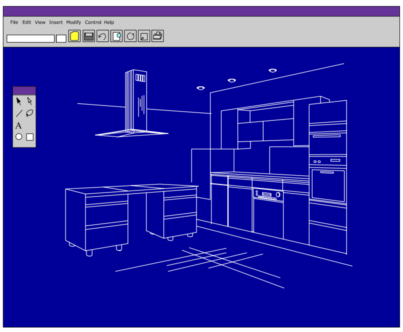
- Virus checker
- Disk editor
- Defragmenter
- Disk backup
- Search
- Sort

There are many more utility programs available, but the one you choose will depend on the access you have to your computer. For example, if you want to defragment the hard disc on your computer, this might prevent anyone else using it for days, possibly! So take advice before you choose.

Once you have decided on which one to use, implement the routine and write up a short report outlining the process. This should include how you got it to run and what the results were. Make a list of what you did and what the utility did. Also print out any results or screen shots of the procedure.

7.4 Graphic applications

There are many applications that use computers to generate images. The images can be an end in themselves, for example using a painting or a drawing package to generate a diagram or picture. These packages can vary in sophistication, from simple 2 dimensional object rendition, or elementary paintings to professional design work used by graphic illustrators.



Computer Aided Design

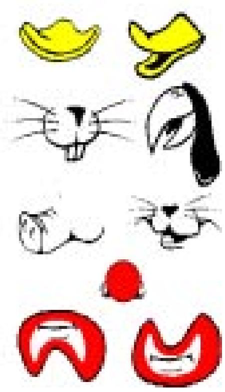
Other packages need to be used as part of a larger design process. For example, using Computer Aided Design (CAD) applications in areas as diverse as kitchen appliance design to electronic circuit design.

The computer is used to produce "blueprints" as part of the design of a project.

Additional CAD functions

CAD applications also allow the designer to perform calculations such as costs of materials, quantitiy of materials etc.

With integrated circuit design, the package may also integrate to other software tools, letting the designer check electrical connectivity or simulate the operations of components within the design.



Animations and simulations

Computer games and simulations rely on graphical environments using exciting animations to underlie their actions. The graphics could be produced in a drawing package and imported into an authoring environment where each image frame can be altered to give an impression of movement. They could also be created directly in the authoring package itself.

These development environments also support "live action" graphics that can be mixed with computer-generated images. These are now routinely used in the television and film industry.

7.4.1 JPEG

JPEG (Joint Photographic Experts Group) is a group of experts that develop and maintain standards for compression algorithms for computer image files. Any graphic image file created or converted using a JPEG standard is given this name.

A JPEG file is created by choosing one of the set of compression algorithms available. For example, similar colours and minor changes in colour, need not be coded separately.

JPEG processing makes image files small by removing detail. This is called a lossy compression. This will reduce the number of colours used to store the image and avoid unnecessary repetition of bit patterns.

The highest quality results in the largest file size so a trade off has to be made between image quality and file size.

JPEG is one of the image file formats supported on the World Wide web suffixed by ".jpg".

7.4.2 GIF

The web also supports GIF (Graphics Interchange Format) images. These images are based on a compression algorithm that creates a codebook or dictionary of particular bit patterns. These in turn, are then substituted resulting in a smaller file. When decoding, the algorithm uncompresses the file to generate the original image.

Significantly, GIFs preserve all the precise shapes in an image i.e. result in lossless compression.

7.4.3 TIFF

TIFF (Tag Image File Format) is a common format for encoding bitmap images e.g. scanner images, and can be any resolution.

TIFF files can be in any of several classes, including gray scale, colour palette, or RGB full colour, and can include files with JPEG and GIF formats.

A TIFF file can be identified as a file with a ".tiff" or ".tif" suffix.

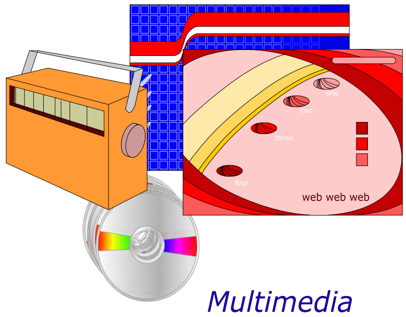
Graphic files

You are required to show that you have created and implemented one of the standard graphics file formats. You will be asked to produce a graphic using the hardware and software available to you, and then asked to describe the format of that graphic. Graphic file formats may or may not be hardware independent but all will require an application package that will open files in that format.

1. First thing is to choose which type of graphic you are going to produce. This could be generating the graphic from a package, scanning an image, or taking a photograph.
2. Using the hardware and software appropriately, create your graphic
3. Save your graphic on your computer and answer the following questions:
    1. Is the graphic a vector graphic a bitmapped graphic, or a combination?
    2. What is the file format? (e.g. .HGL, .DXF, .JPG, .GIF)
    3. What is the resolution of the graphic you have produced?
    4. How many pixels is this?
    5. What is the colour depth (or grey scale)?
    6. How much storage is required for the graphic?
    7. Is the storage requirement appropriate to its use?
    8. If yes, explain why
    9. If not, what has to be done to the graphic?

7.5 Other applications

7.5.1 Multimedia software



Multimedia

Multimedia is the combination of text, graphic artwork, sound, animation and video data types delivered to the user by computer or other electronic means.

If the end user can control what is displayed, then it is called **interactive multimedia**.

When data elements are linked to form a structure through which the user can navigate then it is called **hypermedia**. The World Wide web, for example, is a hypermedia environment.

Applications of multimedia

Multimedia can be used in a variety of application areas. Examples include the development of a World Wide web site, training and education programmes where information is presented for learning, or CD-ROM based encyclopaedias and games.

Multimedia in the home includes anything from cookery to interior design. Although currently CD-ROM based, in future most multimedia software will reach the home via television sets with an in-built mechanism for user interaction. A pay-for play mechanism will allow you to download applications and entertainment packages as you require them.

In public services, such as transport, museums, libraries and shopping centres, multimedia information kiosks designed to provide information and assistance at the touch of a button will become more prevalent.

7.5.1.1 Common features

The software that allows you to put the different data elements together, build interactivity and navigation is called *authorware*. Many different authorware packages exist but most support the following basic authoring operations.

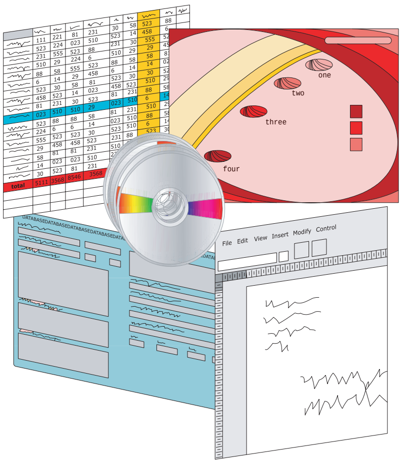
- ability to import media types;
- editing, particularly text and static images;
- organising navigation;
- programming - visually using icons and objects;
- scripting language;
- playback;
- delivery by building a run-time version of the project that does not need the full authoring software to execute.

Examples of authoring software include: Macromedia Director, Apple's Hypercard and Windows Toolbox.

7.5.1.2 Standard file formats

You have already met the standard file formats for text and graphic images. Additional formats that need to be agreed in multimedia include how video and sound will be communicated.

7.5.1.3 Interactive presentations and media types



In modern information systems such multimedia systems, be they software packages (e.g. Powerpoint) or web pages, you have available a number of multimedia elements to enhance your presentation. These include various types of images, video and sound, and more sophisticated elements such as animations and various interactive components. Designing effective presentations with multimedia elements of this kind can be problematic in that you must decide which ones suit your presentation and which ones fit together to enhance the basic information being presented e.g. on a web page. Each element has its own advantages and disadvantages.

Images

The most common multimedia element is the image. Images can take many forms such as simple graphs, showing scientific data or more sophisticated forms such as digital photographs. The old saying "a picture is worth a thousand words" has been shown to be untrue but the choice of a good image can greatly enhance and simplify the material you are trying to present. As with all visual communication the key is to choose the right sort of image to enhance your presentation. The quality of the images van also be a major issue. Some image formats, e.g. JPEG use compression which lowers the quality of the image. The size of the image is also important as this will determine the time the takes to load the image, e.g. across the web.

Video

Video can greatly enhance visual communication in a multimedia system. Here, there is no limit to the kind of detailed information you can present. However, video requires a lot of storage space on hard disks, and a great deal of computer memory to work properly.

Another potential problem is that video formats often involve compression and when you play the video on screen the computer's processor is not powerful enough to decompress and the video becomes quite jerky. Sending video across networks has similar problems as video tends to be big and places pressure on bandwidth. Hence the video can appear jerky.

Sound

The use of various sound effects in presentations is common as is the playing of music using a computer. Most people are aware of MP3 standards and music and the arguments which have arisen over the downloading of music over the Internet. The main limitations of using sound are similar to video. High quality sound such as music takes up a great deal of space on a hard disk, demands much memory and quality can be lost when playing due to decompression problems. On the other hand the use of simple sounds to alert users usually does not suffer from such problems.

Animations

Many interactive presentations use animations. The nature and quality of animations can vary enormously, ranging from simple animated movements, e.g. animated icons to full blown cartoon style animations such as *Disney's Toy Story*. Producing good animations is a specialised skill that requires a great deal of expertise and professional skills, not to mention special equipment such as graphical high powered workstations with very large storage media, fast processors and large amounts of memory. The use of animations in multimedia presentations is therefore limited by the same kinds of problems as those outlined for video and sound.

7.5.1.4 Review questions

Q4: The term *multimedia* means:

a) More than one channel of communication.
b) The combined use of video, text, sound and graphics
c) Many different applications open at the same time.

Q5: What is the difference between *interactive multimedia* and *animations*?

Q6: Which of the following application areas would suit a multimedia presentation?

a) Creating a program in a high level language.
b) Teaching youngsters the Green Cross Code.
c) Writing an operating system utility.

Q7: State three common features of multimedia authorware packages.

7.5.2 web software

In 1993 the World Wide web appeared. Internet usage grew rapidly and it continues to grow daily. In 1994 the first search engine, the World Wide web Worm, had an index of 110,000 pages. By 1997 webCrawler reported 2 million. In November 1997, AltaVista claimed it handled 20 million queries per day. By the year 2000, the top search engines reported in excess of 200 million queries per day. Estimated at around 2.1 billion pages today, the number of web pages is due to double in a year.

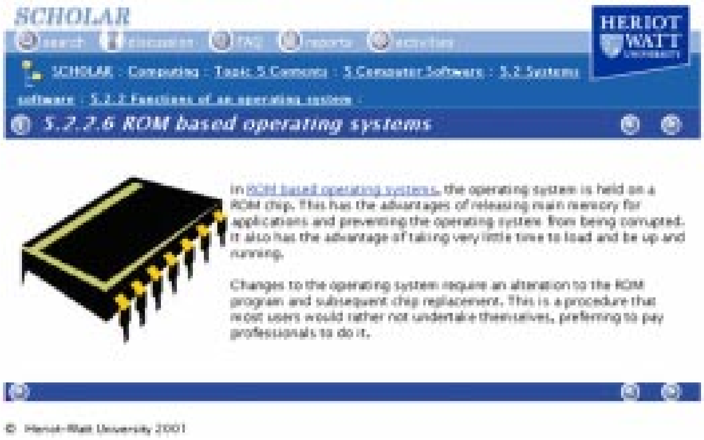
Some people confuse the WWW with the Internet. They are not the same thing. The web is a mass of global resources that can be accessed anywhere in the world using a Uniform Resource Locator (**URL**) and a **browser**.

Resources take many forms, including text documents, static graphic images, video clips or programs. This is known as a multimedia. Resources may also link to other resources, via hypertext links, creating a massive web of associated information.

On the web is a simulation that shows you how a user communicates with computers wordwide via the Internet. You should now look at this animation.

7.5.2.1 websites

A website is a collection of web files that stats with a home page. Most companies, organizations, or individuals that have websites have a single address that locate their site. This is their home page address. From the home page, you can get to all the other pages on their site. web pages are described as HTML but can be built using an authoring package that allows the user to concentrate on the design of the pages. The software will then generate the necessary HTML to be uploaded to the host server.



7.5.2.2 Standard file formats

web pages are described using Hypertext Markup Language **(HTML)**. HTML uses generalised and descriptive element types, which means that documents can be displayed on text screens, under graphical user interfaces or even projected through speakers for the sight impaired.

web authors use HTML tags to describe whether a word should be emboldened, or a paragraph break should be taken. In other words, it is a language that describes how data on a page should *look*.

HTML is interpreted by a browser, such as Netscape Navigator or Intermet Explorer, to render page content on a display screen. A browser does not know what the data on a page means. Hyperlinks that form associations with other web pages are viewed by the browser as demands to fetch and display a new page. An example of how pages are connected across the web using links is shown in Figure 7.4

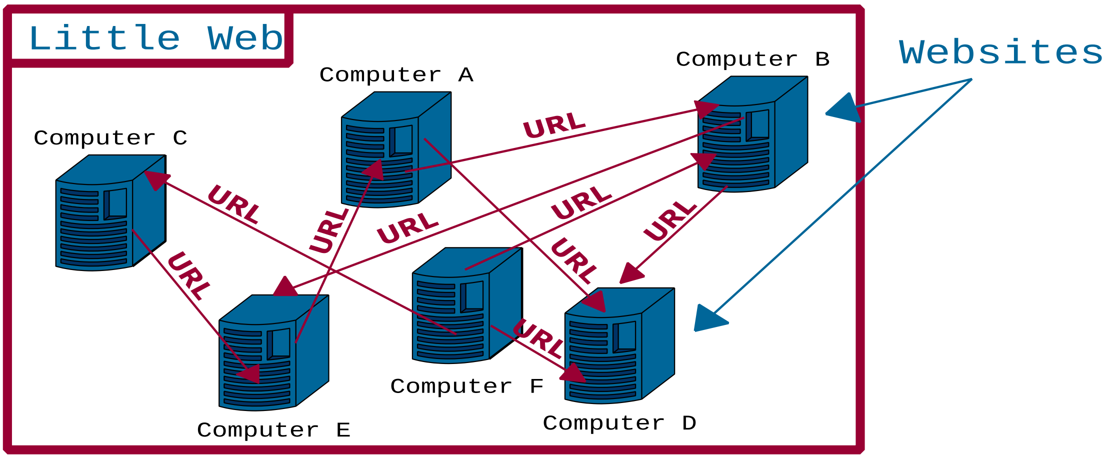


Figure 7.4:

In 1994 the World Wide web Consortium (WC3) was founded. The consortium aims are:

- to develop common protocols that promote evolution of the WWW;
- to ensure WWW interoperability.

Some of the work undertaken by WC3 is listed below.

- The WC3 reacted to proprietary formatting markup by inventing a simple HTML-oriented stylesheet language called Cascading Style Sheet (CSS). This ensured that developers worked to a common markup standard.
- In 1997 HTML 3.2 was published followed by HTML 4 in April 1998 and HTML 4.01 in December 1999. XHTML (Extensible Hypertext Markup Language) 1.0, which features the semantics of HTML 4 using the syntax of XML (Extensible Markup Language), became a recommendation in January 2000.
- XML 1.0 recommendation in 1998 was the first step towards the next generation web, allowing communities to design languages that suited their particular needs and to integrate them into a general infrastructure based on XML.
- Web Accessibility Guidelines 1.0 were published in 1999 explaining how to author web pages and sites that can be accessed by people with disabilities.

7.5.3 Network software

This section returns to the scenario used when considering the hardware requirements for a network. In your revisit you should consider the software issues in light of the work on this unit.

7.5.3.1 A small secondary school using RM Community Connect

Research Machines (RM) provide Community Connect as a networking solution for schools who cannot afford to commit a large number of technical support staff to running their networks. The Community Connect software is based on Windows NT4 as the server operating system, Internet Information Server running the Intranet and Windows 98 on network stations. Schools are unusual network environments in that they require high bandwidth connections, have peak periods when users log on and off, and need a high level of security. They are also unusual in that users do not normally always work from a specific network station, but may log on to several different stations in the course of a day. RM provide a comprehensive suite of network management tools specifically designed for schools.

Hardware: Two servers, one acting as a file and applications server the other acting as a communications server. 150 network stations with a variety of different specifications. Both servers and a router are connected to a 24 port 100Mbps switch. Cabling throughout the school is category 5a UTP cabling connected using 100Mbps hubs. The *router* is a Cisco 2600 series connected to two ISDN lines giving access to four 64Kbps data channels.

Software: Both servers are running Microsoft Windows NT4. The network stations run Microsoft Windows 98. For Internet access the communications server runs Microsoft Proxy server. RM provide a suite of network management applications for adding users and groups to the network and for installing software on stations remotely.

Communications: The network has a Gigabit per second (Gbps) backbone between the servers and the switch and the rest of the cabling runs at 100Mbps. The protocols used are **NetBEUI** and TCP/IP.

Functionality: Printers can be shared throughout the school using dedicated print servers. Internet access and email are available from all network stations. The Community Connect software allows users to access files on their network drive via an Internet log-in using the Point To Point Tunneling protocol. The RM Connect system allows users to access shared areas of the hard disk which the teacher has set up containing class materials and work-sheets. A school Intranet is available containing subject specific web links and pupil projects with a staff area for the school handbook and other administrative documents.

Security: Network security is imposed using NT identities and passwords. Users are made members of NT groups whose permissions determine their access to areas of the Intranet, levels of access to software and access to shared resources.

Station security is imposed using proprietary software which backs up all crucial system files on to a hidden directory on drive C and sets them to read-only status. These files are not visible to users when they are logged on to the network. The software restores these files when a station is re-booted. This means that any configuration changes made to a station are removed by switching it off and on again. In the event of the system files of a station being permanently corrupted, the operating system and other software can be installed remotely and the station "rebuilt" to the same state it was in prior to the problem.

Building a website or multimedia catalogue

Again, this activity returns to the scenarios when considering hardware. Make a list of the software you would need to build a website or a multimedia catalogue stored on a CD. Record your findings in a table format and discuss with your teacher.

If you have time you should consider the doing both tasks.

7.6 Summary

The following summary points are related to the learning objectives in the topic introduction:

- The purpose and main functions of a single user OS when starting and in the management of user commands, files, memory, input/output, resources and processes;
- The use of utility programs including virus checker, disk editor and defragmenter;
- Description and use of standard file formats;
- Software requirements for setting up a multimedia catalogue, a LAN and web-site.

7.7 End of topic test

An online assessment is provided to help you review this topic.

Topic 8

Supporting Software

Contents

Prerequisite knowledge

Before studying this topic you should be able to:

- *Describe a virus and how it operates;*
- *Describe the common symptoms of virus infections and how they can spread;*
- *Explain the need for anti-virus software.*

Learning Objectives

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

- *Describe factors affecting software compatibility including memory, storage and operating system issues;*

- *Classify viruses by type of file infected;*
- *Describe virus code actions;*
- *Distinguish between a virus, a worm and a Trojan horse;*
- *Describe anti-virus software detection techniques.*

Revision

The following exercise tests the prerequisites for this topic. Ensure that you are happy with your responses before progressing.

Q1: A virus is best described as:

a) a disabling code that infects a computer
b) a disabling code written into a program
c) an error written in a code
d) an error that occurs only when a program is running

Q2: Which of the following is the most likely source of a virus?

a) an e-mail message
b) an e-mail attachment
c) a software package
d) an operating system

Q3: To remove a virus you should:

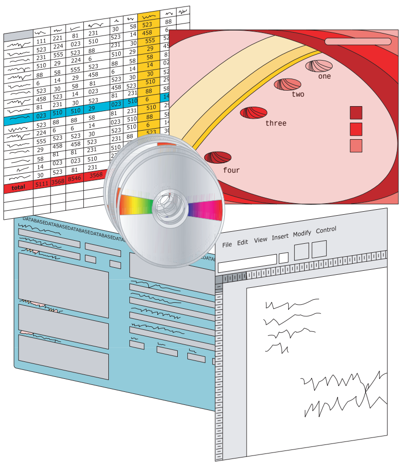
a) install virus checking software
b) reinstall the operating system
c) install a firewall
d) install anti-virus software

8.1 Introduction

This unit on Supporting Software considers the factors affecting the deployment of software on a system. These include compatibility issues between the software and the intended hardware and operating system. Once up and running, the issue of protecting software and systems from viruses is then considered.

8.2 Application software

Application software supports user tasks. Examples of general user tasks are the creation of a report, or the construction a database of names and addresses, or using a spreadsheet to keep track of household accounts. Specific tasks include the work of graphic designers in creating artwork, or scientists using data logging equipment to montior and control an environment or multimedia authors creating educational software.



Many software companies offer the same type of application product. It is important to know enough to distinguish between them in order to correctly match a computer specification to task requirements and to minimise expenditure. Selection of appropriate software can be further complicated by **general purpose packages** (GPPs). These offer support for a range of object types and operations. For example, they may include separate modules for processing text, spreadsheet data, databases, graphics and communications.

GPPs commonly offer less functionality than a special purpose package. For instance, a graphic designer requires more complex image manipulation, supported in packages such as Photoshop or Illustrator, than those available within Microsoft Office or Clarisworks. Architectural work is another example where specialised modeling techniques and image rendering may be needed. Such specialist capability is not supported in drawing or painting modules within GPPs.

It is not necessary, nor realistic for you to be an expert in every software package available, but you should at least be familiar with the general class of functions that different types of application packages support. This will give you a flavour of its capability and provide a basis for you to further explore the package. You should also be aware of standard data formats that facilitate the exchange of data across different package types.

8.2.1 Application requirements

When buying software applications it is important to ensure that they will not conflict with your operating system or your hardware. If there are problems then your software will not run properly if at all! Each application software package includes details of the minimum system specification required to run the application. This will include:

- operating system;
- minimum RAM;
- minimum processor speed;
- minimum hard disk space;
- peripherals required.

8.2.2 Software compatibility

Whether you are doing a fresh installation of or upgrading from a previous piece of software, checking your hardware and software for compatibility is a must. Hardware, operating systems and applications in use must be compatible with one another.

Individual families of processors, operating systems, and applications will attempt to be downwards compatible with previous versions but this claim should be approached with caution when upgrading and installing software.

The main factors to consider are:

- Memory requirements;
- Storage requirements;

- Operating System.

It is important that checks are made to make sure you have the needed software, hardware, and operating system to run the application before installing.

8.2.3 Memory requirements

"How much RAM do I need?" is a question that should be asked of any software.

Software will normally specify a minimum RAM requirement but may also recommend more for enhanced performance. For example, a audio sequencer may require a minimum of 24Mb to run, but may use 256Mb to allow for greater sampling rates for a smoother output or to introduce sound effects.

The individual requirements of the user of the software should be considered along with the amount of memory required.

8.2.4 Storage requirements

For applications that generate large files it is also important to consider the storage requirements. Both the type of storage and speed of file transfer should be considered. For example, multimedia files that require regular updating during their production could be usefully stored on a fast rewriteable CD during that process. This would allow the user/editor fast access and backup facilities.

8.2.5 Operating system

Computers are designed to run a wide variety of applications but you must check the application you want installed on your computer does not conflict with your computer's operating system.

An operating system can be vulnerable to a wide variety of software incompatibilities. The sheer number of operating system versions can cause problems. For example, a program written for an OS from 1995 may not work properly on a computer that is using a more up-to-date version of the OS. Device drivers are another area requiring checks.

Device drivers are programs that allow a device to communicate with the software that uses the device. Most drivers come with the operating system but some may need a new driver installed when you connect the device to your computer. Problems with drivers arise when computer manufacturers use different hardware than others and therefore install different drivers. This may result in a particular software application running on one machine and not another.

Your own system

Answer the following questions about the system you are using:

- What operating system(s) are installed?
- What is the available RAM?
- What is the processor speed?
- What is the disk capacity?

- What peripherals are available?

Using recent computer magazines or software documentation, select some packages and consider their suitability to your system. Use the same questions and match the package to your system, i.e.:

- What operating system is required?
- What RAM is required?
- What processor speed is required?
- How much disk space is required?
- What peripherals are required?

When selecting software packages try to include a specialised package as well as general purpose packages. For example, would your system be able to run a sound editing package? If not, what would have to be added or altered in your system to make this possible?

If you have access to another system with a different specification then repeat this exercise with the same, or other software packages.

8.3 Viruses

A virus is a program code that causes some undesirable and unexpected event to happen in a computer. Viruses are usually disguised as something innocent and are designed so that they automatically spread within or between computer systems. Viruses can enter a system as attachments to an e-mail, a download from the web, or from on a disk or CD. Some viruses take effect as soon as their code is executed and others can wait until circumstances cause their code to be executed by the computer. Viruses can be quite harmful and erase data or close down a system.

8.3.1 Virus types

Viruses are classed by three main types:

- File virus;
- Boot sector virus;
- Macro virus.

8.3.1.1 File virus

File viruses attach themselves to program files such as .exe or .com files. When the program is loaded, the virus is also loaded.

A file viruses can also take the form of a complete program, or script, attached to something else, e.g an e-mail. They then take up residence in the computer ready to cause havoc.

8.3.1.2 Boot sector virus

These viruses infect executable code found in certain system areas on a disk. They attach to the boot sector on disks or the master boot record on hard disks. To infiltrate the boot sector, the virus is read while the system is running and then activated the next time the operating system is loaded.

8.3.1.3 Macro virus

Macro viruses are fairly common viruses, but tend to do the least damage. Macro viruses infect applications and typically cause a sequence of actions within the application e.g. inserting unwanted words or phrases in a document.

8.3.2 Virus code actions

Viruses don't all follow the same course of action. They can, and do, use a combination of the following actions:

- Replication;
- Camouflage;
- Watching;
- Delivery.

8.3.2.1 Replication

A computer virus has the unique ability to replicate. Like a biological virus they can spread quickly and can be difficult to control. They can attach themselves to almost any type of file and spread as files are copied and sent between computer users. A virus can take a long time to replicate itself before activation. This gives it time to be spread over many computers before being discovered.

8.3.2.2 Camouflage

It is possible for a virus to avoid detection by taking on the characteristics that detection software is programmed to look for and ignore. However, detection software has evolved to prevent this happening.

Today's anti-virus software does much more than simply check particular characteristics (or signature) of a virus. They also check the virus code and even checksum the virus code to identify it. With these cross-checks it would be extremely difficult for a virus to camouflage itself and get past detection.

8.3.2.3 Watching

A virus can lie in wait and ambush a computer when something routine is carried out e.g. opening a particular application. The damage routines will activate when certain conditions are met. On a certain date, or when the infected user performs a particular action may trigger the virus.

8.3.2.4 Delivery

Infected disks brought in from the outside used to be the main source of viruses until e-mail provided the ideal delivery vehicle.

Once delivered the virus will wait for the trigger to drop its payload.

8.3.3 Other infections

8.3.3.1 Worm

A worm is a self-replicating virus that does not alter any files but takes up residence in the computer's active memory and duplicates itself. They only become noticeable once their replication consumes the memory to the extent that the system slows down or is unable to carry out particular tasks. Worms tend to use the parts of the computer's operating system that is not seen by the user until it is too late.

8.3.3.2 Trojan horse

A Trojan horse is a program where harmful code is contained inside another code which can appear to be harmless. Once the apparently harmless code is in the computer, it releases the malicious code to do its damage. Trojan horses may even claim to be anti-virus in order to get the user to install it.

The name comes from the deception that the Greek army played on the people of Troy during the Trojan War. They presented Troy with a large wooden horse in which they had secretly hidden their warriors. Once inside the city gates, the warriors emerged from the horse and took control of the city.

8.4 Anti-virus techniques

The best protection against a virus is to know that each file you open from an e-mail, disk or from the web is free from any virus. This requires anti-virus software that can screen e-mail attachments, web downloads, and checks all of your files from time-to-time removing any viruses that are found.

Techniques used by ant-virus software to detect a virus include:

- Checksum;
- Signature;
- Heuristic;
- Memory monitoring.

8.4.1 Checksum

In checksum detection, a checksum of key files (boot an executable) is recorded at source and stored in the system. When these files are called to execute the checksum is calculated and compared with what it should be. If there are any anomalies, then the file about to be run could have be infected and a warning given.

There is also the facility to recalculate the checksum when boot or executable file are legitimately altered by the user

8.4.2 Virus signatures

A virus signature is a unique pattern of bits within a virus. It can be used to detect and identify specific viruses. Once known, the anti-virus software uses the virus signature to scan for the presence of malicious code and removes it.

8.4.3 Heuristic detection

Heuristic detection describes the technique of approaching a problem through past knowledge. The technique is used to find unknown viruses that have not yet been identified by their signatures by looking for characteristics in files that have previously been associated with a known virus.

Heuristics can also detect a virus that has disguised its signature, by recognising a particular characteristic of the virus behaviour.

Heuristic techniques work on the probabilities of a file being infected i.e. it will give a percentage likelihood of a file being infected by weighing up the file behaviour. For example, if a file attempts to access your address book then that might be suspicious. If the same file includes code that checks a date, then the suspicion rises. There will come a point when a warning is issued on the possibility of a virus.

8.4.4 Memory resident monitoring

Some anti-virus software is memory resident and is loaded on start up. It actively monitors the system for viruses whilst the computer is switched on and checks programs for infection every time they run. This will include the boot files on start up, checking any disk as it is accessed, checking any files accessed during operation and checking any files being loaded on the hard drive.

The price to be paid with memory resident programs is that they can cause delays in program loading and execution whilst the checks are being carried out.

8.5 Using a virus information library

Using a virus information library

Literally thousands of viruses have been detected and catalogued using a Virus Information Library (VIL). One such VIL can be found at:

http://vil.nai.com/vil/

Using this or another source, find the details of at least one of each type of virus:

- file
- boot
- macro
- worm
- trojan

For each one, make a note of the following:

1. name:
2. type:
3. symptoms:
4. date discovered (try to find one from this year):
5. medium for infection (e.g. email, website):
6. cure (if any):

Another probem with viruses is that some of them are hoaxes (and some of the hoaxes are themselves hoaxes). Include a hoax (or a virus disguised as a hoax) in your list of virus types and note its details.

8.6 Summary

The following summary points are related to the learning objectives in the topic introduction:

- Software and hardware compatibility issues including memory, storage and operating system;
- Classification of viruses by file type;
- Descriptions of virus code actions;
- Distinction between virus forms;
- Protection against viruses.

8.7 End of topic test

An online assessment is provided to help you review this topic.

Glossary

Accumulator

A special purpose *register* used by the Arithmetic Logic Unit to hold the accumulated results of calculations

Accuracy

A measure of how well a computer representation of an image, sound etc. matches the original

ADC

Analogue to digital converter used to generate a digital signal from analogue input

Address Bus

Used to pinpoint a memory location

Adjustable-Split Keyboard

A three-piece folding keyboard organised into three adjustable sections

ADSL

ADSL stands for Asymmetric Digital Subscriber Line - a modem technology which converts existing twisted-pair telephone lines into access paths for multimedia and high-speed data communications. ADSL can transmit up to 6Mbps to a subscriber, and as much as 832Kbps or more in both directions.

Adware

Adware is software which is free to use but which pays for itself by presenting the user with advertising material.

ALU

The Arithmetic and Logic Unit, ALU, is part of the CPU. It carries out computation.

Arithmetic Logic Unit

The heart of a computer where data is processed and manipulated

ASCII

American Standard Code for Information Interchange

Asynchronous Transmission

A character is sent as soon as it is available rather than using a clock pulse to synchronise transfer

Avi

Audio Video Interleaved

Bandwidth

Bandwidth describes the amount of data which can be transmitted via a network connection, usually measured in bits per second.

Bi-Directional

Transfers data in both directions

Binary number system

A system in which numbers are represented using only the digits 0 and 1.

Bit

A single unit of binary data.

Bit-depth

The number of bits used to represent each pixel.

Bit map

A representation of image data where each bit corresponds to an individual pixel on the screen

Bluetooth

Bluetooth is a wireless networking protocol which allows devices to connect together to provide services like Internet access, printing and data transfer without the intervention of the user.

Boolean

Two-state algebra developed by George Boole.

Boot Program

A small program held in ROM that is executed when the computer is powered up

Broadcasting

Broadcasting over a network uses the User Datagram Protocol which does not require an acknowledgment. This is a more efficient use of bandwidth than individual downloads.

Browser

Software used to render HTML pages

Bubble Jet Printer

Operates by heating the ink until it forms a vapourised bubble which is then squirted as a tiny droplet of ink onto a page

Buffer

A part of memory used to hold data when communicating with a peripheral. Used when a fast acting part of the system (CPU) is exchanging data with a slow acting device (printer)

Bus Network

Each device on the network is directly connected to a single communications line

Byte

Eight units of binary data (bits).

Cache Memory

A small amount of random access memory that sits between the processor and RAM in order to speed up data transfer

Caching

A cache is a local copy of data available over a network. When for example a web page is requested, the network software retrieves the page, but also saves it locally. When the page is requested again, it will make it available from the cache, thus speeding up the process of viewing the page. This only works for web pages which do not have dynamic data on them or which have not altered recently. The system works better for graphics and other elements of a web page which do not change very often.

Capacity

The amount of data that can be stored

CCD

Charge-Coupled Device used to record light intensity. CCDs form the heart of scanners and digital cameras

Central Processing Unit

The (CPU) coordinates and controls the activities of all other units in the computer system. It executes program instructions and manipulates data in accordance with the instructions

Character Printer

A device which prints characters one at a time. Examples include the daisy wheel or golf ball

Character Set

The set of characters that can be represented and displayed by a computer

Chatrooms

Chatrooms are areas on the Internet where users can communicate using text messages in real time. Chat servers can be accessed using an IRC client which allows you to select different discussion areas which you can join or leave.

Client/Server

A network model where data is held and controlled centrally on a file server and accessed by individuals on a workstation

Clipping

Reducing the amplitude of a sound sample to fall within a given sample range

Coercivity

A property of magnetic storage devices which describes how easily magnetisation can be switched from one orientation to another

Collision

A collision occurs when two devices transmit simultaneously on an Ethernet network.

Control Characters

Special non-printing characters in a character set, used for special purposes, e.g. carriage return and end of file

Control Unit

Includes timing/control logic and an instruction decoder. It sends signals to other parts of the computer to direct the fetch and execution of machine instructions

CPU

The CPU, or Central Processing Unit. This is where instructions are processed and computations are carried out. This is the control centre of the computer.

CU

The Control Unit, CU, is part of the CPU. It exerts overall control over the operation of the CPU.

Curie Temperature

The point at which magnetic material ceases to be a permanent magnet

Customised Keypad

A specially designed input device that can be programmed

Cycle Time

The cycle time is the time between clock pulses.

Cylinder

A collection of tracks on each platter that are equidistant from the spindle

Data Bus

Used to transfer data to and from the CPU. The data bus can be common to devices and main memory allowing transfers to take place from and to peripherals or from and to main memory

Data compression

A technique of reducing the storage space occupied by a file

Data encyption

A method of securing transmitted data through encoding

Data Overrun

This occurs when the CPU ignores an interrupt request from a peripheral. The interface merely discards the byte and informs the CPU that data has been lost

Direct Access

Referred to as random access. The retrieval of any disk data given the sector location

Disk Based OS

An operating system that is distributed on and loaded from disk

DRAM

Dynamic Random Access Memory. Slower, volatile store that needs to have a continuous signal to refresh the contents of the chip

EEPROM

Electrically Erasable Programmable Read Only Memory can be selectively reprogrammed

Electronic Forum

An electronic forum is a web based discussion area which require users to join before they can post a message or query. They are often used by software distributors to provide technical support and feedback for their products.

Email

Email is system of communication which allows users on a network to send text messages and attached files to another user. Email addresses are usually of the form: username@domainname.

Email Clients

An email client is a piece of software which can be used to connect to an email server and retrieve electronic mail.

EPROM

Electrically Programmable Read Only Memory can have the contents erased and replaced with new data by the user. Data is erased by shining ultraviolet light onto the chip

Ethernet

A popular local area network commonly using coaxial cable with transmission rates of 10 or 100Mbps

Exponent

Represents the range of a decimal number

External Memory

Holds quantities of data too large to store in main memory. It is also used to keep a permanent copy of programs and data

Fetch-Execute Cycle

The repeated process of fetching instructions from main memory, decoding the instructions and executing them until an instruction to HALT is encountered

File Server

Provides centralised storage and resource management for users of a network

File servers

A file server is a dedicated machine on a network which controls access to resources on a network. A file server will usually be running a network operating system and have enough disk space to give every user on the network space to store files.

Firewall

A firewall is software or hardware which protects a local area network from outside access by monitoring and blocking network traffic.

Flat-bed Scanner

An input device similar to a photocopy machine where documents are scanned, face down on a glass surface

Flop

Floating point operations per second

Freeware

Freeware is software which is distributed free by the programmer, or is a cut down version of a commercial product which is distributed free in the hope that users will purchase the full version.

General Purpose Package

An application package that provides more than a single type of application to solve specific problemse.g. word processing, database, spreadsheet, etc

General Purpose Registers

A set of registers internal to the CPU whose role is not defined at the time of manufacture. Programmers may use these registers as appropriate within their programs

HTML

Hypertext Markup Language

Hub

A hub is a multi-port repeater in an Ethernet network. Hubs are used to distribute a network connection to a number of machines in a room or an area of a building. Hubs typically have 12 or 24 ports.

Hypermedia

When data elements are linked to form a structure through which the user can navigate

Instruction Register

Used to hold the current instruction that is being executed

Integrated Keyboard and Touchpad

A keyboard containing both keys and a touchpad

Interactive Multimedia

Multimedia presentation where the end user can control what is being displayed by activating buttons and hyperlinks on a page

Inter-block Gap

Used to describe the empty space between data blocks on a tape device

Interface

A unit that sits between the CPU and a peripheral device and compensates for the differences in speed, codes, etc

Internal bus

The medium for communicating data and control signals between the component parts of the CPU.

Internet

The Internet is a global internetwork consisting or millions of computers connected together using a variety of high speed communications systems. Home users connect to the Internet using the telephone system.

Internetwork

An internetwork is two or more computer networks connected together.

Intranet

An intranet is a private network which provides similar services within an organisation to those provided by the Internet outside it but which is not necessarily connected to the Internet. An intranet is often used for distribution of information within the organisation. Some organisations give limited access to their intranets to other organisations or the general public. This is known as an "extranet".

IR

The Instruction Register, IR, is a register in the CPU. It holds the instruction currently being executed.

ISDN

ISDN stands for Integrated Services Digital Network. With ISDN, voice and data are carried by bearer channels (B channels) occupying a bandwidth of 64 Kbps (Kilo-bits per second). Some switches limit B channels to a capacity of 56Kbps. A data channel (D channel) handles signaling at 16 Kbps or 64 Kbps, depending on the service type.

Keyboard Controller

A special integrated circuit that is used to receive and interpret keyboard scan codes

LINUX

LINUX is an open source operating system developed by Linus Torvalds.

List-Server

A list-server is an email based electronic conference. Any message sent to the list server will be automatically emailed to everyone on that list.

Local Area Network

A Local Area Network (LAN) is a number of computers connected together within a single building or organisation. Local area networks tend to be characterised by high bandwidth, low error rates and short distances between computers.

Logic gates

A device that performs a logical operation upon its input signals to produce its output signals.

Machine Code Program

A binary code program that is executable by a specific computer processor

Machine Cycle Time

The cycle time for one fetch-execute-cycle, sometimes expressed in MIPS, *millions of instructions per second*

Main Memory

(RAM) stores programs and data while the computer is operating. It is organised so that data can be both read from and written to it. It is a *volatile* store that loses its contents when the machine is switched off

Mantissa

A non-zero value used to represent the precision of a decimal number

MAR

The MAR, or Memory Address Register, specifies the address in memory for the next read or write operation from or to memory.

Matrix Printer

A device which forms characters or graphics out of ink dots

MDR

The MDR, or Memory Data Register, contains the data to be written to memory or receives the data read from memory. The MDR is sometimes known as the Memory Buffer Register (MBR).

Memory Address Register

An internal register of the CPU that is used to hold the address of a location in main memory

Memory Buffer Register

Used to hold data that has just been sent to or from the CPU

Memory Upgrade

Expanding physical memory by adding more RAM modules.

Metropolitan Area Networks

A network which interconnects buildings or other facilities extending over a city-wide area

MIPS

MIPS stands for Millions of Instructions Per Second. This is used as a performance measure.

Modified Keyboard

Keyboards that have been redesigned to suit the needs of the workplace or to improve the health and safety of users

MPEG

A video data standard derived by the Motion Pictures Expert Group

Multiscan Monitor

Display technology able to deal with different scanning frequencies

Multi-tasking

Where several applications are open at the same time, and the user can switch easily between them.

NetBEUI

NetBIOS Extended User Interface (NetBEUI) is a network protocol developed by Microsoft and originally used with LAN Manager.

Network

A network is one or more computers connected together in a way which allows them to communicate or share data and resources.

Network Interface Card

A Network Interface Card (NIC) is an interface fitted inside a personal computer or network terminal which allows it to communicate with other machines over a network. The card technology will vary according to the network used, but every card on a network must have some way of uniquely identifying itself and some means of converting the signals form the computer to a form which can be transmitted over the connection.

Network Manager

Person who controls access to services and shared resources on a local area network.

Network Operating System

An operating system that that manages access to data and resources on a networ, ensuring the privacy and security of data.

Network Topology

The arrangement of computers to form a network

Newsgroups

Newsgroups are electronic bulletin boards for text based discussion on any subject. Un-moderated newsgroups allow any user to post a message. Moderated newsgroups have an administrator who filters messages before they are posted.

News Server

A News Server is a computer which stores and forwards Usenet messages.

Non-Volatile

A characteristic of a memory chip which keeps its contents when power is removed

Page Printer

A device which generates an entire page image. The data is sent to the printer as an entire page which requires it to have a fairly large internal memory. Example include the laser printer.

Parallel Transmission

Each bit of an 8-bit byte is transmitted along a set of parallel wires at the same time

Parity Bit

An additional bit that is transmitted as part of a byte. The parity bit is altered to reflect *even* or *odd* parity. Data transmitted can be checked using the parity bit to ensure there are no errors during transmission

PC

The Program Counter, PC, is a register in the CPU. It holds the address in memory of the next instruction to be executed in the program.

Peer to Peer

A network model where all computers on the network are equal and data may be shared from computer to computer. Any node may set itself up as a file server

Pit

Pits are areas burned on a CD-ROM by a laser - typically 0.5 microns wide and 0.83 to 3 microns long

Pixel

A picture element which is the smallest display element

Pixel Replication

The replication of pixels in the x and y direction by a scale factor to produce an enlarged image

Process

A program in execution.

Program Counter

An internal register of the CPU used to hold the address of the next instruction to be executed

Proxy Server

A Proxy Server is a machine which receives requests for internet pages and forwards them. A proxy server can provide a LAN with a single point of access to the Internet and can act as a filter to block access to unsuitable material.

Quantisation

The rounding of sound samples to the nearest integer

QuickTime

An Apple standard for storing and playing video data. Provides a methodology for interleaving audio data with video data.

RAID

A Redundant Array of Inexpensive hard Disks (RAID) is a method of providing fault tolerant hard disk storage so that if one disk fails, the operating system can recover the data from the other disks and recreate the failed hard disk image when it is replaced.

Raster Graphics

A method of producing an image on a display screen by illuminating horizontal lines of dots on the screen

Redirector

A redirector is a LAN device driver which translates operating system requests into network events and transmits them to the right protocol stack.

Removable Storage

A storage device that is external to the computer and can be disconnected to facilitate data transfer to another machine

Repetitive Strain Injury

A disorder that is caused by awkward posture or movement when using a keyboard

Resolution

The total number of pixels in the width and height of an image

Resolution Independent

A graphical representation that is independent of the display resolution

Ring Network

Each device is connected on the network to a ring communications line around which signals are sent

ROM based OS

An operating system program stored on a ROM chip.

Rotational Speed

The speed at which a disk rotates. Measured in *revolutions per minute* (rpms)

Router

A router is a computer with two network cards which is responsible for routing data from one network to another. A router keeps track of the IP addresses of the computers on the networks connected to its network interface cards and directs IP packets appropriately. A router functions at the Network layer of the OSI model.

Sample Size

The amount of data that is stored per sample

Sampling

An analogue sound value measured every *nth* fraction of a second and stored digitally

Sampling Rate

The number of sound samples that are taken per second

Scan Code

A code that is generated when a key is pressed on the keyboard

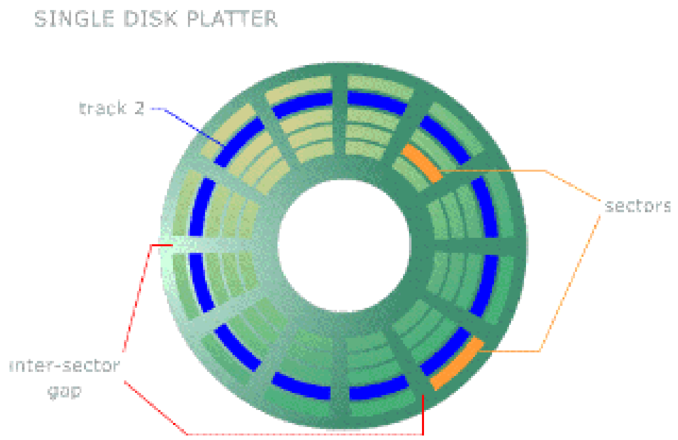
SCSI

Small Computer Systems Interface

Search Engine

A search engine is a service provided on the Internet which indexes web pages. When you enter a term in a search engine it will respond with a list of all the web pages it has indexed where that term occurs. Using a search engine takes practice as it is important to choose the term you search for with care so that the search engine returns a usable number of "hits".

Sector



A simplified representation of data storage is to view the surface of the disk as having 1s, representing a "north" magnetisation and 0s, representing a "south" magnetisation.

Strictly speaking, the data is compressed using *run-length limited* (RLL) encoding, but this method and its purpose are beyond the scope of this topic.

The smallest addressable portion of a track and the smallest unit of data that can be read or written

Serial Transmission

Each bit of the byte is sent out, one at a time over the communications line

Shareware

Shareware is software which requires you to pay a fee to the distributor if you want to continue using it after a certain period of time. Sometimes the trial period is enforced by the software ceasing to function after the period has expired, sometimes the decision is left to the user's conscience.

SHTTP

Secure Hyper Text Transmission Protocol (SHTTP) encrypts data transmitted over the Internet. It is often used to allow users to purchase items over the Internet using e-commerce, or to transfer sensitive information.

Sign and Magnitude

The most sigificant bit of the word is used to represent the sign of the number with the remaining bits representing the magnitude

Skewing

Occurs in parallel transmission where the individual bits of a byte arrive may at their destination at different times

Spam

Spam is a term used to describe electronic junk mail.

Spooling

The temporary storage of input or output data on magnetic disk or tape. Preferred method used when a peripheral is shared across a network or when large data files are being transferred

SRAM

Static Random Access Memory. Very fast *volatile* store suited to cache memory

Star Network

All devices on the network connect though a central connection point or controller

Start Bit

Marks the beginning of a character in asynchronous transmission

Stop Bit

Marks the ends of a character in asynchronous transmission

Stored Program

A series of machine instructions that are held in main memory

Stored program concept

The idea that the sequence of instructions to solve a problem should be stored in the same memory as the data. This concept was proposed by John Von Neumann in 1945.

Switch

A switch is a device to divide an Ethernet network up into separate collision domains. A switch keeps a database of the address of each machine on the network and only transmits a frame of data for a machine on to the segment of cable which that machine is connected to.

Synchronous Transmission

The transmission of data between two devices is timed to a clock pulse

System bus

The medium for communicating data and control signals between the main components of the computer.

Tele-working

Tele-Working is the ability to work from somewhere geographically separate from the company you work for.

Token

A signal that passes around a network and is a carrier for a data packet

Track

A circular section of a disk that is divided into equal-length sections called sectors

Two's Complement

A representation of negative integers that is formed by changing each 1 bit to a 0 and each 0 bit to a 1 and then adding 1

Two-State Machine

Electronic components of a computer that can be in one of only two states. Binary digits 0 and 1 are used to represent these two states

UART

Universal Asynchronous Receiver/Transmitter

Unicode

A 16-bit symbol representation system

Uni-Directional

Transfers data in one direction only

UNIX

UNIX is a network operating system originally developed by AT Bell laboratories.

URL

Uniform Resource Locator

UTP

Unshielded Twisted Pair (UTP) cabling is used in 10baseT and 100baseT Ethernet installations. It consists of thin cables twisted together to avoid interference created by electrical induction. UTP cabling is classified according to the data transfer rate it can support and its immunity to interference.

Vector Graphic

A mathematical representation of a graphical object

Video Conferencing

Video Conferencing is a system using video cameras, and a high bandwidth network connection such as an ISDN line which allows a number of people to communicate with each other using sound, video and to share data such as text and graphics even though they are thousands of miles apart.

Video Controller

Special circuitry that is used to generate the signals needed for a video output device to display data

Video Digitising

The conversion of analogue video data to a digital represenatation

Virtual Memory

Part of the hard disk is used as if it was main memory.

Volatile

Refers to a type of memory which loses its contents when power is no longer supplied to the chip

Von Neumann architecture

Von Neumann architecture is based upon the basic idea that the sequence of instructions to solve a problem should be stored in the same memory as the data.

VRAM

A separate memory, operating at high-speed used to hold screen data that is to be displayed

Web Server

A web Server is a piece of software running on a machine on a network which provides which sends out web pages in response to requests from Internet browsers.

Wide Area Network

A Wide Area Network (WAN) is a number of computers connected together which are geographically remote. Wide are networks tend to be characterised by low bandwidth, high error rates and long distances between computers. A wide are network may not be controlled by any one organisation.

Wide Area Networks

A network linking machines worldwide

Windows 2000 Server

Windows 2000 Server is a network operating system developed by Microsoft as a replacement for Windows NT.

Windows NT

Windows New Technology (NT) is a network operating system developed by Microsoft.

Wireless Application Protocol

Wireless Application Protocol (WAP) is a protocol which runs on mobile phones and provides a universal open standard for bringing Internet content to mobile phones and other wireless devices.

Answers to questions and activities

1 Data Representation

Revision (page 3)

Q1: a) 01001001

Q2: Code

Q3: 200 x 600 pixels requires 128 0000 bits of memory

= 128 000/8 = 16 000 bytes

= 16 000/1024 = **15.6 Kbytes**

Decimal to binary conversion - a ready reckoner (page 7)

Q4: 10000110

Q5: 10010100

Q6: 110001010

Using both methods to convert decimal to binary (page 7)

Q7: 11101

Q8: 10010

Q9: 1001111

Q10: 100010001

Q11: 1111111

Q12: 1011100110

Q13: 111110110111

Q14: 10011000011011

Binary to decimal conversion - practice (page 9)

Q15: 11

Q16: 429

Q17: 275

Q18: 127

Q19: 128

Q20: 1489

Q21: 918

Q22: 1573

Answers from page 11.

Q23: Because it is easy and cheap to manufacture two-state devices.

Q24:

A	B	A OR B
0	0	0
0	1	1
1	0	1
1	1	1

A	B	A NAND B
0	0	1
0	1	1
1	0	1
1	1	0

Q25: As an 8-bit pattern of 0s and 1s.

Answers from page 12.

Q26: b) 0..255

Q27: c) 0..4095

Q28: a) 0..2^n-1

Answers from page 13.

Q29: 10110

Answers from page 13.

Q30: 1011101

Q31: 10111

Q32: 010111

Converting 8-bit binary numbers to Two's Complement - practice (page 16)

Q33: 11010110

Q34: 10001000

Q35: 11111000

Q36: 11001001

Q37: 01010111

Q38: 00000001

Q39: 10111010

Q40: 01101110

Answers from page 16.

Q41: d) all of the above

Answers from page 17.

Q42:

- 00001111
- 11101011
- 11001010111

Q43: There are 2 possible representations of 0 and handling arithmetic is more complex due to the need to take into consideration the sign bit.

Q44: The mantissa affects the precision of the number, while the exponent affects the range.

Q45: Negative numbers are represented using Two's Complement. The positive value is converted to binary, each bit is inverted and the value 1 is added to produce the final result. Alternatively sign and magnitude can be used to represent negative integers, although this has problems in representing the value 0 twice and in the need to consider the sign bit when carrying out arithmetic operations.

Q46:

- 00000011
- 00000100
- 00001110

Q47: 0.101

Q48: 0.0011

Q49: 0.0001100110

Q50: 0.75

Q51: 0.3125

Answers from page 19.

Q52: c) 128 symbols

Q53: c) an extra bit is used to provide encoding of 256 different character symbols

Q54: b) a way of encoding symbols using a 16-bit representation.

Q55: Neither ISO-Latin-1 or ASCII can accommodate international language symbols.

Q56: One example that would need a Unicode representation is Japanese. Others include Chinese, Korean, Arabic, Armenian, Gujarati and Bengali.

Calculating memory requirements - 1 (page 23)

Q57: 16000

Q58: 60000

Q59: 98304

Calculating memory requirements - 2 (page 24)

Q60: 128000

Q61: 480000

Q62: 786432

Calculating memory requirements - 3 (page 25)

Q63: 384000

Q64: 1440000

Q65: 2359296

Answers from page 27.

Q66: b) A vector graphic is a mathematical description of the objects that make up the image

Q67: 60.75

Q68: 65536

Q69:

Assumptions made are that there is no compression and no palette restrictions, therefore:

Each dot (or pixel) requires 24 bits or 3 bytes

There are 300 x 5 x 300 x 4 dots on each frame

Each frame requires 300 x 5 x 300 x 4 x 3= 5 400 000 bytes storage

There are 25 x 60 frames in the 1 minute clip giving 1500 frames in total

Total storage is 1500 x 5 400 000 requiring 8 100 000 000 bytes of storage

Or 8 100 000 000/ 1024 x 1024 Mb

Equals 7724.76 Mb Or 7.54 Gb

Hence the need for compression!

2 Computer Structure

Revision (page 31)

Q1: ROM, or Read Only Memory

Q2: d) Memory

Answers from page 33.

Q3: The Analytical Engine was controlled by a set of instructions entered as punch holes on a set of metal cards. A different program could be carried out by altering the pattern on the metal cards. This was the idea behind a general purpose machine.

Q4: It relates to modern compters as it decribes a machine that can execute a stored program in a finite series of steps, which is what computers do.

Q5: Vacuum tube technology.

Q6: Circuits for many operations were incorporated into a single chip. This made possible the personal computer.

Answers from page 36.

Q7:

- CPU
- main memory
- external memory,
- peripheral devices
- the system bus

Q8: To allow communication between the component parts of the computer. Carries data, memory addresses and control signals.

Q9: The stored-program concept. Made the computer a general-purpose problem-solving tool - load a different program to solve a different problem.

Q10: The main memory holds programs and data.

Main memory is volatile (information disappears when power is switched off) and has not sufficient capacity, hence cannot be used for long-term storage of large amounts of information. Thus external memory, which is non-volatile and large capacity, e.g. computer hard discs.

Q11: It would be transmitted via the system bus to the main memory and loaded from there via the data part of the system bus to the CPU.

Q12: The instruction register and the program counter.

The program counter holds the address in memory of the next instruction to be executed, and the instruction register holds the instruction currently being executed.

Answers from page 40.

Q13: a) machine code is represented in binary

Q14: A stored program is a series of machine instructions that are loaded into main memory from where they are fetched decoded and executed by the CPU.

Q15: c) fetches, decodes and executes machine instructions

Q16: The fetch-execute cycle describes how machine instructions, once loaded into main memory, are repeatedly fetched, decoded and executed, one machine instruction at a time, until an instruction to HALT is encountered.

Q17: c) binary cannot be used to represent a two-state device

Identifying system components used in a task (page 42)

Your discussions may have identified the following:

- **Step 1.** Double click on the WP icon which causes the application to load from **external memory** (hard disk) to **main memory** (RAM).
- **Step 2.** The **CPU** starts to execute the program in RAM and a new document window appears on the **output device** (display screen).
- **Step 3.** Text is entered via an **input device** (keyboard) and transferred to main memory where it is stored and subsequently transferred to the display screen.
- **Step 4.** Editing of text is carried out, whereby the WP application program executed by the **CPU** manipulates data in **main memory**.
- **Step 5.** The file is saved to external storage by transferring the contents of main memory to the hard disk.

Answers from page 44.

Q18: a) RAM is volatile

Q19: The term non-volatile refers to memory that retains its contents when power is not supplied to the chip i.e. the contents are permanent.

Q20: b) EEPROM

Q21: EPROM chips can be reprogrammed by erasing the original contents using ultraviolet light. The main disadvantage of these chips is that the entire contents are erased even when only a small portion needs to be reprogrammed.

Q22: a) RAM cannot be written to

Describing memory maps (page 47)

Q23: Ask your teacher/lecturer to check your answer

Q24: Ask your teacher/lecturer to check your answer

Answers from page 53.

Q25: The CPU coordinates and controls the activities of all other units in the computer system. It executes program instructions and manipulates data in accordance with the instructions.

Q26: b) RAM

Q27: The control unit generates a pulse at a constant frequency. On each pulse, a machine operation is carried out.

Q28: b) successive addition

Q29: General purpose registers are provided within the CPU for programmer use.

Answers from page 55.

Q30: c) 32

Q31: b) 256

Answers from page 58.

Q32: b) carry a memory address from which data can be read or to which data can be written

Q33: The address bus is described as unidirectional as it carries information in only one direction, i.e. from the processor to main memory.

Q34: c) to carry data/instructions from main memory to CPU or to carry data from CPU to main memory

3 Computer Performance

Revision (page 61)

Q1: b) clock

Q2: d) Microcomputer

Q3: c) Clock speed

Answers from page 73.

Q4: The clock rate is not a measurement of the number of instructions that are carried out per clock cycle. A computer may execute twice the number of instructions in 200 clock cycles as another. Other factors such as data bus width, instruction size etc. must also be taken into consideration.

Q5: b) using the hard disk as an extension of main memory

Q6: An increase in the width of the data bus can improve system performance as more data can be transferred to the CPU from memory and vice versa at a time.

Q7: c) 2^{24}

Q8: The width of the address bus defines the number of main memory locations that can be directly addressed. If this is increased then physical main memory can be increased. This will reduce the need to use the hard disk as a virtual store, improving memory access times.

Q9: b) Insertion of additional memory modules

Q10: By raising 2 to the power of 16.

4 Peripherals

Revision (page 77)

Q1: a) DVD

Q2: d) LASER

Q3: c) buffer

5 Networking

Revision (page 105)

Q1: d) LAN

Q2: a) distance

Q3: b) spamming

6 Using Networks

Revision (page 129)

Q1: c) e-commerce

Q2: d) teleworking

Q3: a) Copyright, Design and Patents Act

7 Computer Software

Revision (page 149)

Q1: d) ASCII

Q2: a) operating systems

Q3: c) software to control a computer's resources

Answers from page 173.

Q4: b) The combined use of video, text, sound and graphics

Q5: With interactive multimedia the user provides input to the program that alters the actions that it takes. For example, moving onto a particular page, or displaying additional information when a button is pressed. Whereas with animations the user simply watches a multimedia presentation and has no control over how it behaves.

Q6: b) Teaching youngsters the Green Cross Code.

Q7: Any of the following features are possible answers:

- ability to import media types;
- editing, particularly text and static images;
- organising navigation;
- programming - visually using icons and objects;
- scripting language;
- playback;
- delivery by building a run-time version of the project that does not need the full authoring software to execute.

8 Supporting Software

Revision (page 181)

Q1: a) a disabling code that infects a computer

Q2: b) an e-mail attachment

Q3: d) install anti-virus software