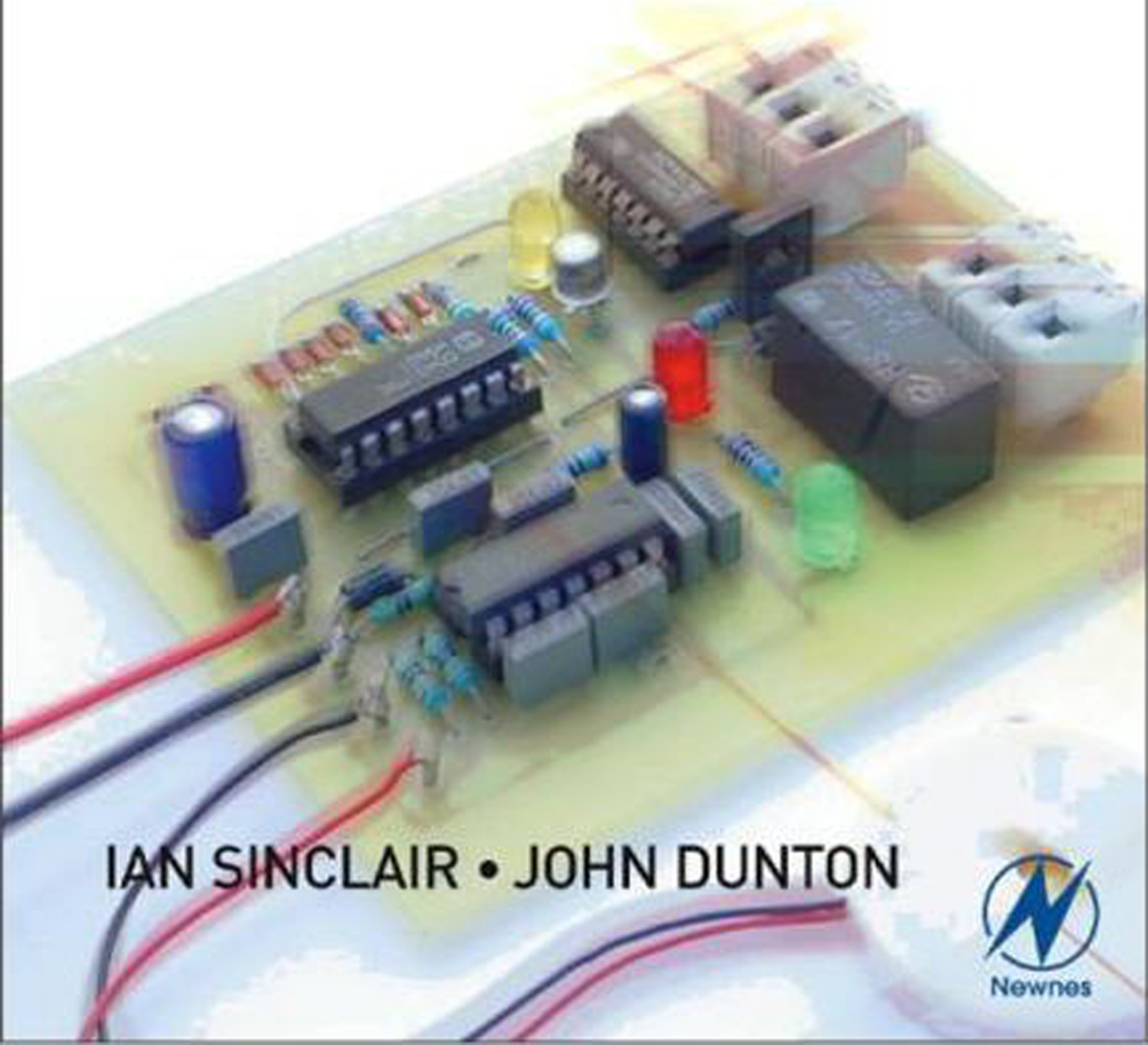
SIXTH EDITION

PRACTICAL ELECTRONICS HANDBOOK



IAN SINCLAIR • JOHN DUNTON

Newnes

Practical Electronics Handbook

This page intentionally left blank

Practical Electronics Handbook

Sixth Edition

Ian R. Sinclair and John Dunton

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK
OXFORD • PARIS • SAN DIEGO • SAN FRANCISCO
SINGAPORE • SYDNEY • TOKYO
Newnes is an imprint of Elsevier

Newnes is an imprint of Elsevier
Linacre House, Jordan Hill, Oxford, OX2 8DP
30 Corporate Drive, Burlington, MA 01803

First edition 1980
Reprinted 1982, 1983 (with revisions), 1987
Second edition 1988
Reprinted 1990
Third edition 1992
Fourth edition 1994
Reprinted 1997, 1998, 1999
Fifth edition 2000
Reprinted 2001
Sixth edition 2007

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

Cover photo by Thomas Scarborough, reproduced by permission of Everyday Practical Electronics. www.epemag.co.uk

ISBN 13: 978-0-75-068071-4
ISBN 10: 0-75-068071-7

Typeset by Cepha Ltd
Printed and bound in Great Britain

07 08 09 10 11 10 9 8 7 6 5 4 3 2 1

CONTENTS

Preface

Component data books are often little more than collections of specifications with little or nothing in the way of explanation or application and, in many cases, with so much information crammed into a small space that the user has difficulty in selecting what is needed. The cost of publishing paper data books, the rate that new products are being brought to market and the ease with which electronic copies of data sheets can be distributed by e-mail or downloaded from websites has begun to deter manufacturers from printing data books at all. This book, now in its sixth edition, has been extensively revised, with a large amount of new material added, to serve the needs of both the professional and the enthusiast. It combines data and explanations in a way that is not served by websites.

Although the book is not intended as a form of beginners' guide to the whole of electronics, the beginner will find much of interest in the early chapters as a compact reminder of electronic principles and circuits. The constructor of electronic circuits and the service engineer should both find the data in this book of considerable assistance, and the professional design engineer will also find that the items brought together here include many that will be frequently useful and which would normally be available in collected form in much larger volumes.

The book has been designed to include within a reasonable space most of the information that is useful in day-to-day electronics together with brief explanations which are intended to serve as reminders rather than full descriptions. In addition, topics that go well beyond the scope of simple practical electronics have been included so that the reader has access to information on the advanced technology that permeates so much of modern electronics.

IAN R. SINCLAIR
JOHN DUNTON

This page intentionally left blank

INTRODUCTION:
MATHEMATICS CONVENTIONS

Quantities greater than 100 or less than 0.01 are usually expressed in the *standard form* of $A \times 10^n$, where A is a number, called the *mantissa*, less than 10, and n is a whole number called the *exponent*. A positive value of n means that the number is greater than unity, a negative value of n means that the number is less than unity. To convert a number into standard form, shift the decimal place until the portion on the left-hand side of the decimal point is between 1 and 10, and count the number of places that the point has been moved. This is the value of n. If the decimal point has had to be shifted to the left the sign of n is positive; if the decimal point had to be shifted to the right the sign of n is negative.

Example: 1200 is 1.2×10^3 and 0.0012 is 1.2×10^{-3}

To convert numbers back from standard form, shift the decimal point n figures to the right if n is positive or to the left if n is negative.

Example: $5.6 \times 10^{-4} = 0.00056$ and $6.8 \times 10^5 = 680\,000$

Note in these examples that a space has been used instead of the more familiar comma for separating groups of three digits (thousands and thousandths). This is recommended engineering practice and avoids confusion caused by the use, in other languages, of a comma as a decimal point. Numbers in standard form can be entered into a calculator by using the key marked Exp or EE – for details see the manufacturer's instructions.

Where formulae are to be worked out, numbers in standard form can be used, but for writing component values it is more convenient to use the prefixes shown in the table below. The prefixes have been chosen so that values can be written without using small fractions or large numbers.

Some variants of standard form follow a similar pattern in allowing numbers between 1 and 999 to be used as the whole-number part of the expression, so that numbers such as 147×10^{-4} are used. A less common convention is to use a fraction between 0.1 and 1 as a mantissa, such as 0.147×10^{7}.

Powers of 10 and prefixes

Prefix	Abbreviation	Power of ten	Multiplier
Giga	G		1 000 000 000
Mega	M		1 000 000
kilo	k		1000
milli	m		1/(1000)
micro	μ		1/(1 000 000)
nano	n		1/(1 000 000 000)
pico	p		1/(1 000 000 000 000)

Note that 1000 pF = 1 nF; 1000 νF= 1μF and so on. In computing, the K symbol means 1024 rather than 1000 and M means 1 048 576– these quantities are the nearest exact powers of two.

Examples: 1 kΩ = 1000 Ω (sometimes written as 1 K0, see pages 7-8)

1 nF = 0.001 μF, 1000 pF or 10^{-9} F

4.5 MHz = 4500 kHz = 4.5×10^{6} Hz

Throughout this book equations have been printed in as many forms as are normally needed so that the reader should not have to transpose the equations. For example, Ohm's law is given in all three familiar forms of $V = IR$, $R = V/I$ and $I = V/R$. The units that must be used with such formulae are shown and must be adhered to – if no units are quoted then fundamental units (amp, ohm, volt) are implied.

For example,the equation $X = 1/(2\pi fC)$ is used to find the reactance of a capacitor in ohms, using C in farads and f in hertz. If the equation is to be used with values given in μF and kHz then values of 0.1 μF and 15 kHz are entered as 0.1×10^{-6} and 15×10^{3}. Alternatively, the equation can be written as $X = 1/(2\pi fC)$ MΩ using values of f in kHz and C in nF.

In all equation multiplication is normally indicated by the use of a dot, such as $f.C$ or by close printing as shown above in $2\pi fC$. Where brackets are used in an equation, the quantities within the brackets should be worked out first, and where there are brackets within brackets, the portion of the

equation in the innermost brackets must be worked out first, followed by the material in the outer brackets. Apart from brackets, the normal order of working out is to carry out multiplication and divisions first followed by additions and subtractions. For example:

$$2(3+5) \text{ is } 2 \times 8 = 16 \text{ and } 2 + (3 \times 5) \text{ is } 2 + 15 = 17$$

Transposing, or changing the subject of an equation, is simple provided that the essential rule is remembered: an equation is not altered by carrying out identical operations on each side.

Example: $Y = (5aX + b)/C$ is an equation that can be transposed so that it can be used to find the value of X when the other quantities are known.

The procedure is to keep changing both sides so that X is left isolated.

Starting with $Y = \dfrac{5aX + b}{C}$, the steps are as follows:

(a) Multiply both sides by C, the result is $CY = 5aX + b$

(b) Subtract b from both sides, the result is $CY - b = 5aX$

(c) Divide both sides by $5a$, the result is $\dfrac{CY - b}{5a} = X$

So that the equation has become $X = \dfrac{CY - b}{5a}$ which is the transposition we required.

This page intentionally left blank

CHAPTER 1

RESISTORS

Passive components

Passive components are those that need no power supply for their operation and whose action will dissipate power, though in some cases the amount of dissipation is negligible. No purely passive component can have an output that supplies more power than is available at the input. Active components, by contrast, make use of a power supply, usually DC, so that the signal power output of an active component can be higher than the signal power at the input. Typical passive components are resistors, capacitors and inductors. Familiar active components are transistors and ICs.

All components, active or passive, require to be connected to a circuit, and the two main forms of connection, mechanical and electrical, used in modern electronic circuits are the traditional wire leads, threaded through holes in a printed circuit board (see Chapter 18) and the more modern surface mounting devices (SMDs) that are soldered directly on to the tracks of a board. Both passive and active components can use either type of connection and mounting.

Components for surface mounting use flat tabs in place of wire leads, and because these tabs can be short the inductance of the leads is greatly reduced. The tabs are soldered directly to pads formed onto the board, so that there are always tracks on the component side of the board as well as on the opposite side. Most SMD boards are two sided, so that tracks and components are also placed on the other side of the board. Multilayer boards are also commonly used, particularly for mobile phones (4 to 6 layers) and computer motherboards.

The use of SMDs results in manufacturers being able to provide components that are physically much smaller, but with connections that dissipate heat more readily, are mechanically stronger and have lower electrical resistance and lower self-inductance. Some components can be made so small that it is impossible to mark a value or a code number onto them. This presents no problems for automated assembly, since the tape or reel need only be inserted into the correct position in the assembly machine, but considerable care needs to be taken when replacing such components manually, and they should be kept in their packing until they are soldered into place. Machine assembly of SMD components is followed by automatic soldering processes, which nowadays usually involve the use of solder paste or cream (which also retains components in place until they are soldered) and heating by blowing hot nitrogen gas over the board. Packaging of SMD components is nowadays normally on tapes or in reels.

Resistors

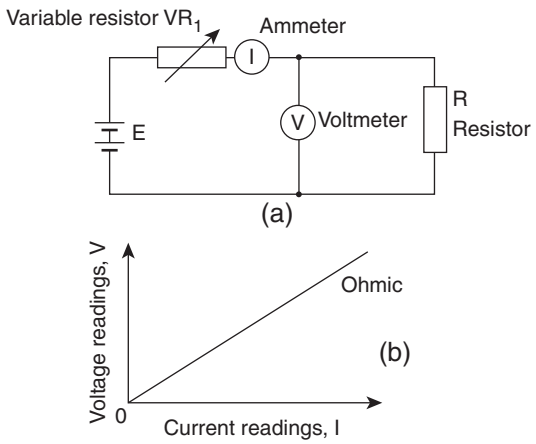
The resistance of a sample of material, measured in units of **ohms** (Ω), is defined as the ratio of voltage (in units of volts) across the sample of material to the current (in units of **amperes**) through the material. The name ampere is usually abbreviated to **amp**. When we draw a graph of voltage *across* the sample (a resistor) plotted against current *through* the material, the value of resistance is represented by the *slope* of the graph. For a metallic material kept at a constant temperature, a straight-line graph indicates that the material is ohmic, obeying Ohm's law (Figure 1.1).

Non-ohmic behaviour is represented on such a graph by a curved line or a line that does not pass through the zero-voltage, zero-current point that is called the **origin**. Non-ohmic behaviour can be caused by temperature changes (as in light bulbs and thermistors), by voltage generating effects (as in thermocouples and cells) and by conductivity being affected by voltage (as in diodes). Typical examples of deviation from linearity are illustrated in Figure 1.2.

A material that is ohmic will have a constant value of resistance (subject to minor alteration with temperature change) and can be used to make resistors. Resistance values will be either colour coded, or have values printed on using the conventions of BS1852: 1970 (see later).

Figure 1.1

(a) A circuit for checking the behaviour of a resistor. **(b)** The shape of the graph of voltage plotted against current for an ohmic resistor, using the circuit in **(a)**.



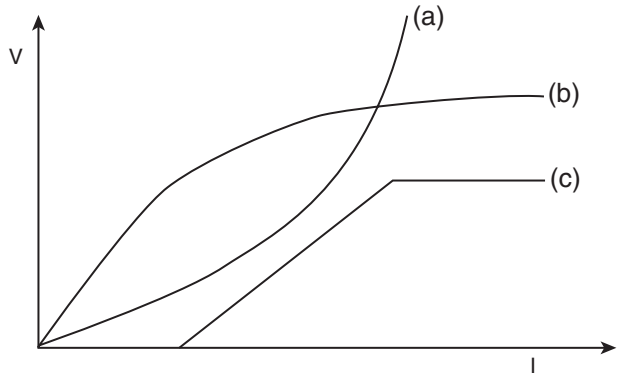


Figure 1.2

Three types of non-ohmic behaviour indicated by graph curves or lines: **(a)** light bulb, **(b)** ntc thermistor, **(c)** diode.

Resistivity

The resistance of any sample of a material is determined by its dimensions and by the value of **resistivity** of the material. Wire drawn from a single reel (with uniform diameter) will have a resistance that depends on its length. For example, a 3 m length will have three times the resistance of

a 1 m length of the same wire. When equal lengths of wire of the same material, but different diameters, are compared, the resistance multiplied by the square of diameter is the same for each. For example, if a given length of a sample wire has a resistance of 12 ohms and its diameter is 0.3 mm, the same length of wire made from the same material but with a diameter of 0.4 mm will have resistance R given by:

$$R \times 0.4^2 = 12 \times 0.3^2 \text{ so } R = \frac{12 \times 0.3^2}{0.4^2} = \frac{12 \times 0.09}{0.16} = 6.75 \text{ ohms}$$

Resistivity measures the effect that the material itself (irrespective of dimensions) contributes to the resistance. The resistivity of the material can be measured by finding the resistance R of a sample, multiplying this by the area of cross-section (assumed uniform) and dividing by the length of the sample.

As a formula this is written:

$$\rho = \frac{RA}{L}$$

ρ = resistivity
R = resistance
A = area of cross-section
L = length

When R is expressed in ohms, A in square metres (m^2) and L in metres, the unit, of ρ (Greek rho) will be ohm-metres (**not** ohms per metre). Since most wire samples are of circular cross-section, $A = \pi r^2 l$ or $\frac{1}{4}(\pi d^2)$ where d is the wire's diameter.

Resistivity calculations

Because the resistivities of commonly used materials are well known and can be looked up in tables, to find the resistance in ohms of a piece of wire of known length and diameter the formula is more useful in the form:

$$R = \frac{\rho L}{A}$$

with ρ in ohm-metres, L in metres and A in square metres (m^2). This formula can be rewritten as

$$R = 1.27 \times 10^{-3} \frac{\rho L}{d^2}$$

with ρ in nano-ohm metres, L in metres, and d (diameter) in millimetres. Table 1.1 shows values of resistivities in nano-ohm metres for various metals, including both elements and common alloys. For some purposes, conductivity is used in place of resistivity. The conductivity, symbol σ (Greek sigma), is defined as 1/resistivity, so $\rho = 1/\sigma$. The unit of conductivity is

Table 1.1 Values of resistivity and conductivity at 0°

Pure elements		
Metal	**Resistivity**	**Conductivity**
Aluminium	27.7	37
Copper	17	58
Gold	23	43
Iron	105	9.5
Nickel	78	12.8
Platinum	106	9.4
Silver	16	62.5
Tin	115	8.7
Tungsten	55	18.2
Zinc	62	16
Alloys		
Carbon-steel (average)	180	5.6
Brass	60	16.7
Constantan	450	2.2
Invar	100	10
Manganin	430	2.3
Nichrome	1105	0.9
Nickel-silver	272	3.7
Monel metal	473	2.1
Kovar	483	2.0
Phosphor-bronze	93	10.7
18/8 stainless steel	897.6	1.11

Notes: The values of resistivity are in nano-ohm metres. The values of conductivity are in megasiemens per metre.

the siemens per metre, S/m. The resistivity formulae, using basic units, can be rearranged in terms of conductivity as:

$$R = \frac{L}{\sigma A} \text{ or } L = RA\sigma$$

Conductivity values are also shown in Table 1.1.

The calculation of resistance for a sample by either formula follows the pattern of the following examples.

Example A: *Find the resistance of* 6.5 m *of wire, diameter* 0.6 mm*, if the resistivity value is* 430 nano-ohm metres (430 nΩm).

$$\text{Using } R = \frac{\rho L}{A} \text{ with:}$$

$\rho = 4.30 \times 10^{-9}$, $L = 6.5$ m, $A = \frac{1}{4}(\pi d^2) = \frac{1}{4}\pi(0.6 \times 10^{-3})^2$ (remembering that 1 mm $= 10^{-3}$ m), $A = 2.83 \times 10^{-7}$ m^2 so

$$R = \frac{4.30 \times 10^{-9} \times 6.5}{2.83 \times 10^{-7}} = 9.88 \text{ ohms, about 10 ohms.}$$

Using the second version of the formula, we get:

$$R = 1.27 \times 10^{-3} \frac{I}{d^2} = \frac{1.27 \times 10^{-3} \times 4.30 \times 6.5}{0.36}$$
$$= 9.88 \text{ ohms about 10 ohms.}$$

Example B: *To find the length of wire that is needed for a given resistance value, the formula is transposed to:*

$$L = \frac{RA}{\rho}$$

using R in ohms, A in square metres and ρ in ohm-metres to obtain L in units of metres. An alternative formula is:

$$L = 785.4 \times \frac{Rd^2}{\rho}$$

using R in ohms, d in millimetres and ρ in nano-ohm metres.

Example C: *To find the diameter of wire needed for a resistance* R *and length* L *metres, using* ρ *in nano-ohm metres, the formula for d in millimetres is:*

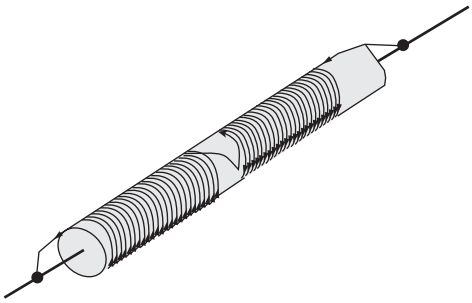
$$d = 3.57 \times 10^{-2}\sqrt{\frac{\rho L}{R}}$$

Resistor construction

The materials used for resistor construction are generally metal alloys, pure metal or metal-oxide films, or carbon (solid or in thin-film form). Wire-wound resistors use metal alloy wire wound onto ceramic formers. The windings must have a low self-inductance value, so that the wire is wound using the method shown in Figure 1.3 with each half of the winding wound in the opposite direction.

Figure 1.3

Non-inductive winding of a wire-wound resistor. The two halves of the total length of wire are wound in opposite directions so that their magnetic fields oppose each other.



Wire-wound resistors are used when very low values of resistance are needed or when very precise values must be specified (for meter shunts, for example). Large resistance values in the region of 20 kΩ upwards need such fine-gauge wire that failure can occur due to corrosion, especially in tropical conditions of high temperature and high humidity, so high-value, wire-wound resistors should not be used for marine or tropical applications unless the wire can be protected satisfactorily.

Carbon composition resistors, once the main type of resistor used for electronics, are now rarely used. They consist of a mixture of graphite and clay whose resistivity depends on the proportion of graphite in the mixture. Because the resistivity value of such a mixture can be very high, greater resistance values can be obtained without the need for physically large components. Resistance value tolerances (see later) are high, however, because

of the greater difficulty in controlling the resistivity of the mixture and the final dimensions of the carbon composition rod after heat treatment. You should **not** specify carbon composition resistors for any new design unless cost is an overriding factor.

Metal film, carbon film and metal-oxide film resistors are more recent types that form the vast majority of resistors used today. They are made by evaporating metals (in a vacuum or an inert atmosphere), or metal oxides (in an oxidizing atmosphere) onto ceramic rods. The resistance value is controlled (1) by controlling the thickness of the film and (2) by cutting a spiral path on the film after it has been deposited. These resistors are considerably cheaper to make than wire-wound types and can be made to much closer tolerances than carbon-composition types. The costs of such resistors are now almost the same as those of composition types. Figure 1.4 shows typical fixed resistor shapes.

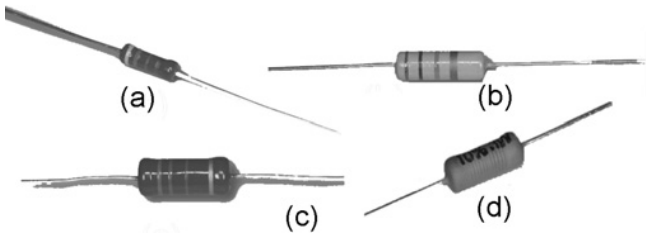


Figure 1.4

Typical resistors: **(a)**, **(b)**, **(c)** carbon film, **(d)** wirewound. (Original photos by Alan Winstanley.)

Variable resistors and potentiometers can be made using all the methods that are employed for fixed resistors. The component is termed a **potentiometer** when connections are made to both ends of the resistive track and also to a sliding connection; a **variable resistor** uses only one connection to one end of the track and one to the sliding connector. By convention, both are wired so that the quantity that is being controlled will be increased by clockwise rotation of the shaft as viewed by the operator. A **trimmer** is a form of potentiometer, often miniature, that is preset on test and not normally alterable by a user of equipment. Figure 1.5 illustrates the variety of potentiometer and trimmer shapes.

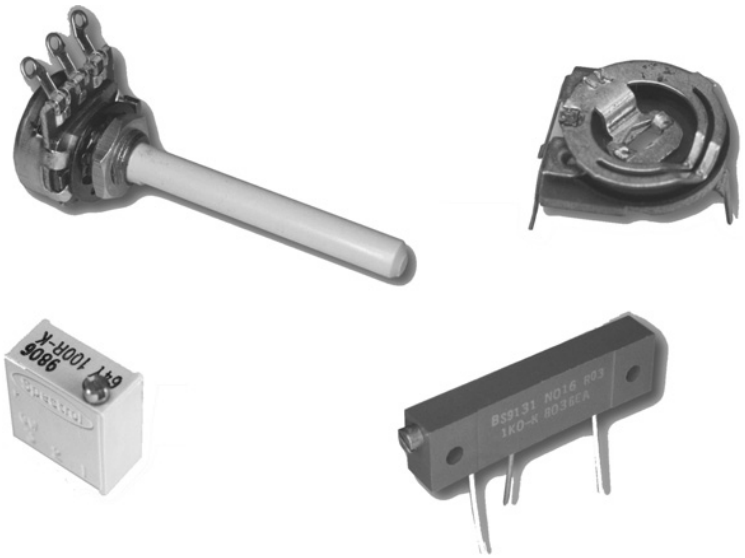


Figure 1.5

Typical potentiometer and trimmer shapes. (Original photos by Alan Winstanley.)

Tolerances and E-series

Any mass-production process that is aimed at producing a target value of a measurable quantity will inevitably produce a range of values that are centred around the desired value and for which a maximum **tolerance** can be specified. The tolerance is the maximum difference between any actual value and the target value, usually expressed as a percentage. For example, a 10 kΩ 20% resistor may have a value of:

$$10\,000 + \left(\frac{20}{100} \times 10\,000\right) = 12\ \text{k}\Omega \text{ or}$$

$$10\,000 - \left(\frac{20}{100} - 10\,000\right) = 8\ \text{k}\Omega$$

Tolerance series of preferred values, shown in Table 1.2, are ranges of target values chosen so that no component can be rejected on grounds of incorrect value. They also allow a designer to specify a component whose variation will not be more than that allowed for in calculations. The mathematical basis of these preferred values is the sixth root of ten ($\sqrt[6]{10}$) for the E6 20%

series (there are six steps of value between 1 and 6.8), and the twelfth root of ten ($\sqrt[12]{10}$) for the E12 10% series. The **E-figure** indicates the number of values in each decade (1–10, 10–100, 100–1000, etc.) of resistance value. The figures produced by this series are rounded off. For example:

$$\sqrt[6]{10} = 1.46 \quad \left(\sqrt[6]{10}\right)^2 = 2.15 \quad \left(\sqrt[6]{10}\right)^3 = 3.16 \quad \left(\sqrt[6]{10}\right)^4$$
$$= 4.64 \quad \left(\sqrt[6]{10}\right)^5 = 6.8$$

These figures are rounded to the familiar 1.5, 2.2, 3.3, 4.7 and 6.8 that are used in the 20% series, and similar rounding is used for the 10%, 5%, 1% and other series, with the 5% series using values based on the 18th root of ten. A simple view of the tolerance series is that, taking the 20% series as an example, 20% up on any value will overlap with 20% down on the next higher value.

Example: 4.7 + 20% *of* 4.7 = 5.64 *and* 6.8 − 20% *of* 6.8 = 5.44, *allowing an overlap.*

After manufacture, resistors are graded with the 1%, 5% and 10% tolerance values removed, and the remaining resistors are sold as 20% tolerance values. Because of this it is pointless to sort through a bag of 20% 6K8 resistors, for example, hoping to find one that will be of exactly 6K8 value. Such a value will have been removed in the grading process by the manufacturer. When close-tolerance components are specified it will be for a good reason and 20% tolerance components cannot be substituted for 10% or 5% types. Nowadays it is more common to find that the highest tolerance that is sold is of 10%, reflecting the diminished number of carbon composition resistors being manufactured.

Resistance value coding

Values of resistors (and capacitors) that use conventional wire mounting are usually indicated by a set of coloured bands (Figure 1.6). At one time,

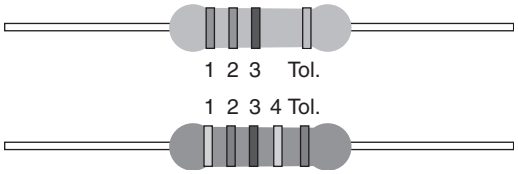
Table 1.2 Preferred values tolerance series

						1.0	1.5	2.2	3.3	4.7	6.8						
E6 series 20% tolerance																	
			1.0	1.2	1.5	1.8	2.2	2.7	3.3	3.9	4.7	5.6	6.8	8.2			
E12 series 10% tolerance																	
1.0	1.1	1.2	1.3	1.5	1.6	1.8	2.0	2.2	2.4	2.7	3.0	3.3	3.6	3.9	4.3	4.7	5.1
						5.6	6.2	6.8	7.6	8.2	9.1						
E24 series 5% tolerance																	
	1.00	1.02	1.05	1.07	1.10	1.13	1.15	1.18	1.21	1.24	1.27	1.30	1.33	1.37	1.40	1.43	
	1.47	1.50	1.54	1.58	1.62	1.65	1.69	1.74	1.78	1.82	1.87	1.91	1.96	2.00	2.05	2.10	
	2.15	2.21	2.26	2.32	2.37	2.43	2.49	2.55	2.61	2.67	2.74	2.80	2.87	2.94	3.01	3.09	
	3.16	3.24	3.32	3.40	3.48	3.57	3.65	3.74	3.83	3.92	4.02	4.12	4.22	4.32	4.42	4.53	
	4.64	4.75	4.87	4.99	5.11	5.23	5.36	5.49	5.62	5.76	5.90	6.04	6.19	6.34	6.49	6.65	
	6.81	6.98	7.15	7.32	7.50	7.68	7.87	8.06	8.25	8.45	8.66	8.87	9.09	9.31	9.53	9.76	
E96 series 1% tolerance																	

The numbers then repeat, but each (taking the E96 set as an example) multiplied by ten, up to 97.6 Ω, then multiplied by 100 up to 976 Ω and so on.

Figure 1.6

Coloured bands for coding value on a resistor.



three bands were used, allowing two significant digits and one multiplier figure, but because of the widespread use of close-tolerance components, it is now more common to use four or five bands with one band used for tolerance. The use of the tolerance band is a useful guide to the order of bands, because there is often no other indication of which end of the resistor band 1 is located. In the absence of other clues, you have to assume that the correct order of bands is the one that gives value in a valid tolerance set.

The coding (Table 1.3) can use three bands for value and one for tolerance for components in the tolerances from E6 to E24 (one place of decimals). In this scheme, the first band shows the first significant figure, the second band the second significant figure, and the third band the multiplier (the power of ten), with the fourth band indicating tolerance. For the E96 values, an additional significant figure band is added, so that the tolerance band is the fifth. Resistors manufactured for some specialized purposes can use an additional band to indicate temperature coefficient.

Resistance values on components and in component lists are often coded according to BS 1852. In this scheme, no decimal points are used and a value in ohms is indicated by **R**, kilohms by **K** (not k), and megohms by **M**. The letter R, K or M is used in place of the decimal point, with a zero in the leading position if the value is less than 1 ohm. This scheme avoids two sources of confusion:

1. the appearance of a dot due to a dirty photocopy being taken as a decimal point.

2. the continental practice of using commas and points in the opposite sense to UK practice.

Example: 1K5 = 1.5 k or 1500 ohms; 2M2 = 2.2 M; 0R5 = 0.5 ohms.

Table 1.3 Resistor colour code

Figure	Colour	
0	black	
1	brown	
2	red	
3	orange	
4	yellow	
5	green	
6	blue	
7	violet	
8	grey	
9	white	
0.01	silver	used as multiplier colours
0.1	gold	
Tolerance:		
10%	silver	
5%	gold	
2%	red	
1%	brown	

No tolerance band is used if the resistor has 20% tolerance.

Surface mounted resistors

Two forms of coding are used for surface mounted resistors (and capacitors). The three-symbol code uses two digits for the significant figures and one as multiplier, so that $471 = 47 \times 10 = 470\,\Omega$ and $563 = 5\text{K}6$. Values below 10 are indicated in BS1852 form, so that $2\text{R}2 = 2.2\,\Omega$. The alternative marking, which is better suited to E96 resistors makes use of letter codes for the significant figures, and a number to indicate the multiplier. The codes are indicated in Table 1.4.

Resistor characteristics

Important characteristics of resistor types include resistance ranges, usable temperature range, stability, noise level, and temperature coefficient. Wirewound resistors are available in values that range from fractions of an ohm (usually 0R22) up to about 10 kΩ (though higher values up to 100 kΩ

Table 1.4 Letter and number codes for SM components. Resistance values are in ohms, and the same coding is used for capacitors in units of picofarads

A = 1	B = 1.1	C = 1.2	D = 1.3	E = 1.5	F = 1.6	G = 1.8	H = 2	J = 2.2	K = 2.4	L = 2.7
M = 3	N = 3.	P = 3.6	Q = 3.9	R = 4.3	S = 4.7	T = 5.1	U = 5.6	V = 6.2	W = 6.8	X = 7.5
Y = 8.2	Z = 9.1	a = 2.5	b = 3.5	d = 4	e = 4.5	f = 5	m = 6	n = 7	t = 8	y = 9
		0 = ×1	1 = ×10	2 = ×100	3 = ×1 k	4 = ×10 k	5 = ×100 k	6 = ×1 M		

are available). Carbon composition resistors can be obtained in ranges of around 2R2 to 1M0 and film resistors normally range from 1R0 to 1M0. Typical usable temperature ranges are –40°C to +105°C for composition and –55°C to +150°C for metal oxide. Wire-wound resistors can be obtained that will operate at higher temperatures (up to 300°C) depending on construction and resistance value.

The **stability of value** means the maximum change of value that can occur during shelf-life, on soldering, and in use in adverse conditions such as operating in high temperatures in damp conditions. These changes are in addition to normal tolerances. Composition resistors have the poorest figures for stability of value, with typical shelf-life change of 5%, soldering change of 2% and 'damp-heat' change of 15%. Metal-oxide resistors can, typically, have shelf-life changes of 0.1%, soldering changes of 0.15% and damp-heat changes of 1%. The noise level of a resistor is specified in terms of microvolts (μV) of noise signal generated per volt of DC across the resistor. Such noise levels range from 0.1 μV/V for metal oxide to a minimum of 2.0 μV/V for composition, and for composition resistors the value increases for higher values of resistance. The formula that is used to find the noise level of composition resistors is:

$$2 + \log_{10}\left(\frac{R}{1000}\right)\mu V/V$$

so, for example, a 680 kΩ resistor would have a predicted noise level of

$$2 + \log_{10}\left(\frac{680\,000}{1000}\right) = 4.8\ \mu V/V$$

The **temperature coefficient of resistance** measures the change of resistance value as the surrounding temperature changes. The basic formula is:

$$R_\theta = R_0\,(1 + \alpha\theta)$$

R_θ is resistance at θ°C
R_0 is resistance at 0°C
α is the temperature coefficient
θ is the temperature in °C

The value of temperature coefficient is usually quoted in parts per million per °C (abbreviated to ppm/°C) and this has to be converted to a fraction, by dividing by one million, to be used in the formula above.

Example: *What is the value of a* 6k8 *resistor at* 95°C *if the temperature coefficient is* +1200 ppm/°C?

Converting +1200 ppm/°C into standard fractional form gives:

$$\frac{1200}{1\,000\,000} = 1.2 \times 10^{-3} = 0.0012$$

Using the formula, $R_{95} = 6.8\,(1 + 0.0012 \times 95) = 7.57\ \text{k}\Omega$

- Remember that the multiplication must be carried out before the addition.

Note that if the resistance at some temperature ϕ°C other than 0°C is given, the formula changes to:

$$R_\theta = R_\phi \left(\frac{1 + \alpha\theta}{1 + \alpha\phi} \right)$$

- Remember that you cannot cancel the 1s in this equation.

Example: *If a resistor, temperature coefficient* 1.5×10^{-3}*, has a value of* 10 Ω *at* 20°C*, its resistance at* 80°C *can be found by:*

$$R_{80} = 10 \times \frac{1 + (80 \times 1.5 \times 10^{-3})}{1 + (20 \times 1.5 \times 10^{-3})} = 10 \times \frac{1.12}{1.03} = 10.87\ \Omega$$

Temperature coefficients may be **positive**, meaning that the resistance will *increase* as the temperature rises, or **negative**, meaning that the resistance will *decrease* as the temperature rises. Carbon composition resistors have temperature coefficients of typically +1200 ppm/°C and metal oxide types have the lowest temperature coefficient values of ±250 ppm/°C. Note that tables of temperature coefficients normally quote temperature coefficient of *resistivity* rather than resistance. For all practical purposes, the two coefficients are identical.

Dissipation and temperature rise

The power dissipation rating (P), measured in watts (W), for a resistor indicates how much power can be converted to heat without damage to the resistor caused by its rise in temperature. The rating is closely linked to the physical size of the resistor, so that ¼ W resistors are much smaller than 1 W resistors of the same resistance value. These ratings assume 'normal' surrounding (*ambient*) temperatures, often 70°C, and for use at higher ambient temperatures derating must be applied according to the manufacturer's specification. For example, a ½ W resistor may need to be used in place of a ¼ W when the ambient temperature is above 70°C. In Figure 1.7 is shown the graph of temperature rise plotted against dissipated power for average ½ W and 1 W composition resistors. Note that these figures are of temperature rise **above the ambient** level. If such a temperature rise takes the resistor temperature above its maximum rated temperature permitted for its type, a higher wattage rating of resistor must be used. Resistors with high ohmic values may need to be **derated** (run at a lower dissipation) when they are used in hot surroundings.

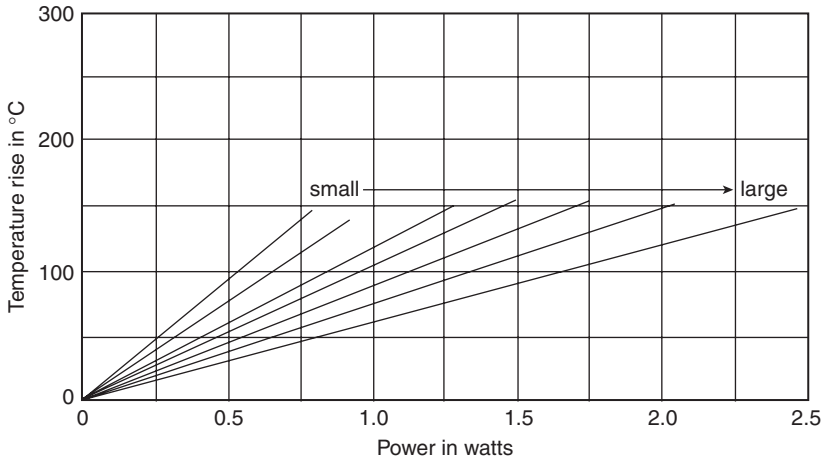


Figure 1.7

Temperature rise and power dissipation for typical resistors. The temperature scale is in °C above the surrounding (ambient) temperature. For example, in a room at 20°C, a ½ W resistor dissipating 0.1 W will be at a temperature of 40°C.

The power dissipation in watts is given by $P = VI$, with V the voltage across a conductor in volts and I the current through the conductor in amps. When current is measured in mA and V in volts, VI gives power dissipation in *milliwatts*, often more useful for electronics components. This expression for dissipated power can be combined with Ohm's law when the resistance R of the conductor is constant, giving:

$$P = \frac{V^2}{R} \text{ or } P = I^2R$$

The result will be in watts for V in volts and R in ohms, or I in amps and R in ohms. When R is given in kΩ, V^2/R gives P in milliwatts; when I is in mA and R in kΩ then P is also in milliwatts.

Note that power is defined as the amount of energy (also called work, *W*) transformed (from one form to another) per second. The unit of energy is the joule (J), and the number of joules dissipated is found by multiplying the power in watts by the time in seconds for which the power has been dissipated, so

$$W = \frac{V^2t}{R} \text{ or } W = I^2Rt.$$

- Be careful not to confuse the abbreviations *W* (work or energy) and W (watts of power). The abbreviation *p* is used for pressure and P for power.

Variables and laws

The law of a variable resistor or potentiometer must be specified in addition to the quantities that are specified for any fixed resistor. The potentiometer law (called *taper* in the USA) describes the way in which resistance between the slider and one contact varies as the slider is rotated; the law is illustrated by plotting a graph of resistance against shaft rotation angle (Figure 1.8).

A linear law potentiometer (Figure 1.8a) produces a straight-line graph, hence the name *linear*. Logarithmic (log) law potentiometers are extensively used as volume controls and have the graph shape shown in Figure 1.8b.

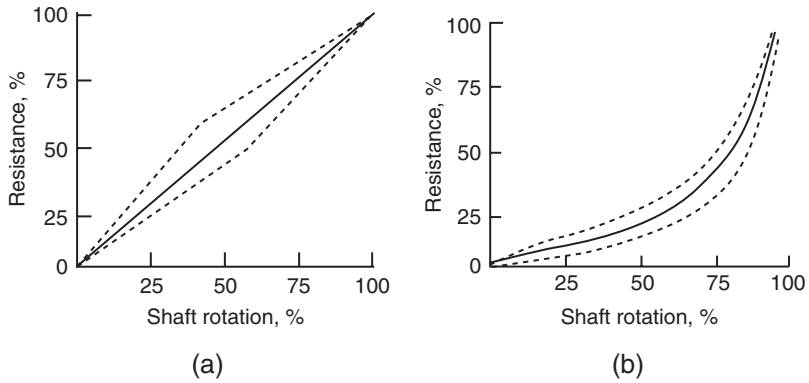


Figure 1.8

Potentiometer laws: **(a)** linear, **(b)** logarithmic. In the USA the word 'taper' is used in place of 'law', and 'audio' in place of 'log'. Broken lines show tolerance limits.

Less common laws are anti-log and B-law, and specialized potentiometers with sine or cosine laws are also available.

Resistors in circuit

Resistors in a circuit obey Ohm's circuit law (not really a law in this sense) and Kirchoff's laws. Ohm's circuit law is written in its three forms as:

$$V = RI, \text{ or } R = V/I \text{ or } I = V/R$$

where V is voltage across two points, I is the current flowing between the points and R is the (constant) resistance between the points. The units of these quantities are as shown in Table 1.5. These equations can be applied even to materials that do not obey Ohm's law if the value of R for some stated set of conditions can be found.

Materials that do not obey Ohm's law do not have a *constant* value of resistance, but the relationships shown above (which simply state a definition of resistance) still hold. The equations are most useful when the resistance values are constant, hence the use of the name Ohm's law to describe the relationships.

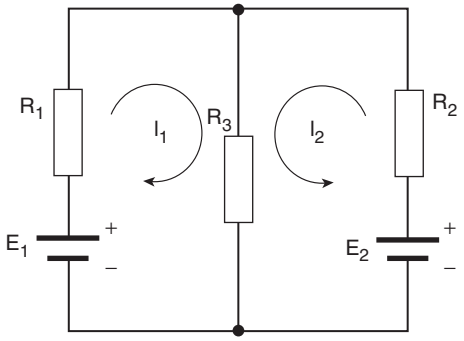
Table 1.5 Ohm's law and units

Forms of the law: V = RI, R = V/I, I = V/R

Units of V	Units of R	Units of I
Volts, V	Ohms, Ω	Amps, A
Volts, V	Kilohms, kΩ	Milliamps, mA
Volts, V	Megohms, MΩ	Microamps, μA
Kilovolts, kV	Kilohms, kΩ	Amps, A
Kilovolts, kV	Megohms, MΩ	Milliamps, mA
Millivolts, mV	Ohms, Ω	Milliamps, mA
Millivolts, mV	Kilohms, kΩ	Microamps, μA

• KIRCHOFF'S LAWS

Kirchoff's laws relate to the conservation of energy, which states that energy cannot be created or destroyed, only changed into different forms. This can be expanded to laws of conservation of voltage and current. In any circuit, the voltage across each series component (carrying the same current) can be added to find the total voltage. Similarly, the total current entering a junction in a circuit must equal the sum of current leaving the junction. These laws are illustrated in Figure 1.9.



Current law:
current in $R_3 = I_1 + I_2$

Voltage law:
$E_1 = R_1I_1 + R_3(I_1 + I_2)$
$E_2 = R_2I_2 + R_3(I_1 + I_2)$

Figure 1.9

Kirchhoff's laws. The current law states that the total current leaving a circuit junction equals the total current into the junction – no current is 'lost'. The voltage law states that the driving voltage (or EMF) in a circuit equals the sum of voltage drops $I \times R$ around the circuit.

In Figure 1.10 are shown the rules for finding the total resistance of resistors in series or in parallel. When a combination of series and parallel connections is used, the total resistance of each series or parallel group must be found first before finding the grand total.

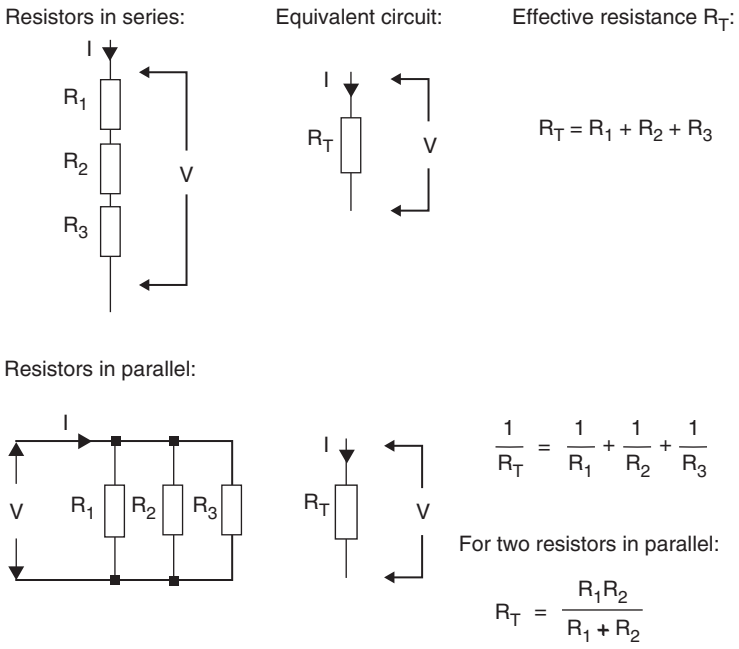
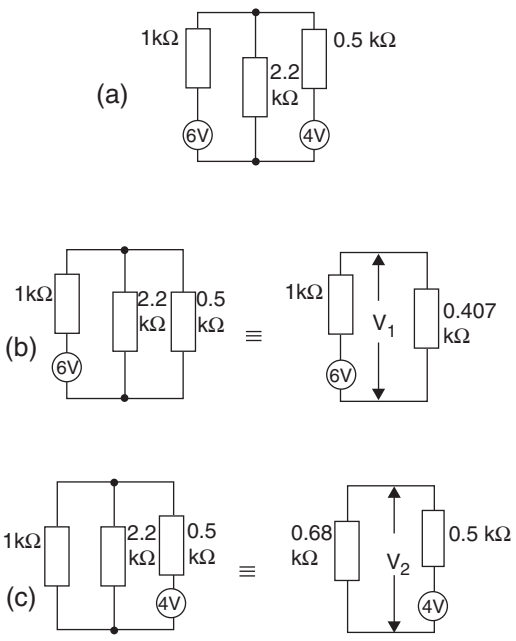


Figure 1.10

Resistors in series and in parallel.

- **THE SUPERPOSITION THEOREM**

The superposition theorem is very useful for finding the voltages and currents in a circuit with two or more sources of supply, and is usually easier to use than Kirchoff's law equations. Figure 1.11 shows an example of the theorem in use. One supply is selected and the circuit is redrawn to show the other supply (or supplies) short-circuited (leaving only the internal resistance of each supply). The voltage and current caused by the first supply



In this network, there are two generators and three resistors. The generators might be batteries, oscillators, or other signal sources.

To find the voltage caused by the 6 V generator, replace the 4 V generator by its internal resistance of 0.5 kΩ. Using Ohm's law, and the potential divider equation: V = 1.736 V.

To find the voltage caused by the 4 V generator, the 6 V generator is replaced by its 1 kΩ internal resistance. In this case: V = 2.315 V.

Now the total voltage in the original circuit across the 2.2 kΩ resistor is simply the sum of these: 4.051 V.

Figure 1.11

Using the superposition theorem. This is a simple method of finding the voltage across a resistor in a circuit where more than one source of EMF is present.

can then be calculated, using $V = RI$ methods together with the rules for combining series and parallel resistors. Each supply is treated in turn in the same way, and finally the voltages and currents caused by each supply are added.

The superposition principle: this states that in any linear network, the voltage at any point is the sum of the voltages caused by each generator in the circuit. To find the voltage caused by a generator replace all other generators in the circuit by their internal resistances, and use Ohm's law. A linear network means an arrangement of resistors and generators with the resistors obeying Ohm's law, and the generators having a constant voltage output and constant internal resistances.

Example: *In the network shown, find the voltage across the* 2.2 kΩ *resistor.*

- **THEVENIN'S THEOREM**

Thevenin's (pronounced Tay-venin) theorem is, after Ohm's circuit law, one of the most useful electrical circuit laws. The theorem states that any network of linear components, such as resistors and batteries, can be replaced in its effect by an equivalent circuit consisting only of a voltage source and a resistance in series. The size of the equivalent voltage is found by taking the open-circuit voltage between two points in the network, and the series resistance is found by calculating the resistance between the same two points assuming that the voltage source is short-circuited. (See later for examples of Thevenin's theorem, illustrated in Figure 1.13.)

- There is a corresponding theorem, Norton's theorem, which states that any network of linear components can also be considered to consist only of a constant current source and a resistor in parallel.

Figure 1.12 shows two important networks, the potential divider and the bridge. When no current is taken from the potential divider, its output voltage V is given by:

$$V = \frac{R_2 E}{R_1 + R_2}$$

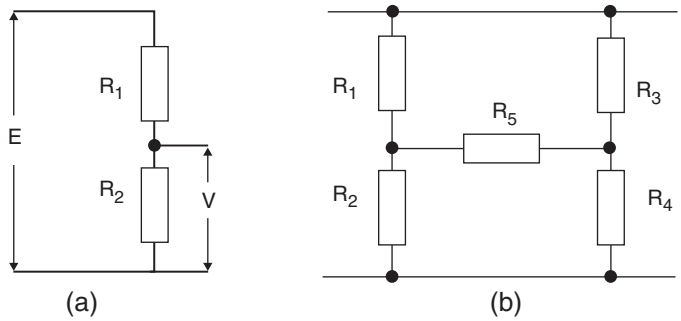


Figure 1.12

A potential divider **(a)** and bridge **(b)** circuit.

as shown, but when current is being drawn, as is the case when a transistor is being biased by this circuit, the equivalent circuit, using Thevenin's theorem as shown in Figure 1.13, is more useful. The bridge circuit, when no current is drawn, is said to be balanced when there is no voltage across R (which is usually a galvanometer or microammeter). In this condition:

$$\frac{R_1}{R_2} = \frac{R_3}{R_4}$$

If the bridge is **not** balanced, the equivalent circuit derived from using Thevenin's theorem is, once again, more useful.

Thermistors

Thermistors are resistors made from materials that have large values of temperature coefficient. Both PTC and NTC types are produced for applications that range from temperature measurement to transient current suppression. Figure 1.14 shows some representative types. Miniature thermistors either in bead form or in glass tubes are used for temperature measurement, using a bridge circuit (Figure 1.15), and are also used for timing circuits and in stabilizing the amplitude of sine wave oscillators (see Chapter 7).

Thermistors are self-heating if the current through them is allowed to exceed the limits laid down by the manufacturers, so the current flowing in a bridge-measuring circuit must be carefully limited. Larger thermistor types, with lower values of cold resistance (measured at 20°C) are used for current regulation, such as circuits for degaussing colour TV tubes, controlling the surge current through filament light bulbs, or reducing the speed of fan when a set temperature is reached. The general form of graphs of resistance plotted against temperature is that shown in Figure 1.16, and the formula for finding the resistance at any temperature is shown in the following section and example. The graph shows the ratio R_T/R_{25} (where R_T is the resistance at any temperature and R_{25} is the resistance at 25°C) plotted against temperature, and is a curve. A logarithmic plot of resistance against (1/temperature) can be more useful than one of resistance against temperature because it gives straight line characteristics that are easier to interpret and extrapolate, but such graphs are more difficult to extract useful information from.

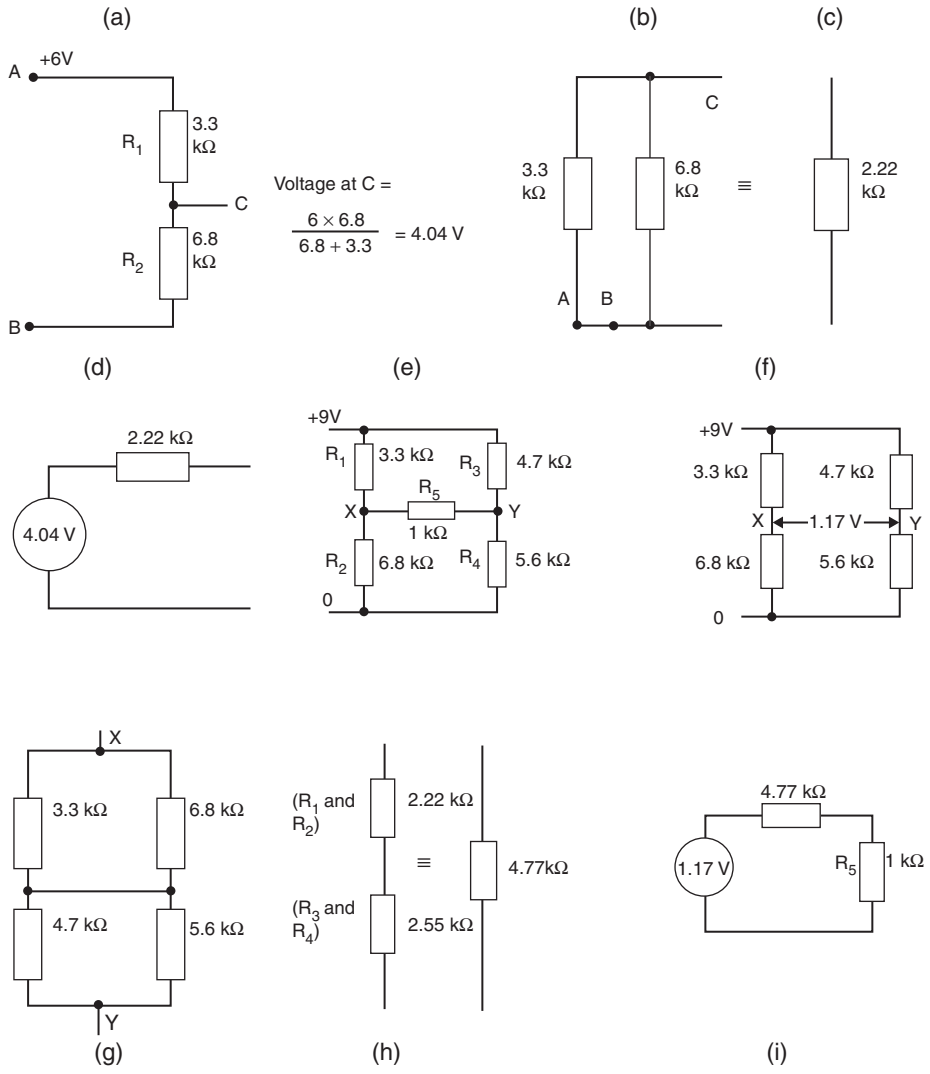


Figure 1.13

Using Thevenin's theorem. The potential divider **(a)** has an output voltage, with no load, of 4.04 V. It is equivalent to a 4.04 V source whose internal resistance is found by imagining the voltage supply short-circuited **(b)** and **(c)**, so the equivalent is as shown in **(d)**. This makes it easy to find the output voltage when a current is being drawn. Similarly the bridge circuit **(e)** will have an open-circuit voltage, with R_5 removed, of 1.17 V across X and Y **(f)**, and the internal resistance between these points is found by imagining the supply short-circuited **(g)**.The combination of resistors in **(g)** is resolved **(h)** to give the single equivalent **(i)**.

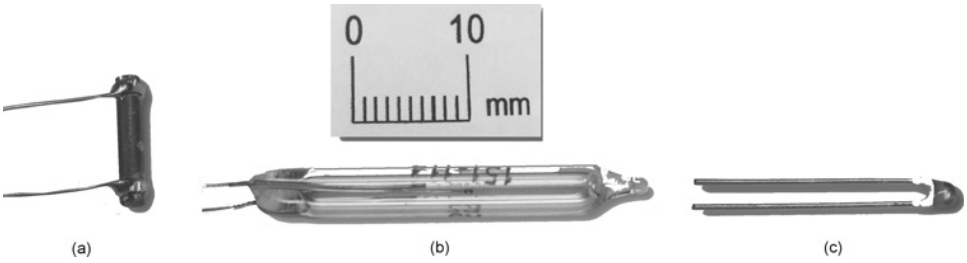


Figure 1.14

Some thermistor shapes: **(a)** rod, **(b)** glass bead, **(c)** bead. (Original pictures by Alan Winstanley.)

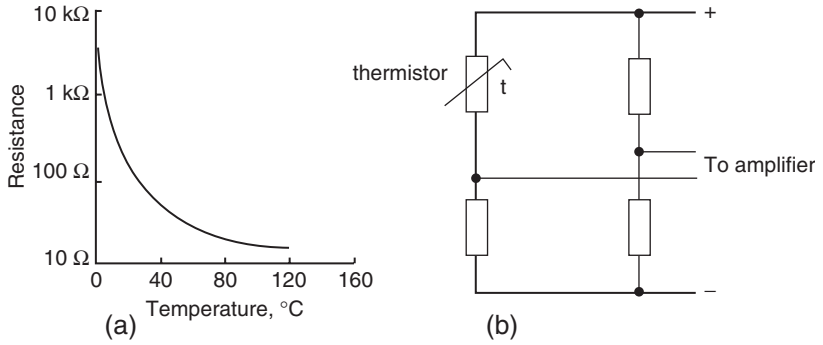


Figure 1.15

Thermistor bridge for temperature measurement. **(a)** Thermistor typical characteristics, **(b)** thermistor bridge for temperature measurement. Note the symbol for a thermistor. (Note the symbol for a thermistor.)

Variation of resistance with temperature

For a thermistor, the temperature variation of resistance is generally of the form:

$$R_{\theta_1} = R_{\theta_2} \exp\left(\beta\left[\frac{1}{\theta_2} - \frac{1}{\theta 1}\right]\right)$$

R_{θ_1} = resistance at temperature θ_1
R_{θ_2} = resistance at temperature θ_2
β = thermistor constant

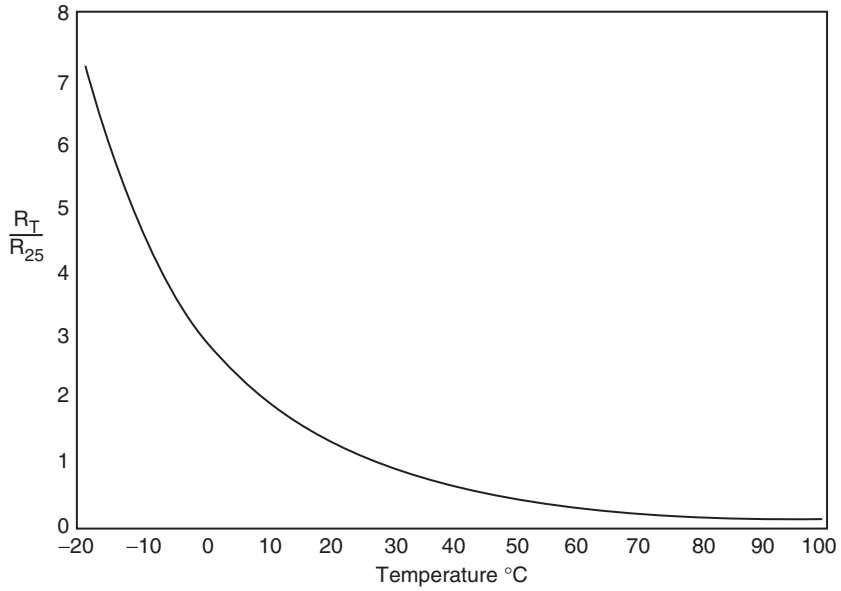


Figure 1.16

Graph of resistance ratio plotted against temperature for a typical thermistor.

- Temperatures are in the Kelvin scale, equal to °C + 273. The expression exp (x) is an alternative method of writing e^x which is better suited for printed equations, and is also used in calculators.

Example: *A thermistor has a resistance of* 47 kΩ *at* 20°C. *What is its resistance at* 100°C *if its* β *value is* 3900*?*

Using the above equation:

$$R_{100} = 47 \times \exp\left(3900\left[\frac{1}{373} - \frac{1}{293}\right]\right)$$
$$= 47 \times \exp(-2.8548) = 2.7 \text{ (resistances in k}\Omega\text{)}$$

Calculator procedure

Enter value of known temperature $\boldsymbol{\theta_1}$ then press keys **1/x** =

Enter value of θ_2 then press keys [1/x] [=] [×]. Enter value of **b** and press keys [=] [e^x] [×]

Enter value of R_{θ_2}

Press [=] key and read the answer.

CHAPTER 2

CAPACITORS

Capacitance

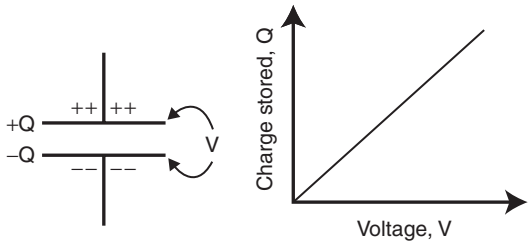
Two conductors that are not connected and are separated by an insulator constitute a capacitor. When a source of EMF such as a cell is connected to such an arrangement, current flows momentarily, transferring change (in the form of electrons) from one conducting plate (the + plate) to the other (Figure 2.1). When a quantity of charge Q (measured in units of coulombs) has been transferred, the voltage across the plates equals the voltage V across the voltage source. For a fixed arrangement of conductors and insulator, the ratio Q/V is a constant called the capacitance, C. The relationship can be written in the three forms:

$$Q = CV \quad C = Q/V \quad V = Q/C$$

with V in volts, Q in coulombs and C in farads.

Figure 2.1

Basic principles of the capacitor. The relationship Q/V shown in the graph, is defined as the capacitance, C.



The parallel-plate capacitor

The parallel-plate capacitor is the simplest theoretical (and practical) arrangement and its capacitance value is, for ideal conditions, easy

to calculate. For a pair of parallel plates of equal area A, separation d, the capacitance is given by:

$$C = \frac{\varepsilon_r \varepsilon_0 A}{d}$$

The quantity $\varepsilon\varepsilon_0$ is a universal constant called the **permittivity of free space**, and it has the fixed value of 8.84×10^{-12} farads per metre. Air has approximately this same value of permittivity also, but other insulating materials have values of permittivity that are higher by the factor ε_r, a pure number with no units, which is different for each material. Values of this quantity, now called **relative permittivity** (formerly called **dielectric constant**), are shown in Table 2.1 for some common materials. The formula can be recast, using units of cm^2 for area, mm of spacing, and result in pF, as:

$$C = \frac{0.88 \times \varepsilon_r \times A}{d} \text{ pF}$$

Table 2.1 Typical values of relative permittivity at 20°C

Material	Relative permittivity value
Aluminium oxide	8.8
Araldite resin	3.7
Bakelite	4.6
Barium titanate	600–1200 (varies with voltage)
FR4 (fibreglass PCB material)	4.5
Magnesium silicate	5.6
Nylon	3.1
Polystyrene	2.5
Polythene	2.3
PTFE	2.1
Porcelain	5.0
Quartz	3.8
Soda glass	6.5
Titanium dioxide	100

These units are more practical for small plate sizes but some allowance must be made for edge effects (the capacitance is slightly less than the predicted

value) and for stray capacitance between any conductor and the metal that surrounds it. Even a completely isolated piece of metal will have some capacitance and in some circumstances this may be significant.

Example 1: *Find the capacitance between two parallel plates* 2 cm × 1.5 m, *spaced by a* 0.2 mm *layer of material of relative permittivity value* 15.

Using $C = \varepsilon_1\varepsilon_0 A/d$, with $\varepsilon_r = 15$, $\varepsilon_0 = 8.84 \times 10^{-12}$ F/m, $A = 0.02 \times 1.5$ m^2 and $d = 0.2 \times 10^{-3}$ m

$$C = \frac{15 \times 8.84 \times 10^{-12} \times 0.02 \times 1.5}{0.2 \times 10^{-3}}$$
$$= 1.989 \times 10^{-8}\,\text{F} \approx 0.02\ \mu\text{F}$$

which is about 2×10^{-8} F or 0.02 μF.

Example 2: *Find the capacitance between two parallel plates* 2 cm × 1 cm *spaced* 0.1 mm *apart by a material with relative permittivity* 8.

Using $C = 0.88 \times \varepsilon_r \times A/d$ with $A = 2 \times 1 = 2$ cm^2 and $d = 0.1$ mm we get, in pF:

$$C = \frac{0.88 \times 8 \times 2}{0.1} = 140.8 \text{ or } 141\ \text{pF}$$

Construction

Like resistors, capacitors can be obtained in the older wire-connected style, or, more commonly now, as SMD components. Small-value capacitors can be made using thin plates of insulating material (a *dielectric*) metallized on each side to form the conductors. Thin plates can be stacked and interconnected (Figure 2.2c), to form larger capacitance values up to 1000 pF or more.

Silvered mica (also called silver mica) types were formerly used where high stability of value is important, as in oscillators, but are now quite rare, having been replaced by porcelain types or for some purposes by the C0G/NP0 (see later) types. Porcelain has come into use because, unlike mica (a natural material whose specifications can vary wildly) the materials can be

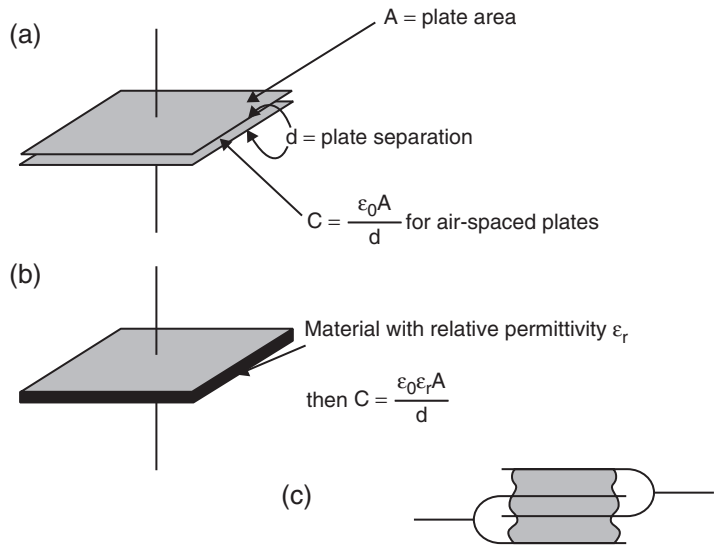


Figure 2.2

The parallel-plate capacitor, **(a)** air-spaced, **(b)** using a dielectric, **(c)** with multiple plates.

manufactured to a tight specification. Porcelain capacitors are found mainly in SMD form, and are used extensively in RF and microwave circuits.

Ceramics are used generally for less critical applications such as RF coupling and decoupling. **Ceramic tubular** capacitors make use of small ceramic tubes that are silvered inside and outside. Ceramic capacitors have, typically, values that range from 1 pF to 0.22 μF for ceramic disks, and up to 10 μF for multi-layer types (ceramic chips). The scale of values usually follows the E12 values of 1.0, 1.2, 1.5, 1.8, 2.2, 2.7, 3.3, 3.9, 4.7, 5.6, 6.8, 8.2.

Ceramic capacitors are graded into types referred to as **C0G**, **X7R**, and **Y5V**. The letter-number-letter references are used to identify temperature characteristics, using codes that depend on the classification of the dielectric. There are four classes, and the lower the class number the better the performance of the capacitor; the Class 4 types are virtually obsolete. The code for Class 1 dielectrics uses the first symbol to indicate the significant figures of the temperature coefficient in ppm/°C, the second figure is the

multiplier, and the third is the tolerance in ppm/°C. Typical of this class is the C0G type; the complete coding is shown in Table 2.2.

Table 2.2 Letter-number coding of ceramic capacitors

Significant figure in ppm/°C	Multiplier	Tolerance in ppm/°C (25–85°C)
C = 0.0	0 = −1	G = ±30
B = 0.3	1 = −10	H = ±60
L = 0.8	2 = −100	J = ±120
A = 0.9	3 = −1000	K = ±250
M = 1.0	4 = +1	L = ±500
P = 1.5	6 = +10	M = ±1000
R = 2.2	7 = +100	N = ±2500
S = 3.3	8 = +1000	
T = 4.7		
V = 5.6		
U = 7.5		

For Class 2 and 3 dielectrics (which include the popular X7R, X5R, Z5U and Y5V types), the code is different (Table 2.3). The first symbol indicates the lower limit of the operating temperature range, the second indicates the upper limit of the operating temperature range, and the third indicates the maximum capacitance change allowed over the operating temperature range.

Table 2.3 Class 2 and 3 coding system

Low temperature limit:

X = −55°C Y = −30°C Z = +10°C

High temp. limit:

4 = +65°C 5 = +85°C 6 = +105°C 7 = +125°C 8 = +150°C 9 = +200°C

Capacitance change over temperature range:

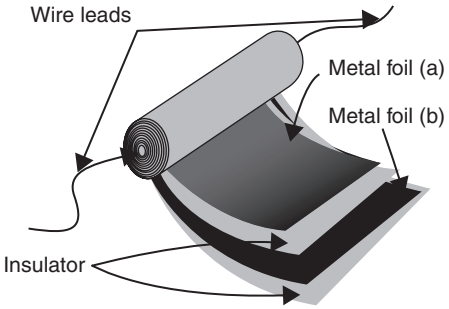
A = ±1.0%	B = ±1.5%	C = ±2.2%	D = ±3.3%	E = ±4.7%	F = ±7.5%
P = ±10%	R = ±15%	S = ±22%	T = +22 to −33%	U = +22% to −56%	V = +22% to −82%

The popular C0G types have zero temperature coefficient (usually ±30 ppm/°C) and have the highest stability and lowest loss of all the ceramic types. The X7R ceramics have higher losses, but are small and cheap, and are obtainable as multilayer types (particularly in SM form). Ceramic chip capacitors use ceramic dielectric materials which have been formed into thin layers with metal film electrodes alternately exposed on opposite edges of the set of laminates. This assembly is then fired at high temperature in absence of oxygen to produce a single block of ceramic, to which metal connections can be made at the opposite edges. The film chip type can be made in high values (up to 4.7 μF), intended particularly for power supply filtering applications where a low effective series resistance (ESR, see later) is desirable.

Rolled capacitors use strips of insulating material as their dielectric. Paper was formerly used, but because the characteristics of paper are so variable, it is much more common to use polyethylene (polythene), polyester, polycarbonate, polypropylene or other plastics films which are metallized and then rolled up (Figure 2.3), with another insulating strip to prevent the metallizing on one side shorting against the metallizing on the other side. Using this construction, quite large capacitance values can be achieved in a small volume and values of up to several μF are common.

Figure 2.3

The rolled construction used for capacitors which makes use of sheet dielectrics such as paper, polyester, polystyrene or polycarbonate.



Electrolytic capacitors are used when very large capacitance values are needed; the more common type is the aluminium electrolytic. One plate is of aluminium foil in contact with an aluminium perborate solution in the form of a jelly or paste; the other plate is an aluminium container.

The insulator is a film of aluminium oxide which forms on the positive plate when a voltage, called the forming voltage, is applied during manufacture. Because the film of oxide can be very thin, only a few molecules thick, and the surface area of the aluminium foil can be very large, especially if the surface is roughened, very large capacitance values (up to several farads) can be achieved.

The disadvantages of aluminium electrolytics include leakage current (which is high compared to other capacitor types), the need for polarization (the + and − markings must be observed and DC applied) and comparatively low-voltage operation (less significant in transistor and IC circuits, but ruling out the use of electrolytics in high-voltage transmitter circuits). Incorrect polarization can cause the oxide layer to break down and if large currents then flow, as is likely if the capacitor is used as the reservoir in a power supply unit, the capacitor will explode, showering its surroundings with corrosive jelly. **Tantalum electrolytics** use a solid dielectric and can be used unpolarized (but not necessarily **reverse** polarized) and have much lower leakage currents than aluminium types, making them more suitable for some applications.

One factor that is quoted for electrolytics to a greater extent than for other types is the **ESR**, effective series resistance. The ESR is the pure resistance of a capacitor to an AC signal. The significance of this is that if the reactance of a capacitor is very low, its capability for carrying current is reduced, and the heating caused by current is much greater. High ESR values can cause many problems with power supplies, particularly switching power supplies, and can also present problems with time constants and circuit loading. Ultra-low ESR electrolytics are quoted as having ESR of 0.025 Ω or less, but some electrolytic types can have values of more than 1 Ω, some even more than 10 Ω. Many modern applications call for low-ESR capacitors to be used.

The ESR is related to the loss factor for the capacitor. The loss angle (δ) for a capacitor is defined as the phase angle between signal current and signal voltage, and the loss factor is the (trigonometric) tangent of the angle, tan δ. The relationship between δ and ESR is:

$$\text{Tan } \delta = 2\pi f(\text{esr})$$ where f is the frequency of the applied current

Low-impedance electrolytics can be specified for critical tasks, and the choice is between the Sanyo OS-CON aluminium type, which uses an organic semiconducting electrolyte, and the low-ESR type of tantalum electrolytic.

In Figure 2.4 are shown the shapes of a variety of capacitor types.

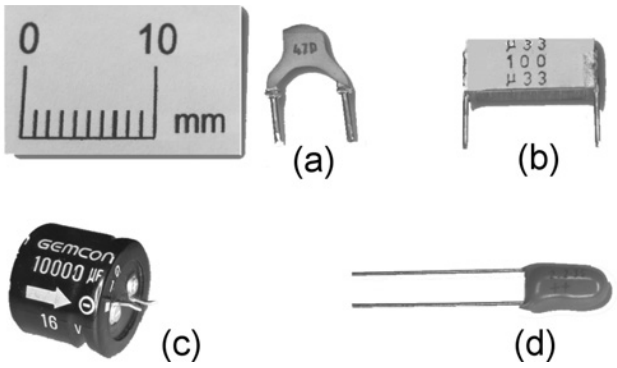


Figure 2.4

Capacitor selection: **(a)** ceramic, **(b)** polyester stacked film, **(c)** electrolytic, **(d)** tantalum electrolytic. (Original photos by Alan Winstanley.)

Other capacitor characteristics

The same series of preferred values (usually 20% and 10%) that are used for resistors are applied also to values of capacitance other than electrolytics. Some old components will still be found with values such as 0.02 μF, and can be replaced with the preferred value of 0.022 μF. Some capacitor manufacturers mark the values in pF only, using the prefix k (confusingly) to indicate thousands of pF (equal to nanofarads) (Table 2.4). Colour-coded values are always in pF units. For SM capacitors, the two- and three-symbol codes shown earlier for resistors are used with values in pF.

Electrolytic aluminium capacitors are always subject to very large tolerance values, of the order of −50% + 100%, so the actual capacitance value may

Table 2.4 Colour coding for small block or bead capacitors

Band	1	2	3	4
Black	–	0	×1	10 V
Brown	1	1	×10	
Red	2	2	×100	
Orange	3	3	–	
Yellow	4	4	–	6.3 V
Green	5	5	–	16 V
Blue	6	6	–	20 V
Violet	7	7	–	
Grey	8	8	×0.01	25 V
White	9	9	×0.1	3 V
Pink	–	–	–	35 V

range from half the marked value to twice the marked value. The insulation resistance between the plates is often so low that capacitance meters are unable to make accurate measurements. Capacitance values marked in circuit diagrams can use the BS1852 method of 6n8, 2μ2, etc., but are often marked in μF or pF. Quite commonly, fractional values refer to values in μF and whole numbers to pF unless marked otherwise, so values of 0.02, 27, 1000 and 0.05 mean 0.02 μF, 27 pF, 1000 pF (= 1 nF) and 0.05 μF respectively.

For all capacitors, the working voltage limits (abbreviated as *VW*) must be carefully observed. Above this voltage limit, sparking between the conductors can break down the insulation causing leakage current and the eventual destruction of the capacitor. The maximum voltage that can be used is much lower at high ambient temperatures than at lower temperatures. Some types have a limited self-heal capability following sparking. Note that the lower temperature limit is vitally important for electrolytics because when the jelly freezes the electrolytic action ceases. Working voltage values as low as 3 V may be found in high-value electrolytics such as are used as voltage backup in digital circuits, and values as high as 20 kV can be used for ceramic capacitors intended for TV EHT circuits and for transmitter circuits. Capacitors for higher working voltages can be constructed to special order. In Table 2.5 are shown the common working voltages used in semiconductor circuits.

Table 2.5 Capacitors – common working voltages

10 V	16 V	20 V	25 V	35 V	40 V
63 V	100 V	160 V	250 V	400 V	1000 V

Changes of temperature and of applied DC voltage both affect the value of capacitors because of changes in the dielectric. Both PTC and NTC types can be obtained, and the two are often mixed to ensure minimal capacitance change in, for example, oscillator circuits. Paper and polyester capacitors have, typically, positive temperature coefficients of around 200 ppm/°C, but silver mica types have much lower positive temperature coefficients.

Aluminium electrolytics have large positive temperature coefficients with a considerable increase in leakage current as temperature increases. In addition, as noted above, such electrolytics cannot be used below −20°C because the electrolyte paste or jelly freezes. The normal working range for other types is −40°C to +125°C, though derating may be needed for the higher voltages. The specified voltage ratings generally apply up to 70°C ambient temperature. A few types of capacitors, notably the High-K ceramics, change value as the applied voltage is varied. Such capacitors are quite unsuitable for use in applications such as tuned circuits, and should be used only for non-critical decoupling applications.

Variable capacitors can make use of the variation of overlapping area, or of variation of spacing between parallel plates. Air dielectric is used for the larger types (360 pF or 500 pF) but miniature variables make use of mica or plastic sheets between the plates. Compression trimmers are manufactured mainly in the smaller values, up to 50 pF. In use, the moving plates are always earthed, if possible, to avoid changes of capacitance due to stray capacitance when the control shaft is touched. This capacitance change has been used deliberately in the famous Theremin (the first musical synthesizer) and is also used in proximity detectors.

- Derating must be applied to all capacitors, according to the manufacturer's instructions, for extremes of temperature, voltage or frequency.

Energy and charge storage

The amount of **charge** stored by a capacitor is given by $Q = CV$, and when C is given in μF and V in volts, charge Q is then in microcoulombs (μC).

Example: *How much charge is stored by a* 0.1 μF *capacitor charged to* 50 V*?*

Using $Q = CV$ with C in μF, V in volts then $Q = 0.1 \times 50 = 5\ \mu C$

The amount of **energy**, in units of joules, stored by a charged capacitor is most conveniently given by $W = \frac{1}{2}CV^2$. Other equivalent expressions are:

$$W = \frac{Q^2}{2C} \quad \text{or} \quad W = \frac{QV}{2}$$

Example: *How much energy is stored in a* 5 μF *capacitor charged to* 150 V*?*

Using $W = \frac{1}{2}CV^2$

with $C = 5 \times 10^6$ and $V = 150$ then $W = \frac{1}{2} \times 5 \times 10^{-6} \times (150)^2 = 0.056$ J

This calculation is used in connection with the use of capacitors to fire flash bulbs or in capacitor discharge car ignition systems. In circuits, the laws concerning the series and parallel connections of capacitors are the inverse of those for resistors:

For capacitors in parallel, the voltage across each capacitor is equal, so:

$$C_{total} = C_1 + C_2 + C_3 + \ldots$$

For capacitors in series, the charge on each capacitor is equal, so:

$$\frac{1}{C_{total}} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \ldots$$

- **TIME CONSTANTS**

The charging and discharging of a capacitor is never instant. When a sudden step of voltage is applied to one plate of a capacitor, the other plate voltage

will step in voltage by the same amount. If a resistor is present that connects the second plate to a different voltage level the capacitor will then charge or discharge to this other voltage level. The time needed for this change is about 4 time constants, as shown in Figure 2.5 – theoretically the time is infinite but a time of four times the time constant allows the charge to reach 98% of the final amount. A figure of 3 times the time constant is sometimes used, representing 95% charging. The mathematical basis for these figures is illustrated later in this chapter.

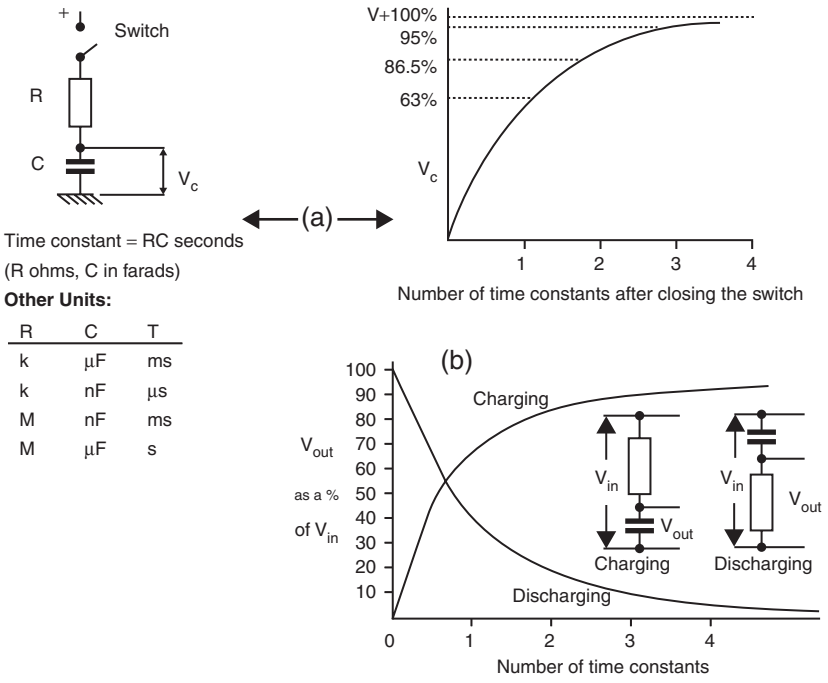


Figure 2.5

Capacitor charging and discharging: **(a)** principles of charging, **(b)** universal charge discharge curves.

The quantity called **time constant**, T, is measured by R × C where R is the resistance of the charge or discharge resistor and C is the capacitance. For C in farads and R in ohms, the time constant T is in seconds. For the

more practical units of μF and kΩ, T is in milliseconds (ms); and for C in nF and R in kΩ, T is in microseconds (μs).

Example: *In the circuit of Figure* 2.6*a; how long does the voltage at the output take to die away?*

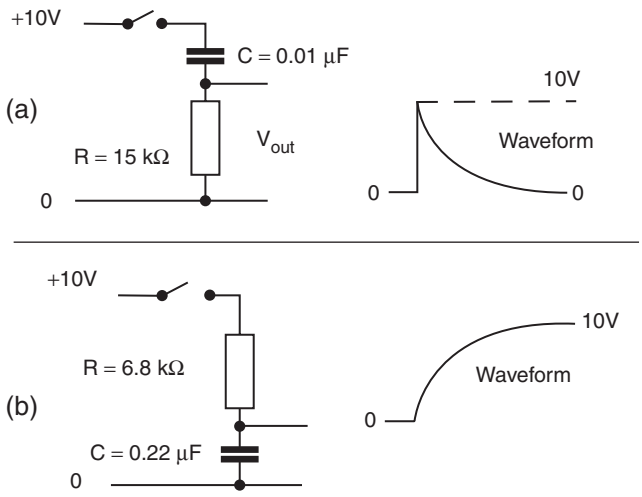


Figure 2.6

Time constant: **(a)** differentiating circuit, **(b)** integrating circuit.

Solution: With C = 0.01 μF (equal to 10 nF) and R = 15 kΩ, T = 150 μs. Four time constants will be 4 × 150 μs = 600 μs, so we can take it that the output voltage has reached zero after 600 μs.

Example: *In the circuit of Figure* 2.6*b, how long does the capacitor take to charge to* 10 V*?*

Solution: With C = 0.22 μF (or 220 nF) and R = 6.8 kΩ, T = 6.8 × 200 which is 1496 μs. Four time constants will be 4 × 1496 = 5984 μs or 5.98 ms, approximately 6 ms of charging time.

These calculations of charging and discharging times are important in determining the shape of the output when a step voltage is applied to a capacitor–resistor combination.

The bases of these time constant calculations are the exponential charging and discharging formulae for a capacitor. For a capacitor discharging from an initial voltage V_0 through a time constant RC, the voltage V across the capacitor changes in time t is described by the equation:

$$V = V_0 \exp\left(\frac{-t}{RC}\right)$$

For a capacitor being charged to a voltage V_0 through a time constant, the equation becomes:

$$V = V_0\left(1 - \frac{t}{RC}\right)$$

If we assume that $t = 4RC$, and rearrange the first equation, we get:

$$V/V_0 = \exp(-4)$$

which gives $V/V_0 = 0.0183$, so the voltage has reached 1.8% of the initial voltage, well discharged. For the charging of a capacitor the four time constants give $(1 - 0.0183) = 0.9817$, around 98% of final voltage.

We can also rearrange the equations to find the time required to charge or discharge a capacitor to a required level. For a discharge:

$$T = -RC\ln\left(\frac{V}{V_0}\right)$$

with the symbol meanings as before, and ln meaning natural logarithm. For example, if you want to find that time is needed to discharge from 10 V to 4 V with time constant 10 μs, the formula becomes:

$$T = 10 \times 10^{-6}\ln\left(\frac{4}{10}\right) = 10^{-5} \times 0.916 = 9.16 \times 10^{-6}\text{ or } 9\,\mu\text{s}.$$

The negative sign is needed because of the negative values of the natural logarithm (ln).

For charging, the formula becomes $T = -RC\ln\left(1 - V/V_0\right)$.

- **REACTANCE**

The reactance of a capacitor for a sine wave signal is given by:

$$X_C = \frac{1}{2\pi fC} \quad (2\pi = 6.28)$$

where C is the capacitance, in farads and f is frequency, in hertz. Reactance is measured in units of ohms and is defined by the ratio:

$$\tilde{V}/\tilde{I} \quad \text{or} \quad v/i \quad \text{using the convention noted below,}$$

where $\tilde{V}$ is the AC voltage across the capacitor and $\tilde{I}$ is the AC current through the circuit containing the capacitor. Unlike resistance, reactance is not a constant but, for a capacitor, varies inversely with frequency (Figure 2.7a). In addition, the sine wave of current is ¼ cycle (90°) ahead of the sine wave of voltage across the capacitor plates.

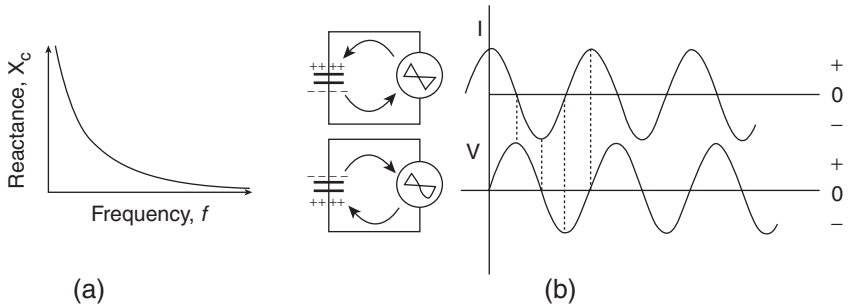


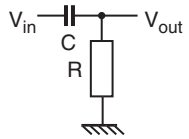
Figure 2.7

Capacitive reactance to AC signals: **(a)** graph showing how capacitive reactance varies with frequency of signal, **(b)** phase shift. As the capacitor charges and discharges, current flows alternately in each direction. The maximum current flow occurs when the capacitor is completely uncharged (zero voltage), and the maximum voltage occurs when the capacitor is completely charged (zero current). The graph of current is therefore one-quarter of a cycle (900) ahead of the graph of voltage.

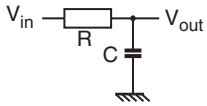
NOTE: From this point on we shall use the convention that electrical quantities in uppercase (such as V, I) mean steady (DC) or mains AC values, and quantities in lowercase (such as v, i) mean signal values

Table 2.6 Amplitude and phase tables for RC circuits

$G = v_{out}/v_{in}$; φ = phase angle; the time constant $T = CR$, $\omega = 2\pi \times$ frequency (f)



ωT	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.5	2.0	3.0	4.0	5.0
$\varphi°$	84.3	78.7	73.3	68.2	63.4	59.0	55.0	51.34	48.0	45.0	33.7	26.6	18.4	14	11.3
G	0.099	0.196	0.287	0.37	0.45	0.51	0.57	0.62	0.67	0.707	0.83	0.9	0.95	0.97	0.98



ωT	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.5	2.0	3.0	4.0	5.0
$\varphi°$	−5.7	−11.3	−16.7	−21.8	−26.5	−31	−35	−38.6	−41.8	−45	−56	−63	−72	−76	−79
G	0.99	0.98	0.96	0.9	0.89	0.85	0.82	0.78	0.74	0.707	0.55	0.45	0.32	0.24	0.2

- **CR CIRCUITS**

A CR circuit is one that contains both capacitors and resistors (either in series or in parallel). The action of a CR circuit upon a sine wave is to change both the amplitude and the phase of the output signal as compared to the input signal. For the action of inductive-resistive circuits see Chapter 3. Universal amplitude/phase tables can be prepared, using the time constant of the CR circuit and the frequency **f** of the sine wave, These tables are shown, with examples, in Table 2.6.

This page intentionally left blank

CHAPTER 3

INDUCTIVE AND TUNED CIRCUIT COMPONENTS

Inductors

An inductor is a component whose action depends on the magnetic field that exists around any conductor when a current flows through that conductor. When the strength of such a magnetic field (or magnetic *flux*) changes, a voltage is induced between the ends of the conductor. This voltage is termed an **induced EMF**, using the old term of EMF (**electromotive force**) to mean a voltage that has not been produced by a current flowing through a resistor.

At one time inductors were invariably fairly large components and were used in domestic radios as well as in a variety of other applications, but modern inductors for signal use are often SMD components and though used to a much lesser extent in domestic radio are extensively employed in other devices. Inductors intended for 50 Hz AC mains are invariably large components, but the extensive use of switch-mode power supplies has reduced the need for these items, though they are still made in large quantities.

If we confine our attention to static devices such as coils and transformers rather than moving devices such as electric motors, the change of magnetic field or flux can only be due to a change in the current through one conductor. The induced EMF is then in such a direction that it **opposes** this change of current, and the faster the rate of change of current the greater is the opposing EMF. Because of its direction, the induced EMF is called a **back-EMF**. The laws governing these effects are Faraday's laws and Lenz's law, summarized in Figure 3.1.

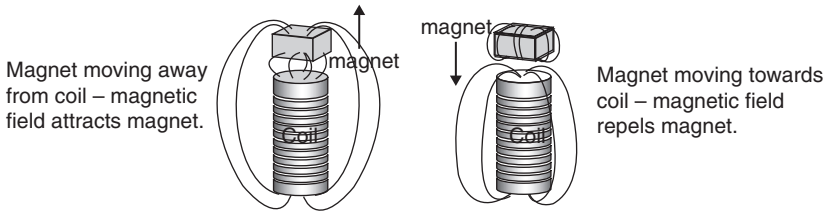


Figure 3.1

Faraday's and Lenz's laws. Faraday's laws relate the size of the induced (generated) voltage in a coil to the strength, speed of the magnet and the size of the coil. Lenz's law is used to predict the direction of the voltage.

Faraday's laws: *Voltage induced depends on rate of change of magnetic flux. For an EMF caused by a moving magnet, change of flux is proportional to strength of magnet, speed of magnet or coil, number of turns of coil and area of cross-section of coil.*

Lenz's law: *The direction of induced EMF is such that it always opposes the change (movement in this example, or change of current through a coil) that causes it.*

The size of the back-EMF can be calculated from the rate of change of current through the conductor and the details of construction of the conductor such as straight wire or coil, number of turns of coil, use of a magnetic core and so on. These constructional factors are constant for a particular conductor and can be lumped together as one quantity called **inductance** or, more correctly, **self-inductance**, symbol L.

L is defined in the equation:

$$E = L\frac{dI}{dt} \quad \text{where}$$

E is the amount of back-EMF
L is inductance
$\frac{dI}{dt}$ is the rate of change of current

The symbol 'd' is used here to mean a small change of the quantity that follows – the notation of calculus. If E is measured in volts and dI/dt is the rate of change of current in amps per second, then L is in units of

henries (H), named after the US pioneer Joseph Henry whose discoveries parallel those of Faraday.

Example 1: *What back-EMF is developed when a current of* 3 A *through a* 0.5 H *coil is reduced to zero in* 20 ms*?*

The amount of back-EMF is found from:

$$E = L\frac{dI}{dt} = 0.5 \times \frac{3}{20 \times 10^{-3}} = 75\,\text{V}$$

Note that this 75 V back-EMF will exist only for as long as the current is changing at the quoted rate. The back-EMF may be much greater than the normal DC voltage drop across the resistance of the inductor when a steady current is flowing. The rate of change of current is seldom (almost **never**) uniform like this, so the back-EMF is usually a pulse waveform whose maximum value can be found by measurement.

The existence of self-inductance in a circuit causes a reduction in the rate at which current can increase or decrease in the circuit. For a coil with inductance L and resistance R, the time constant for the circuit is L/R seconds, with L in henries and R in ohms. In Figure 3.2 it is shown how the current at a time t after switch-on varies in an inductive circuit – once again we can take the time of 4 time constants to represent the end of

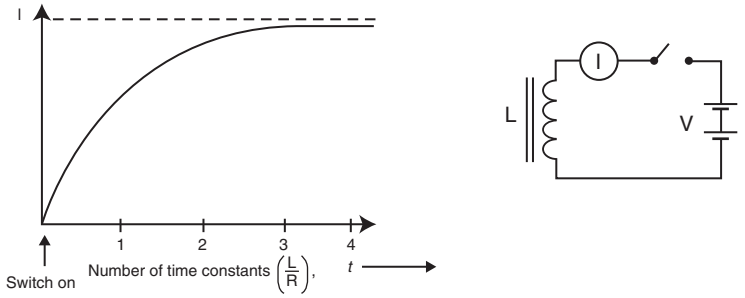


Figure 3.2

The growth of current in an inductive circuit.

the process. Mathematically, the current at time t is given by:

$$I = I_{max}\left(1 - \exp\left[\frac{-Rt}{L}\right]\right)$$

where I_{max} is the final value of current (equal to E/R where E is the voltage and R is the total circuit resistance). When a current I_{max} is switched off, the equation for current becomes:

$$I = I_{max} \exp\left[\frac{-Rt}{L}\right]$$

The large EMF (equal to LdI/dt) which is generated when current is suddenly switched off in an inductive circuit can have destructive effects, causing sparking at contacts or breakdown of semiconductor junctions. Figure 3.3 shows the commonly used methods of protecting switch contacts and semiconductor junctions from these switching transients.

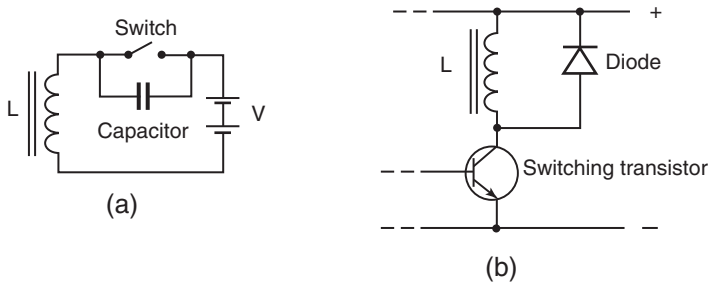


Figure 3.3

Protection against voltage surges in inductive circuits: **(a)** using a capacitor across switch contacts, **(b)** using a diode across the inductor.

The changing magnetic field around one coil of wire (or any other conductor) will also affect other windings nearby. The two windings are then said to have **mutual inductance**, symbol M, and units henries. This is the principle of the **transformer**.

By definition:

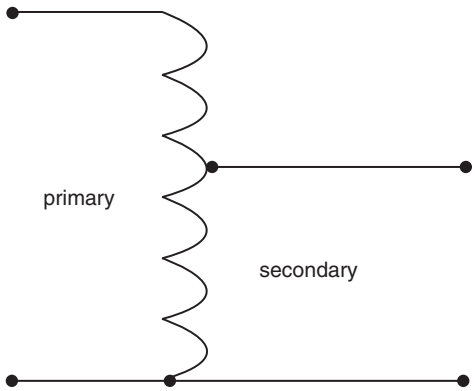
$$M = \frac{\text{back-EMF induced in second winding}}{\text{rate of change of current in first winding}}$$

Transformers

Transformers make use of the effect of **mutual induction**, whether they are the multiple winding type of transformer or the **autotransformer**, in which one single winding is used, with connections tapped for different connections (Figure 3.4). The main types of transformers that are used in modern electronics circuits are:

Figure 3.4

Principle of an autotransformer.



1. mains transformers, used in power supplies, requiring a large core size;

2. matching transformers used for feeding lines;

3. tuned transformers, used in signal amplifiers to achieve a specified bandwidth.

Of these, the older forms of tuned transformers are seldom used now, having been replaced by combinations of wideband ICs and electromechanical filters, so we shall confine our attentions to the mains and the signal-matching types.

The transformer – other than the autotransformer type – has at least two windings, one of which is designated as the **primary** winding, the other as the **secondary**, and the action consists of an alternating voltage applied to the primary winding so causing an alternating voltage to appear at the secondary. Unless the transformer is intended only for purposes of isolation, the primary and secondary voltage levels are usually different. The conventional style of transformer consists of a bobbin on which both primary and secondary windings are formed, usually with a metal foil layer between the windings to act as an electrostatic screen. The core is then assembled by putting E and I sections of thin steel alloy into place, with the bobbin lying in the arms of the E section (see Figure 3.5).

There are, however, several other forms of construction. When twin bobbins are used side by side, the electrostatic screening can often be dispensed with,

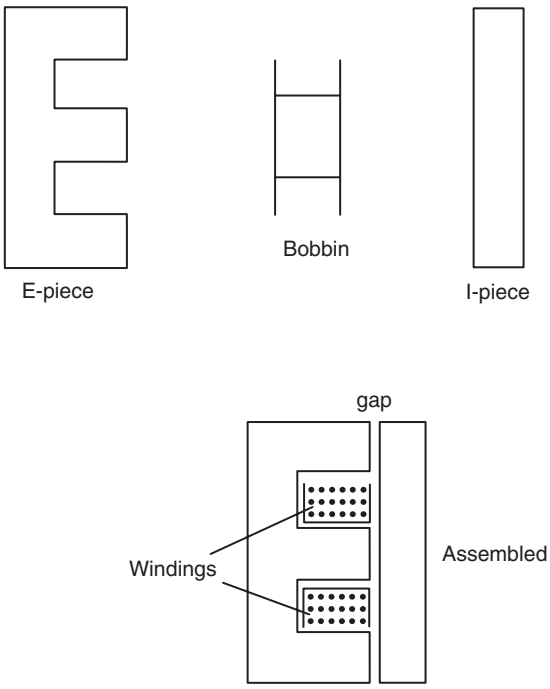


Figure 3.5

The E and I form of core for small transformers.

and some transformers make use of a **C-core** – a pair of C-shaped metal pieces – rather than the E and I structure. Another form, very common now, is the **toroidal** transformer, in which both windings are placed over a ring of magnetic material. The toroidal type has in the past been very expensive to produce because of the difficulty of winding the turns into place, but development of toroidal winding machinery has made these transformers much more readily available. Their main advantage is that they have a very low external magnetic field, so they are often specified for use in equipment where hum pickup levels must be kept as low as possible. Another type, used for audio and radio frequencies, is the **pot-core** variety in which the coils are not only wound over a core but surrounded by a magnetic casing.

A **perfect transformer** can be defined as one in which no power is dissipated, so the power supplied to primary winding (equal to primary voltage × primary current) is exactly equal to the power taken from the secondary (secondary voltage × secondary current). Only very large transformers approach this state of perfection, and for the sizes that are encountered in electronics the efficiency of a transformer, defined as:

$$\frac{\text{power output at secondary}}{\text{power input at primary}}$$

will be of the order of 80% to 90%. For many purposes, however, the power loss in a transformer is not particularly important provided it does not cause the transformer to overheat.

- The role of efficient mains-frequency transformers in electronic equipment is now much less because of the prevalence of switch-mode supplies (see Chapter 7).

Another equivalent definition of perfection in a transformer is that all of the magnetic flux of the primary winding will cut across the secondary winding. This leads to another way of defining transformer losses in terms of **leakage inductance**, meaning the portion of the primary inductance which has no inductive effect on the secondary. Leakage inductance is more commonly used to define losses in a signal-carrying transformer than for a mains type, particularly since irregularities in the response of a transformer to wide-band signals are usually caused by leakage inductance and its resonance with stray capacitances.

As a result of the zero-power-loss definition of a perfect transformer, there is a simple relationship between the voltages and a number of turns at primary and secondary respectively, which is:

$$\frac{\text{primary voltage}}{\text{secondary voltage}} = \frac{\text{primary number of turns}}{\text{secondary number of turns}}$$

assuming that the self-inductance of the primary winding is enough to form a reasonable load in itself, because if the primary self-inductance is too low, the efficiency of the transformer will also be very low. In general, the higher the ratio of reactance to resistance for the primary winding, the more efficient the transformer is likely to be.

For all types of transformers other than autotransformers, the isolation between primary and secondary windings is important. Transformers that are specifically designed for isolation will include a DC voltage isolation test as part of the specification, and for such purposes it is normal for the insulation to be able to withstand several kilovolts DC between the primary and secondary windings without measurable leakage. The insulation from each winding to ground (usually the core or casing of the transformer) should also be of the same order.

Some types of transformers can be used with direct current flowing, and for such transformers the maximum amount of DC is stated, because excessive current could cause saturation. Saturation of the core means that the relative permeability will be reduced almost to the value for air, so transformer action will be almost lost; this is usually avoided by having an air-gap in the core (see Figure 3.5), thus restricting the amount of flux. A few transformers are designed such that the core will saturate on overload to prevent excessive signal being passed to the secondary circuit, and some types of transformer use the same principle to distort signals for wave-shaping purposes.

Signal-matching transformers

Many types of signal-matching transformers can be bought ready-made. These include 600 Ω line **isolating transformers** which are used to isolate telephone users' equipment, particularly mains-connected equipment such as facsimile (fax) equipment and computers, from the telephone lines

in order to ensure that it would be impossible for mains voltage ever to be connected to the telephone line. Such transformers must be obtained from sources who can guarantee that they are constructed to standards approved by the telephone company; in the UK the relevant specification is HED 25819. These telecommunications isolating transformers are of 1:1 turns ratio, and an overload on the consumer's side of the winding will fuse the winding rather than cause high voltages to be passed to the telephone lines. This happens because an overload saturates the core so that it becomes totally inefficient as a magnetic coupling between primary and secondary.

A very common type of signal-matching application is for '**100 V line**' transformers for public-address systems. Because the power loss of audio signals on long cables is proportional to the square of current, the output of an amplifier for public-address use is usually at a standard level of 100 V for full rated power, so the current is comparatively low. Since loudspeakers generally have impedances in the 3 Ω to 15 Ω range, a matching transformer is needed for each loudspeaker (Figure 3.6). Matching transformers of this type have a selection of secondary tapping points to allow the use of loudspeakers of various impedance ratings, and the power handling can be from 1 W to several kW.

- US readers should note that the use of the word 'line' in this context has no connection with power lines.

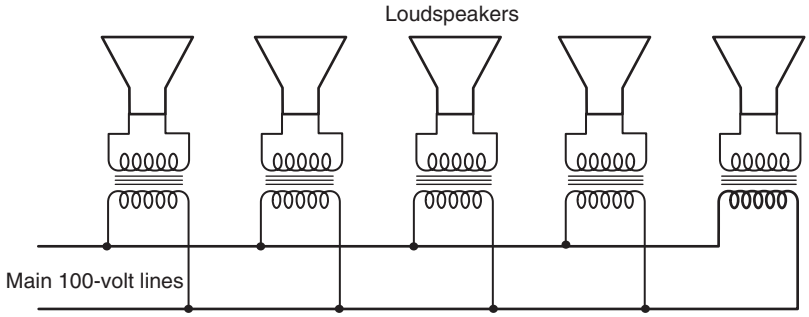


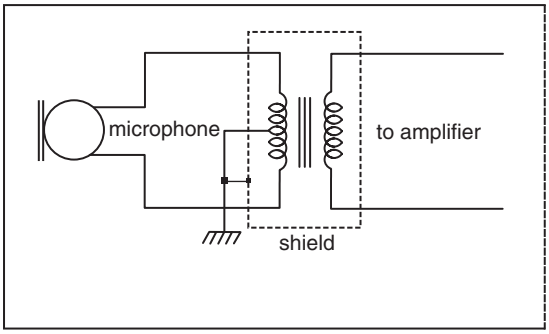
Figure 3.6

Using matching transformers for a PA system.

Another audio application for a matching transformer is the **microphone transformer** which is intended to match a low-impedance microphone into a high impedance amplifier. General-purpose matching transformers of this type are designed for moving-coil microphones in the impedance range 20 Ω to 30 Ω, or dynamic microphones in the 200 Ω to 600 Ω region, and more specialized types can be obtained for ribbon microphones, usually from the manufacturers of the microphones. The primary winding of a microphone is usually centre-tapped so that the microphone cable can be balanced around ground, as illustrated in Figure 3.7, greatly reducing hum pickup, and the whole transformer is encased in metal shielding to minimize hum pickup in the transformer windings.

Figure 3.7

Using a centre-tapped transformer for a microphone lead to minimize hum pickup.



The other standard forms of signal transformer are **pulse transformers**, which are intended to transmit pulse waveforms between circuits that may be at very different AC or DC levels, such as thyristor circuits. There is no requirement for such transformers to carry low frequency signals, and their leakage inductance also is of little importance, so very small units can be used, subject to the insulation resistance being sufficient. A typical requirement is for a voltage test to 2.8 kV peak for a transformer intended to work in the bandwidth of 3 kHz to 1 MHz.

A factor that is often quoted for these pulse transformers is the **voltage–time product**, meaning the product of output pulse amplitude (in volts) and pulse duration (in microseconds). This product, typically 200 Vμs is a way of ensuring that the transformer does not suffer from excessive dissipation from pulse signals. Pulse transformers of this type can be obtained with

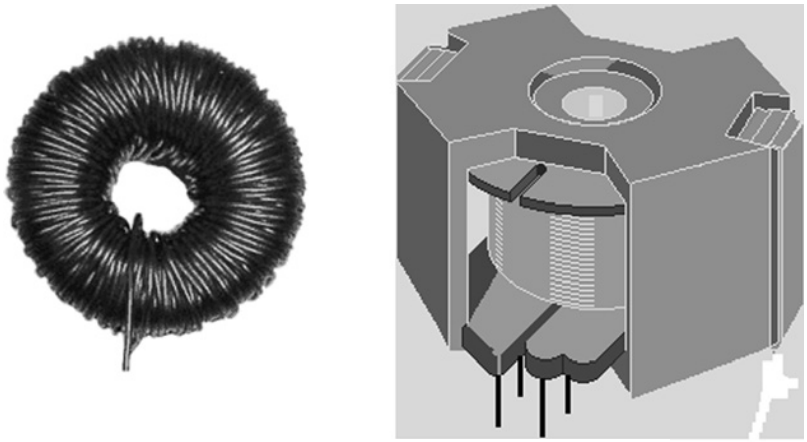
1:1 windings, 1:1 + 1 (two secondaries, or centre-tapped secondary) or 2:1 + 1 ratios. Primary inductance levels are in the range 3–12 mH with leakage inductance values of 8–30 μH. These transformers can be obtained as open or fully encapsulated units according to requirements.

For other requirements, particularly RF line to amplifier matching, the transformers have to be constructed to specification. In some cases, a simple tapped winding (autotransformer) will be sufficient; for other applications a transformer may have to be made to a very strict specification. Some of the most useful information on such transformers and on wound components generally is contained in amateur radio handbooks, obtainable from either the RSGB in the UK or the ARRL in the USA. The US manuals have the advantage of containing information on circuits that operate at frequencies and power levels which cannot legally be used by amateurs in the UK.

Mains transformers

Mains transformers for power supplies (other than **switch-mode** supplies) conform to a fairly standard pattern. These transformers use laminated cores, and the older types use the familiar I and E shaped core pieces which can be fitted together with an air gap. The size of this air gap is a very important feature of the transformer, and is the reason for the difficulties that many users experience when they rebuild a transformer for another purpose, such as rewinding the secondary for a different voltage. The air gap acts for the magnetic circuit of the transformer as a high resistance would in a current circuit, and its magnetic effect is to restrict the magnetic flux in the core. This greatly reduces the likelihood of saturating the core with the large amounts of current that flow in the windings. An air gap is particularly important for mains frequency chokes in smoothing circuits which are likely to carry DC as well as AC ripple, but the use of chokes for this purpose is by now rare.

The traditional I and E, or C, core, however, is not ideally suited to all types of transformer requirements, particularly those which demand a low level of magnetic field around the transformer. A simple solution to the requirement for low external magnetic field is the toroidal transformer (Figure 3.8a), which has become much more generally available thanks to the development of efficient toroid-winding machines in the last 20 years.



(a) (b)

Figure 3.8

(a) and **(b)** The principle of the toroidal winding, which is much more efficient for concentrating flux. **(a)** simple toroid, **(b)** pot core.

The main point to note about toroidal transformers is that it can be only too easy to ruin their performance by incorrect mounting, because it is possible to make the mounting form a metal path which is in effect a shorted secondary turn that will dissipate a large part of the energy of the transformer. The other efficient solution is the use of a pot core which completely surrounds the inductor with a material of high permeability (see Figure 3.8b).

The specifications for mains transformers reflect the normal use of such transformers with rectifiers and capacitors to form power supplies. The most important rating is the **volt–amp rating** (VA) for each secondary winding, expressing the maximum current that can be drawn at the winding voltage. The term volt-amp is used rather than watt because the use of watts would imply a power factor of unity. Because the transformer is not 100% efficient, the volt-amps at the primary will be greater than the sum of the volt-amps at the secondary windings and, in part, though seldom stated directly, this is often implied in a figure for **magnetizing current**, meaning the current which flows in the primary when no load is connected to any secondary winding.

- Modern power supplies make use of active circuits with the aim of keeping the load current in phase with the load voltage and minimizing spikes and harmonics. These techniques are beyond the scope of this book.

- A very common practice now is to provide mains transformers with two primary windings rated at 110 V so that the transformer can be used with paralleled inputs on 110 V supplies or with series connections on 220 V.

The **regulation** of a transformer is an important factor in its use for power supply circuits. When the transformer is loaded by a rectifier and smoothing circuit, and full rated current is being drawn from the secondary (or from each secondary if there are several windings), the regulation is then the fractional drop in voltage, defined as:

$$\frac{\text{open circuit voltage} - \text{full-load voltage}}{\text{open circuit voltage}}$$

and expressed as a percentage. The regulation percentages can be very large for small transformers, typically 20% for a 3 VA type, falling to 5% or less for the larger transformers of 200 VA or more. Some manufacturers quote open-circuit and full-load voltage levels rather than regulation. One important point to note is that many manufacturers quote the full-load figure for secondary voltage output. This means that for a small transformer with poor regulation, the open-circuit voltage can be as much as 20% higher, and allowance must be made for this in the circuits which are connected to the transformer. Unless voltage stabilization is used, this order of voltage change between no-load and full-load may be unacceptable for applications that involve the use of ICs.

For any transformer, it is important to have some knowledge of the likely temperature rise during full-load operation. This figure is not always quoted, and an average for the larger transformers is 40°C above ambient for each winding (though most of the temperature rise originates in the secondary windings). Smaller transformers can have greater temperature rise figures, typically 60°C. The maximum acceptable temperature of a transformer is often not quoted and should not exceed 90°C unless the manufacturer specifies another figure. Transformers which use class E insulation can be run at a maximum working temperature of 120°C,

but this figure is exceptional among the usual range of transformers for power supplies. The full rating for a transformer implies a 25°C ambient, and the manufacturers should be consulted if higher ambient temperatures are likely.

Since transformers are subject to high peak voltages – the sum of AC and DC voltages – there is a figure of proof voltage (otherwise known as flash test voltage) for each transformer type which is at least 2 kV. This measures voltage breakdown between windings and also between each winding and the metal core. The higher grades of transformers will be tested to higher proof voltages, typically at 5 kV sustained for one minute, and transformers that are intended for special purposes such as heater supplies to cathode ray tubes whose cathodes are operated at very high voltage (negative voltages) will have to be tested to considerably higher voltages. The low-voltage requirements of modern instrument CRTs, however, imply that such transformers are seldom required now other than for servicing of old instruments.

The **winding resistance** of a transformer is not often quoted, though secondary winding resistance is an important factor when designing a power supply whose regulation (before the use of a stabilizing circuit) needs to be known. Note that transformers intended for 60 Hz supplies should not be used in 50 Hz applications because of the risk of overheating. Where winding resistance values are quoted, both primary and secondary will be quoted, and a typical primary resistance for a 240 VA transformer is 4 Ω, with higher values for the smaller transformers. Secondary resistances for low-voltage windings are much lower, of the order of 0.05 Ω for a winding rated at 10 A, higher for windings of lower current rating or for high-voltage windings.

For unusual secondary voltage requirements, it is possible to buy transformer kits, in which the primary winding is supplied on its bobbin, but the secondary has to be wound, and the bobbins then assembled on to the core. These transformer kits are usually of the conventional E and I core type, but several manufacturers supply toroidal cores with a primary winding already provided, and these are particularly useful for very low voltage supplies which require only a few turns of secondary winding. For each size of core, the manufacturer will quote the number of secondary turns per volt of output, typically from two turns per volt for the 200 VA size to six turns per volt for the 20 VA size.

The wire provided in these kits is the conventional enamelled copper, and the range of diameters is around 0.2 mm to 2.0 mm. When you select a wire gauge for a secondary winding you should bear in mind the power dissipation heating that you can expect at full rated current. For applications needing more than 10 A you will need to use wire of more than 2.0 mm diameter.

- Remember the rule of thumb that you need at least 1000 μF of reservoir capacitor per ampere of output current from a power supply. Remember also that the RMS current in the transformer windings is substantially more than the output DC current.

For details of transformer kits in the UK, see the ElectroComponents catalogue, or the International Web site at http://www.rs-components.com. UK users can go directly to http://rswww.com. You can register at the Web site to receive information about components, and updates of the product list and technical information.

Other transformer types

Mains isolation transformers use a 1:1 winding ratio and are intended to permit isolation from the mains supply. One important application is in the servicing of the older type of TV receivers, in which one mains lead was connected to the metal chassis. Though this ought to be the neutral lead, there can be no certainty of this, particularly when a twin-lead is used with no colour-coding. By using an isolating transformer, the whole chassis can, if required, be grounded, or it can be left floating so that there is no current path through the body of anyone touching any part of the circuit unless another part of the circuit is touched at the same time. Isolation transformers are also used for operation of power tools in hazardous situations (outdoors and in very humid surroundings), and some types can be bought already fitted with the standard form of splashproof socket for outdoor use, along with a ground-leakage contact breaker.

Autotransformers consist of a single tapped winding, so they offer no isolation, unlike the double-wound form of transformer. Fixed ratio autotransformers are intended to allow the use of electrical equipment on different mains voltages, for example the use of US 110 V equipment on European 220 V supplies. The demand for this type of transformer

in the UK tends to be localized around US air bases, but there is a large amount of test equipment in use which demands a 115 V supply and which has to be supplied by way of an autotransformer. It is important to ensure that any such equipment cannot under any circumstances be accidentally plugged into 220 V mains, and a fuse should be incorporated to prevent damage in the case of an accidental overload. Autotransformers can also be used to provide 220 V for European equipment being used in a country where the supply voltage is 110 V AC.

The more common type of autotransformer is the variable type, such as the well-established Variac (trade-mark of Claude Lyons Ltd.). This consists of a single toroidal winding with the mains supply connected to one end and to a suitable tap (the taps provide for different mains voltage levels), and an output terminal which is connected to a carbon brush whose position on the winding can be varied by rotating a calibrated knob. This allows for an output to be obtained whose voltage can be smoothly varied from zero to a voltage greater than the mains supply voltage, typically 270 V. Current ratings range from 0.5 A to 8 A depending on the size of toroidal core that is used. Variable autotransformers can be obtained either in skeleton form, with virtually no protection from the windings or connections, or in various degrees of enclosure. Since these are autotransformers there is no mains isolation, and if isolation is needed, it must be provided by a separate isolating transformer used to feed the autotransformer.

Single inductors are constructed in the same way as transformers but, of course, with a single winding. The use of cores for the lower frequencies is essential, but for RF use the coil is used either without a core or with a core that consists of a low-loss magnetic material such as ferrite, often referred to as a dust-core because it consists of magnetic particles that are not in electrical contact.

- **SURFACE-MOUNTED INDUCTORS**

Inductors were the last type of components to appear in SM format but are now easily available. Typically these are of solenoidal form, making them small and light. The inductance range is typically up to 1.2 mH, with current carrying capability to around 8 amps DC. Typical applications include switch-mode power supplies, digital camcorders, car navigation systems, and notebook PCs.

Other inductor topics

- For purposes of switching circuits, the energy stored in an inductor winding may need to be calculated, in units of **volt-seconds**.
- Planar inductors and transformers are formed in spiral or helical windings deposited on silicon and used for microwave applications.
- PDMA (plastic deformation magnetic assembly) inductors are new devices, allowing vertical standing inductors to be fabricated on a chip.
- FR cores (Figure 3.9), are formed from high-permeability materials and shaped like a flattened napkin ring. They are used for transient suppression over flat cables, particularly ribbon cable. The effect is to add to the impedance of a cable for a range of frequencies.

Figure 3.9

Typical FR cores.

- The **skin effect** at high frequencies causes the apparent resistance of a coil to increase above its DC value. This is because the electron path (causing the current flow) in the conductors avoids regions where the magnetic field is strongest. This effect also results in a decrease of inductance value for these signals. A rule of thumb, due to Terman, gives the diameter of an isolated wire that will have its apparent resistance increased by 10% for an operating frequency f as:

  $$D = 200/\sqrt{f} \text{ with } D \text{ in mm and } f \text{ in Hz}$$

  so the effect is to make the use of thick wires undesirable. For example, at 10 kHz, the formula shows that the wire diameter should be less than 2 mm, and at 1 GHz the 10% increase will apply to wires whose diameter is only 6/1000 mm!

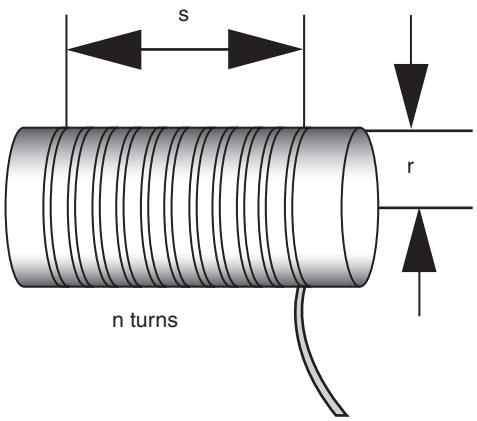
- The **proximity effect** (no simple rule of thumb here!) will also apply for a wire in a coil, because of the other turns of the coil.

- **Leakage inductanc**e in RF transformers can be greatly reduced by interleaving the primary and secondary windings to provide closer coupling, together with an efficient core (if the frequency range permits the use of a core).

- Large amplitudes of current can reduce the effective inductance of a cored inductor because the permeability (see later) of the core is not constant.

Inductance calculations

Of all electronics calculations those of inductance are the least precise. When an air-cored coil is used, the changing magnetic field does not affect all of the turns equally, so only the central turns are fully affected with the outer turns receiving a lesser amount. Using a magnetic core concentrates the field so as to even out the effect, but also makes the size of induced EMF less predictable unless the effect of the core in terms of its relative permeability can be precisely measured. In addition, the relative permeability of the core changes if DC flows in the windings. We have seen also that effects such as skin effect and proximity effect will also require other corrections to calculations. Any equations for the inductance of a coil are



$$L = \frac{n^2 r^2}{228(r + 1.11s)}$$

r = coin radius, in cm
L = inductance in μH
n = number of turns
s = length of coil, in cm

Figure 3.10

Calculating the (approximate) inductance of an air-cored, single-layer coil (a solenoid).

therefore very approximate and should be used only as a starting point in the construction of an inductor. Figure 3.10 shows a formula for the number of turns of a single-layer, close-wound coil, a solenoid to achieve a given inductance. The length of the coil is assumed to be more than the radius (otherwise a much more complex formula is needed).

This approximate formula gives reasonable results for single-layer air-cored solenoids of the values that are used for tuning radio circuits at frequencies up to VHF, though a skin-effect correction may be needed at the higher frequencies. The addition of a core of ferrite material will cause an increase in inductance which could be by a factor as high as the relative permeability (Table 3.1) of the ferrite.

Try http://www.vwlowen.demon.co.uk/java/coil.htm for finding number of turns of specified wire gauge to provide a stated inductance value, given diameter and length.

The total permeability of a material is given by $\mu_0 \times \mu_r$, where μ_0 is a universal constant called the **permeability of free space**, units henries per metre, and μ_r is **relative permeability**, a pure number with no units. These quantities are analogous to the permittivity of free space and relative permittivity as used for capacitor calculations.

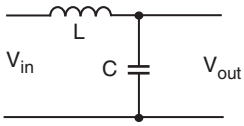
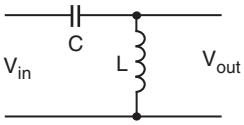
Table 3.1 Relative permeability values

Relative permeability, $\mu_r = \dfrac{\text{Inductance of coil with core}}{\text{Inductance of coil without core}}$

Alternatively, inductance value with core $= \mu_r \times$ inductance value without core.

Material	Relative permeability, maximum value
Silicon–iron	7000
Cobalt–iron	10 000
Permalloy 45	23 000
Permalloy 65	600 000
Mumetal	100 000
Supermalloy	1 000 000
Dustcores	10 to 100
Ferrites	100 to 2000

Table 3.2 Reactive circuit response

Circuit	V_{out}/V_{in}	Phase angle
	$\dfrac{1}{1-(f^2/f_0^2)} \approx -\dfrac{f_0^2}{f^2}$	0° when $f < f_0$ 180° when $f > f_0$
	$\dfrac{1}{1-(f_0^2/f^2)} \approx -\dfrac{f^2}{f_0^2}$	0° when $f > f_0$ 180° when $f < f_0$

Notes: f_0; is the frequency of response $= 0.16/\sqrt{LC}$
f is the frequency at which response is to be found.
$>$ means 'greater than', $<$ means 'less than' and $\approx$ means 'approximately equal to'.

The multiplying effect on inductance of using a core is seldom as large as the figure of relative permeability because for most cores the magnetic material does not completely enclose the coil. Manufacturers of ferrite cores (pot cores) that completely enclose a coil former provide winding data appropriate for each type and size of core. The following example shows how inductors can be adjusted for a different inductance value using

the principle that inductance is proportional to the square of the number of turns.

Example: *The inductance of a* 120*-turn coil is measured as* 840 μH. *How many turns need to be removed to give* 500 μH*?*

Solution: Since $L \propto n^2$ (L = inductance and n = number of turns)

$$\frac{L1}{L2} = \frac{n_1^2}{n_2^2} \text{ giving } \frac{840}{500} = \frac{120^2}{n_2^2} \text{ so that } n_2^2 = \frac{120^2 \times 500}{840} = 8571$$

$n_2 = \sqrt{8571} = 93$ approximately; requiring 27 turns to be removed.

Untuned transformers

An untuned transformer (for signals) consists of two windings, primary and secondary, neither of which is tuned by a capacitor, on a common core. For low frequency use, a massive core made from laminations (thin strips) of transformer steel alloy such as silicon-iron must be used. Transformers that are used only for higher audio frequencies can make use of considerably smaller cores. At radio frequencies the losses caused by transformer steels make such materials unacceptable and ferrite materials are used as cores. For the highest frequencies no form of core material is suitable and only self-supporting, air-cored coils, usually of thick silver-plated wire, can be used. In the higher UHF bands, inductors can consist of straight wire or metal strips. High frequency signals flow mainly along the outer surfaces of conductors, so tubular conductors are as efficient as solid conductors but use less metal and allow the use of water-cooling. In addition, a plated coating can considerably improve the efficiency of the conductor, hence the use of silver plating on UHF conductors. For an untuned transformer with 100% coupling between primary and secondary, the ratio of AC voltages $\tilde{V}_s/\tilde{V}_p$ is equal to the ratio of winding turns N_s/N_p with **s** meaning secondary and **p** primary. When an untuned transformer is used to transfer power between circuits of different impedance, Z_p and Z_s, the best match for optimum power transfer is obtained when:

$$\frac{N_s}{N_p} = \sqrt{\frac{Z_s}{Z_p}}$$

Inductive reactance

The reactance of an inductor for a sine wave signal of frequency f hertz is $2\pi fL$, where L is the self-inductance in henries. The reactance is defined, as before, as the ratio of AC voltage across the inductor to AC current through it, and is measured in ohms. For a coil whose reactance is much greater than its resistance the voltage sine wave is 90° (¼ cycle) ahead of the current sine wave (Figure 3.11).

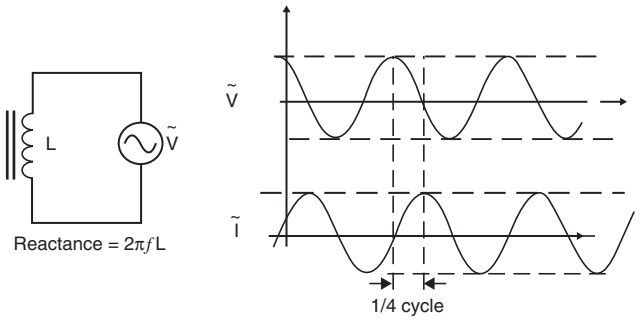


Figure 3.11

Reactance and phase shift of a perfect (zero-resistance) inductor.

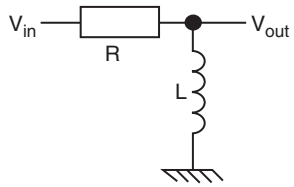
LCR circuits

The action of both CR and LR circuits upon a sine wave is to change both the amplitude and the phase of the output signal as compared to the input signal. Universal amplitude/phase tables can be prepared, using the time constant of the CR or LR circuit and the frequency f of the sine wave. The tables for an inductor–resistor circuit (LR) are shown, with examples, in Table 3.3.

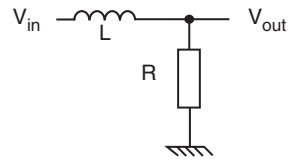
When a reactance (L or C) is connected in circuit with a resistance R, the general formulae for the total impedance Z are as shown in Table 3.4. Impedance is defined, like reactance, as AC volts across the circuit divided by AC current through the circuit, but the phase angle between voltage

Table 3.3 Amplitude and phase tables for LR Circuits

$G = V_{out}/V_{in}$; φ = **phase angle; the time constant** $T = L/R$, $\omega = 2\pi \times$ **frequency** (f)

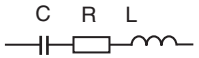
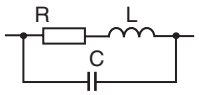
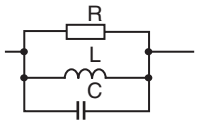


ωT	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.5	2.0	3.0	4.0	5.0
$\varphi°$	84.3	78.7	73.3	68.2	63.4	59.0	55.0	51.34	48.0	45.0	33.7	26.6	18.4	14	11.3
G	0.099	0.196	0.287	0.37	0.45	0.51	0.57	0.62	0.67	0.707	0.83	0.9	0.95	0.97	0.98



ωT	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.5	2.0	3.0	4.0	5.0
$\varphi°$	−5.7	−11.3	−16.7	−21.8	−26.5	−31	−35	−38.6	−41.8	−45	−56	−63	−72	−76	−79
G	0.99	0.98	0.96	0.9	0.89	0.85	0.82	0.78	0.74	0.707	0.55	0.45	0.32	0.24	0.2

Table 3.4 Impedance Z and phase angle ϕ

Circuit	Z	ϕ	
	$Z = \sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}$	$\phi = \tan^{-1}\left(\frac{\omega L - 1/(\omega C)}{R}\right)$	(a)
	$Z = \sqrt{\frac{R^2 + \omega^2 L^2}{(1 - \omega^2 LC) + \omega^2 C^2 R^2}}$	$\phi = \tan^{-1}\left(\frac{\omega[L(1 - \omega^2 LC) - CR^2]}{R}\right)$	(b)
	$Z = \frac{1}{\sqrt{\left(\frac{1}{R}\right) + \left(\omega C - \frac{1}{\omega L}\right)^2}}$	$\phi = \tan^{-1}\left(R\left[\frac{1}{\omega L} - \omega C\right]\right)$	(c)

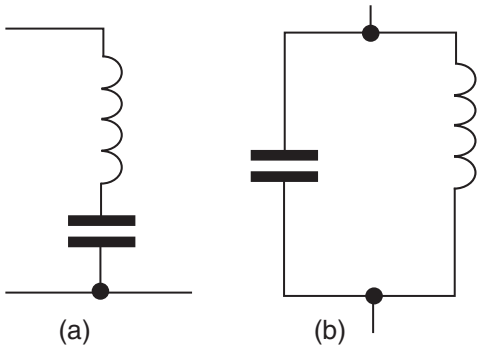
Notes: $\omega = 2\pi \times$ frequency (f)
$\tan^{-1}$ = angle whose tangent is equal to.

and current will not usually be 90°, and will be 0° only at resonance (see later).

The combination of inductance and capacitance produces a **tuned circuit** which may be series (Figure 3.12a) or parallel (Figure 3.12b) connected. Each type of tuned (or *resonant*) circuit has a frequency of resonance, f_0, at which the circuit behaves like a pure resistance (with no inductance or capacitance) so that there is no phase shift; the current wave is in phase with the voltage wave.

Figure 3.12

Tuned circuits: **(a)** series, **(b)** parallel.



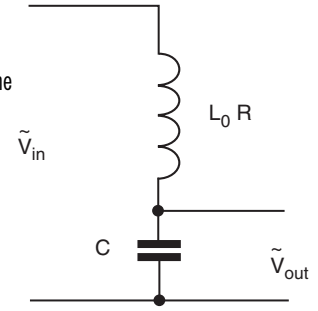
At all other frequencies the circuit will behave either as a resistor–inductor or a resistor–capacitor circuit, with the appropriate direction of phase shift. Below the frequency of resonance, the parallel circuit behaves like an inductor–resistor circuit and the series circuit behaves like a capacitor–resistor circuit. Above the frequency of resonance, the parallel circuit behaves like a capacitor–resistor circuit and the series circuit behaves like an inductor–resistor circuit. At resonance, the parallel circuit behaves like a large value pure resistor and the series circuit as a low-value pure resistor. In other words, at resonance there is no phase shift between current and voltage in the circuit.

The series resonant circuit can provide voltage amplification of the resonant frequency when the circuit shown in Figure 3.13 is used.

The level of voltage amplification at the frequency of resonance is given by $(2\pi fL)/R$ or $1/(2\pi fCR)$, and this quantity is termed **the circuit magnification factor**, symbol Q. There is no *power* amplification because

Figure 3.13

Voltage amplification of a tuned series circuit. The amplification is of the resonant frequency only, and can occur only if the signal source is of comparatively low impedance.



the voltage step-up is achieved by increasing the current through the circuit, assuming a constant input voltage. Table 3.3 shows typical phase and amplitude response formulae in universal form for the series resonant circuit. Note that the tuning capacitance may be stray capacitance or include stray capacitance as a significant portion. A back-biased diode (a varactor) can also be used so that the capacitance can be voltage-variable. The parallel-tuned circuit is used as a load that is a pure resistance with no phase shift at the frequency of resonance, f_0. The size of this equivalent resistance is called dynamic resistance (R_d) and is calculated from the formula:

$$R_d = \frac{L}{CR}$$

L = inductance in henries
C = capacitance in farads

The effect of adding a resistor in parallel with such a tuned circuit is shown in Figure 3.14; this reduces the dynamic resistance value, but the shape of the dynamic resistance vs. frequency graph changes, so that the (relatively) higher resistance is maintained over a larger frequency range. This effect, called *damping*, is used to extend the bandwidth of tuned amplifiers. Table 3.3 shows the amplitude and phase response of a parallel-tuned circuit in general form. Once again, the capacitance can be mainly or entirely due to stray capacitance.

The impedance of a series circuit is given by the formula shown, with an example, in Table 3.4a. Note that both amplitude (in ohms) and phase angle (in radians) are given. The corresponding expression for a parallel circuit in which the only resistance is that of the coil is also shown in

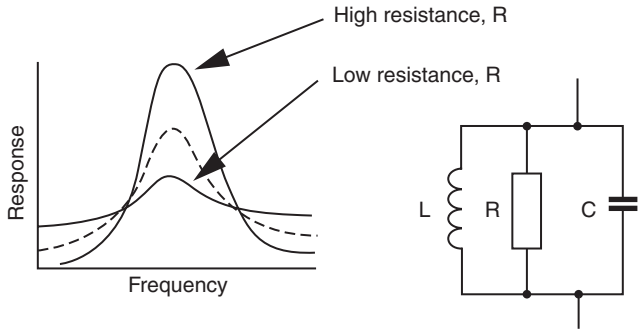


Figure 3.14

The effect of damping resistance on the resonance curve.

Table 3.4b. When a damped parallel circuit is used, the resistance of the coil has generally a negligible effect compared to that of the damping resistor, and the formula of Table 3.4c applies.

Coupled tuned circuits

When two tuned circuits are placed so that their coils have some mutual inductance, M, the circuits are said to be **coupled**. The size of the mutual inductance is not easy to calculate; one approximate method using a nomogram is shown in Figure 3.15.

When the mutual inductance (M) between the coils is small compared to their values of self-inductance (L_1, L_2) then the coupling is said to be **loose** and the response curve shows a sharp peak. When the mutual inductance between the coils is large compared to the self-inductance values the coupling is **tight** (or **overcoupled**) and the response curve shows twin peaks. For each set of coupled coils there is an optimum amount of coupling at which the peak of the response curve is flattened and the sides steep. This type of response is an excellent compromise between **selectivity** (choice of a desired frequency) and **sensitivity** (maximum Q for the resonant frequency). Figure 3.16 shows typical graphs of response for loose, tight and optimum coupling.

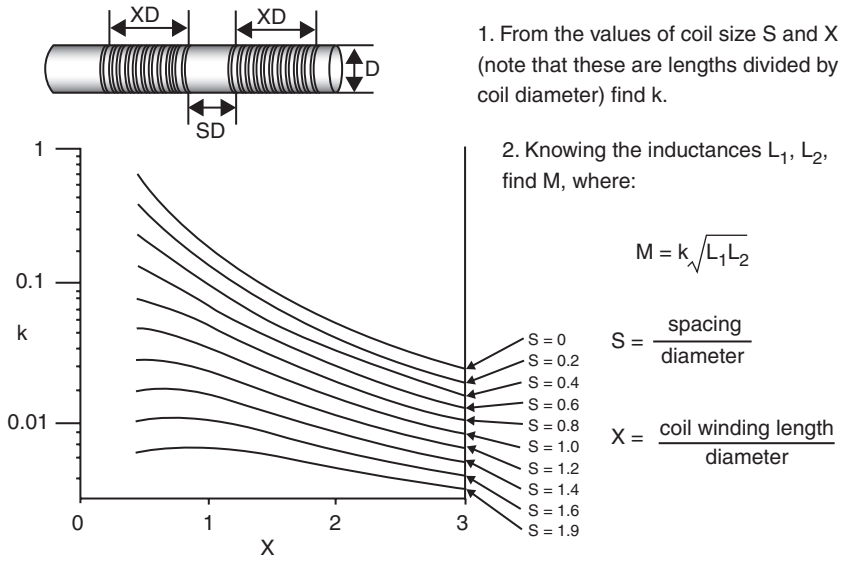


Figure 3.15

Calculating mutual inductance.

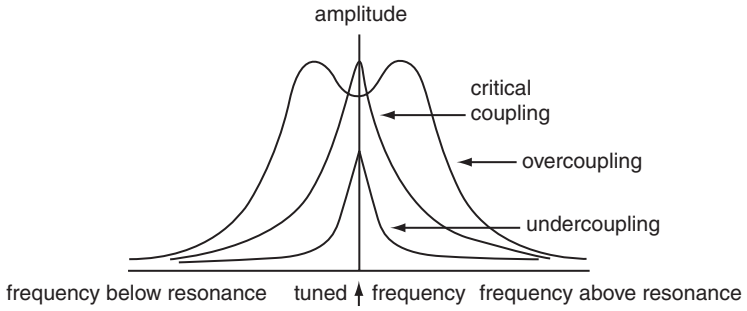


Figure 3.16

The effect of a damping resistance on the coupling of inductors.

The **coefficient of coupling**, k, is defined by the equation:

$$k = \frac{M}{\sqrt{L_1 L_2}}$$

which reduces to M/L if both coils have the same value of L. **Critical coupling** occurs when $k = 1/Q$ assuming that both coils have the same value of Q factor – if they do not, then the figure $Q = \sqrt{Q_1 Q_2}$ can be used. The size of the coefficient of coupling depends almost entirely on the spacing between the coils and no formulae are available to calculate this quantity directly. Other types of coupled circuit, along with some design data, are shown in Figure 3.17. These make use of a common impedance or reactance for coupling and are not so commonly used with passive components.

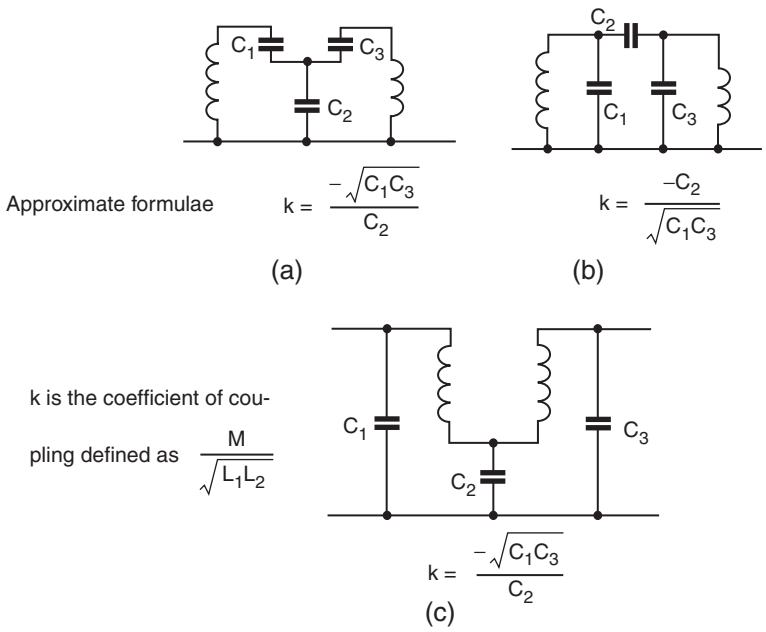


Figure 3.17

(a)–(c) Other methods of circuit coupling, and their design formulae.

Quartz crystals

Quartz crystals, cut into thin plates and with electrodes plated onto opposite flat faces, can be used as resonant circuits with Q values ranging from 20 000 to 1 000 000 or more. They are all piezoelectric and can therefore be used as transducers (sender or receiver) for ultrasonic waves. The equivalent circuit of a crystal is shown in Figure 3.18.

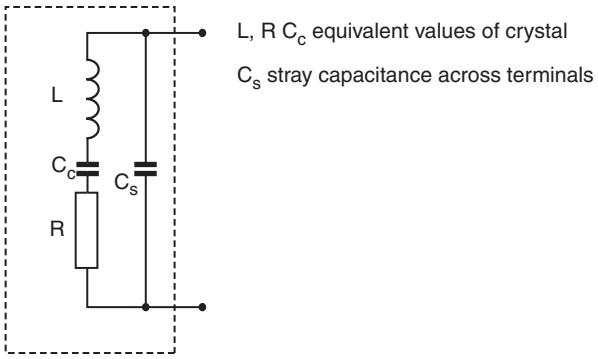


Figure 3.18

Equivalent circuit of a quartz crystal.

The L and C values in this equivalent circuit are referred to as **motional inductance** and **motional capacitance**, and values will be specified by the manufacturer. These values, with a very high ratio of L to C, could not be provided by any assembly of separate components, and it is that which provides the very high Q-factor for a crystal.

The crystal by itself acts as a series resonant circuit with a very large inductance, small capacitance and fairly low resistance (a few thousand ohms). The stray capacitance across the crystal will also permit parallel resonance to occur at a frequency that is slightly higher than that of the series resonance. Figure 3.19 shows how the reactance and the resistance of a crystal vary as the frequency is changed – the reactance is zero at each resonant frequency and the resistance is maximum at the parallel resonant frequency. Usually the parallel or the series resonant frequency will be specified when the crystal is manufactured.

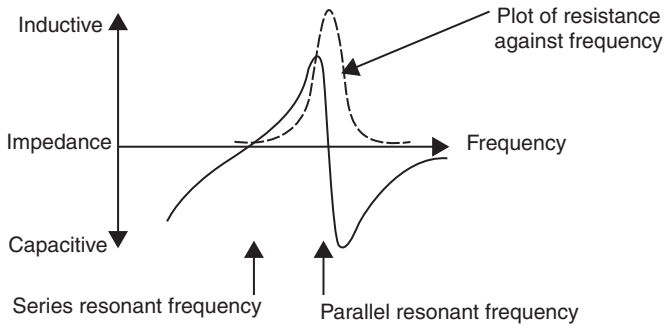


Figure 3.19

Variation of reactance and resistance of a crystal near its resonant frequencies.

A crystal can be used at its fundamental (lowest) frequency or at odd harmonics (called **overtones**), usually 3rd or 5th, sometimes 7th. For frequencies below about 30 MHz, it is usual to excite the fundamental frequency of the crystal. The overtones are not precisely integer multiples (like 3, 5, 7) of the fundamental, and if you buy a crystal to use at a frequency higher than 30 MHz you will be expected to use it in **overtone mode** (tuned to the overtone), and it will have been calibrated at this overtone, not at the fundamental.

Some suppliers can now provide inverted mesa crystals that will resonate at fundamental frequencies higher than 30 MHz.

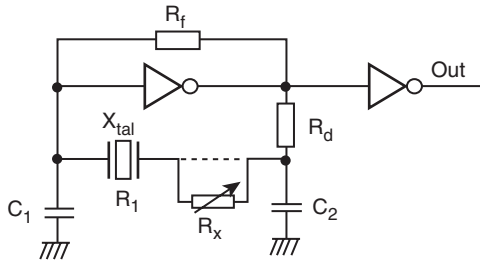
Because a crystal can be used either in parallel or series resonance, you need to calibrate an oscillator circuit for the mode you intend to use; you cannot calibrate to both series and parallel modes.

When a parallel oscillation mode is to be used, the load capacitance becomes important, because in parallel resonance the crystal has inductive reactance, and this inductance in parallel with the load capacitance forms the oscillating circuit. For each crystal intended to be used in parallel mode, an optimum range of load capacitance will be quoted. When the crystal is used in series mode, the load capacitance is relatively unimportant.

Effective series resistance (**ESR**) for a crystal is normally in the range 25 Ω to 100 Ω, and is specified by the supplier. The oscillator design

Figure 3.20

Crystal oscillator circuit for adjusting drive level.



should ensure that minimal added resistance exists, otherwise there may be difficulty in starting the crystal oscillating. The crystal oscillator circuit, shown in Figure 3.20, has effectively negative resistance and can be used for adjusting the drive level in the course of designing an oscillator circuit. For easy starting the value of negative resistance should be around 5–10 times the resistance at resonance.

The ESR is generally higher with the SMD type of crystals, and can cause a problem if the oscillator circuit does not have sufficient loop gain. The outstanding examples are the crystals for quartz watches, cut to oscillate at 32.768 kHz. These can have very high ESR values, typically 10–30 kΩ, and the oscillator circuit must be designed to overcome this difficulty – a crystal circuit intended for radio use will not work with a watch crystal.

The connections of a crystal to a printed circuit board introduce a stray shunt capacitance in parallel with the crystal's LC model, C_s, as shown in Figure 3.18. This holder capacitance is typically in the region 2 pF to 6 pF. Some oscillator circuits will not tolerate excessive holder capacitance, particularly at higher frequencies.

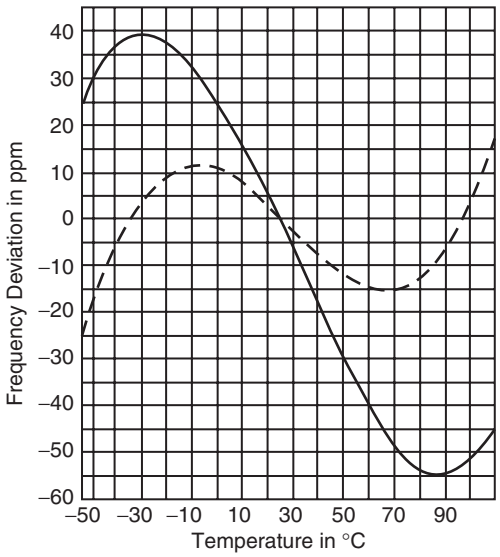
The crystal oscillator circuit must be designed so as to prevent excessive power dissipation in the crystal, otherwise the crystal can be destroyed by excessive vibration. In addition, driving a crystal too hard will cause changes in characteristics because of non-linearity. The power dissipated is given by RI^2, where R is the crystal equivalent resistance and I the RMS driving current. For a parallel resonant oscillator, the crystal current is found by dividing the RMS voltage across the load capacitor by the reactance of the load capacitor at resonant frequency. For a series circuit, the crystal current is found by dividing the RMS voltage across the crystal by the internal series resistance of the crystal.

- **TEMPERATURE EFFECTS**

Crystal oscillators are not immune from temperature effects; Figure 3.21 shows graphs of frequency plotted against temperature that are typical of AT-cut crystals. In this diagram, the dashed line is typical of the general run of AT crystals, and the solid line refers to VHF/UHF crystal running in parallel resonance mode. Figure 3.22 shows the frequency variation plotted against temperature for typical AT and X cut (for 32.768 kHz watch crystals) quartz crystals that are designed for use over a wide temperature range.

Figure 3.21

Frequency vs. temperature for AT cut crystals in normal circuit (dotted line) and in parallel resonance mode (solid line).



Wave filters

Wave filter circuits are networks that contain reactive components (typically L and C) that accept or reject frequencies above or below stated cut-off frequency limits which are calculated from the values of the filter components. Output amplitude and phase vary considerably as the signal frequency approaches a cut-off frequency and the calculations that are involved are beyond the scope of this book. The use of computer simulation (see Chapter 17), is advisable when designing such circuits. Much more

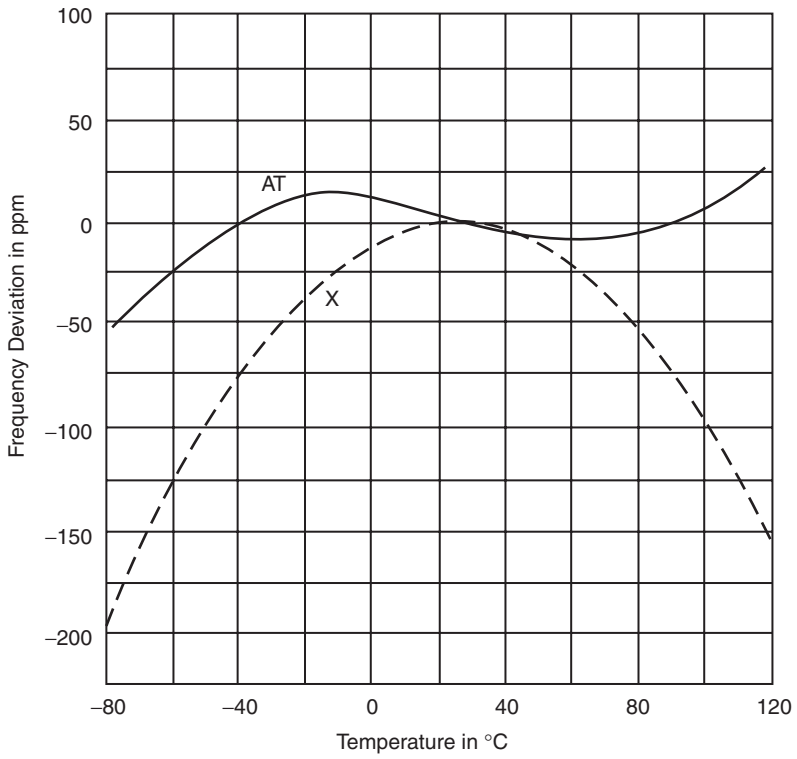


Figure 3.22

Frequency variation vs. temperature for typical AT and X cut crystals.

easily predictable responses can be obtained, for audio frequencies at least, by using active filters (see Chapter 6).

Quartz crystals are used extensively in filter circuits to provide very sharp cut-off points, and in these applications the frequency–temperature characteristic of the devices is the most important parameter. This leads to the use of AT-cut crystals as the preferred type, providing very good frequency stability over a wide temperature range.

Ceramic resonators, using materials such as lead zirconate titanate (PZT) are extensively used in filter circuits, and in microprocessor timing applications. These materials are piezoelectric, and can resonate in several modes

depending on their resonance frequency. Their precision of oscillation is lower than that of quartz crystals, but very much better than a discrete LC circuit, with a temperature coefficient of around $10^{-5}/°C$ in a temperature range of, typically, $-10°C$ to $+80°C$. They are considerably lighter and smaller than quartz crystals, and relatively immune to alterations on loading or in power supply voltage. Ceramic resonators in SM format often have load capacitors built in; other configurations may require load capacitors to be added.

This page intentionally left blank

CHAPTER 4

CHEMICAL CELLS AND BATTERIES

Introduction

Chemical cells were the original source of DC, and have always been an important form of power supply for electronic equipment. Historically, cells and batteries have been in use for over two hundred years, and the problems that are encountered with one of the simplest and oldest types of cell are a good introduction to the reasons why so many diverse battery types exist nowadays, and to the technology that is used. Strictly speaking, a **battery** is an assembly of single cells, so the action of a cell is the subject of this Chapter.

Any type of chemical cell depends on chemical action which is usually between a solid (the **cathode** plate) and a liquid, the **electrolyte**. The use of liquids makes cells less portable, and the trend for many years has been to using jellified liquids or moist solids, and also to materials that are not strong acids or alkalis. The voltage that is obtained from any cell depends on the amount of energy liberated in the chemical reaction, but only a limited number of chemical reactions can be used in this way, and for most of them, the energy that is liberated corresponds to a voltage of between 0.8 V and 2.3 V per cell with one notable exception, the 3+ V **lithium cell**. This range of voltage represents a fundamental chemical action that cannot be altered by refining the mechanical or electrical design of the cell.

The current that can be obtained from a cell is, by contrast, determined by the area of the conducting plates and the resistance of the electrolyte material, so there is a relationship between physical size and current capability. The limit to this is purely practical, because if the cell is being used for a portable piece of equipment, a very large cell makes the equipment less portable and therefore less useful. Hundreds of types of

cells have been invented and constructed since 1790, and most of them have been forgotten, not even being mentioned in school textbooks. By the middle of the 20th century, only one type of cell was commonly available, the Leclanché cell, which is the familiar type of 'ordinary' torch cell. The introduction of semiconductor electronics, however, has revolutionized the cell and battery industry, and the requirements for specialized cells to use in situations calling for high current, long shelf-life or miniature construction have resulted in the development and construction of cells from materials that would have been considered decidedly exotic in the earlier part of the 20th century.

Primary and secondary cells

A **primary cell** is one in which the chemical reaction is not readily reversible. Once the cell is exhausted, because the electrolyte has dissolved all of the cathode material or because some other chemical (such as the **depolarizer**, see later) is exhausted, recharging to the original state of the cell is impossible, though for some types of primary cell, a limited extension of life can be achieved by careful recharging under microprocessor control. In general, attempts to recharge a primary cell without using a control circuit will usually result in the internal liberation of gases which will eventually burst explosively through the case of the cell.

A **secondary cell** is one in which the chemical reaction is one that is designed to be reversible. Without getting into too much detail about what exactly constitutes reversibility, reversible chemical reactions are not particularly common, and it is much more rarely that such a reaction can be used to construct a cell, so there is not the large range of cells of the secondary type such as exists for primary cells. The nickel–cadmium secondary cell which is used so extensively nowadays in the form of rechargeable batteries is a development of an old design, the nickel–iron cell due to Edison in the latter years of the 19th century, and is now almost completely superseded by the nickel–hydride type and the lithium-ion rechargeable cell.

There is a third type of cell, the fuel cell, which despite very great research efforts for some 30 years has not become as common as was

originally predicted. A fuel cell uses for its power a chemical reaction which is normally combustion, the burning of a substance, and is an efficient method of generating an EMF from a fuel.

Battery connections

When a set of cells is connected together, the result is a **battery**. The cells that form a battery could be connected in series, in parallel, or in any of the series-parallel arrangements, but in practice the connection is nearly always in series. The effect of both series and parallel connection can be seen in Figure 4.1. When the cells are connected in series, the open-circuit

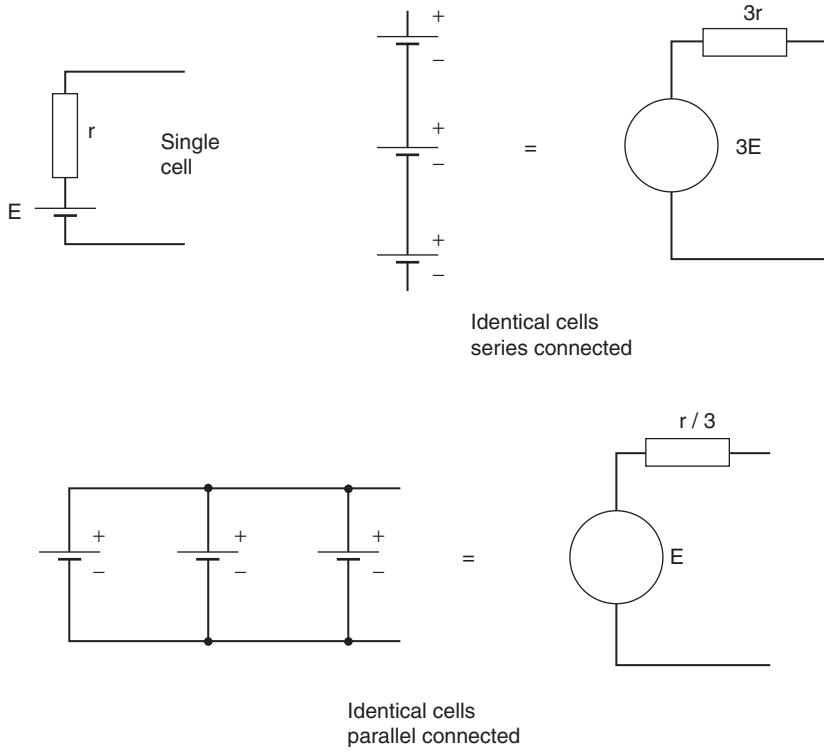


Figure 4.1

Connecting cells in series and in parallel.

voltages (EMFs) add, and so do the internal resistance values, so the overall voltage is greater, but the current capability is the same as that of a single cell.

When the cells are connected in parallel, the voltage is as for one cell, but the internal resistance is much lower, because it is the resultant of several internal resistances in parallel. This allows much larger currents to be drawn, but unless the cells each produce exactly the same EMF value, there is a risk that current will flow between cells, causing local overheating. For this reason, primary cells are never used connected in parallel, and even secondary cells, which are more able to deliver and to take local charging current, are seldom connected in this way.

Higher currents are obtained by making primary cells in a variety of sizes, with the larger cells being able to provide more current, and having a longer life because of the greater quantity of essential chemicals. The limit to size is portability, because if a primary cell is not portable it has a limited range of applications. Secondary cells have much lower internal resistance values, so if high current capability is required along with small volume, a secondary cell is always used in preference to a primary cell. One disadvantage of the usual type of nickel–cadmium secondary cell in this respect, however, is a short 'shelf-life', so if equipment is likely to stand for a long time between periods of use, secondary cells may not be entirely suitable, because they will always need to be recharged just before use.

The important parameters for any type of cell are its open-circuit voltage (the EMF), its 'typical' internal resistance value, its shelf-life, active life and energy content. The **internal resistance** is the resistance of the electrolyte and other conductors in the cell, and its value limits the amount of current that a cell can provide because it causes the output voltage of the cell to drop when current flows. The **shelf-life** indicates how long a cell can be stored, usually at a temperature not exceeding 25°C, before the amount of internal chemical action seriously decreases the useful life. The **active life** is less easy to define, because it depends on the current drain, and it is usual to quote several figures of active life for various average current drain values. The **energy content** is defined as EMF × current × active life, and will usually be calculated from the most favourable product of current and time. The energy content is more affected by the type of chemical reaction and the weight of the active materials than by details of design.

Simple cell

All of the cells that are used today can trace their origins to the voltaic pile that was invented by Alessandro Volta (after whom the volt unit was named) around 1782. Each portion of this device was a sandwich of cloth soaked in brine, and laid between one plate of copper and one plate of zinc. When sufficient of the sandwich cells were assembled into a battery, the voltage was enough to cause effects such as the heating of a thin wire, or the twitching of the leg of a (dead) frog – the effect discovered by Luigi Galvani.

The next step was to the simple cell, as we now call it, which used as metal zinc (the **cathode**) and as the liquid, sulphuric acid, to provide the chemical reaction, and the other contact, the **anode**, that was needed, was provided by a copper plate which also dipped into the acid (Figure 4.2). The action

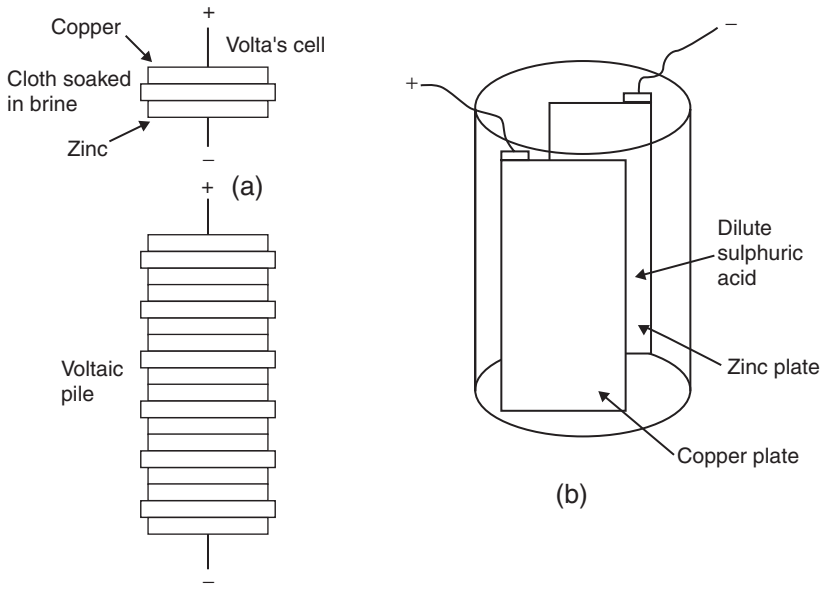


Figure 4.2

(a) and **(b)** The original form of wet simple cell.

is that, when the zinc dissolves in the acid, electrons are liberated. These electrons can flow along a wire connected to the zinc, and back into the chemical system through the copper plate, so meeting the requirement for a closed path for electrons.

In terms of conventional current flow, a decision made long before the existence of electrons was suspected, a current flows from the positive copper plate, the **anode**, to the negative zinc plate, the **cathode**. All cells conform to this pattern of a metal dissolving in an acid or alkaline solution and releasing electrons that return to the cell by way of an inert conductor (not affected by the electrolyte) which is also immersed in the solution. The original zinc–sulphuric acid type of cell is known as the simple cell to distinguish it from the many types that have followed.

The simple cell has several drawbacks that make it unsuitable for use other than as a demonstration of principles. The use of sulphuric acid in liquid form makes the cell unsuitable for any kind of portable use, since acid can spill and even at the dilution used for the simple cell it can cause considerable damage. The cell cannot be sealed, because as the zinc dissolves it liberates hydrogen gas which must be vented.

There are more serious problems. The sulphuric acid will dissolve the zinc, though at a slower rate, even when no circuit exists, so the cell has a very short shelf-life and not much active life. In addition, the voltage of the cell, which starts at about 1.5 V, rapidly decreases to zero when even only a small current is taken because the internal resistance rises to a large value as the cell is used. This makes the cell unusable until the zinc is removed, washed, and then re-inserted.

The efforts that were made to understand the faults of the simple cell have led to the development of considerably better cells, because by understanding principles we are better able to design new products. The problem of the zinc dissolving even with no circuit connected was solved by using very pure zinc or by coating the zinc with mercury. The problem is one of *local action*, meaning that the impurities in the zinc act like anodes, forming small cells that are already short-circuited. By using very pure zinc, this local action is very greatly reduced, but in the 18th century purification of metals had not reached the state that we can expect nowadays. Mercury acts to block off the impurities without itself acting as an anode, and this was a much easier method to use at the time.

The use of mercury in cells is now strongly discouraged on environmental grounds.

The rapid increase in internal resistance proved to be a more difficult problem, and one that could not be solved other than by redesigning the cell. The problem is that dissolving zinc in sulphuric acid releases hydrogen gas, and this gas coats the surface of the anode as it is formed, an action that was originally called **polarization**. The gas appears at the anode because of the action of the electrons entering the solution from the external circuit. Because hydrogen is an insulator, the area of the anode that can be in electrical contact with the sulphuric acid is greatly reduced by this action, so the internal resistance increases. When local action is present, the internal resistance will increase from the moment that the cell is assembled, though for the pure-zinc cell or the type in which the zinc has been coated (amalgamated) with mercury, the internal resistance increases only while the cell is used. The insulation provided by hydrogen gas is used as the dielectric for electrolytic capacitors whose construction closely mirrors that of a cell.

The problem can be solved only by removing the hydrogen as it forms or by using a chemical reaction that does not generate any gas, and these are the solutions that have been adopted by every successful cell type developed since the days of Volta. The removal of hydrogen is achieved by using an oxidizing material, the **depolarizer**, which has to be packed around the anode. The depolarizer must be some material that will not have any chemical side-effects, and insoluble materials like manganese (II) oxide have been used very successfully in the past and are still widely used.

The Leclanché cell

The cell that was developed by the French chemist Leclanché in the 19th century has had a remarkably long history, and in its 'dry' form is still in use, though now grandified by the title of *carbon–zinc cell*. In its original form (Figure 4.3) the electrolyte was a liquid, a solution of ammonium chloride. This is mildly acid, but not fiercely corrosive in the way that sulphuric acid is, and one consequence of using this less acidic electrolyte is that the zinc, even if not particularly pure, does not dissolve in the solution to the same extent when no current is passing in the external circuit. Local action is still present, but greatly reduced as compared to a

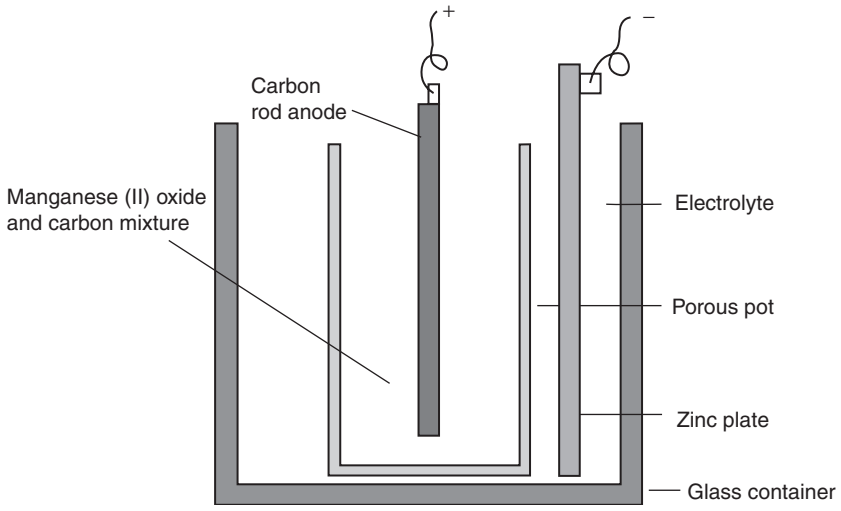


Figure 4.3

The original form of the Leclanché wet cell.

zinc–acid type of cell. The anode for the cell is a rod of carbon, a material that is chemically inert and therefore not attacked by the electrolyte. The carbon rod is surrounded by a paste of manganese dioxide, all contained inside a porous pot so that the electrolyte keeps the whole lot wet and conducting. The action when current flows is that zinc dissolves in the mildly acid solution, releasing electrons which then travel through the circuit.

At the anode, the electrons would normally react with the water in the liquid to produce hydrogen, but the action of the manganese dioxide is to absorb electrons in preference to allowing the reaction with the water to proceed, producing a different oxide of manganese (a reduced state). As the cell operates, the zinc is consumed, as also is the manganese dioxide, and when either is exhausted the cell fails. The open-circuit voltage is about 1.5 V, and the internal resistance can be less than one ohm. The older form of the Leclanché cell was in service for operating doorbells and room indicators from mid-Victorian times, and some that had been installed in those days were still working in the late 1940s.

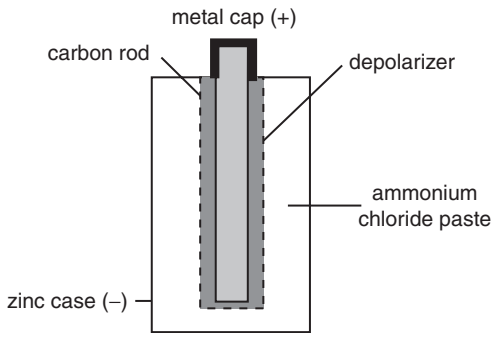
The reason for this is that the Leclanché cell was quite remarkably renewable. The users could buy spare zinc plates, spare ammonium chloride (which could also be used for smelling salts) and spare manganese dioxide,

so that the cell could be given an almost indefinite life on the type of intermittent use that it had, provided that the liquid level was topped up at intervals. Some worked for well over 20 years without any attention at all, tucked away in a cool cupboard on a high shelf.

The 'dry' form of the Leclanché cell is the type that until quite recently was the only familiar form of primary cell. The construction (Figure 4.4) follows the principles of the older wet type of cell, but the ammonium chloride electrolyte is in jelly form rather than liquid, and the manganese oxide is mixed with graphite and with some of the jelly to keep it also moist and conducting. The action is the same, but because the dry cell is usually smaller than the wet variety and because its jelly electrolyte is less conductive, this form of the cell has generally a higher internal resistance than the old wet variety. The advantage of portability, however, totally overrules any disadvantages of higher internal resistance, making this the standard dry cell for most of the twentieth century.

Figure 4.4

The modern form of dry carbon–zinc cell.



The carbon–zinc dry cell, as it is more often called now, fails totally either when the zinc is perforated or when the manganese dioxide is exhausted. One of the weaknesses of the original design is that the zinc forms the casing for the cell, so when the zinc becomes perforated, the electrolyte can leak out, and countless users of dry cells will have had the experience of opening a torch or a transistor radio battery compartment to find the usual sticky mess left by leaking cells. The term 'dry' cell never seems quite appropriate in these circumstances. The problem cannot be dealt with simply by using a thicker zinc casing and by restricting the amount of manganese dioxide as the cell will eventually fail because of high internal resistance before the zinc is used up.

The carbon–zinc cell does not have a particularly long shelf-life and once it has been used, the electrolyte starts to dissolve the zinc at a slow but inexorable rate. This corresponds to an internal current within the cell, called the self-discharge current. Perforation will therefore invariably occur when an exhausted cell is left inside equipment, and the higher the temperature at which the cell is kept, the faster is the rate of attack on the zinc.

This led to the development of leakproof cells with a steel liner surrounding the zinc. Leakproofing in this way allowed a much thinner zinc shell to be used, thus cutting the cost of the cell (though it could be sold at a higher price because of the leakproofing) and allowing the cell to be used until a much greater amount of the zinc had been dissolved. Leakproofing is not foolproof, and even the steel shell can be perforated in the course of time, or the seals can fail and allow electrolyte to spill out. Nevertheless, the use of the steel liner has considerably improved the life of battery-operated equipment.

The alkaline primary cells

A different group of cell types makes use of alkaline rather than acid electrolytes, so although the principle of a metal dissolving in a solution and releasing electrons still holds good, the detailed chemistry of the reaction is quite different. On the assumption that the reader of this book will be considerably more interested in the electrical characteristics of these cells rather than the chemistry, we will ignore the chemical reactions unless there is something about them that requires special notice. One point that does merit attention is that the alkaline reactions do not generate gas, and this allows the cells to be much more thoroughly sealed than the zinc–carbon type. It also eliminates the type of problems that require the need of a depolarizer, so the structure of alkaline cells can, in theory at least, be simpler than that of the older type of cell. Any attempt to recharge these cells other than by well-designed (microprocessor-controlled) circuitry will generate gas and the pressure will build up until the container fractures explosively.

The best-known alkaline type of cell is the Manganese Alkaline, whose construction is illustrated in Figure 4.5. This was invented by Sam Ruben in the USA in 1939 and was used experimentally in some wartime equipment,

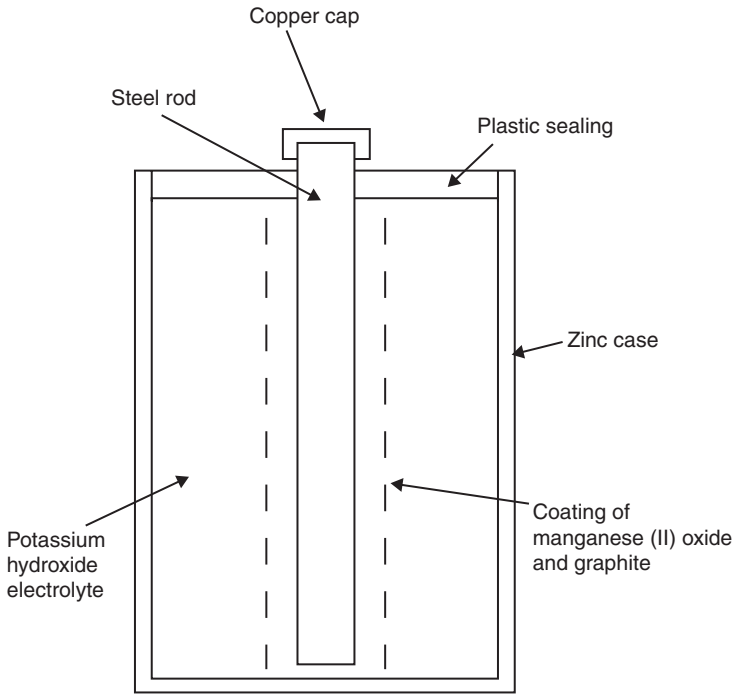


Figure 4.5

Typical cross-section of a manganese alkaline cell.

but the full-scale production of manganese alkaline cells did not start until the 1960s. The cell uses zinc as the cathode, with an electrolyte of potassium hydroxide solution, either as liquid or as jelly, and the anode is a coating of manganese (II) oxide mixed with graphite and laid on steel. The cell is sealed because the reaction does not liberate gas, and the manganese (II) oxide is used for its manganese content rather than for its oxygen content as a depolarizer.

The EMF of a fresh cell is 1.5 V, and the initial EMF is maintained almost unchanged for practically the whole of the life of the cell. The energy content, weight for weight, is higher than that of the carbon–zinc cell by a factor of 5–10, and the shelf-life is very much better owing to an almost complete lack of secondary action. All of this makes these cells very

suitable for electronics use, particularly for equipment that has fairly long inactive periods followed by large current demand. Incidentally, though the cells use alkali rather than acid, it must be remembered that potassium hydroxide is a caustic material which will dissolve the skin and is extremely dangerous to the eyes. An alkaline cell must never be opened, nor should any attempt ever be made to recharge it other than with specialized charging equipment.

Miniature (button) cells

Miniature cells are the types specified for deaf-aids, calculators, cameras and watches, but they are quite often found in other applications, such as for backup of memory in computing applications and for 'smart-card' units in which a credit-card is equipped with a complete microprocessor and memory structure so that it keeps track of transactions. The main miniature cells are silver oxide and mercury, but the term mercury cell can be misleading, because metallic mercury is not involved.

The mercuric oxide button cell, to give it the correct title, uses an electrolyte of potassium hydroxide (Figure 4.6) which has had zinc oxide dissolved in it until saturated, so the cell can be classed as an alkaline type. The cathode is the familiar zinc, using either a cylinder of perforated zinc foil or a sintered zinc-powder cylinder fastened to the button-top of the cell and insulated from the bottom casing. The anode is a coating of mercury (I) oxide mixed with graphite to improve conductivity and coated on nickel-plated steel or stainless steel which forms the casing of the cell. The EMF of such cells is low, 1.2–1.3 V, and the energy content is high, with long shelf-life owing to the absence of local action.

The silver oxide cell is constructed in very much the same way as the mercuric oxide cell, but using silver (I) oxide mixed with graphite as the anode. The cathode is zinc and the electrolyte is potassium hydroxide as for the mercuric oxide cell. The EMF is 1.5 V, a value that is maintained at a steady level for most of the long life of the cell. The energy content is high and the shelf-life long.

All of these miniature cells are intended for very low current applications, so great care should be taken to avoid accidental discharge paths. If the cells

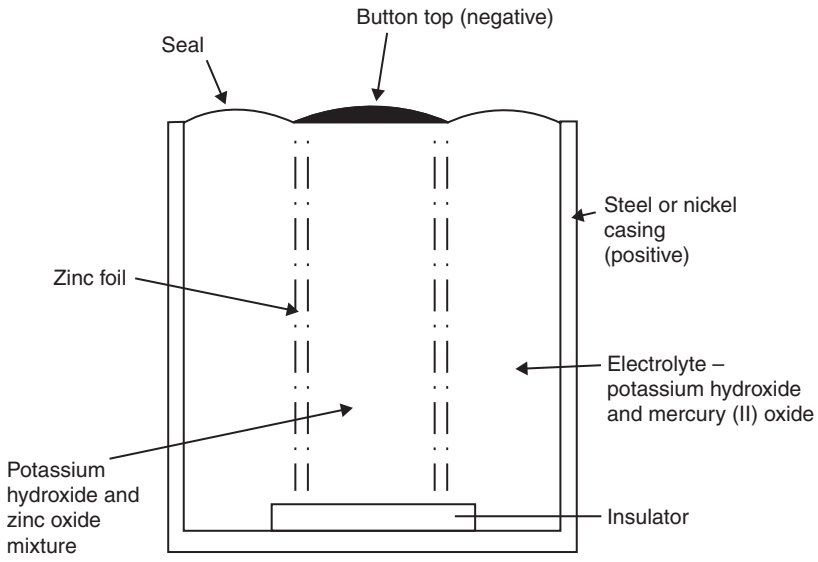


Figure 4.6

The construction of a mercuric oxide cell.

are touched by hand, this will leave a film of perspiration that is sufficiently conductive to shorten the life of the cell drastically. When these cells are fitted, they should be moved and fitted with tweezers, preferably plastic tweezers or with dry rubber gloves if you need to use your hands. These cells should not be recharged, nor disposed of in a fire. The mercury type is particularly hazardous if mercury compounds are released, and it should be returned to the manufacturer for correct disposal if this is possible; otherwise it should be disposed of by a firm that is competent to handle mercury compounds.

Lithium cells

Lithium is a metal akin to potassium and sodium which is highly reactive, so much so that it cannot be exposed to air and reacts with explosive violence with water. The reactive nature of lithium metal means that a water solution

cannot be used as the electrolyte and much research has gone into finding liquids which ionize to some extent but which do not react excessively with lithium. A sulphur–chlorine compound, thionyl chloride, is used, with enough dissolved lithium salts to make the amount of ionization sufficient for the conductivity that is needed. The Lithium cell is the most recently developed type of cell, remarkable for its unexpectedly high EMF.

The lithium (Figure 4.7) is coated on to a stainless-steel mesh which is separated from the rest of the cell by a porous polypropylene container. The anode is a mixture of manganese (II) oxide and graphite, also coated on to stainless-steel mesh. The whole cell is very carefully sealed. The reaction can be used to provide a cell with an exceptionally high EMF of 3.7 V, very long shelf-life of 10 years or more, and high energy content. The EMF is almost constant over the life of the cell, and the internal resistance can be low.

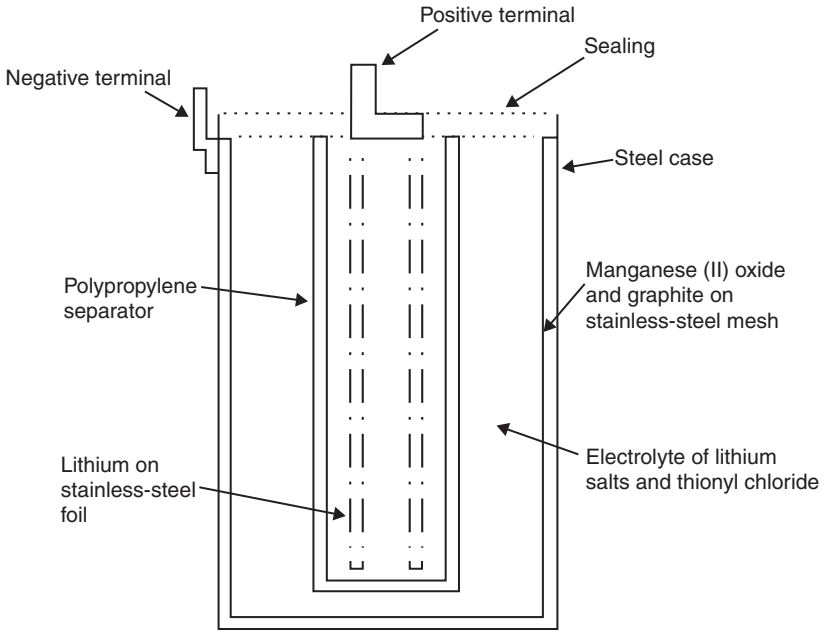


Figure 4.7

The construction of a lithium cell.

Lithium cells are expensive, but their unique characteristics have led to them being used in automatic cameras where focusing, film wind, shutter action, exposure and flash are all dependent on one battery, usually a one- or two-cell lithium type. For electronics applications, lithium cells are used mainly for memory backup, and very often the life of the battery is as great as the expected lifetime of the memory itself. The cells are sealed, but since excessive current drain can cause a build-up of hydrogen gas, a 'safety-valve' is incorporated in the form of a thin section of container wall which will blow out in the event of excess pressure. Since this will allow the atmosphere to reach the lithium, with risk of fire, the cells should be protected from accidental over-current, which would cause blow-out. A recommended protection circuit is illustrated in Figure 4.8.

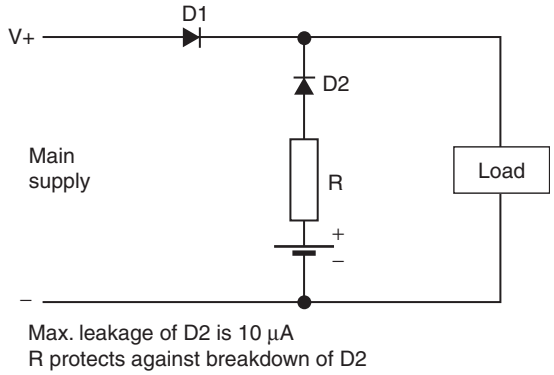


Figure 4.8

A recommended reverse-current protection circuit for a lithium cell in a simple backup application.

This is for use in applications where the lithium cell is used as a backup, so that D1 conducts during normal memory operation and D2 conducts during backup. Short-circuit failure of D2 would cause the lithium cell to be charged by the normal supply, and the resistor R will then limit the current to an amount which the cell manufacturer deems to be safe. If the use of a resistor would cause too great a voltage drop in normal backup use, it could be replaced by a quick-blowing fuse, but this has the

disadvantage that it would cause loss of memory when the main supply was switched off.

Lithium cells must **never** be connected in parallel, and even series connection is discouraged and limited to a maximum of two cells. The cells are designed for low load currents, and in Figure 4.9 is shown a typical plot of battery voltage, current and life at 20°C. Some varieties of lithium cells exhibit voltage lag, so that the full output voltage is available only after the cell has been on load for a short time – the effect becomes more noticeable as the cell ages. Another oddity is that the capacity of a lithium cell is slightly lower if the cell is not mounted with the +ve terminal uppermost. See later for rechargeable lithium cells.

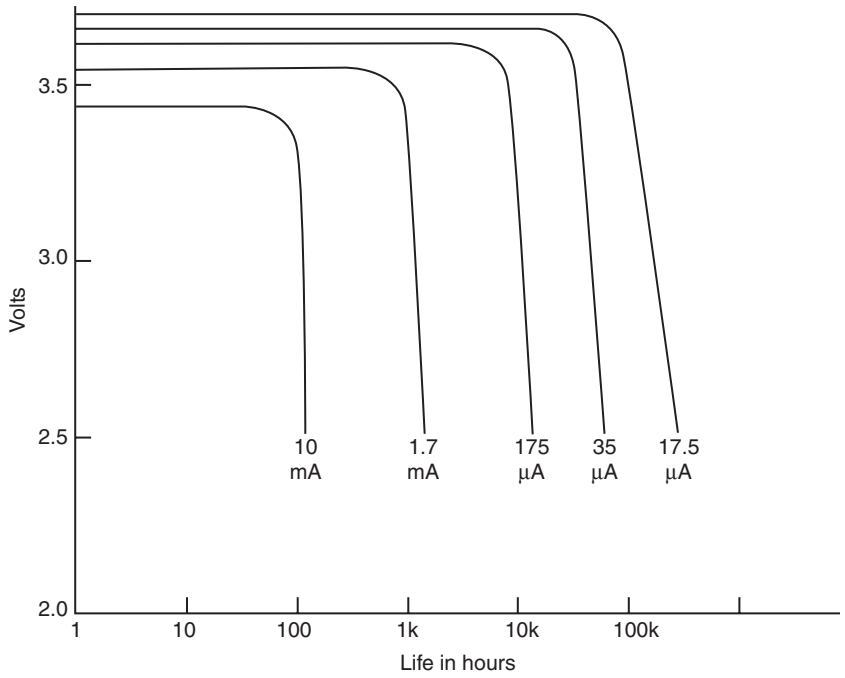


Figure 4.9

Typical plots of lithium cell voltage, current and life at normal room temperature levels.

Secondary cells

A secondary cell makes use of a reversible chemical process, so that when the cell is discharged, reverse current into the cell will recharge it by restoring the original chemical constitution. Unlike primary cell reactions, reversible reactions of this type are unusual and for many years only two basic types were known, the lead–acid type and the alkali–metal type.

The lead–acid cell construction principle is illustrated in Figure 4.10. Both plates are made from lead and are perforated to allow them to be packed

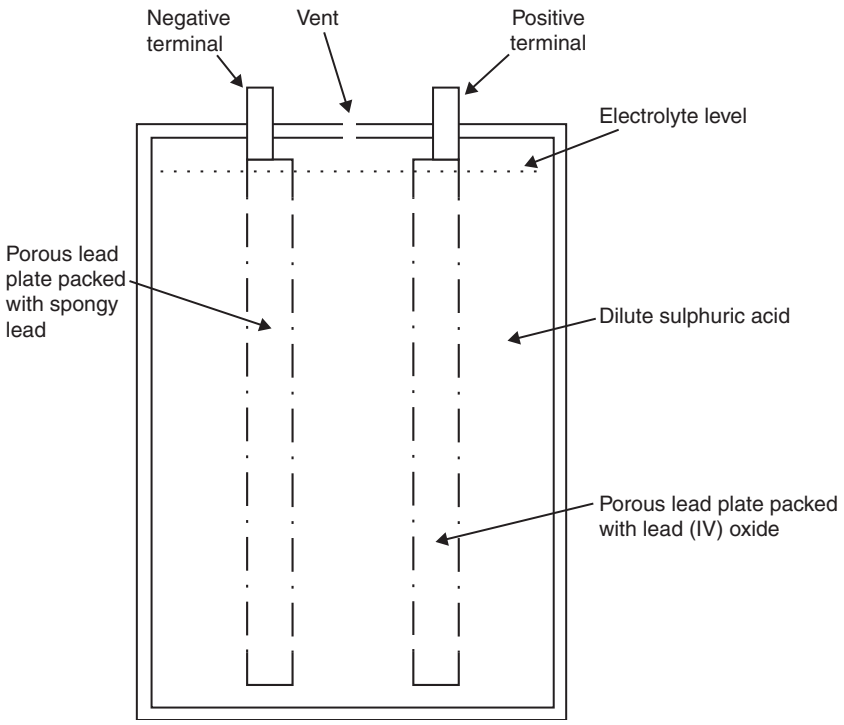


Figure 4.10

Principles of the lead–acid secondary cell.

with the active materials. One, the positive plate (anode), is packed with lead (IV) oxide, and the negative plate (cathode) is packed with spongy or sintered lead which has a large surface area. Both plates are immersed in sulphuric acid solution. The acidity is much greater than that of the electrolytes of any of the acidic dry cells, and very great care must be taken when working with lead–acid cells to avoid any spillage of acid or any charging fault that could cause the acid to boil or to burst out of the casing. In addition, the recharging of a vented lead–acid cell releases hydrogen and oxygen as a highly explosive mixture which will detonate violently if there is any spark nearby.

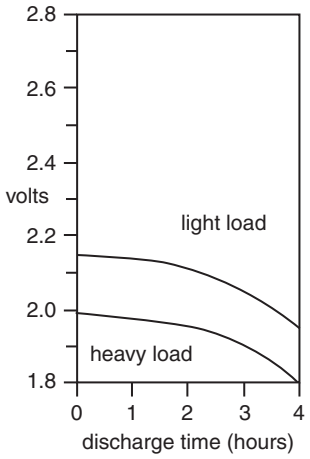
Never connect a lead of a complete circuit to a lead–acid battery; any connection should be to a circuit that has a switch, with the switch off. Making a connection to a complete circuit is likely to cause a spark at the time of connection, which can cause an explosion.

The fully charged EMF is 2.2 V (nominally 2.0 V), and the variation in voltage is quite large as the cell discharges. Figure 4.11 shows typical discharge graphs for light-load and heavy-load respectively.

The older vented type of lead–acid cell is now a rare sight, and modern lead–acid cells, as used in cars, are sealed, relying on better control of charging equipment to avoid excessive gas pressure. The 'dry' type of cell uses electrolyte in jelly form, so these cells can be used in any operating position.

Figure 4.11

Voltage drop on discharging a lead–acid cell.



Cells that use a liquid electrolyte are constructed with porous separator material between the plates so that the electrolyte is absorbed in the separator material, and this allows these cells also to be placed in any operating position. Since gas pressure build-up is still possible if charging circuits fail, cells are equipped with a pressure-operated vent that will reseal when pressure drops again.

Lead–acid cells are used in electronics applications mainly as backup power supplies, as part of uninterruptible power systems, where their large capacities and low internal resistance can be utilized. Capacity is measured in ampere-hours, and sizes of 9 Ah to 110 Ah are commonly used. Care should be taken in selecting suitable types – some types of lead–acid cells will self-discharge considerably faster than others and are better suited to applications where there is a fairly regular charge/discharge cycle than for backup systems in which the battery may be used only on exceptional occasions and charging is also infrequent. Figure 4.12 shows the self-discharge

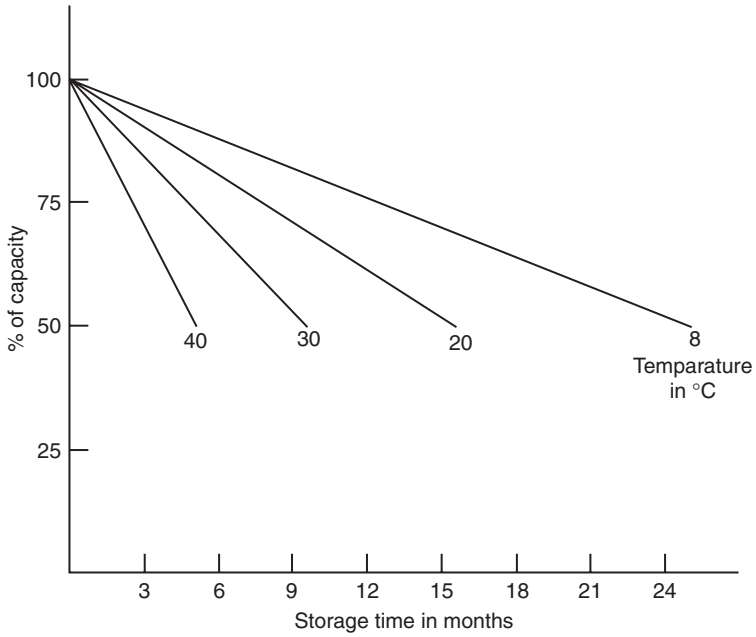


Figure 4.12

Self-discharge plots for a jelly type of lead–acid cell.

rates of jelly-electrolyte cells at various temperatures, taking the arbitrary figure of 50% capacity as the discharge point.

Lead–acid batteries need to be charged from a constant-voltage source of about 2.3 V per cell at 20°C – Figure 4.13 shows the variation of charging voltage per cell with ambient temperature of the cell. Cells can be connected in series for charging provided that all of the cells are of the same type and equally discharged. A suitable multi-cell charger circuit, using a variable-voltage regulator such as the LM317T, is illustrated in Figure 4.14.

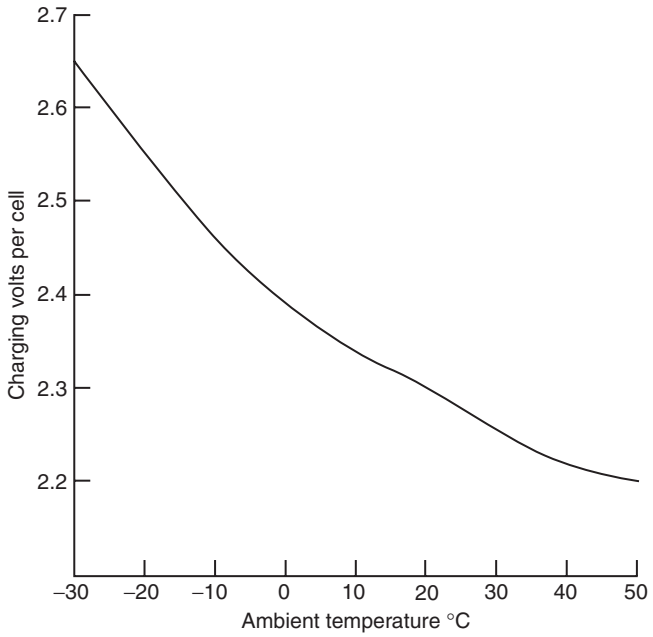
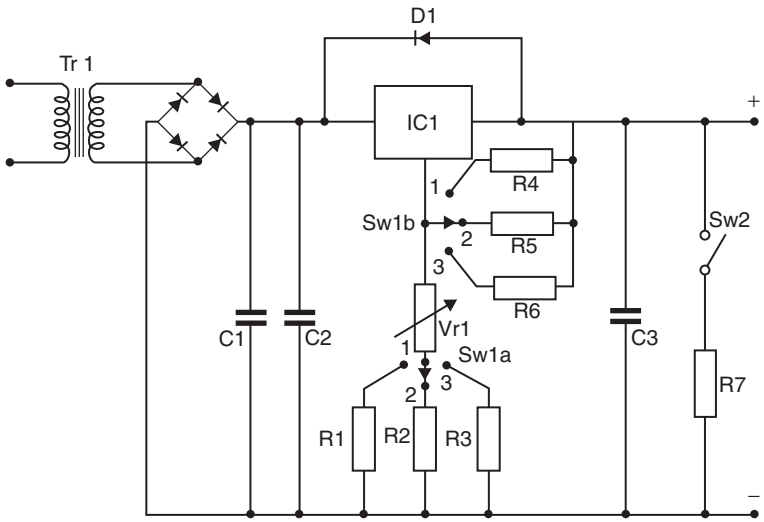


Figure 4.13

Temperature variation of charging voltage for a lead–acid cell.

For batteries of more than 24 V (12 cells) the charging should be in 24 V blocks, or a charging system used that will distribute charging so that no single cell is being over-charged. Parallel charging can be used if the charger can provide enough current. The operating life of a lead–acid cell



Tr1 30 V 1.6 A IC1 1.5 A var. stab
D1 IN4001 Vr1 100 R lin
C1 470μ 25 V C2 100n C3 1μ 25 V
R1 33R R2 300R R3 360R R4 91R
R5 68R R6 39R R7 220R 2.5W

To set voltage close S2 adjust Vr1
S1 settings: 1..1 cell 2..3 cells
3..6 cells

Figure 4.14

A circuit for a multi-cell charger. (Courtesy of RS Components.)

is usually measured in terms of the number of charge/discharge cycles, and is greater when the cell is used with fairly high discharge currents – the worst operating conditions are of slow discharge and erratic recharge intervals, the conditions that usually prevail when these cells are used for backup purposes.

One condition that must be avoided is **deep discharge**, when the cell has been left either on load or discharged for a long period. In this state, the terminal voltage falls to 1.6 V or less and the cell is likely to be permanently damaged unless it is immediately recharged at a very low current over a long period. Typical life expectancy for a correctly operated cell is of the order of 750–6000 charge/discharge cycles.

Nickel–cadmium cells

The original type of alkaline secondary cell, invented by Edison at the turn of the 19th/20th century, was the nickel–cathode iron–anode type, using sodium hydroxide as the electrolyte. The EMF is only 1.2 V, but the cell can be left discharged for long periods without harm, and will withstand much heavier charge and discharge cycles than the lead–acid type. Though the nickel–iron alkaline secondary cell still exists, powering milk-floats and fork-lift trucks, it is not used in the smaller sizes because of the superior performance of the nickel–cadmium and nickel–hydride types of cells which are now the most common type of secondary cell used for cordless appliances and in electronics uses.

Nickel–cadmium (Ni–Cd) cells can be obtained in two main forms, mass-plate and sintered plate. The mass-plate type used nickel and cadmium plates made from smooth sheet; the sintered type has plates formed by moulding powdered metal at high temperatures and pressures, making the plates very porous and of much greater surface area. This makes the internal resistance of sintered-plate cells much lower, so larger discharge currents can be achieved. The mass-plate type, however, has much lower self-discharge rates and is more suitable for applications in which recharging is not frequent. Typical life expectancy is from 700 to 1000 charge/discharge cycles.

One very considerable advantage of the nickel–cadmium cell is that it can be stored for 5 years or more without deterioration. Though charge will be lost, there is nothing corresponding to the deep discharge state of lead–acid cells that would cause irreversible damage. The only problem that can lead to cell destruction is reverse polarity charging. The cells can be used and charged in any position, and are usually supplied virtually discharged, so they must be fully charged before use. Most nickel–cadmium cell types have a fairly high self-discharge rate, and a cell will on occasion refuse to accept charge until it has been 're-formed' with a brief pulse of high current. Cells are usually sealed but provided with a safety-vent in case of incorrect charging.

In use, the nickel–cadmium cell has a maximum EMF of about 1.4 V, 1.2 V nominal, and this EMF of 1.2 V is sustained for most of the discharge time. The time for discharge is usually taken arbitrarily as the time to reach an

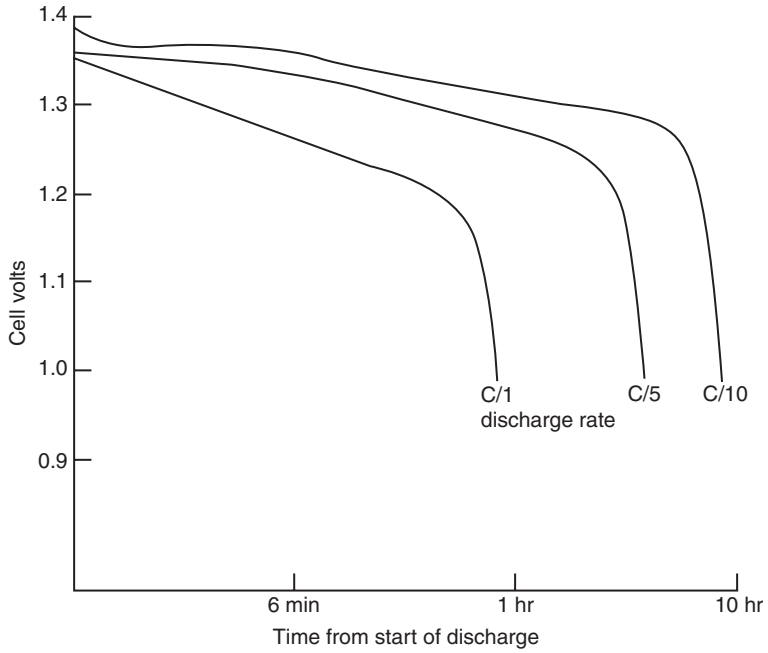


Figure 4.15

Typical discharge characteristics for small nickel–cadmium cells.

EMF of 1 V per cell, and Figure 4.15 shows typical voltage–time plots for a variety of discharge rates. These rates are noted in terms of capacity, ranging from one fifth of capacity to five times capacity, when capacity is in ampere-hours and discharge current in amps. For example, if the capacity is 10 Ah, a C/5 discharge rate means that the discharge current is 2 A.

Charging of nickel–cadmium cells must be done from a **constant-current** source, in contrast to the constant-voltage charging of lead–acid types. The normal rate of charge is about one tenth of the Ah rate, so for a 20 Ah cell, the charge rate would be 2 A. Sintered types can be recharged at faster rates than the mass-plate type, but the mass-plate type can be kept on continuous trickle charge of about 1% of capacity (for example, 10 mA for a cell of 1 Ah capacity). At this rate, the cells can be maintained on charge for an extended period after they are fully charged, but this over-charge period is about three times the normal charging time. Equipment such

as portable and cordless phones which would otherwise be left on charge over extended intervals such as Bank Holiday weekends and office holidays should be disconnected from the charger rather than left to trickle-charge. This means that a full charge will usually be needed when work resumes, but the life of the cells can be considerably extended if the very long periods of charging can be avoided. Another option is to leave the equipment switched on so as to discharge the cells, and fit the mains supply with a timer so that there will periodic recharging.

In Figure 4.16 is shown a recommended circuit for recharging. This uses a 7805 regulator to provide a fixed voltage of 5 V across a resistor, so the value of the current depends on the choice of resistor and not on the voltage of the cell. The value of the resistor has to be chosen to suit the type of cell being recharged; values from 10 Ω to 470 Ω are used depending on the capacity of the cell. Because the regulator system is floating with respect to ground, this can be used for charging single cells or series sets of a few cells. Ready-made chargers are also available which will take various cells

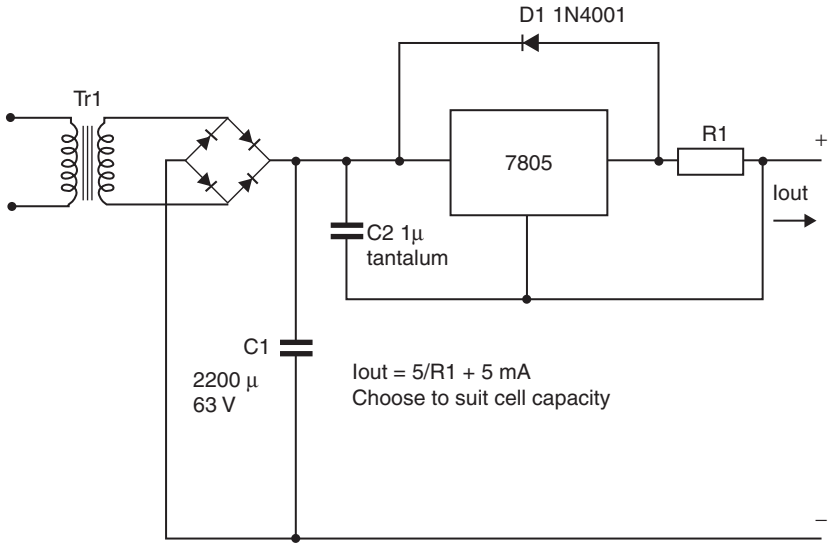


Figure 4.16

A recommended charging circuit for nickel–cadmium cells. (Courtesy of RS Components.)

singly or in combination, with the correct current regulation for each type of cell.

A major disadvantage of Ni–Cd cells is the **memory effect**. If a cell is frequently recharged before its voltage has appreciably fallen, its capacity is reduced, and it eventually has to be recharged much more frequently to be kept in service. Some users recommend the use of dischargers, a load that will fairly rapidly discharge a Ni–Cd cell to a level that does not cause damage, so recharging is always carried out on a cell that is almost completely discharged. This can considerably extend the life of a cell, but is not a practical proposition for Ni–Cd cells that are embedded in equipment, such as in cordless phones. The use of nickel–metal hydride (Ni–MH) cells greatly reduces this effect, and many portable applications, such as mobile phones, digital cameras and camcorders, now make use of lithium-ion rechargeable cells.

A form of silver cell has also been used in rechargeable form. This uses an anode of porous zinc, usually a sintered component, with a silver (I) oxide and graphite cathode. The electrolyte is potassium hydroxide solution that has been saturated with zinc hydroxide. The cell can take a limited number of recharging cycles, but is now uncommon.

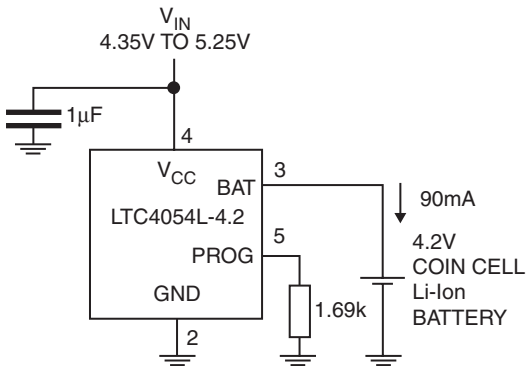
Lithium-ion rechargeable cells

The **lithium-ion cell**, as distinct from the lithium cell, does not contain metallic lithium and so does not present the hazard of lithium if it is broken. The cell consists of three layers; a porous insulating separating film sandwiched between a carbon anode and a cathode in sheet form that is coated with alloy of lithium with cobalt, nickel or manganese (so three forms of cell are possible). The cell is filled with an electrolyte which is a salt of lithium dissolved in an organic liquid (not water), often nowadays in a gel form. The EMF is of the order of 4.0 V after charging dropping to 2.6 volts for a discharged cell, so these batteries are often used along with a built-in voltage regulator to maintain an output voltage of around 3.0 V for as long as the cell EMF is above this level.

This type of cell can be only half the weight and size of a Ni–Cd cell of the same capacity, with none of the memory effect that causes so

Figure 4.17

A simple Li-ion charger circuit for a single button cell. (Courtesy of Linear Technology.)



many problems with Ni–Cd cells. They also have high thermal stability and resistance to overcharging. The charging circuit is specialized, and chargers intended for other cell types must never be used for Li-ion cells. In Figure 4.17 is shown a simple one-cell charger, delivering a maximum current of 90 mA to a cell.

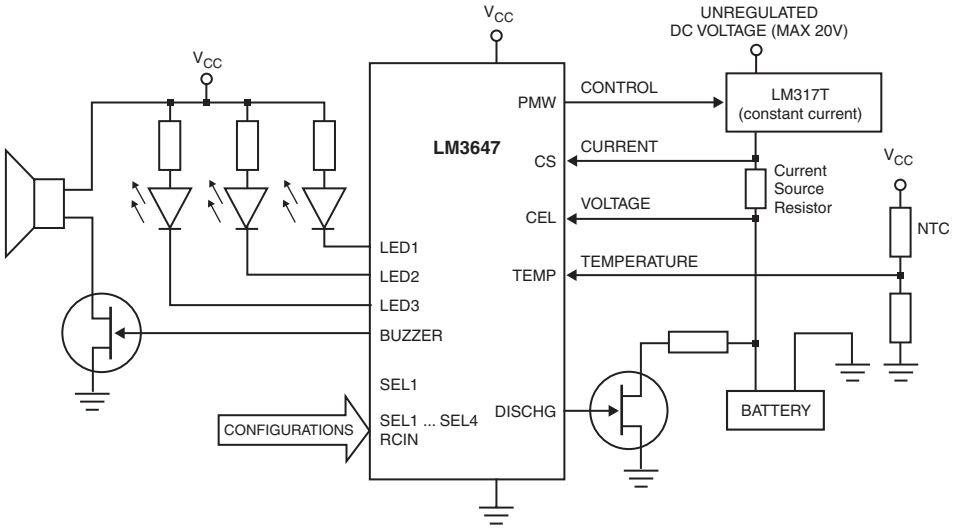


Figure 4.18

A charger circuit for all types of rechargeable batteries. (Courtesy of National Semiconductor.)

In Figure 4.18 is shown a universal charger circuit, for lithium-ion (Li-ion), nickel–metal hydride (Ni–MH) and nickel–cadmium (Ni–Cd) batteries. This circuit can be configured to use either pulsed-current or constant-current charging methods, or to discharge before charging. The charging time is regulated by monitoring voltage, temperature and time.

Nothing's perfect, and some types of Li-ion cells have been reported as being mechanically fragile, with suggestions that some pieces of equipment will work only with Li-ion cells from the same manufacturer. This is a rapidly developing area of cell technology, and these early reports are now out of date. The use of Li-ion cells is almost universal in products such as digital camcorders, laptop computers and portable DVD players. Modern Li-ion cells incorporate internal protection circuits to prevent explosions resulting from overcharging, and some manufacturers, worried by 'gray' imports, now incorporate hologram labels to show that their battery is genuine and contains protection. There is currently a suggestion to extend this and incorporate a processor that can ensure that only genuine batteries will work in devices such as mobile phones.

This page intentionally left blank

CHAPTER 5

ACTIVE DISCRETE COMPONENTS

Diodes

Semiconductor diodes can use two basic forms of construction, point-contact or junction. Point-contact diodes are still available and used for small-signal purposes where a low value of capacitance between the terminals is of primary importance – their main use has been for RF demodulation, but even in this use they are now seldom encountered because the diode action is usually incorporated as part of an IC that combines several functions (such as IF amplification, demodulation and signal processing). Most of the applications for point-contact diodes can be more usefully carried out by devices such as the BAT85 Schottky diode.

Junction diodes are obtainable with a much greater range of voltage and current applications, and are used for most other purposes. Apart from diodes intended for specialized purposes, such as light-emitting diodes, the fabrication materials are silicon or (less commonly) germanium, with germanium used almost exclusively for point-contact diodes. An **ideal diode** would form a short circuit for current in one direction (the forward direction) and an open circuit for current in the reverse direction.

Practical diodes have a low forward resistance (whose value is not constant) and a high reverse resistance; and they conduct when the anode voltage is a few hundred millivolts more positive than the cathode voltage. Semiconductor diodes conduct using minority carriers, meaning that the electrons carry current through the P-region and holes carry the current through the N-region.

The diode can be destroyed by excessive forward current, which causes high power dissipation at the junction or point-contact, or by using excessive

reverse voltage which also causes junction or point-contact breakdown, allowing conduction in the reverse direction. This in turn may result in an open circuit caused by excessive current. For any diode, therefore, the published ratings of peak forward current and peak reverse voltage should not be exceeded, and should not be approached if reliable operation is to be achieved. If both peak forward current and peak reverse voltage are together near their limits, some derating should be applied. Characteristics for a typical (old) point-contact germanium diode and a typical small-signal silicon junction diode are also shown in Figure 5.1.

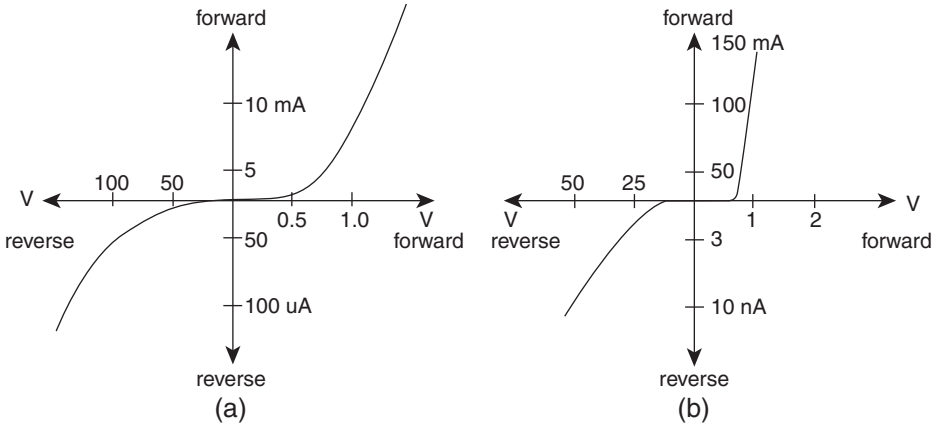


Figure 5.1

Characteristics of real diodes: **(a)** germanium point diode, **(b)** silicon junction diode. Note the different scales which have to be used to allow the graphs to be fitted into a reasonable space.

Comparing these two extremes:

(a) Germanium point-contact diodes, seldom now used, have lower reverse resistance values, conduct at a lower forward voltage (about 0.2 V) but have higher forward resistance because of their small junction area. They also have rather low peak values of forward current and reverse voltage.

(b) Silicon junction diodes have very high values of reverse resistance, conducting at a forward voltage of around 0.55 V, can have fairly

low forward resistance values, and can have fairly high peak values of forward current and reverse voltage.

The **forward resistance** of a diode is not a fixed quantity, but is, very approximately, inversely proportional to current, so the resistance is high when the current is low and vice versa. Another approximation that is useful for small currents is that the forward voltage of a silicon diode increases by only 60 mV for a tenfold increase in current. The effect of temperature change on a silicon diode is to change the forward voltage across the conducting diode at any fixed value of current. A change of about 2.5 mV per °C is a typical figure, with the voltage reducing as the temperature is raised. The reverse (leakage) current is much more dependent on temperature and a useful rule of thumb is that the leakage current doubles for each 10°C rise in temperature.

Zener diodes are used with reverse bias, making use of the breakdown that occurs across a silicon junction when the reverse voltage causes a large electrostatic field to develop across the junction. This breakdown limit occurs at low voltages (below 6 V) when the silicon is very strongly doped, and such breakdown is termed Zener breakdown, from Clarence Zener who discovered the effect. For such a true Zener diode, the reverse characteristic is as shown in Figure 5.2.

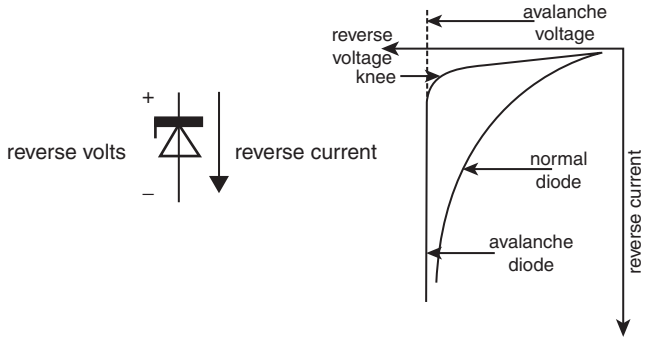


Figure 5.2

Zener diode. The true Zener effect causes a 'soft' breakdown at low voltages; the avalanche effect causes a sharper turnover.

As the graph in Figure 5.2 illustrates, the reverse current does not suddenly increase at the Zener voltage, and the voltage across the diode is not truly stabilized unless the current is more than a few milliamps. This type of characteristic is termed a **soft breakdown** characteristic. In addition to this, a true Zener diode has a negative temperature coefficient – the voltage across the reverse-biased diode (at a constant current value) decreases as the junction temperature is increased.

Avalanche breakdown occurs in diodes which have lower doping levels, at voltages above about 6 V. The name is derived from the avalanche action in which electrons are separated from holes by the electric field across the junction and these electrons and holes then cause further electron-hole separation by collisions. These diodes have hard characteristics (Figure 5.2), with very little current flowing when the reverse voltage is below the avalanche limit, and large currents above this limit. In addition, the temperature coefficient of voltage across the diode increases as the junction temperature is raised.

Both types of diodes are, however, known as **Zener diodes** and those with breakdown voltages in the range of 4 V to 6 V can combine both effects. At a breakdown voltage of about 5.6 V the opposing temperature characteristics balance with the result that the breakdown (stabilized) voltage of a 5.6 V (usually written as 5V6) diode is practically unaffected by temperature changes. The stabilization of a diode is measured by its **dynamic resistance**, defined as the ratio:

$$\frac{dV}{dI} \quad \text{meaning} \quad \left(\frac{\text{voltage change}}{\text{current change}}\right)$$

whose units are ohms when dV is the change of voltage across the diode caused by a change of current dI through the diode under stabilized conditions. This ratio should be below 50 ohms, and reaches a minimum value of about 4 ohms for a diode with a breakdown voltage of about 8 V.

The types that are termed **reference diodes** are doped to an extent that makes the breakdown voltage practically constant despite changes in ambient temperature. Voltages of 5 V to 6 V are used and temperature coefficients ranging from +0.01% per degree down to +0.0005% per degree can be achieved. These reference diodes are used for very precise voltage stabilization.

Another method of obtaining a very stable reference voltage makes use of a **band-gap circuit**. This uses transistors operated with different emitter current density figures to produce a stable 60 mV difference between the base-emitter voltages. This 60 mV is amplified by a factor of ten and added to a V_{be} voltage to give a stable output of 1.25 V, and this can be used as a reference voltage.

Varactor diodes

All junction diodes have a measurable capacitance between anode and cathode when the junction is reverse biased, and this capacitance varies with the size of the reverse voltage, being least when the reverse voltage is high (which could mean voltage levels of 6 V or less). This variation, caused by the removal of charge carriers from the junction at high reverse voltages, is made use of in varactor diodes, in which the doping is arranged so as to provide the maximum possible capacitance variation consistent with high resistance. A typical variation is of 10 pF at 10 V bias to 35 pF at 1 V reverse bias. Varactor diodes are used for electronic tuning applications and a typical circuit is illustrated in Chapter 7 (Figure 7.42). The symbols (all four that you will find in circuit diagrams) and a typical characteristic are illustrated in Figure 5.3.

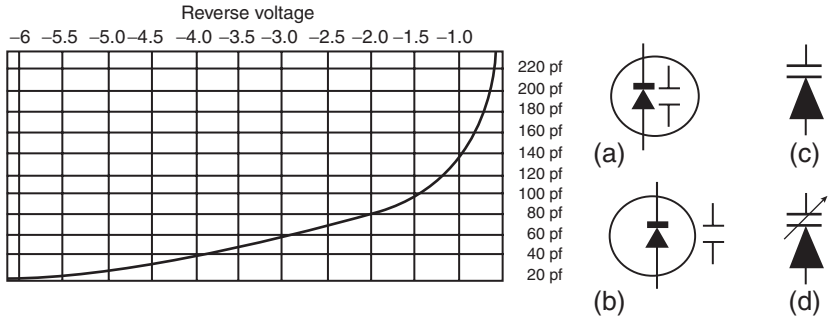


Figure 5.3

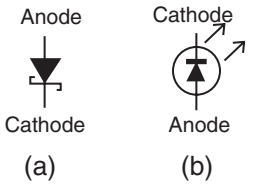
Varactor diode typical characteristic and symbols: **(a)** and **(b)** US, **(c)** and **(d)** UK. The official UK symbol is shown in **(c)**.

Schottky diodes

Schottky diodes are named for their discoverer, the physicist Walter Schottky. A Schottky diode consists of a metal-semiconductor junction, in which the semiconductor is usually silicon, and the metal can be, typically, silver, aluminium, gold, chromium, nickel, platinum or tungsten, or alloys of exotic metals. The diode conducts using majority carriers, so that the forward drop is small, only about 0.2 V compared to the 0.6 V of a silicon diode. In addition, the diodes have very fast switching times, meaning that when the voltage is switched off the current also turns off with only a very small delay. This feature makes the Schottky diode useful in RF applications such as RF demodulation and in high-frequency switch-mode power supplies. Because of the low voltage drop, the diodes also make excellent power rectifiers, particularly for high-frequency supplies, though the reverse current is too high for some applications. Figure 5.4a shows the relevant symbol.

Figure 5.4

Symbols: **(a)** Schottky diode, **(b)** LED.



Schottky diodes are also used embedded into ICs (see later) in logic circuits, and as part of complex devices ranging from photodiodes to MOSFETs. Silicon carbide Schottky diodes are now being used for high-current diodes with very high voltage ratings (up to 1200 V).

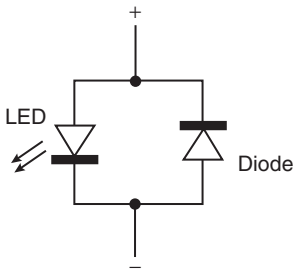
LEDs

Light-emitting diodes (LEDs) use **compound semiconductor** materials such as gallium arsenide or indium phosphide. The relevant symbol is illustrated in Figure 5.4b. When forward current passes, light is emitted from the junction. The colour of the light depends on the semiconductor

material used for the diode and the brightness is approximately proportional to the size of forward current. LEDs have higher forward voltages when conducting; around 1.6 V to 2.2 V as compared to the 0.5 V to 0.8 V of a silicon junction. The maximum permitted reverse voltages are very low, typically only 3 V, so a silicon diode must be connected across the LED as shown in Figure 5.5 if there is any likelihood of reverse voltage (or an AC signal) being applied to the diode. A series resistor must always be used to limit the forward current unless pulsed operation is used.

Figure 5.5

Protecting an LED from reverse voltage.



In Table 5.1 is shown some of the current range of LEDs with output colour and forward voltage drop. Note that the infra-red types emit little or no visible light; typical applications include remote controls and short-range signalling. In addition to the types noted in the table, all-white outputs can be achieved by combination structures either (1) using a combination of red, green and blue or (2) combining a blue/UV diode with a white phosphor coating (notably from Marl Optosource Ltd.).

Photodiodes

A photodiode can be regarded as a high-impedance non-ohmic photosensitive device whose current is almost independent of applied voltage. The incident light falls on a reverse-biased semiconductor junction, and the separation of electrons from holes will allow the junction to conduct despite the reverse-bias. Photodiodes are constructed like any other diodes, using silicon, but without the opaque coating that is normally used on signal

Table 5.1 LED materials and characteristics

Material	Colour	V_f (volts)	I_{typ} (mA)	Notes
GaAs	Infrared	1.2–1.3	50	Original type, launched in 1980s
GaAlAs	Infra-red–red	1.4	50	Faintly visible
GaAsP/GaAs	Red	1.6–1.75	20	Very low efficiency
GaP	Red/orange	1.9	30 mA max	Non-linear characteristic
GaAlAsP	Red	1.8–1.9	20	Bright
GaAsP/GaP	Red/orange	1.9	5–20	High efficiency
InGaAlP	Red/orange/ yellow/green	1.9–2.3	20	Bright
GaAsP	Yellow	2	20	First yellow type developed
GaP	Green	2.1	20	First green type
InGaN	Blue/green	3.6	20	Efficiency improving now
GaN	Blue/white	3.6	20	Sensitive to voltage/current overloads
SiC	Blue	3.5	30	Low efficiency
GaN/SiC	Blue/violet	3.8–5	20	
GaN	Ultraviolet	3.9	10	Faintly visible

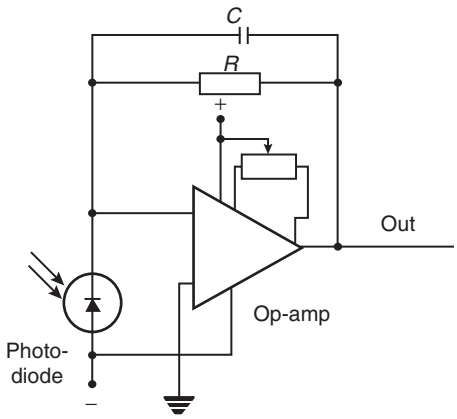
Note: Abbreviations for materials: Al, aluminium; As, arsenic; C, carbon; Ga, gallium; In, indium; N, nitrogen; P, phosphorus; Si, silicon. Oblique stroke indicates one semiconductor on a substrate of another; for example GaAsP/GaAs means gallium–arsenic–phosphorus on gallium arsenide.

and rectifier diodes. The junction area may be quite large, so the photodiode may have more capacitance between electrodes than a conventional signal diode. This can be compensated by using a feedback capacitor in the circuit, illustrated in Figure 5.6, which shows a typical circuit for using a photodiode along with an operational amplifier for a voltage output. The feedback resistor R will determine the output voltage, which will be RI, where I is the diode current.

- Some LEDs can be used as photodiodes with peak sensitivity values in the infra-red or in the visible spectrum, and in some circuits it can be convenient to use the same device as both a receiver and an indicator.

Figure 5.6

The output of the photodiode is normally very small, and amplification is almost always needed. Note the diode symbol, which is like the LED symbol but with the arrow's direction reversed.



Characteristics for photodiodes specify the output current into a short circuit, and the current will be much lower into a resistance of appreciable value. The sensitivity can be quoted in terms of incident light measurements, but Table 5.2, shows, more usefully, the output of some types when the incident light is provided by various typical sources.

Table 5.2 Photodiode output from various sources

Part number	A (mA)	B (μA)	C (μA)	D (mA)
OSD1-5T	0.47	0.45	0.32	0.71
OSD5-5T	1.80	2.10	1.70	1.00
OSD15-5T	4.50	5.60	2.60	1.00
OSD35-5T	11.00	14.00	3.80	1.10
OSD60-5T	28.00	39.00	7.20	1.10

Note: A, noon sunlight; B, room light; C, Super-red LED 1 cm distant; D, laser pointer 1 metre. (Part numbers from Centrovision list.)

In Figure 5.7 is shown typical photodiode spectral response, meaning the sensitivity, in terms of amps per watt (A/W), at different wavelengths of light. The response of a photodiode is not the same as that of a human eye, but the addition of light filters can bring the response closer. The linearity, in terms of output current plotted against strength of light input, is very good.

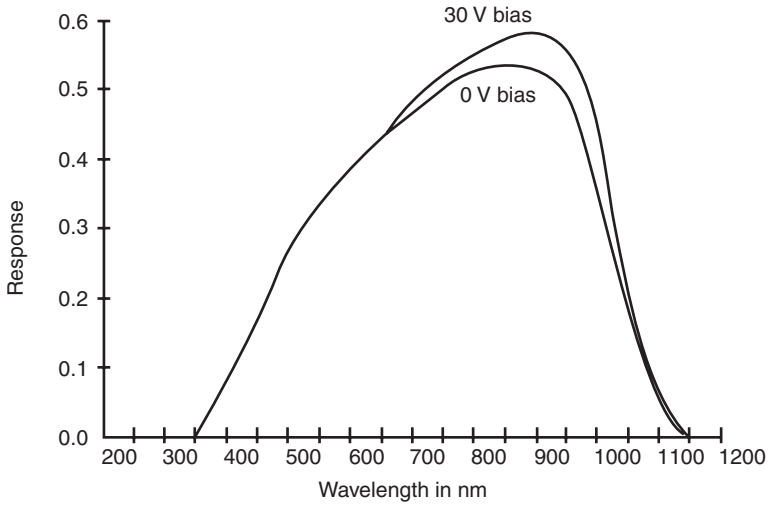


Figure 5.7

Spectral response of a silicon photodiode. (Courtesy of Centrovision.)

Figure 5.8 shows typical circuits using a photodiode using an operational amplifier (see Chapter 6) as a load. The circuit in (a) is used for high sensitivity and operation down to DC levels. The circuit in (b) is preferred when speed of response is preferred to operation at very low frequencies.

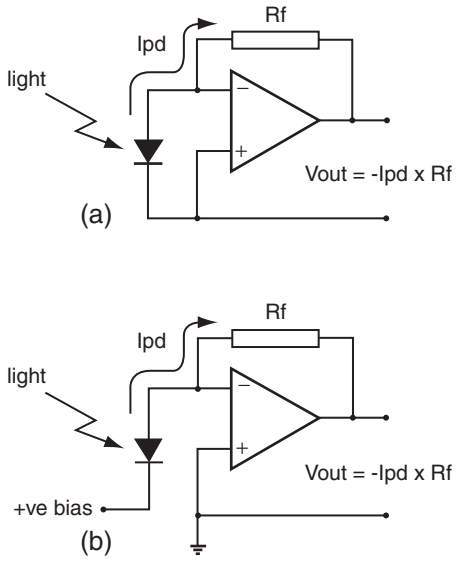
Transient voltage suppressors (TVS)

The transient voltage suppressor is a form of semiconductor device related to diodes. Its purpose is to protect circuits against transients, of either voltage or current. The two main forms are the silicon avalanche junction type and the metal-oxide varistor type.

The **silicon avalanche junction** types use a Zener diode construction with a larger cross-section to achieve higher surge power ratings. The response time is fast and the impedance is low when the avalanche effect starts. They are available either as unidirectional (for DC surges) or bidirectional (for AC surges). The packaging can be in the form of chips, surface mount or axial leads. Figure 5.9 shows a typical characteristic for a bipolar TVS.

Figure 5.8

(a) A photodiode circuit for high sensitivity;
(b) a circuit with better response time.



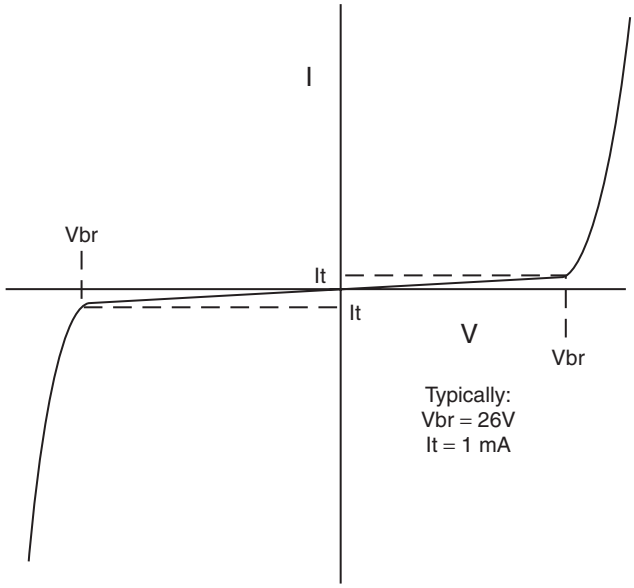


Figure 5.9

Typical silicon TVS characteristic.

Metal-oxide **varistors** are conventionally made using grains of zinc oxide embedded in a mixture of other metal oxides (notable bismuth oxide), and each grain of zinc oxide where it is in contact with another oxide acts as a bidirectional semiconductor junction with a breakdown voltage of around 2–3 V, so the assembly can be thought of as a set of many diodes in series–parallel. These can be packaged in various sizes (corresponding to power dissipation) ranging from single chip form to large finned casings intended to work at thousands of volts and/or thousands of amps.

Typical diode circuits

In Figure 5.10 are shown some common application circuits for diodes, with approximate design data where appropriate. Diode types should be selected with reference to the manufacturer's data sheets, having decided on the basic reverse voltage and load current requirements of the circuit. Note that these circuit are nowadays more likely to be embedded in an integrated circuit rather than existing in separate component (discrete) form.

Transistors

Like signal diodes, transistors can be constructed using either silicon or germanium, but virtually all transistors other than exotic types use silicon; the exotic types use compound semiconductors such as gallium arsenide. The design data in this section refer mainly to silicon transistors. Though you may seldom see transistors used as separate components in modern circuits, it is important to know how they work, because they form the basis of the integrated circuits (ICs) that are used in virtually all electronic circuits today. In addition, experimental circuits for which there is no existing IC available have to be made from a combination of ICs and discrete transistors and diodes. See later for a note about **digital transistors** which have built-in bias resistors.

Figure 5.11 shows a schematic outline of the *bipolar junction* transistor (BJT), one of the two important types of transistor. This is a device that makes use of two junctions in a crystal with a very thin layer between the junctions. The thin layer is called the *base*, and the type of BJT depends on whether this base layer is made from P-type or from N-type material.

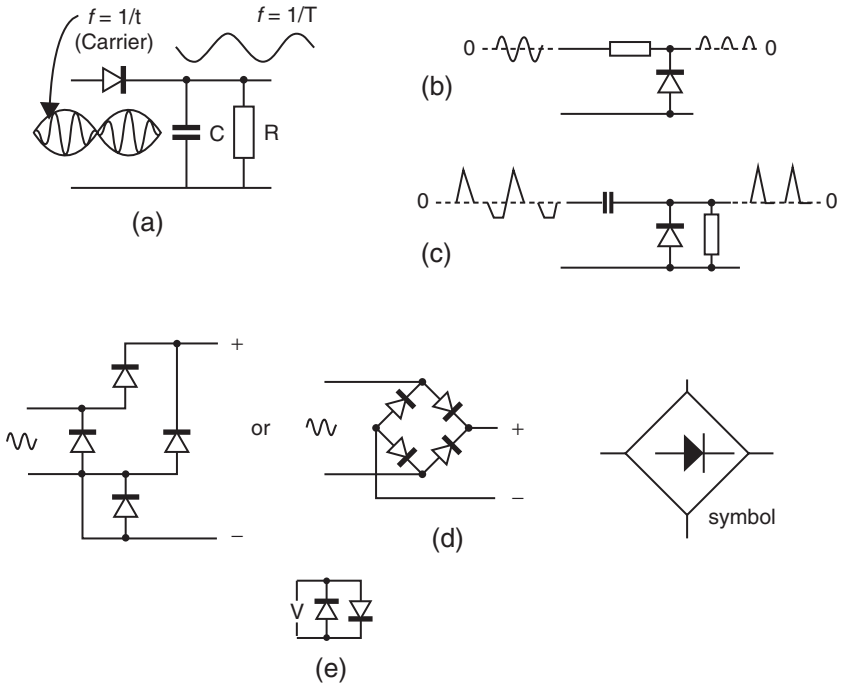


Figure 5.10

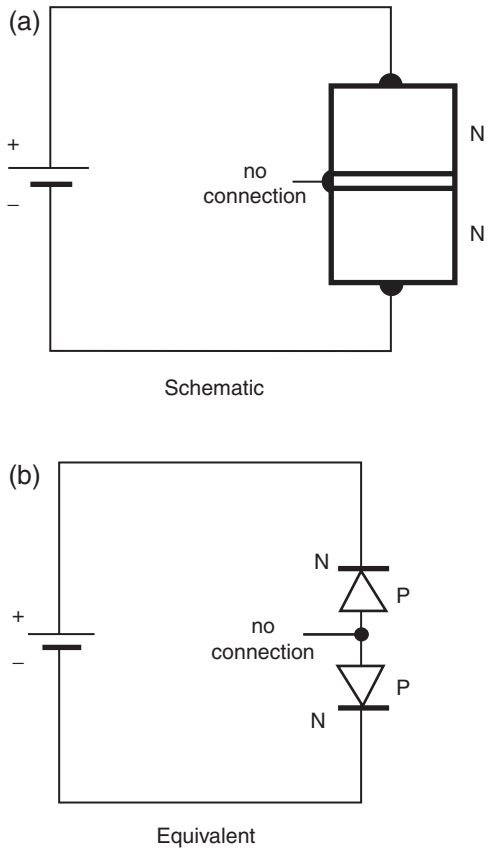
Some diode applications: **(a)** amplitude demodulation, **(b)** and **(c)** signal clipping, **(d)** bridge rectification showing alternative ways of drawing, and the symbol for a bridge rectifier assembly, **(e)** anti-parallel diode assemblies for over-voltage protection or clipping.

If the base layer is of N-type material, the transistor is a P-N-P type, and if the base layer is of P-type material, the transistor is an N-P-N type. The differences lie in the polarity of power supplies and signals rather than in the way that the transistors act. For most of this chapter, we shall concentrate on the N-P-N type of transistor, simply because it is more widely used. The figure also shows an equivalent circuit (b) of two back-to-back diodes, which represents the way that the terminals of a transistor respond to DC measurements.

Consider an NPN transistor connected as shown in Figure 5.11a. With no bias voltage, or with reverse bias, between the base and the emitter connections, there are no carriers in the base-emitter junction, and the voltage between the collector and the base makes this junction reverse-biased, so no current can flow in this junction either. The transistor behaves as if

Figure 5.11

(a) Schematic connection of NPN BJT, **(b)** equivalent diode circuit.



it were two diodes connected anode-to-anode (Figure 5.11b). No current could flow in the circuit even if the battery connections were to be reversed.

When the base-emitter junction is forward-biased, however, electrons will move across this junction. Because the collector–base junction is physically so thin, the collector potential will cause most of the electrons to be swept across this junction to provide collector current even if the junction is reverse-biased.

With both junctions conducting, most of the current will flow between the collector and the emitter, since this is the path of lower resistance. The transistor no longer behaves like two back-to-back diodes

because the electrons passing through the base–emitter junction make the collector–base junction conduct despite the reverse bias between collector and base.

The current flowing between the collector and the emitter is much greater (typically 25 to 800 times greater) than the current flowing between the base and the emitter. If the base is now unbiased or reverse-biased again, no current can flow between the collector and the emitter. Thus the current in the base-emitter junction controls the amount of current passing through the collector–base junction. The word *bipolar* is used because both holes and electrons play their parts in the flow of current. Figure 5.12 shows a typical form of NPN transistor construction.

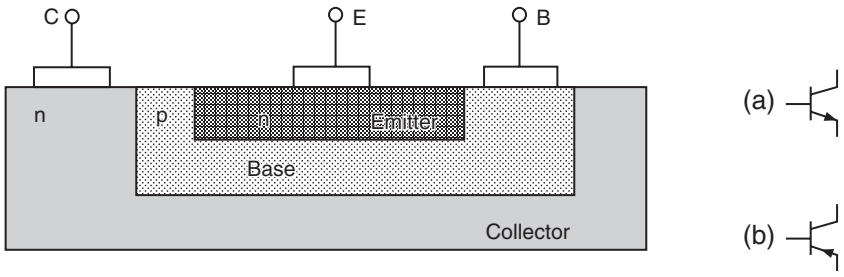


Figure 5.12

Typical NPN transistor construction with the symbols for the NPN transistor **(a)** and its opposite, the PNP type **(b)**.

The working principle of a BJT, then, is that current flows between the collector and the emitter only when current is flowing between the base and the emitter terminals. The ratio of these currents is called the **forward current transfer ratio**, symbol h_{fe}. For the arrangement of Figure 5.13 the ratio is defined as:

$$h_{fe} = \frac{i_c}{i_b}$$

In databooks, a distinction is made between h_{FE} for which I_c and I_b are steady DC values, and h_{fe}, for which i_c and i_b are small-current AC values. The two quantities are, however, generally close enough in value to be

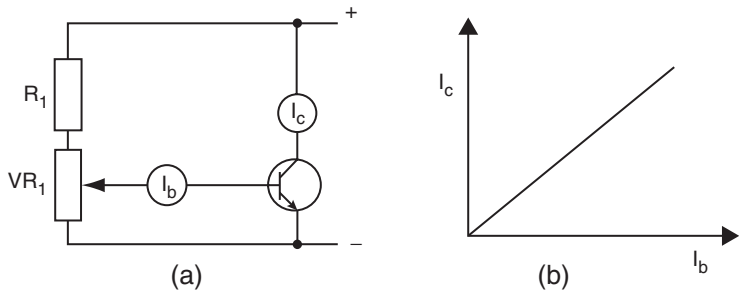


Figure 5.13

Forward current transfer ratio: **(a)** measuring circuit, **(b)** graph. The slope of the graph (I_c/I_b) is equal to the forward current transfer ratio h_{fe}.

interchangeable, and the symbol h_{fe}, will be used here to mean either value. The size of h_{fe} for any transistor can be measured in the circuit shown in Figure 5.13; a simpler method, used in many transistor testers, is shown in Figure 5.14. Values vary from about 25 (power transistors operating at high current levels) to over 1000 for some high-frequency amplifier types.

Base current will not flow unless the voltage between the base and the emitter provides a suitable amount of base current. The precise voltage at which base current starts to be measurable varies from one specimen of transistor to another (even of the same type), but for silicon transistors it is around 0.55 V; we often assume 0.6 V. The PNP type of transistor will require the emitter to be at a more positive voltage than the base; the NPN type will require the base to be more positive than the emitter. When the transistor has the correct DC currents flowing, with no signal applied, it is said to be correctly biased in a **quiescent** state. Amplification is carried out by adding a fluctuating signal voltage to the steady bias voltage at the input of the transistor. The vast majority of transistor circuits use the base as the input terminal, with a minority using the emitter (a configuration called **common base**, because the base is at AC earth and is therefore the common terminal for both input and output).

Any one of a transistor's three electrodes can be connected to perform in this common role, so there are three possible configurations: common emitter, common-collector and common-base. These three basic bipolar transistor

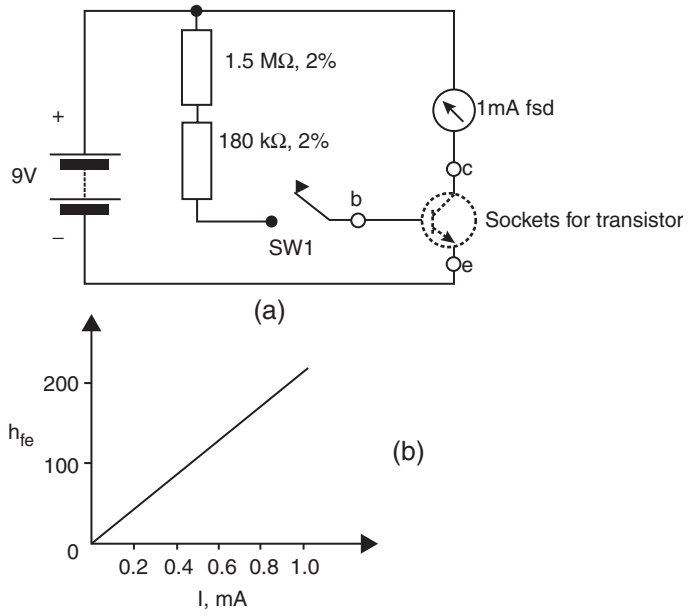
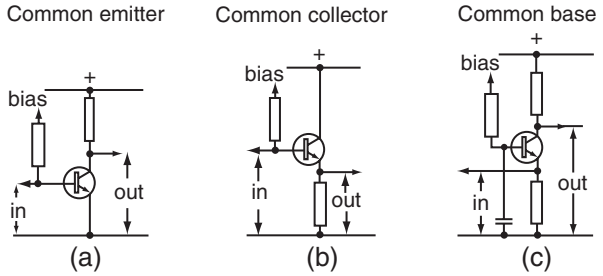


Figure 5.14

A simple transistor tester **(a)** and its calibration graph **(b)**.

circuit connections are shown in Figure 5.15, with applications and values of typical input and output resistances given below each. Figure 5.15a shows the normal common-emitter amplifying connection used in most transistor circuits. The common-collector connection in Figure 5.15b, with signal into the base and out from the emitter, is used for matching impedances, since it has a high input impedance and a low output impedance. The common-base connection, with signal into the emitter and out from the collector, shown in Figure 5.15c is nowadays used mainly for UHF amplification.

The normal function of a transistor when the base-emitter junction is forward biased and the base-collector junction reverse-biased, is to act as a current amplifier. Voltage amplification is achieved by connecting a load resistor (or impedance) between the collector lead and the supply voltage (see Figure 5.15a). Oscillation is achieved when the transistor is connected



	Common emitter	Common collector	Common base
Voltage gain	High (~ 100)	Unity (1)	Medium (10--50)
Current gain	High (50--800)	High (50--800)	Unity (1)
Input resistance	Medium (~ 5K)	High (several K)	Low (~ 50R)
Output resistance	High (~ 40K)	Low (a few ohms)	High (~ 1M)

Figure 5.15

The three circuit connections of a bipolar transistor: **(a)** common-emitter, **(b)** common-collector (or emitter-follower), **(c)** common-base.

as an amplifier with its output fed back, in phase, to its input. The transistor can also be used as a switch or relay when the base-emitter junction is switched between reverse bias and forward bias.

Note that the base-emitter junction of many types of silicon transistors will break down by avalanche action at voltages ranging from 7 V to 20 V reverse bias, though this action does not necessarily cause collector current to flow. The base-emitter junction can be protected by connecting an antiparallel diode between the base and emitter.

Bias for linear amplifiers

A linear amplifier produces at its output (usually the collector of a transistor) a waveform which is a perfect copy, but of greater amplitude, of the waveform applied at the input (usually the base). The voltage gain of such an amplifier is defined as:

$$G = \frac{v_{out}}{v_{in}}$$

where v indicates an AC signal voltage measurement. If the output waveform is not a perfect copy of the input waveform the amplifier is

exhibiting **distortion** of one form or another. One type of distortion is **non-linear distortion** in which the shape of the output waveform is not identical to the shape of the input waveform. Such non-linear distortion is caused by the inherent non-linear characteristics of the transistor and can be minimized by careful choice of transistor type (see later) and by correct bias.

A transistor is correctly biased for linear operation when the desired amount of gain can be obtained with minimum non-linear distortion. This is easiest to achieve when the peak-to-peak output signal from the transistor is much smaller than the DC supply voltage. For very small output signals, the DC voltage level at the collector can be set to almost any reasonable level between zero and the supply voltage, but the preferred value is half-way between supply positive and the voltage level at the emitter. This allows for unexpected overloads and usually places the operating conditions in the most linear portion of the characteristics of the transistor (or, more correctly, the least non-linear region).

When the value of collector resistor has been chosen, bias is applied by passing current into the base so that the collector voltage drops to the desired value of around 0.5 V_{ss} where V_{ss} is the supply voltage. For any bias system, the desired base current must be equal to:

$$\frac{0.5\,V_{cc}}{R_L \times h_{fe}}$$

with V_{cc} in volts, R_L (the load) in kΩ, h_{fe} as a ratio. Figure 5.16 shows three basic bias systems for a single transistor along with design data for obtaining a suitable bias voltage.

The method of Figure 5.16a is the most difficult and least satisfactory because a different resistor value will have to be used for each different transistor. The resistance value is critical, and will usually consist of series–parallel connected resistors because no single resistor will be of the correct value. In addition, the bias will be correct for only one temperature, and will alter drastically as the temperature changes. The method of Figure 5.16b is a considerable improvement over that of Figure 5.16a because of the use of DC feedback. The bias system can be designed around an 'average' transistor (with an average value of h_{fe} for that type) and can be used without modification for other specimens of that transistor and even for

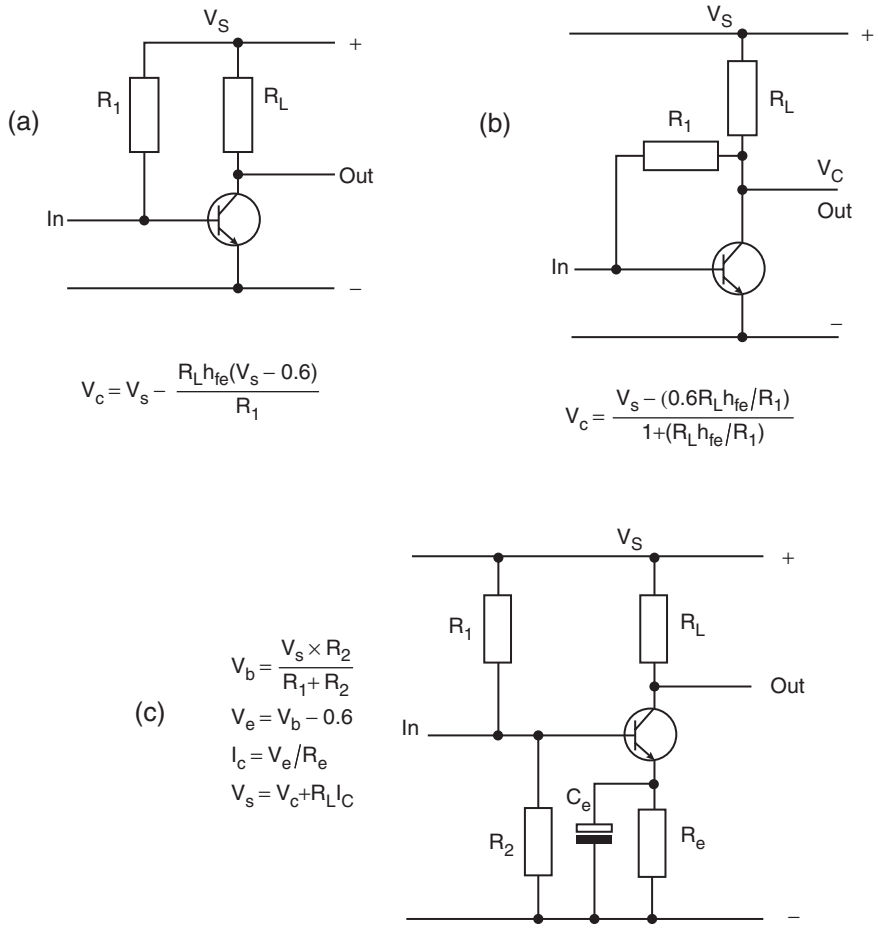


Figure 5.16

Transistor bias circuits: **(a)** simple system, usually unsatisfactory, **(b)** using negative feedback of bias and signal, **(c)** potential divider method

a range of similar transistor types. The collector voltage will change as temperature changes, but to a much smaller extent than that for the circuit in Figure 5.16a, and the bias will normally remain acceptable even for fairly large temperature changes.

The bias system of Figure 5.16c is the most commonly used. It is a bias method that can be used for any transistor provided that the current flowing

through the two base bias resistors R_1 and R_2 is much greater than the base current drawn by the transistor. Unlike the other two systems, the design formula does not require the h_{fe} value for the transistor to be known if the standing current through the transistor is to be only a few milliamps. For power transistors, the quantities that are needed are the V_{be} and I_{be} values at the required collector bias current. This system does not, however, stabilize the collector voltage so effectively against bias changes caused by changes of temperature.

Calculating how stable a bias system will be is needed only for relatively advanced designs, and for a large number of amplifier uses two simple rules can be relied upon:

- Never fix V_{be}, because this will cause large changes in collector current as temperature changes.

- Never fix I_b, because $I_c = h_{fe} \times I_b$, and h_{fe} varies from one transistor another and also with temperature.

If this is not enough, you can calculate the stability factor S. The formula you need depends on which biasing method you use, and we'll look only at the one used for the bias method that uses a voltage divider to set the base voltage, along with an emitter resistor.

The stability factor S can be calculated for changes in transistor parameters caused by changes in temperature or by substituting one transistor for another. The lower the value of S, the more stable your bias system will be. For most purposes, you can simplify the formula by making the reasonable assumption that R_b/R_e is much less than h_{fe}, where R_b and R_e are respectively the resistances in the base and emitter leads. In this case:

$$S \approx 1 + \frac{R_b}{R_e}$$

Note: You can use bias and other calculators in the set downloaded from http://www.angelfire.com. These are very useful, and provide bias calculations and graphical illustration of the effects of external changes on bias. Some of the other facilities are of more interest to designers working

with valves in transmitter circuits, but there is sufficient data included on transistor circuits to be useful.

Transistor parameters and linear amplifier gain

Transistor parameters are measured quantities that describe the action of the transistor. The term **parameter** is used to distinguish these quantities from constants (which would maintain the same value for all transistors). Transistor parameters vary from one transistor specimen to another, even of the same type, and from one value of bias current to another. One such parameter, the common-emitter current gain, h_{fe}, has already been described. Of the parameters needed to design linear amplifiers, g_m is probably the most useful. The **mutual conductance**, g_m, is measured in units of millisiemens (mS, equivalent to milliamps per volt) and is the same parameter as was once used in amplifier design using valves. Note that the use of the capital 'S' distinguishes the Siemens unit from seconds (s). The value of g_m is defined by the equation:

$$g_m = \frac{i_c}{v_{be}} \quad \text{where} \quad i_c = \text{AC signal current, between collector and emitter}$$
$$v_{be} = \text{AC signal voltage between base and emitter}$$

which is the ratio of AC signal current in the collector to AC signal voltage between the base and the emitter. The usefulness of g_m as a parameter arises from the fact that the voltage gain of a transistor amplifier for small signals is given by:

$$A_v = g_m R_L$$

where R_L is the load resistance for signal frequencies. If g_m is measured in mS (equivalent to milliamps per volt) and R_L in kΩ the gain will be correctly stated (gain has no units). Note that R_L will generally be of a lower value than that of the resistor that is connected between the collector and the supply because this resistor value will be shunted by any other load that is connected through a capacitor (Figure 5.17).

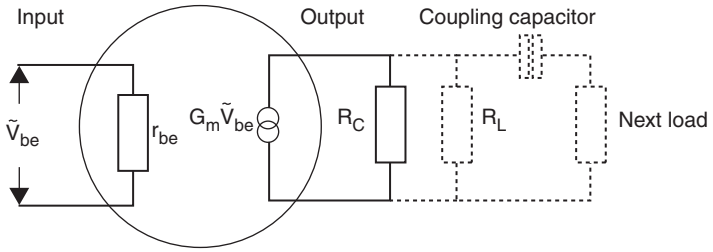


Figure 5.17

A useful equivalent circuit for the transistor. The signal voltage v_{be} between the base and the emitter causes an output signal current $g_m v_{be}$. This current flows through the parallel combination of resistor R_c (the transistor output resistance), R_L, the load resistor, and any other load resistors in the circuit.

This load will usually be the input resistance of the next transistor in a multistage amplifier. A graph of collector current plotted against base emitter voltage is not a straight line, because the transistor is inherently a non-linear device, so g_m does not have a constant value. A useful rule of thumb for small bias currents is that the average value of g_m (in mS) is equal to 40 times the bias current in milliamps. The shape of the g_m against I_c graph is always curved at low current values, but straightening out at higher currents. For a few transistor types the I_c–V_{be} graph has a noticeably straight portion which makes these transistor types particularly suitable for linear amplification applications. It is this (comparative) straightness of the g_m characteristic that makes some types of power transistor much more desirable (and more costly) for use in audio output stages. To take advantage of these linear characteristics, of course, the bias must be arranged so that the working point is at the centre of the most linear region when no signal input is applied, and the signal input must not be so large as to extend into a severely non-linear portion. The 'working point' in this context means the combination of collector current and base voltage that represents a point on the characteristic. Two other useful parameters for silicon transistors, used in common-emitter circuits, are the input and output resistance values. The input resistance is defined as:

$$\frac{\text{signal voltage at base input}}{\text{signal current into base}}$$

common emitter, symbol h_{ie}, and measured with no voltage on the collector. The output resistance is:

$$\frac{\text{signal voltage at collector}}{\text{signal current at collector}}$$

common emitter, symbol h_{oe}, and is measured by applying a signal to the collector with no signal at the base. The output resistance h_{oe}, has about the same range of values, 10 kΩ to 50 kΩ for a surprisingly large number of transistors irrespective of operating conditions, provided that these operating conditions are on the flat portion of the I_c–V_{be} characteristic (Figure 5.18).

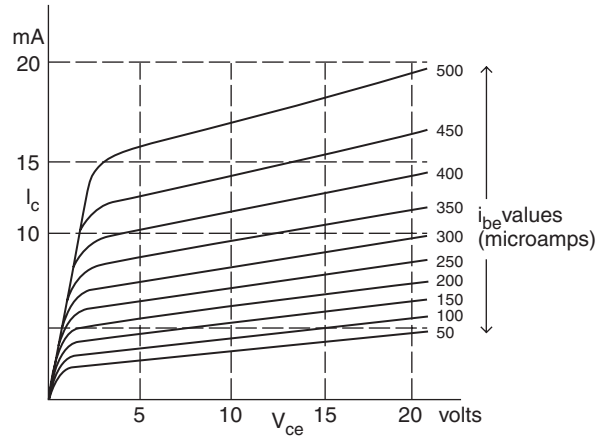


Figure 5.18

The I_c vs. V_{be} characteristic. The flat portion is the operating part. The small amount of slope indicates that the output resistance, R_c, is high, usually 40 kΩ or more.

An average value of 30 kΩ can usually be assumed for small-signal amplifier transistors, though much higher values can be found for some RF types. The input resistance is not a constant because the input stage of a transistor is the base-emitter junction which is a diode with an exponential characteristic. The value of input resistance, h_{ie}, is related to the steady bias current and

to the other parameters by the equation:

$$h_{ie} = \frac{h_{fe}}{G_m} \quad \text{and since } G_m \text{ is about } 40 \times I_c, \; h_{ie} = \frac{h_{fe}}{40\, I_c}$$

where I_c is the steady no-signal bias collector current. For example, if a transistor has an h_{fe} value of 120 and is used at a collector bias current of 1 mA, its input resistance (in kΩ) is:

$$h_{ie} = \frac{120}{40 \times 1} = 3\,\text{k}\Omega$$

We can also use the graph shown in Figure 5.18 as a way of calculating amplification, by drawing a **load line**. A load line is a line whose voltage–current plot represents a load resistance, and it is drawn over the I_c–V_{ce} graph as shown in Figure 5.19. Where the load line cuts any of the I_b lines a value of V_{ce} can be read off, so the output voltage swing for a given input current swing can be calculated. The load line can give a reasonable guide to how linear the amplification will be (equal distances between the points at which the load line meets the I_b lines), but is seldom used nowadays to

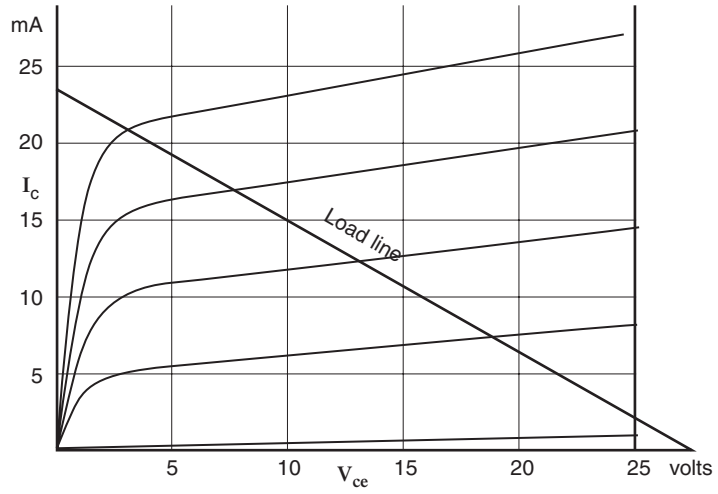


Figure 5.19

A typical load-line drawing.

calculate voltage gain, because you need the additional information to find the relationship between I_b and V_{be}.

- **TRANSISTOR PACKAGING**

Transistors are packaged in a large variety of forms, some of which are now little used. Many are now appearing in SMD form, but there are still many found that have the older traditional metal or plastic cases with wire leads, and the higher-dissipation types are packaged in metal cases that bolt on to a heat sink. A few typical packages are shown in Figure 5.20.

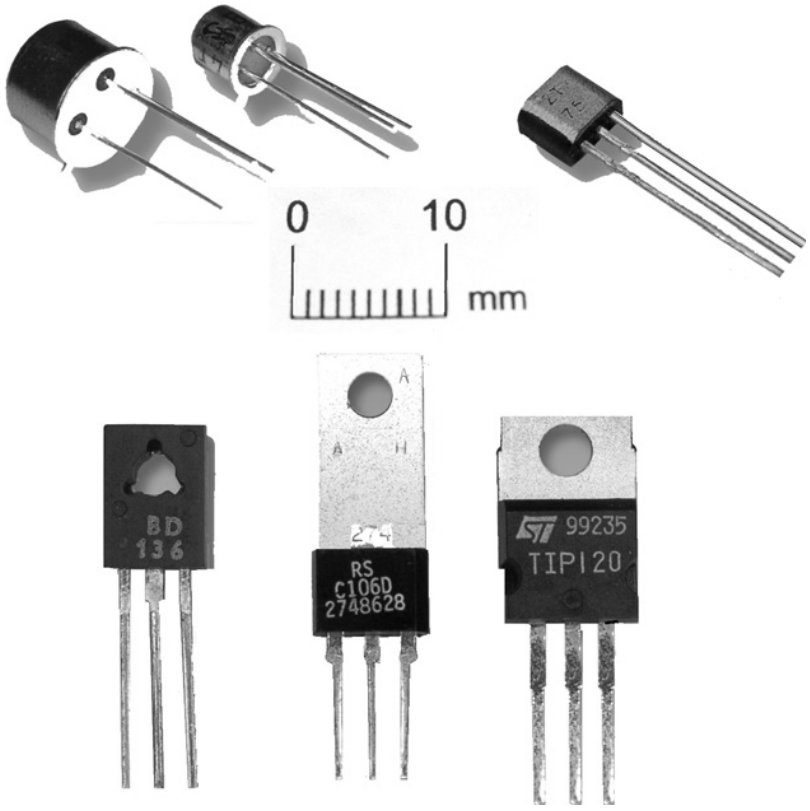


Figure 5.20

Some typical transistor packages showing small-signal transistors above and power transistors below. (Original photos courtesy of Alan Winstanley.)

Noise

Any working transistor generates electrical noise, and the greater the current flowing through the transistor the greater the noise. For bipolar transistors the optimum collector current for low-noise operation is given approximately, in milliamps, by:

$$I_c = \frac{28\sqrt{h_{fe}}}{R_g}$$

where R_g is the signal source resistance in ohms

Low-noise operation is most important for the first stages of audio preamplifiers and for RF tuners and early IF stages. The noise that is generated by large-value resistors is also significant, so the resistors used for small-signal input stages should be of fairly low values if possible and of high stability film types. Variable resistors must not be used in the signal path of any low-noise stage. The greatest contribution to noise, however, is that of the transistor itself, and a good choice of type can be of considerable benefit as regards noise level. Types such as the BC549, BC559, MRF2947AT2, BD437 and BD438 are often specified for audio circuits. For RF use, some typical types are BF495, 2SC2413KP and TSDF1205 (up to 25 GHz).

Voltage gain

The voltage gain of a simple single-stage silicon BJT voltage amplifier can be found from a simple rule of thumb. If V_{bias} is the steady DC voltage across the collector load resistor, and I_c is the collector current and R_L the load, the voltage gain is given by:

$$A_v = 40 \times I_c \times R_L$$

For a single-stage amplifier, the signal is attenuated both at the input and at the output by the potential dividing action of the transistor input and output resistance values with the resistance of the devices that are connected at input and output (microphones, tape heads, other amplifying stages).

If the resistance of the signal source is R_s and the resistance of the next stage is R_{load}, the measured gain of a transistor stage will be:

$$A \times \frac{h_{ie}}{R_s + h_{ie}} \times \frac{R_{load}}{R_{load} + R_C}$$

where R_c is the collector output resistance as shown in Figure 5.21 and A is the value of gain given by $40 \times V_{bias}$. This method gives gain values that are precise enough for most practical purposes. When precise values of gain are needed, negative feedback circuits (see later) must be used. For a multistage amplifier, the gains of individual stages are multiplied together and multiplied also by the attenuation action of each R_s and R_{load} in the circuit.

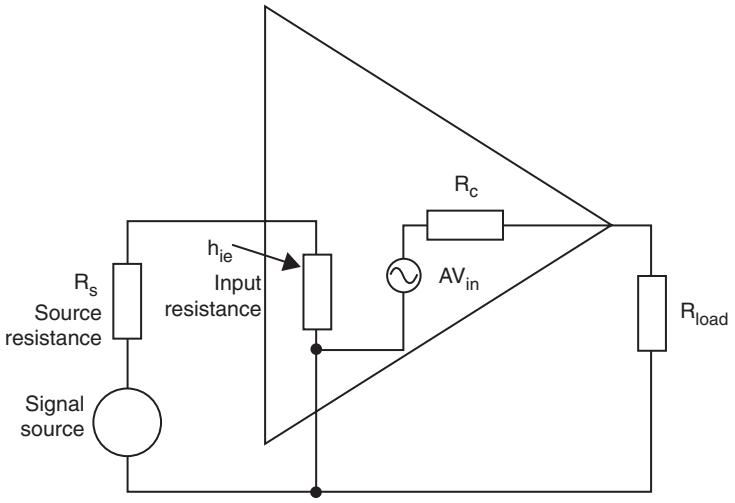


Figure 5.21

The voltage signal equivalent. The voltage gain A is reduced by the potential divider action of the networks at the input and output.

Other bipolar transistor types

The bipolar phototransistor uses the same construction as any normal BJT, but with a window that allows light to strike the base-emitter junction. If this transistor is operated with connections to collector and emitter,

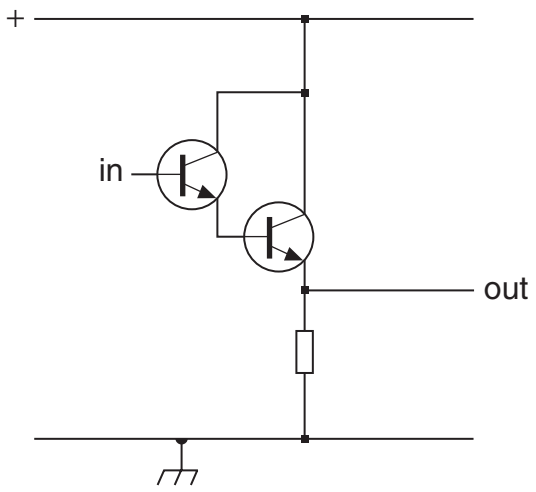
the collector current will be controlled by the amount of light striking the base-emitter junction. The advantage as compared to a photodiode is that transistor action greatly increases the current output for a given amount of light energy. The disadvantage is that the response is slower. Phototransistors are now used to a lesser extent as separate components because ICs are available that combine the phototransistor action with an operational amplifier.

- **DARLINGTON PAIR CIRCUIT**

A Darlington pair, eponymously named after the inventor, is a pair of transistors connected in common-collector mode, with the emitter of the first connected to the base of the second (Figure 5.22). The effect of this is to make the pair behave like a single transistor with current gain equal to $h_{fe1} \times h_{fe2}$, the product of the h_{fe} values of the two transistors. For example, if each transistor has a h_{fe} value of 500, the pair will provide an effective h_{fe} value of $500 \times 500 = 250\,000$. This circuit is used extensively when very high current gain is required, such as in skin resistance detectors, and in conjunction with photodiodes.

Figure 5.22

A typical Darlington pair circuit.



Field-effect transistors

The bipolar transistor relies for its action on making a reverse-biased junction conductive by injecting current carriers (electrons or holes) into it from

the other junction. The principles of the field-effect transistor (FET) are entirely different. In any type of FET, a strip of semiconductor material of one type (P or N) is made either more or less conductive because of the presence of an electric field pushing carriers into the semiconductor or pulling them away.

Field-effect transistors (FETs) are constructed with no junctions in the main current path between the drain and source electrodes, which correspond in function to collector and emitter respectively of a bipolar transistor. The path between these contacts, called the channel, may be P-type or N-type silicon, so FETs may be classed as P-channel or N-channel. Control of the current flowing in the channel is achieved by varying the *voltage* on a third electrode, the gate.

There are two types of field-effect transistor, the junction FET and the metal-oxide-silicon FET, or MOSFET. Both work by controlling the flow of current carriers in a narrow channel of silicon. The main difference between them lies in the method used to control the flow. In a junction FET (JFET or JUGFET), the gate is a contact to a junction formed on the channel and usually reverse biased. This type of FET is not common now. Figure 5.23 shows the construction for an N-channel JFET and the symbols for both N-channel and P-channel JFETs.

A tiny bar of silicon of either type has a junction formed near one end. Connections are made to each end of the bar, and also to the material at the junction; P-type in this example. The P-type connection is called the *gate*, the end of the bar nearest the gate is the *source*, and the other end of the bar is the *drain*. A junction FET of the type illustrated is normally used with the junction reverse-biased, so that few moving carriers are present in the neighbourhood of the junction. This way of using a JFET is also termed *depletion mode*. This, therefore, is an N-channel depletion mode JFET.

The junction, however, forms part of the silicon bar, so if there are few carriers present around the junction, the bar itself will be a poor conductor. With less reverse bias on the junction, a few more carriers will enter the junction and the silicon bar will conduct better; and so on as the amount of reverse bias on the junction decreases.

When the voltage is connected between the source and the drain therefore, the amount of current flowing between them depends on the amount of

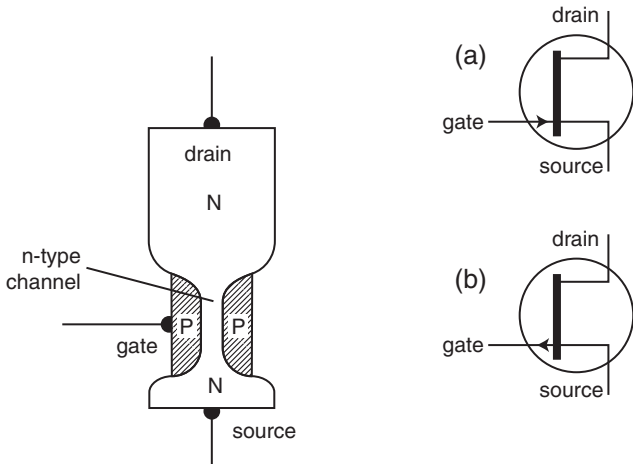


Figure 5.23

Structure of a JFET and symbols **(a)** N-channel, **(b)** P-channel.

reverse bias on the gate; and the ratio is, as for a BJT, the mutual conductance, whose symbol is g_m. This quantity, g_m, is a measure of the effectiveness of the FET as an amplifier of current flow.

For most FETs, gm values are very low, only about 1.2 to 3 mA/V, as compared with corresponding values for a bipolar transistor of from 40 mA/V (at 1 mA current) to several amperes/volt at high levels of current flow. Because the gate is reverse-biased however, practically no gate current flows, so the resistance between gate and source is very much higher than the resistance between base and emitter of a working bipolar transistor.

The predominant type of FET nowadays, particularly for digital circuits, is the **MOSFET**. The drawing of Figure 5.24 shows the basic construction of the metal-oxide-silicon FET or MOSFET. A silicon layer, called the *substrate* or *base*, is used as a foundation on which the FET is constructed. The substrate may have a separate electrical connection, but it takes no part in the FET action and if a separate electrical connection is provided it is usually connected either to the source or the drain. Two regions which are both doped in the opposite polarity to the substrate are then laid on the substrate and joined by a thin channel. In the illustration the substrate

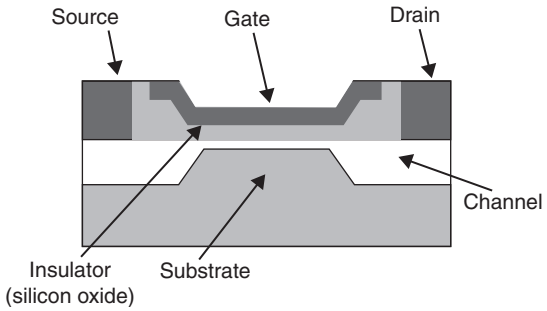


Figure 5.24

MOSFET structure.

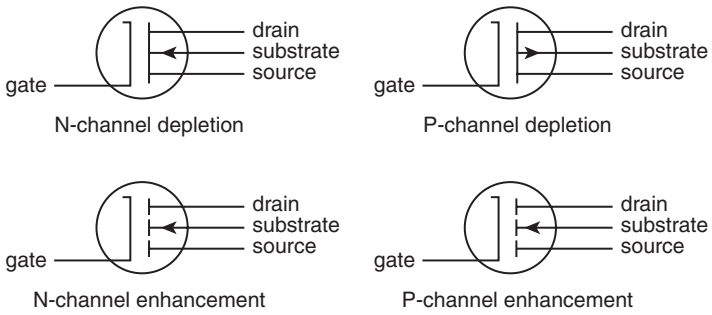


Figure 5.25

Symbols for **(a)** N-channel and **(b)** P-channel MOSFETs.

is of P-type silicon and the source, drain and channel are of N-type so that there is a conducting path between the N-type source and drain regions. Figure 5.25 shows symbols for N- and P-channel MOSFETs used in the two modes (see later) of enhancement or depletion.

The gate is insulated from the channel by a thin film of silicon oxide, obtained by oxidizing some of the silicon of the channel, and a metal film is deposited over this insulating layer to form the gate itself. A positive voltage applied to the gate has the effect of attracting more electrons into the

channel, and so increasing its conductivity. A negative potential so applied would repel electrons from the channel and so reduce its conductivity.

Both N-channel and P-channel devices can be made. In addition, the channel can be either doped or undoped (or very lightly doped). If the channel is strongly doped there will be a conducting path of fairly low resistance between the source and the drain when no bias is applied to the gate. Such a device is usually operated with a bias on the gate that will reduce the source-drain current, and is said to be used in **depletion mode**. When the channel is formed from lightly-doped or undoped material it is normally non-conducting, and its conductivity is increased by applying bias to the gate in the correct polarity, using the FET in **enhancement mode**.

Enhancement mode is more common. With the gate-to-source voltage equal to zero, the device is cut off. When a gate voltage that is positive with respect to the channel is applied, an electric field is set up that attracts electrons towards the oxide layer. These now form an induced channel to support a current flow. An increase in this positive gate voltage will cause the drain-to-source current flow to rise.

- **FET HANDLING PROBLEMS**

Junction FETs cause few handling problems provided that the maximum rated voltages and currents are not exceeded. MOSFETs, on the other hand, need to be handled with great care because the gate must be completely insulated from the other two electrodes by the thin film of silicon oxide. This insulation will break down at a voltage of 20 V to 100 V, depending on the thickness of the oxide film. When it does break down, the transistor is destroyed.

Any insulating material which has rubbed against another material can carry voltages of many thousands of volts; and lesser electrostatic voltages are often present on human fingers. There is also the danger of induced voltages from the AC mains supply. Voltages of this type cause no damage to bipolar transistors or junction FETs because these devices have enough leakage resistance to discharge the voltage harmlessly. The high resistance of the MOSFET gate, however, ensures that electrostatic voltages cannot be discharged in this way, so damage to the gate of a MOSFET is always possible.

To avoid such damage, all MOS gates that are connected to external pins are protected by diodes which are created as part of the FET during manufacture and which have a relatively low reverse breakdown voltage. These protecting diodes will conduct if a voltage at a gate terminal becomes too high or too low compared to the source or drain voltage level, so avoiding breakdown of the insulation of the gate by electrostatic effects.

The use of protective diodes makes the risk of electrostatic damage very slight for modern MOS devices, and there is never any risk of damage to a gate that is connected through a resistor to a source or drain unless excessive DC or signal voltages are applied. Nevertheless, it is advisable to take precautions against electrostatic damage, particularly in dry conditions and in places where artificial fibres and plastics are used extensively. These precautions are:

- Always keep new MOSFETs with conductive plastic foam wrapped round their leads until after they have been soldered in place.
- Always short the leads of a MOSFET together before unsoldering it.
- Never touch MOSFET leads with your fingers.
- Never plug a MOSFET into a holder when the circuit is switched on.

By altering the geometrical shape of a FET, power output FETs of the VFET type can be constructed. The 'V' (of VFET) in this case means 'vertical', describing the construction which is arranged so that the drain can be large and easily put into contact with a heatsink. Matched complementary pairs of VFETs have been used to a considerable extent as the power output stage in high-quality audio amplifiers. The input resistance of either type of FET is high, almost infinite for the MOS type, and low noise levels can be achieved even when using high source resistance values of the order of 1 MΩ.

Negative feedback

Feedback means using a fraction of the output voltage of a circuit to add to the input. When the signals at the input and the output are oppositely

phased (the output is a mirror image of the input), the feedback signal is said to be negative. **Negative feedback** has the effect of subtracting the fed-back signal from the input signal so that it reduces the overall gain of the amplifier. The effect on the gain is as follows:

Let A_o = gain of amplifier with no feedback (also known as the **open-loop gain**)

β = feedback fraction (or loop gain), so that V_{out}/β is fed back

Then the gain of the amplifier when negative feedback is applied is:

$$\frac{A}{1 + A/\beta} = \text{the closed-loop gain}$$

For example, if the open-loop gain is 100 and $\beta = 20$ (so 1/20 of the output voltage is fed back in opposite phase), the closed-loop gain is:

$$\frac{100}{1 + 100/20} = \frac{100}{6} = 16.7$$

A very useful approximation is that if the open-loop gain A_o is very much larger than the feedback fraction (loop gain), the closed-loop gain is simply equal to β. This is because A/β is large, much larger than unity, so the 1 in the equation can be neglected. This makes the expression become:

$$A/(A/\beta) = \beta$$

Negative feedback, in addition to reducing gain, also reduces noise signals that originate in the components of the amplifier if these components are within the feedback loop. It will also reduce distortion provided that the distortion does not cause a serious loss of open-loop gain such as might be caused by an excessive voltage swing.

Input and output resistance values are also affected. If the feedback signal shunts the input (Figure 5.26), input resistance is reduced, often to such an extent that the input terminal is practically at earth potential for signals (a **virtual earth**). If the feedback is in series with the input signal (Figure 5.27), the input resistance of the amplifier is increased, often very considerably. When the feedback network is driven by the voltage signal

Figure 5.26

A feedback circuit in which the feedback signal is in shunt with the input signal. At the output, the feedback resistor is connected in series with the output load.

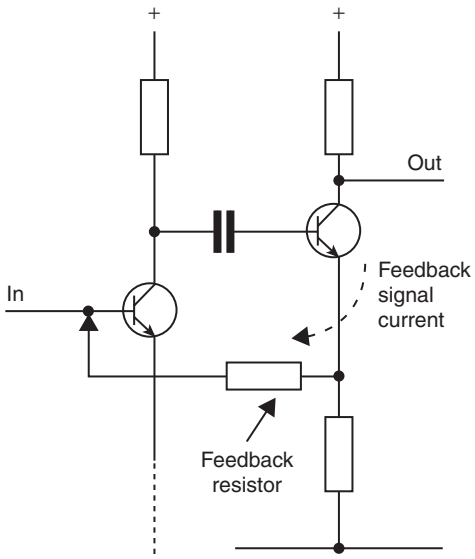
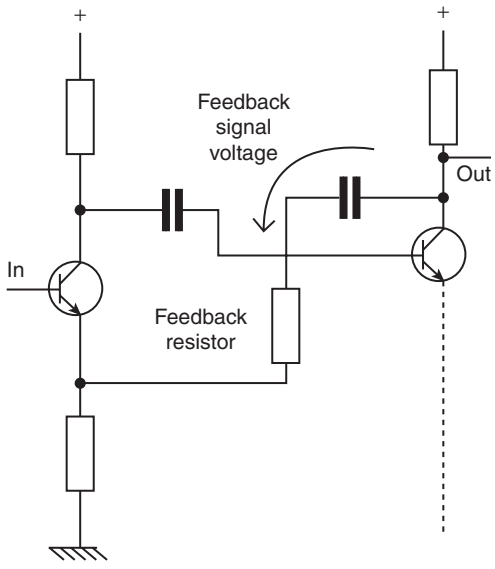


Figure 5.27

A feedback circuit in which the feedback signal is in series with the input signal through the emitter junction of the first transistor. The feedback resistor is also connected in parallel with the output load.



at the output (Figure 5.26), the effect is to reduce output resistance, and when the feedback is driven by the current at the output stage the effect is to increase the output resistance. The effects on output resistance are generally small compared to the effects on input resistance.

Negative feedback has to be applied with caution in circuits that contain time constants, because time constants will inevitably cause phase shift. The effect of a set of time constants, due to coupling components and stray capacitances, can, at some frequency cause a phase shift approaching 180°. This will have the effect of making the feedback signal positive instead of negative, causing instability.

The precise details of designing negative feedback amplifiers that are unconditionally stable are beyond the scope of this book, but in general the requirement is that the feedback should never become positive while the amplifier has enough gain to oscillate. In design practice this condition is examined by drawing Bode plots (Figure 5.28a), which are graphs of the magnitude and phase of amplifier gain. These indicate that a system (which need not be an electronic system) will be stable if the gain margin (gain at

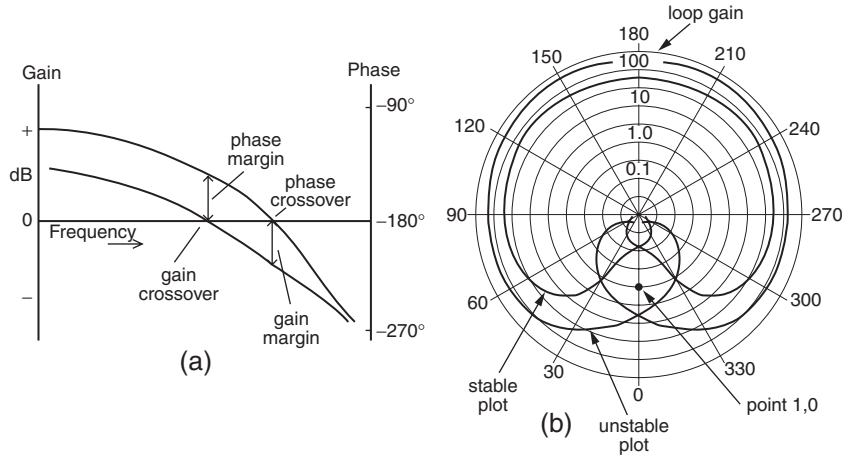


Figure 5.28

(a) A Bode plot of amplitude and phase, **(b)** two superimposed Nyquist plots showing stable and unstable designs.

180° phase) and phase margin (phase at zero gain) are both positive. From these, it is possible to construct a Nyquist diagram (Figure 5.28b) that will more certainly indicate stability. The way the diagram is constructed, a plot that lies outside the point –1,0 indicates instability, but one that lies within this point indicated stability, and one that passes through the –1,0 point will be only marginally stable. Software is available for calculating and displaying Bode plots and Nyquist diagrams. In many examples it is only necessary to ensure that the amplifier has very low gain at a frequency that would cause phase reversal.

Heatsinks

A transistor passing a steady (or average) current I amps and with a steady or average value of voltage V volts between the collector and the emitter will dissipate a power of VI watts. This electrical power is converted to heat at the collector-base junction (where most of the resistance is situated) and unless this heat can be removed the temperature of the junction will increase until the junction fails irreversibly. Heat is removed in two stages, by conduction between the collector junction and the casing of the transistor, and into metal heatsinks if fitted, and then by convection into the air. The temperature of the junction will become stabilized when the rate of removing heat, measured in watts, is exactly equal to the electrical power dissipation – but this may happen only when the temperature of the junction is too high for reliable, continuous operation. The power dissipation of a power transistor is limited therefore mainly by the rate at which heat can be removed.

For practical purposes, the resistance to heat transfer is measured by the quantity called **thermal resistance**, symbol θ, whose units are °C/W. The same measuring units are also used for convection, so all the figures for thermal resistance from the collector-base junction to the air can be added together, as for resistor values in series. The temperature difference between the junction and the air surrounding the heatsink is then found by multiplying the total thermal resistance by the number of electrical watts dissipated, so that:

$$T^{o} = \theta \times W \quad \text{where W is electrical power in watts.}$$

This latter figure of T° is a temperature **difference**, the difference between air temperature (also called **ambient temperature**) and the junction temperature. To find the junction temperature, the temperature of the surrounding air must be added to the figure for T. An ambient temperature figure of 30°C for domestic equipment and 70°C for industrial equipment can be used for estimates. If this procedure results in a value for junction temperature that is higher than the manufacturer's rated values (120°C to 200°C for silicon transistors), or too close for comfort, then the dissipated power must be reduced, a large heatsink used, or a water-cooled heatsink used. Large power transistors are designed so that the transfer of heat from the junction to the casing is efficient, with a low value of thermal resistance, and the largest value of thermal resistance in the heat circuit is that of the heatsink to the air. Small transistors generally have much higher internal thermal resistance values, so heatsinking is less effective.

To ensure low thermal resistance for power transistors, the collector of such transistors is usually connected directly to the metal case or to a metal tab. To prevent unintentional short-circuits, the heatsink may have to be insulated from other metalwork, or the transistors insulated from the heatsink by using thin mica washers. Such washers used along with silicone heatsink grease can have thermal resistance values of less than 1°C/W and are available from transistor manufacturers or components stockists. The use of mica washers makes it possible to use a metal chassis as a heatsink, or to mount several transistors on the same heatsink. The calculation of thermal resistance values for heatsinks is not simple, but for a single metal fin of length L and width D, an approximate formula for thermal resistance is:

$$\theta = \frac{250}{L \times D} \quad \text{with L and D in centimetres, and } \theta \text{ in °C/watt.}$$

Finned heatsinks bought from component suppliers will have been measured, so an average value can be quoted. The measurement of thermal resistance can be carried out by bolting a 25 W wire-wound resistor of the metal-cased type to the heatsink. A value of around 2.2 Ω is suitable, dissipating 4 W at 3 V and 16.4 W at 6 V. The temperature of the heatsink surface is measured when conditions have stabilized (no variation in temperature in 5 minutes) and the electrical power divided by the temperature difference between the heatsink and the ambient temperature gives the

thermal resistance. This method is not precise but it provides values that are well suited for practical work.

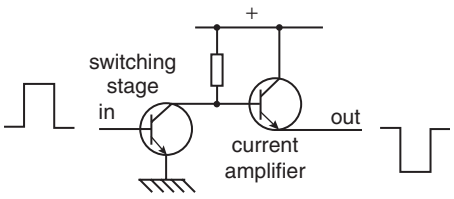
Switching circuits

A linear amplifier circuit creates an 'enlarged' copy of a waveform. By contrast, the output of a pulse (or logic) switching circuit changes rapidly from one value of voltage of current to another in response to a small change at the input. The output waveform need not be similar in shape to the input waveform, but the change in voltage or current should take place with only a small time delay (measured in nanoseconds) after the change at the input.

The BJT has a good switching action because of its large g_m figure. A useful rule of thumb is that the collector current of a transistor will be changed tenfold by each 60 mV change of voltage at the base, provided that neither cut-off nor saturation occurs. Current switching can easily be implemented and a stage of current amplification can be added if larger current swings are needed (Figure 5.29); voltage amplification can also be added if required. Special transistors that incorporate base and emitter resistors are available. These are termed resistor-equipped transistors (**RET**) or **digital transistors**. One common use is in driving the keypad LEDs of mobile phones.

Figure 5.29

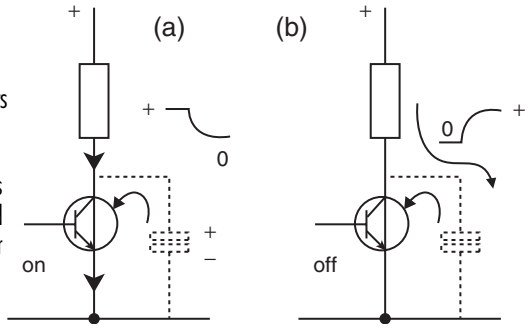
Adding a current-amplifying stage to a simple switching transistor.



A voltage switching stage must use some form of load to convert the current changes at the collector into voltage changes. If this load is a resistor, the switch-on will be faster than the switch-off, because of stray capacitances. At switch-on the stray capacitances are discharged rapidly by the current flowing through the transistor, but when the transistor switches off the capacitances must charge through the resistor, following the usual exponential CR pattern (Figure 5.30).

Figure 5.30

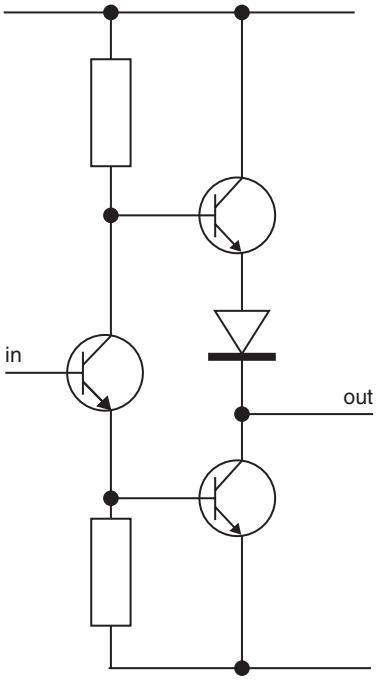
Charging and discharging stray capacitances. When the transistor conducts **(a)** the stray capacitance is rapidly discharged, and the voltage drop at the collector is sharp. When the transistor cuts off **(b)**, the stray capacitance is recharged through the load resistor, causing a slower voltage rise.



If the rise time of the wave does not need to be short the problem of charging can be dealt with by using a low-value resistor of 1 kΩ or less. A better alternative is to use series-connected transistors (Figure 5.31), switching positively in each direction. This type of circuit is extensively used within switching ICs.

Figure 5.31

Using a two-transistor output circuit so that the switching is equally rapid in both directions. This type of output stage is used in digital ICs, either in BJT or MOS format.



For fast switching applications, the **stored charge** of transistors becomes significant. During the time when a BJT is conducting the emitter is injecting charges (holes or electrons) into the base region; a MOSFET is forcing carriers into its channel. These charges cannot disappear instantly when the bias is reversed so that the transistor will conduct momentarily in the reverse direction. As a result, the circuit of Figure 5.31 can suffer from excessive dissipation at high switching speeds because for short intervals both transistors will be conducting. Manufacturers of switching transistors at one time quoted figures of stored charge Q in units of picocoulombs (pC), but nowadays often quote the more useful turn-on and turn-off times in nanoseconds (ns) under specified conditions. Stored charge figures are useful if you need to know the amount of current that will be needed to charge or discharge the base or gate capacitance in a given switching time. An approximate value for turn-off time can be obtained from the equation:

$$t = \frac{Q}{I}$$

where t is the turn-off time in nanoseconds (10^{-9} seconds), Q is stored charge in pC, and I is the current in mA that is to be switched off. This can be rewritten as:

$$I = \frac{Q}{t}$$

to calculate current drive, so that the charge or discharge 10 nC (10 000 pC) in 5 ns will need 2000 mA, 2 A.

BJT switch-off times are improved by reverse biasing the base but some care has to be taken not to exceed the reverse voltage limits since the base-emitter junction will break down at moderate values of reverse voltage, sometimes as low as –5 V.

A considerable improvement in switch-off times is also obtained if the BJT is not allowed to saturate during its switch-on period. This has to be done by clamping the base voltage, and is not easy because of the considerable variation of switch-on voltage between one transistor and another. The fastest switching times are achieved by current-switching circuits in which the transistor is never saturated or cut off. A circuit called the 'Baker clamp' is often used, and the usual alternative is to use Schottky diodes. Typical circuits are illustrated in Figure 5.32.

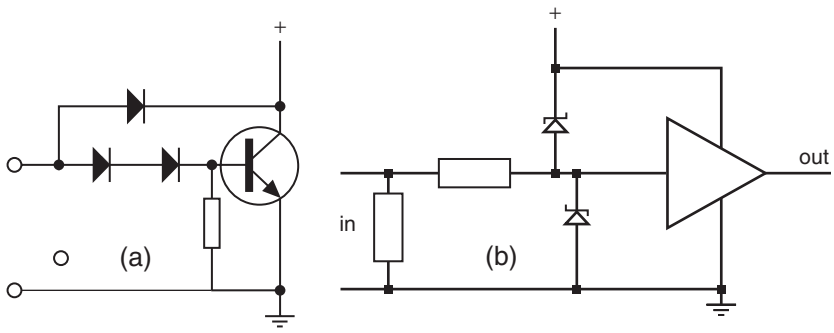


Figure 5.32

(a) a simple Baker clamp circuit; **(b)** using Schottky diodes for clamping at the input to a switching stage.

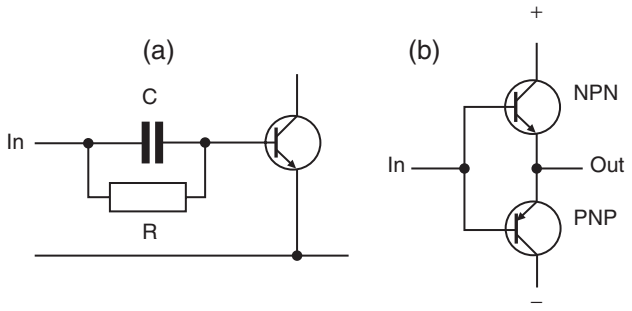


Figure 5.33

Two common switching circuit tricks: **(a)** use of a base-compensation capacitor, **(b)** using a complementary emitter-follower output circuit with no load resistor.

Some circuits commonly used for switching circuits are shown in Figure 5.33, where circuit (a) shows the use of a time constant RC in series with the base of the transistor. The value of C is adjusted for the best shape of leading and trailing edges for a square pulse input. Figure 5.33b shows the familiar complementary double-emitter-follower circuit which uses transistors both to charge and to discharge stray capacitances.

Gating of analogue signals is an action similar to that of pulse or logic switching, but the switch may be a series component rather than a shunt component, with the added restriction that it should not distort the analogue signal while in the ON state. Diodes, bipolar transistors and FETs have all in the past been used in bridge gating circuits. Figure 5.34 shows the outline of a circuit using complementary MOS devices. When a switching voltage pulse is applied to the control terminal, the MOS switching transistors are fully conducting so that there is a current path in either direction between the input and the output. The circuit illustrated here is normally one of a set of four units in a single IC such as the CD4016. The disadvantage of the FET is that its resistance when switched ON is much higher (up to 1 kΩ) than that of a bipolar transistor.

Another very common gating circuit is the long-tailed pair shown in Figure 5.35 which is, however, useful only when the offset voltages (V_{be}) and the voltage change caused by switching are both unimportant.

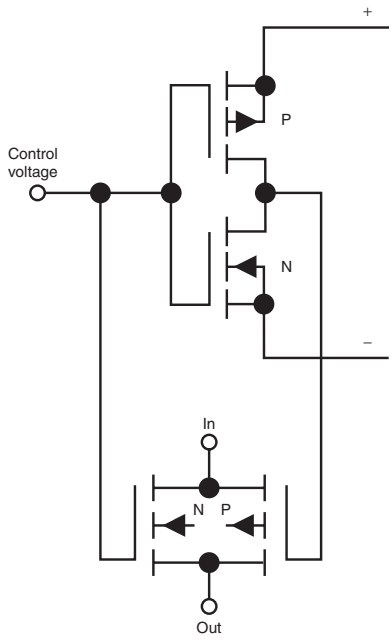


Figure 5.34

A bilateral MOS switching circuit.

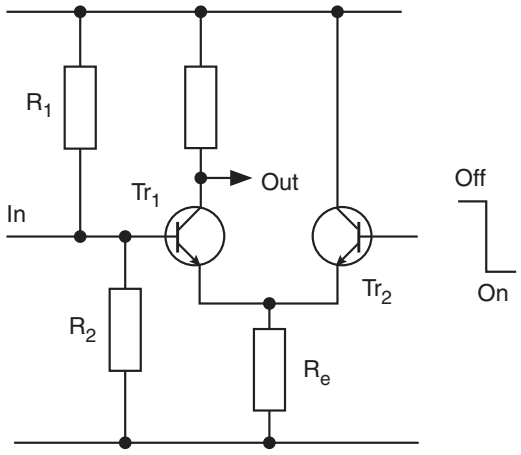


Figure 5.35

The long-tailed pair gate. When Tr_2 is switched off Tr_1 is normally biased by R_1, R_2, and acts as an inverting amplifier. When Tr_2 is switched on, with its base voltage several volts higher than the normal bias voltage of Tr_1, Tr_1 is biased off.

Other switching devices

Unijunction transistors have two base contacts and an emitter contact, forming a device with a single junction which does not conduct until the voltage between the emitter and base contact 1 (Figure 5.36a) reaches a specified level. At this level, the whole device becomes conductive. The unijunction is used to generate short pulses, using circuits such as that shown in Figure 5.36b. The frequency of operation of this circuit is not noticeably affected by changes in the supply voltage because the point at which the unijunction fires (becomes conductive) is a constant fraction of the supply voltage, determined at the time of manufacturing by the position of the emitter junction.

The **intrinsic standoff ratio**, n, for a unijunction is defined as:

$$\frac{\text{firing voltage } (e - b_1)}{\text{supply voltage } (b_2 - b_1)}$$

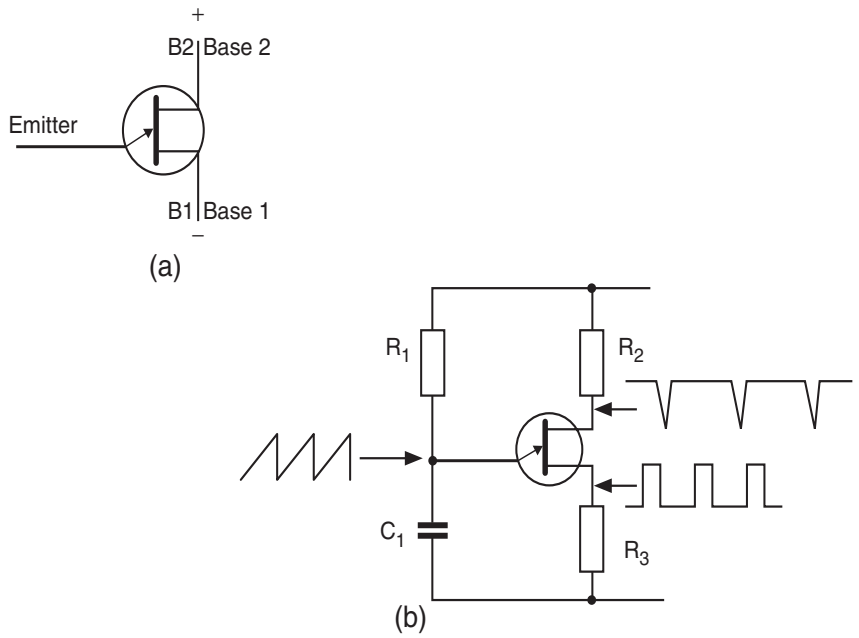


Figure 5.36

Unijunction symbol **(a)** and typical oscillator circuit **(b)**. R_2, R_3 are about 100 ohms each, and the frequency of oscillation is determined by the time constant R_1C_1.

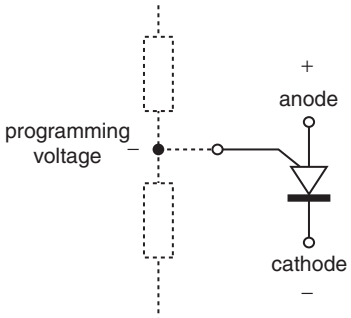
and has values ranging typically from 0.5 to 0.86. Pulse repetition rates up to 1 MHz are obtainable.

Programmable unijunction transistors (**PUT**s) have three terminals, one of which is used to set the value of intrinsic standoff ratio, n, by its connection to a potential divider (Figure 5.37). Firing will occur at the programmed voltage, with a frequency range up to 10 kHz for typical devices.

Thyristors (also called **silicon controlled rectifiers** or **SCR**s) are controlled silicon diodes which are non-conductive in the reverse direction, and do not conduct in the forward direction until they are triggered by a brief pulse or a steady voltage applied between the gate and the cathode terminals.

Figure 5.37

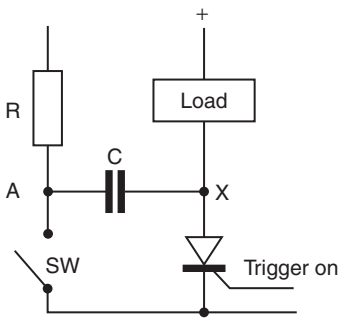
A programmable unijunction transistor (PUT). The firing voltage between anode and cathode is selected by the voltage applied to the third electrode.



A voltage of 0.8 V to 1.5 V and currents ranging from a few μA up to as high as 30 mA are needed at the gate, depending on the current rating of the thyristor. The thyristor ceases to conduct only when the voltage between the anode and the cathode falls to a low value (about 0.2 V), or when the current between the anode and the cathode becomes very low (typically 1 mA or less). DC switching circuits need some form of capacitor discharge circuit (Figure 5.38) to switch off the load.

Figure 5.38

A capacitor turn-off circuit for a thyristor. When the circuit is momentarily closed, the sudden voltage drop at A will cause an equal drop at X, turning off the thyristor until it is triggered again.



AC thyristor switching circuits, using raw AC or full-wave rectified waveforms, are switched off by the waveform itself as it passes through zero on each cycle. A few typical thyristor circuits are shown in Figure 5.39. Note that the gate signal may have to be applied through a pulse transformer, particularly when the thyristor is used to switch mains currents, to avoid connecting the firing circuits to the gate. Triacs are two-way thyristors whose terminals are labelled MT1, MT2 and Gate – examples of circuits are shown in Figure 5.40.

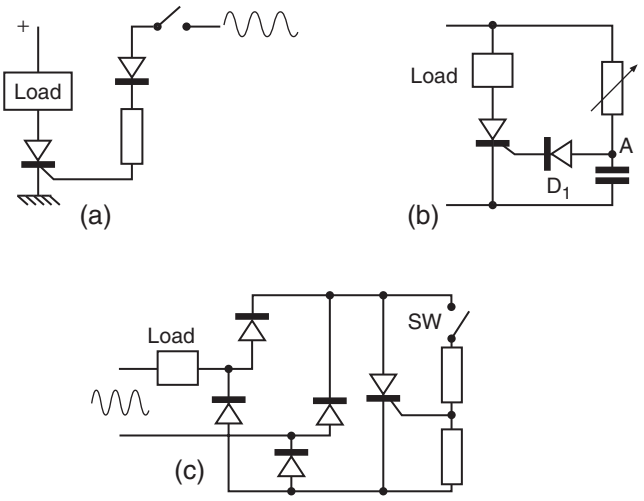


Figure 5.39

AC thyristor circuits: **(a)** Basic half-wave AC relay circuit. **(b)** Basic phase-control circuit. In the half-cycle during which the thyristor can conduct, the gate is activated only when the voltage at A has risen enough to cause the trigger diode D_1, to conduct. The time in the cycle at which conduction starts is controlled by the setting of the variable resistor. **(c)** A full-wave control circuit.

For reliable firing, the pulse at the triac gate should be of the same polarity as MT2. Firing pulses for thyristors and triacs can be obtained from unijunctions or from other forms of trigger devices such as diacs, silicon bidirectional switches, four-layer diodes or silicon unidirectional switches. The **diac**, or bidirectional trigger diode, is non-conductive in either direction until its breakdown voltage is exceeded, after which the device conducts readily until the voltage across its terminals (in either direction) is low. Firing voltages of 20 to 36 V are typical, and the 'breakback' voltage at which the device ceases to conduct is typically 6 V. Brief peak currents of 2 A are possible. The **silicon bidirectional switch** also uses a gate electrode, but operates with one polarity only. **Four-layer diodes** have lower firing and breakback voltages than the other diodes, but essentially similar characteristics.

The **silicon controlled switch** (**SCS**) is a useful device with four electrodes which can be used, according to connections, either as a programmable unijunction or as a low-power thyristor. The connections are referred to

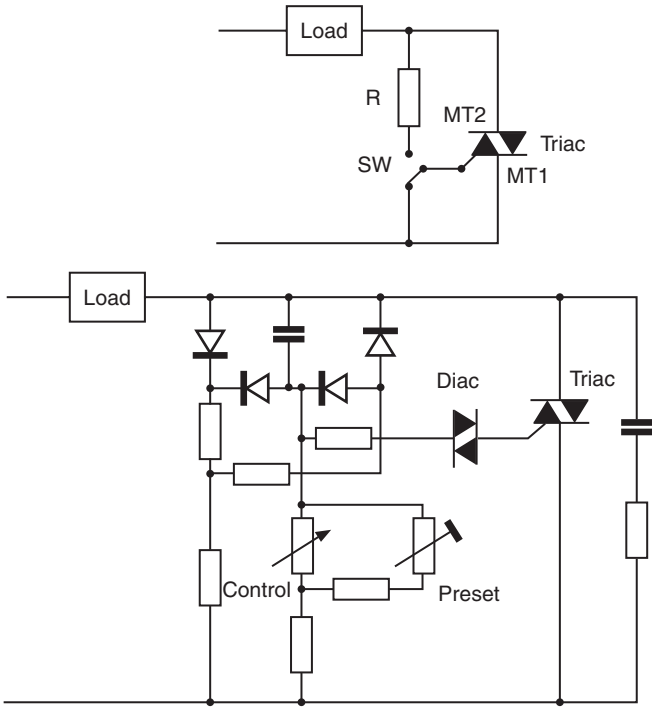


Figure 5.40

Triac circuits: **(a)** basic full-wave relay circuit, **(b)** power regulator circuit, using a diac trigger diode, and radio interference suppression circuit across the triac.

as anode, cathode, gate-anode and gate-cathode. If the gate-cathode is used along with the anode and cathode, low-current thyristor operation is obtained. If the gate-anode is used the device behaves as a PUT. The unused electrode is generally left open-circuited.

The **opto-isolator** or **opto-coupler** is, at its simplest, a combination of an LED and a photo-transistor in a single package arranged so that only the light from the LED can affect the photo-transistor (Figure 5.41). The electrical isolation of the two parts of the device can be almost complete, so the main application is in transferring signals across circuits that have large voltage differences or which must be kept separate. One example is the use in a modem to ensure that the computer is totally isolated from voltage changes on the telephone line and vice versa.

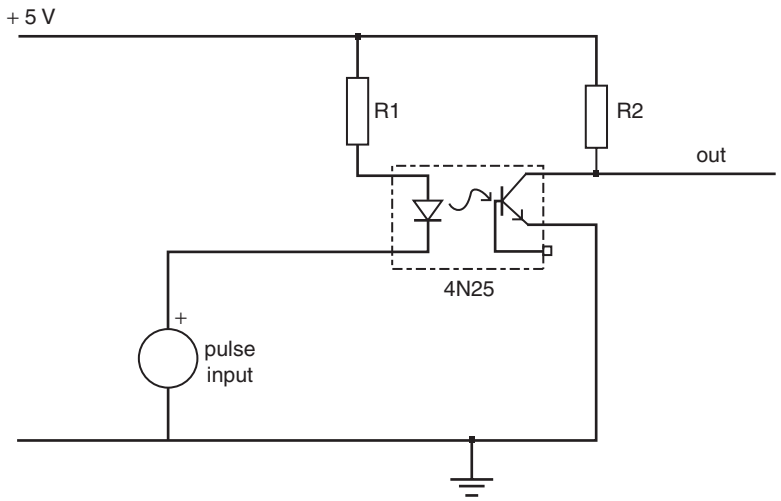


Figure 5.41

Symbol for an opto-isolator or opto-coupler.

The **opto-triac** is a development of the opto-isolator that allows a triac to be gated by signals that are electrically isolated; a typical application is in flashing light shows wherein the lights are triggerd by audio signals. The use of an opto-triac ensures that the low-voltage audio signals are totally isolated from the higher voltages used for the lights.

Diode and transistor coding

In Table 5.3 is shown the European Pro-Electron coding used for semiconductor type numbers. The US JEDEC 1N, 2N and 3N numbers are registration numbers only, and therefore the function of a semiconductor cannot be determined from these type numbers. The Pro-Electron system provided more information but, like so many good ideas, has been overshadowed by the less useful but more prevalent JEDEC system.

The Japanese system also uses registration numbers, but the lettering denotes the purpose of the device, so the coding conveys more information than do the JEDEC numbers. The codings currently used are shown in Table 5.4.

Table 5.3 Pro-Electron coding

The first letter indicates the semiconductor material used:

A Germanium
B Silicon
C Gallium arsenide and similar compounds
D Indium antimonide and similar compounds
R Cadmium sulphide and similar compounds

The second letter indicates the application of the device:

A Detector diode, high speed diode, mixer diode
B Variable capacitance (varicap) diode
C AF (not power) transistor
D AF power transistor
E Tunnel diode
F RF (not power) transistor
G Miscellaneous
L RF power transistor
N Photocoupler
P Radiation detector (photodiode, phototransistor, etc.)
Q Radiation generator
R Control and switching device (such as a thyristor)
S Switching transistor, low power
T Control and switching device (such as a triac)
U Switching transistor, high power
X Multiplier diode (varactor or step diode)
Y Rectifier, booster or efficiency diode
Z Voltage reference (Zener), regulator or transient suppressor diode.

The remainder of the code is a serial number. For consumer applications, such as TV and hi-fi, this has three figures. For industrial and telecommunications use W, X, Y or Z along with two figures.

Table 5.4 Japanese transistor coding system

Code	Device type
2SA	PNP transistor
2SB	PNP Darlington
2SC	NPN transistor
2SD	NPN Darlington
2SJ	P-channel MOSFET or JFET
2SK	N-channel MOSFET or JFET
3SK	Dual-gate N-channel FETs

This page intentionally left blank

CHAPTER 6

LINEAR ICS

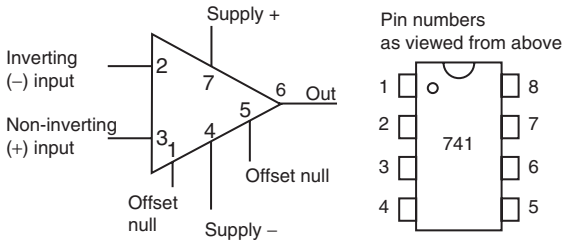
Overview

Linear ICs are single-chip arrangements of amplifier circuits that are intended to be biased and operated in a linear way. This definition is usually extended to include ICs that have a comparatively slow switching action controlled by an approximately linear charge and discharge of a capacitor, such as the 555 timer.

The most important class of linear amplifier IC is the operational amplifier (**op-amp**) which features high-gain, high-input resistance, low-output resistance and DC coupling internally. Such amplifiers whose typical pinout and symbol are illustrated in Figure 6.1 are almost invariably used in negative feedback circuits, and make use of a balanced form of internal circuit so that power supply hum and noise picked up by stray capacitance are both discriminated against.

Figure 6.1

The 741 operational amplifier outline, with pin numbering and the connections shown. The offset-null pins are used only for DC amplifier applications.



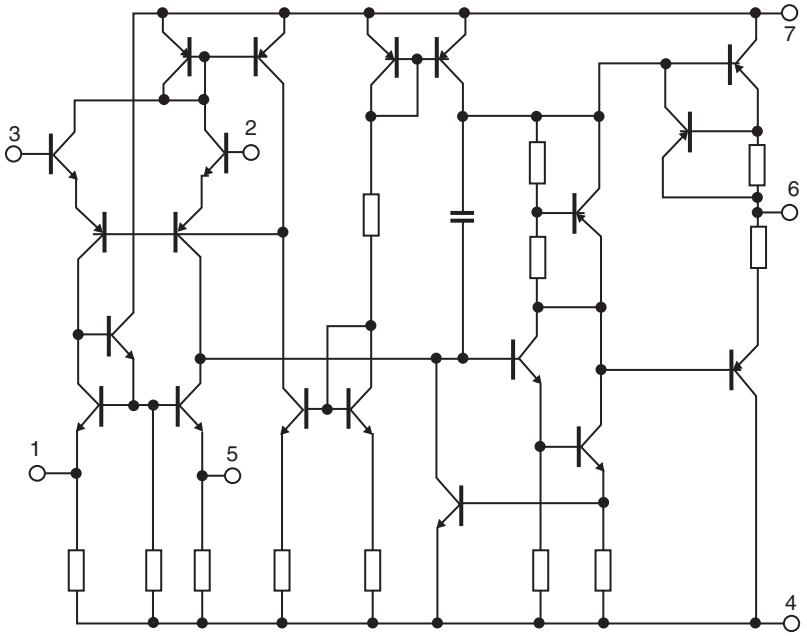


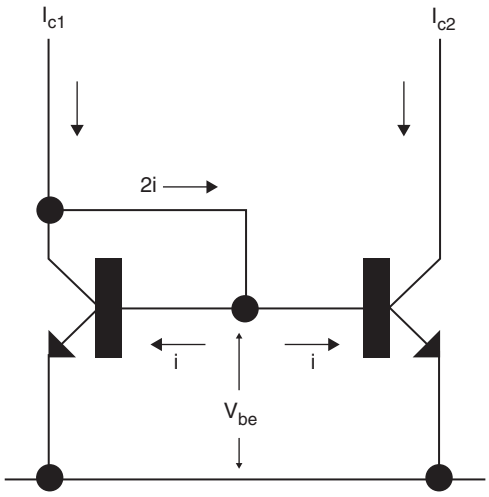
Figure 6.2

The internal circuitry of a 741 op-amp.

As an illustration of the internal circuit techniques that are used, Figure 6.2 shows the internal circuitry of a typical old model of linear IC, the 741 operational amplifier, which is still in production. The circuit is basically that of an elaborate balanced DC coupled amplifier using 20 transistors. One feature, very common in linear ICs, is the use of a current mirror as part of the circuit. The principle of a current mirror is that a current fed in at the input of the current mirror circuit will produce an identical value of current in the second. The circuit is used as a current source to endure that identical currents flow in the balanced amplifier circuits. Figure 6.3 shows a simple current mirror in which a current i flows into each base. This current is taken from the collector lead of one transistor, whose collector current is therefore ih_{fe}. The current in the other collector lead is ih_{fe} and, though these two are not identical, they are very close if h_{fe} has a typical value of around 500. More elaborate circuits can provide much closer matching of currents.

Figure 6.3

A simple current-mirror circuit as used in op-amp circuitry.



The 741 op-amp

The 741 is an old design, but it is still used in large volumes, and it is still typical of many operational amplifiers generally, so the design methods, circuits and bias arrangements which are used for this IC can be used, with small modifications, for other types. Referring to the pinout diagram/symbol of Figure 6.1, the 741 uses two inputs marked (+) and (−). These signs refer to the phase of the output signal relative to each input, so that feedback directly from the output to the (+) input is positive, and feedback directly from the output to the (−) input is negative. The important features of all operational amplifiers are summarized, below, with reference to the 741 as an example.

Gain and bandwidth

An ideal op-amp would have infinite gain and very high bandwidth, an unrealizable dream. Though the gain at DC can be very high (100 000 or more), this does not hold for frequencies significantly above DC. The 741, used as a typical example, has an open-loop gain that is constant only to about 6 Hz, and reduces at the rate of –6 dB/octave until the gain is

unity (0 dB). The frequency for unity gain is written as f_T. Figure 6.4 shows a typical gain–frequency graph.

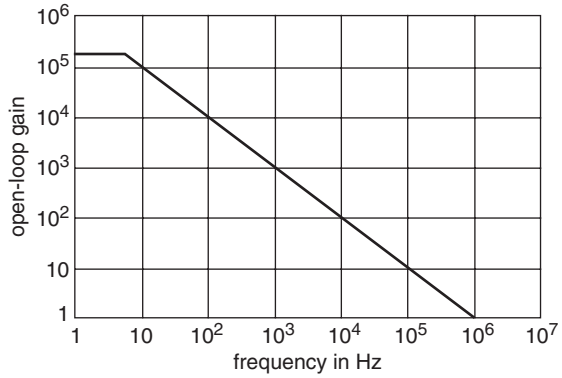


Figure 6.4

Gain vs frequency graph for the 741 op-amp. Note the logarithmic scales.

The graph indicates that the product of gain and bandwidth is constant, with units of Hz or MHz, and for the 741 is around 1 MHz. The closed-loop gain, when the op-amp is being used as a feedback amplifier, should be in the range 0.1 to 0.2 of the open-loop value at the maximum frequency for which the op-amp will be used. This assumes slowly changing signals. For step signals, the **slew rate** (transient response) limits the performance of an op-amp for such signals in a closed-loop circuit. Slew rate, taken as the time for the output to go from 10% to 90% of its final value for a pulse input, can be related to bandwidth by the expression:

$$SR = \frac{0.35}{B} \text{ where } B = \text{bandwidth}$$

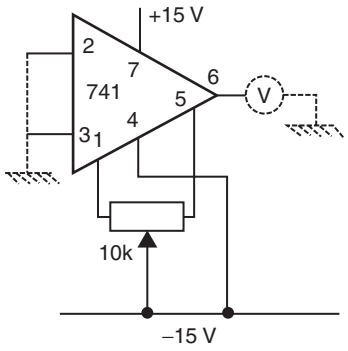
Offset

The circuit arrangement of the 741 is such that, using balanced power supplies, the DC level at the output ought to be at zero volts when both

inputs are connected to zero volts. This does not generally happen because of slight differences in internal components, so an input offset voltage is needed to restore the output to zero voltage. Alternatively, the offset can be balanced out by a potentiometer connected as shown in Figure 6.5. Once set in this way so that the output is at zero volts (with the inputs earthed), the output voltage will then slowly change (drift). The drift may be caused by temperature changes, by supply voltage changes, or simply by old age. Drift is a problem which mainly affects high-gain DC coupled amplifiers and long time-constant integrators; AC amplifier circuits and circuits which can use DC feedback bias are not affected by drift.

Figure 6.5

Using an offset-null control. With the inputs both earthed (balanced power supplies) and a voltmeter connected to the output (dotted lines), the 10 kΩ potentiometer is adjusted so that the output voltage is zero.



Bias methods

For linear amplification, both inputs must be biased to a voltage which lies approximately halfway between the supply voltages. The output voltage can then be set to the same value by:

1. making use of an offset-balancing potentiometer, or

2. connecting the output to the (−) input through a resistor, so making use of DC feedback.

Method (a) is very seldom used, and, since the use of DC feedback is closely tied up with the use of AC feedback, the two will be considered together. The power supply may be of the balanced type, such as the ±15V supply, or unbalanced, provided that the bias voltage of input and output

is set about midway between the limits (+15 and −15, or +V and 0) of supply voltages.

Bias voltages should not be set within three volts of supply voltage limits, so that when a +15 V supply is used, the input or output voltages should not exceed +12 V or −12 V. This limitation applies both to bias (steady) voltage and to instantaneous voltages. If a single-ended 24 V power supply is used, the input and output voltages should not fall below 3 V or rise above 21 V. Beyond these limits, the amplifying action may suddenly collapse because there is not sufficient bias internally.

Basic circuits

Figure 6.6 shows the circuits for an inverting amplifier, using either balanced or unbalanced power supplies. The DC bias conditions are set by connecting the (+) input to mid-voltage (which is earth voltage when balanced power supplies are used) and using 100% DC feedback from the output to the (−) input. The gain, G, is given by:

$$G = -\frac{R_1}{R_2} \text{ where the } (-) \text{ sign indicates inversion.}$$

Note that a capacitor C_1 is needed when a single-ended power supply is used to prevent the DC bias voltage from being divided down in the same ratio as the AC bias. When balanced power supplies are used, direct coupling is possible provided that the signal source is at zero DC volts.

The input resistance for these circuits is simply the value of resistor R_2, since the effect of the feedback is to make the input resistance at the (−) input almost zero; this point is referred to as a **virtual earth** for signals. The output resistance is typically about 150 ohms. Circuits for non-inverting amplifiers are shown in Figure 6.7. Non-inverting amplifiers also make use of negative feedback to stabilize the working conditions in the same way as the inverting amplifier circuits, but the signal input is now to the (+) input terminal, The gain is:

$$G = \frac{R_1 + R_2}{R_2}$$

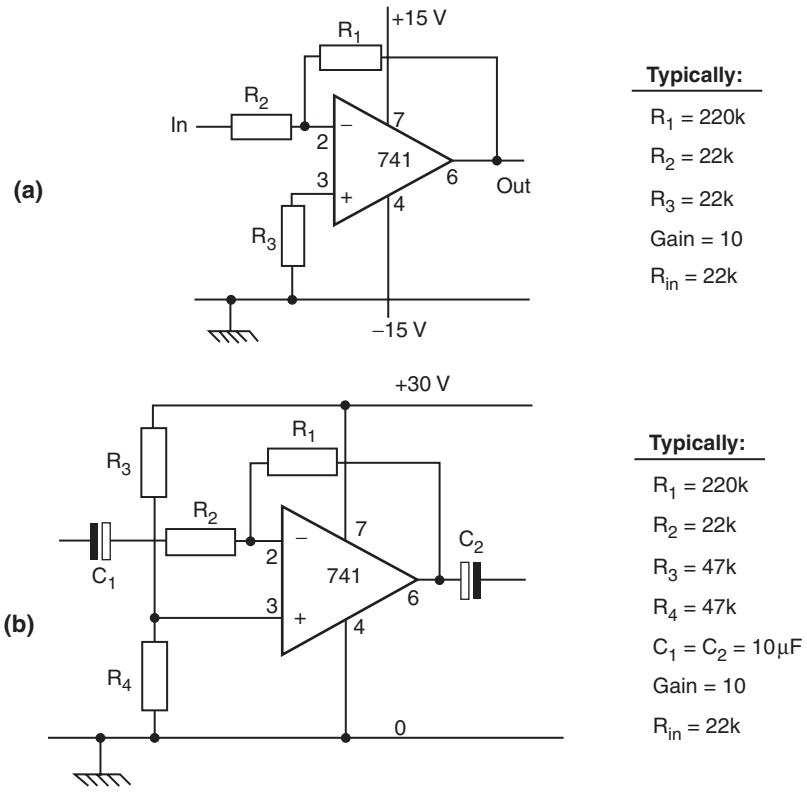


Figure 6.6

Inverting amplifier configuration. Balanced power supplies are used in **(a)**. Ideally, R_3 should equal R_2 though differing values are often used. The gain is set by the ratio R_1/R_2 and the input resistance is equal to R_2. Using an unbalanced power supply **(b)** the (+) input is biased to half the supply voltage (15 V in this example) by using equal values for R_3 and R_4. The gain is again given by R_1/R_2. Coupling capacitors are needed because of the DC bias conditions.

and the circuit is sometimes referred to as the **voltage-follower** with gain. The input resistance is high, usually around 1 MΩ, for the dual supply version, though the bias resistors reduce this to a few hundred kΩ.

Figure 6.8 shows the 741 used as a differential amplifier, though with a single-ended output. The gain is set by the ratio R_1/R_2 as before – note the use of identical resistors in the input circuits to preserve balance.

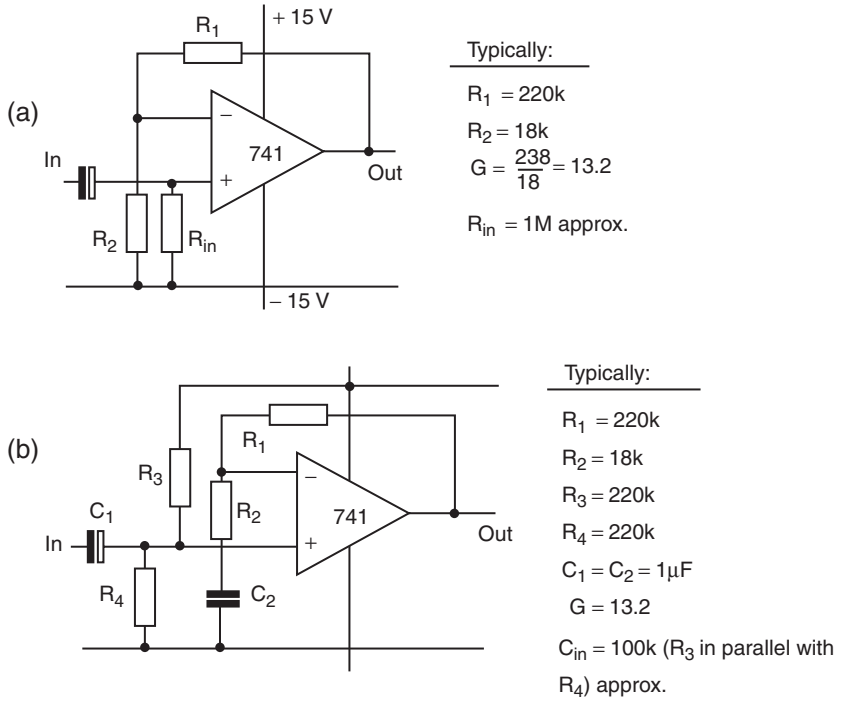


Figure 6.7

Non-inverting amplifiers. Using a balanced power supply **(a)**, only two resistors are needed, and the voltage gain is given by $(R_1 + R_2)/R_2$. The input resistance is very high. When an unbalanced supply **(b)** is used, a capacitor C_2 must be connected between R_2 and earth to ensure correct feedback of signal without disturbing bias. The input resistance is now lower because of R_3 and R_4 which, as far as the signal voltage is concerned, are in parallel.

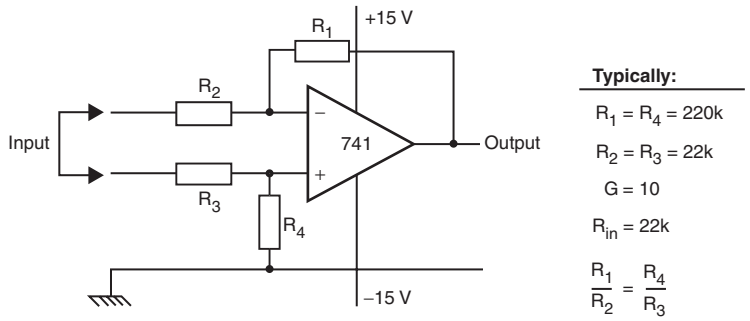


Figure 6.8

Differential amplifier application. Both inputs are used for signals which must be in antiphase (balanced about earth). Any common-mode signals (in phase at both inputs) are greatly attenuated.

General notes on op-amp circuits

The formulae for voltage gain hold for values of gain up to several hundred times, because the gain of the op-amp used in open-loop conditions (without feedback) is very high, of the order of 100 000 (100 dB). The maximum load current is about 10 mA, and the maximum power dissipation 400 mW. The 741 circuit is protected against damage from short circuits at the output, and the protection circuits will operate for as long as the short-circuit is maintained.

The frequency range of an op-amp depends on two factors, the **gain-bandwidth product** for small signals, and the **slew rate** for large signals. The gain-bandwidth product is the quantity, A × B, with A equal to voltage gain (not in dB) and B the bandwidth upper limit in Hz. For the 741, the GB factor is typically 1 MHz so, in theory, a bandwidth of 1 MHz can be obtained when the voltage gain is unity, a bandwidth of 100 kHz can be attained at a gain of 10, a bandwidth of 10 kHz at a gain of 100 times, and so on. This trade-off is usable only for small signals, and cannot necessarily be applied to all types of operational amplifiers.

Large-amplitude signals are further limited by the slew rate of the circuits within the amplifier. The slew rate of an amplifier is the maximum value of change of output voltage that can be achieved at unity gain. Units are usually volts per microsecond. Because this rate cannot be exceeded, and feedback has no effect on slew rate, the bandwidth of the op-amp for large signals, sometimes called the **power bandwidth**, is less than that for small signals. The slew rate limitation **cannot** be corrected by the use of negative feedback; in fact negative feedback acts to increase distortion when the slew rate limiting action starts, because the effect of the feedback is to increase the rate of change of voltage at the input of the amplifier whenever the rate is limited at the output. This accelerates the overloading of the amplifier, and can change what might be a temporary distortion into a longer-lasting overload condition. The relationship between the sine wave bandwidth and the slew rate, for many types of operational amplifier, is:

$$\text{maximum slew rate} = 2\pi E_{peak} f_{max}$$

where slew rate is in units of volts per second (**not** V/μs), f_{max} is the maximum full-power frequency in Hz, and E_{peak} is the peak voltage of

the output sine wave. This can be modified to use slew rate figures in the more usual units of V/μs, with the answer in MHz. For example, a slew rate of 1.5 V/μs corresponds to a maximum sine wave frequency (at 10 V output) of:

$$f_{max} = \frac{1.5}{2\pi \times 10} \text{ MHz } = 0.023 \text{ MHz or } 23 \text{ kHz}$$

Slew rate limiting arises because of internal stray capacitances which must be charged and discharged by the current flowing in the transistors inside the IC: improvement is obtainable only by redesigning the internal circuitry. The 741 has a slew rate of about 0.5 V/μs, corresponding to a low value of power bandwidth of about 6.6 kHz for 12 V peak sine wave signals. The slew rate limitation makes op-amps unsuitable for applications which require fast-rising pulses, so a 741 should not be used as a signal source or feed (interface) with digital circuitry, particularly TTL circuitry, unless a Schmitt trigger stage is also used. Higher slew rates are obtainable with more modern designs of op-amps; for example, the Fairchild LS201 achieves a slew rate of 10 V/μs.

Modern op-amps

The 741 serves as an example, despite its age, because it is still in use and because it is the prototype for most of the op-amp designs that have followed. Nevertheless, much better performance can be obtained by using more modern designs, and in Table 6.1 are summarized some

Table 6.1 Characteristics of four modern op-amps

Type	LT1077	LP324	LMC6084	LM6172
V_{supply}	+5 V	+5 V	+5 VV	15 VV
I_{supply}	50 μA	48 μA	75 μA	2.3 μA
CMRR	100 dB	90 dB	85 dB	110 dB
GB	230 kHz	1.8 MHz	1.3 MHz	100 MHz
Slew rate	0.05 V/μs	8 V/μs	1.5 V/μs	3000 V/μs

Notes: CMRR, common mode rejection ratio; GB, gain × bandwidth

of the more interesting characteristics of a selection of modern designs. These op-amps have been selected from the large range manufactured by National Semiconductor. In most cases these are packaged with four op-amps per package, and the supply current values given in the table apply to a single unit.

Other operational amplifier circuits

Figures 6.9 to 6.12 illustrate circuits other than the straightforward voltage amplifier types. Figure 6.9 shows two versions of a follower circuit with voltage gain, but with useful characteristics, subject to slew rate limitations.

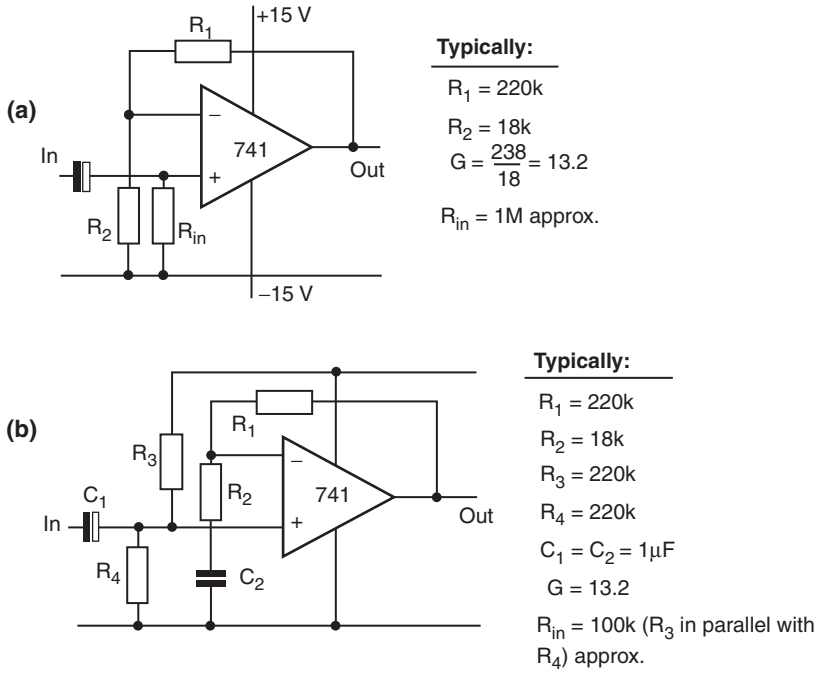


Figure 6.9

The voltage follower. The gain is determined by the values of R_1 and R_2, with high input resistance and low output resistance. The input resistance is determined by the values of resistors R_{in} or R_3 and R_4.

The non-inverting circuit, or voltage follower, performs the same action as the familiar emitter follower, having a very high input resistance and a low output resistance. For this type of circuit, the action of the feedback causes both inputs to change voltage together, as a common-mode signal would, so that any restrictions on the amplitude of common-mode signals (see the manufacturer's sheets) will apply to this circuit. If the resistor R_2 (Figure 6.9a) is omitted, the gain is unity, as for the cathode follower.

Figure 6.10 shows two examples of a 741 as it is used in a variety of 'shaping' circuits in which the gain/frequency or gain/amplitude graph is intended to be non-linear. The use of op-amps for switching circuits is limited by the slew rate, but the types of circuits shown in Figures 6.10 and 6.11 are useful if fast-rising or falling waveforms are not needed.

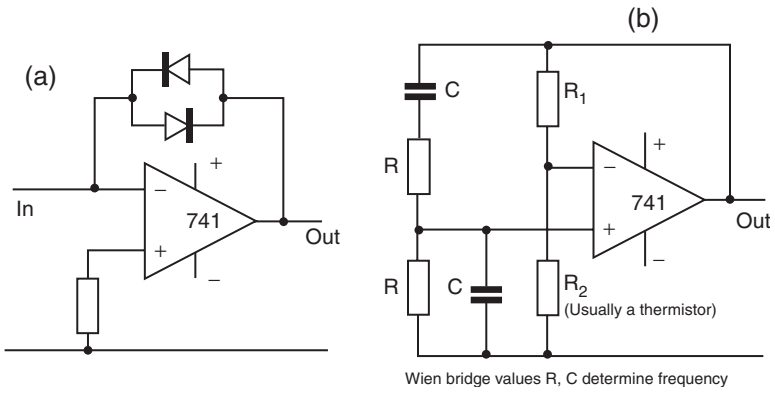
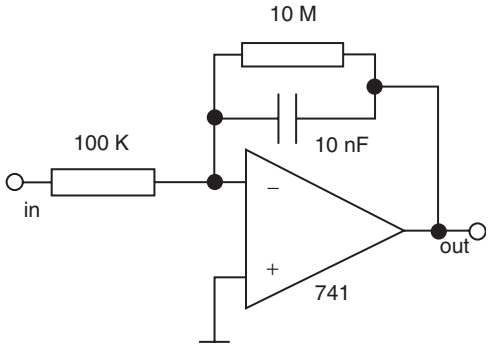


Figure 6.10

Using the 741 in circuits that are not linear amplifiers. **(a)** A limiting amplifier. Because the diodes will permit feedback of voltages whose amplitude is enough to allow the diodes to conduct, the output voltage is limited to about this amplitude but without excessive clipping. The gain is very large for small input signals and very small for large input signals. **(b)** The Wien bridge in the feedback network causes oscillation. The waveform is a sine wave only if the gain is carefully controlled by making $R_1/R_2 = 3$, and this is done usually by making R_2 a thermistor whose resistance value decreases as the voltage across it increases. The frequency of oscillation is given by $f = 1/(2\pi RC)$.

Figure 6.11

A 741 used as a simple integrator.



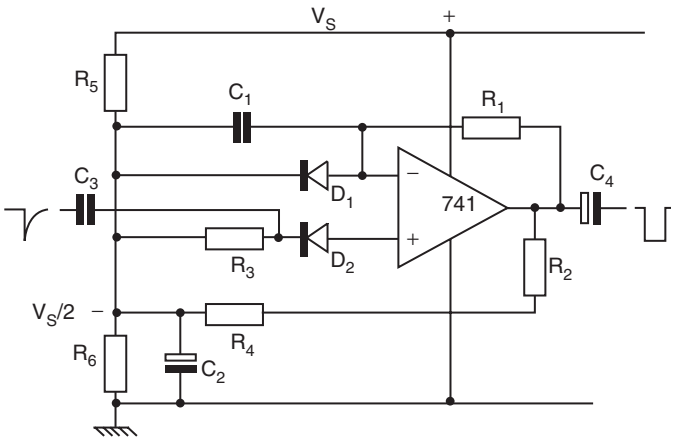


Figure 6.12

A 741 monostable circuit. With no input, the output voltage is high, which causes the (+) input voltage to be higher than the voltage level $V_s/2$ set by R_5 and R_6 (equal values). Because of D_1, the (−) input cannot rise to the same value as the (+) input. A negative pulse at the (+) input causes the output voltage to drop rapidly, taking the (+) input voltage low. The (−) input voltage then drops at a rate determined by the time constant $C_1 \times R_1$. When the (−) input voltage equals the (+) input voltage, the circuit switches back, and the diode D_1 conducts to 'catch' the (−) input voltage and so prevent continuous oscillation.

Current differencing amplifiers

A variation on the op-amp circuit uses current rather than voltage input signals, and is typified by the National Semiconductor LM3900. This also is an old (1972) design, but is still in production and use, though more modern versions such as LM359 and LM3301 are available. In the LM3900 IC, which contains four identical op-amps, the (+) and (−) inputs are **current** inputs, whose voltage is generally about +0.6 V when correctly biased. A single-ended power supply is used, and the output voltage can reach to within a fraction of a volt of the supply limits. The output voltage is proportional to the difference between the currents at the two inputs, so bias conditions are set by large-value resistors.

Figure 6.13 shows the pinout of the chip and a typical amplifier circuit, in which the current into the (+) input is set by R_1, whose value is 2.2 MΩ. Because the ideal bias voltage for the output is half of supply voltage, a 1 MΩ resistor is used connected between the output and the (−) input. In this way, the currents to the two inputs are identical, and the amplifier is correctly biased. Though National Semiconductor pioneered this type of op-amp, similar types are obtainable from other manufacturers.

Other linear amplifier ICs

A very large variety of ICs intended for AF, IF and RF amplifiers can be obtained. For any design work, the full manufacturer's data sheet pack (usually obtainable in PDF format from the manufacturer's website) must be consulted, but a few general notes can be given here. AF IC circuits use direct coupling internally, because of the difficulty of fabricating capacitors of large value onto silicon chips, but the high gains which are typical of operational amplifiers are not necessary for most AF applications. Faster slew rates and greater open-loop bandwidths can therefore be attained than is practicable using op-amps.

Many AF ICs use separate chips for preamplifier and for power amplifier uses, with separate feedback loops for each. Frequency correcting networks composed of resistors and capacitors are usually needed to avoid oscillation,

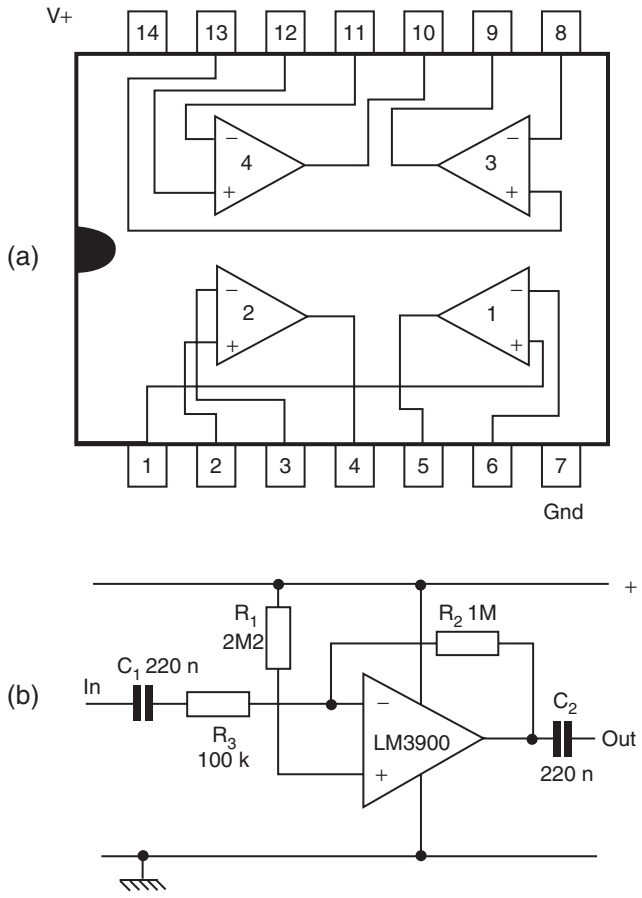


Figure 6.13

The current-differencing amplifier, or Norton op-amp. **(a)** Pinout for the LM3900, which contains four amplifiers in a single fourteen-pin package. **(b)** Typical amplifier circuit. Note the high resistor values.

and heatsinks will be needed for the larger power amplifier ICs. The need for external volume, stereo balance, and bass and treble controls, along with feedback networks, makes the circuitry rather more involved than several other IC applications.

Figure 6.14 shows two examples of AF circuits. Note that the stability of these audio ICs is often critical, and decoupling capacitors, as specified by

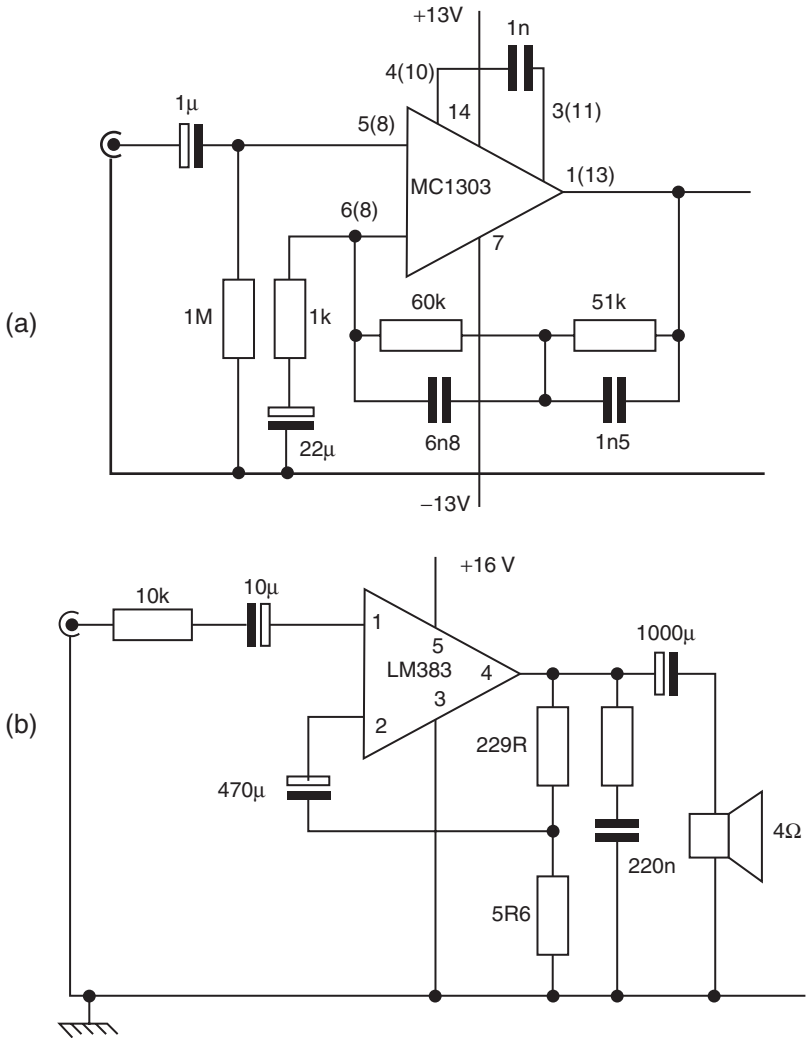


Figure 6.14

Audio amplifier ICs. **(a)** The MC1303 preamplifier is a dual unit for stereo use – the pin numbers in brackets are for the second section. Inputs up to 5 mV can be accepted, and the circuit here is shown equalized for a magnetic pickup. The output is 250 mV with a 5 mV input at a distortion level of about 0.1%. **(b)** The LM383 power amplifier uses a five-pin TO220 package. The power output is 7 W into 4 ohms, with a distortion level of 0.2% at 4 W output. The maximum power dissipation is 15 W when a 4°C/W heatsink is used.

the manufacturers, must be connected as close to the IC pins as possible. For stability reasons also, stripboard construction is extremely difficult with some IC types, and suitable printed-circuit boards should be used.

IF and RF amplifier circuits contain untuned wideband amplifier circuits to which tuning networks, which may be LC circuits or transfilters, may be added. It is possible to incorporate RF, mixer, IF and demodulator stages into a single IC, but generally only when comparatively low frequency RF and IF are used. At one time a very common scheme for FM radio receivers was to use a discrete component tuner along with IC IF and demodulator stages, using the usual 10.7 MHz IF. In Figure 6.15 an example of such an IF stage is shown. Once again, when a large amount of gain is attained in one IC, stability is a major problem, and the manufacturer's advice on decoupling must be carefully followed. At the higher

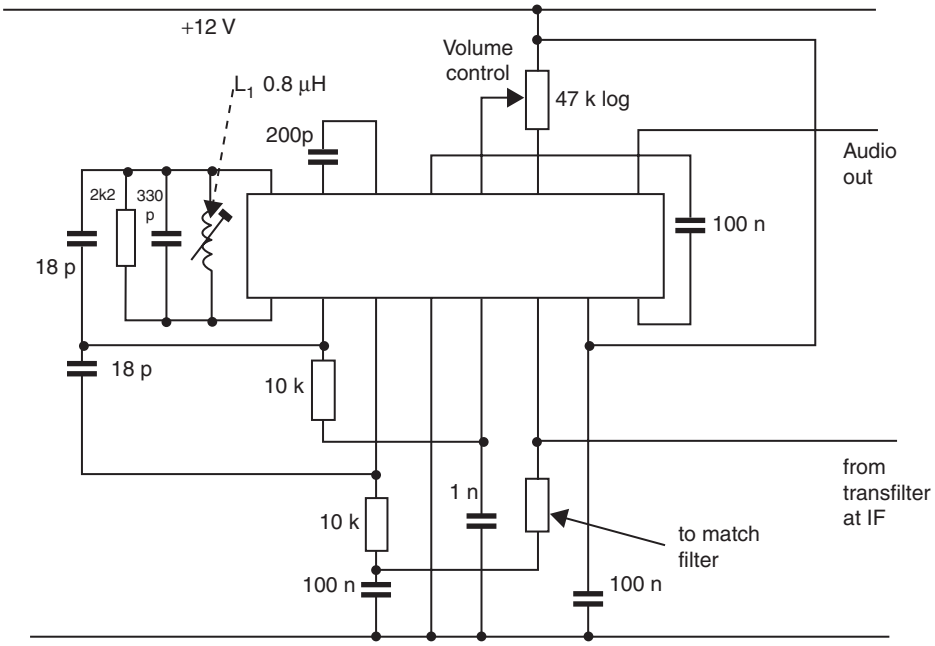


Figure 6.15

An IF/detector IC for use in 10.7 MHz stereo FM IF stages. The minimum input for limiting is 100 µV, and the volume control range (operating on DC) is 80 dB. The audio output is 1.4 V rms with a signal of 15 kHz deviation.

frequencies, the physical layout of components is particularly important, so PCBs intended for the TBA750 IC (and similar) should be used rather than stripboards.

It is much more common now for a complete FM receiver circuit to be integrated into one single IC, and Figure 6.16 shows the Philips TEA5880TS chip block diagram and connections. This is a very modern chip (2004) and some of its salient features are:

- no alignment actions needed;
- stereo decoder needs no adjustment, and no external crystal required;
- adjacent channel rejection built in;
- very high sensitivity;
- RF automatic gain control (AGC) circuit;
- standby mode for power-down, and no power switch circuitry required;
- 2.7 V minimum supply voltage;
- MPX output for RDS;
- covers all Japanese, European and US bands.

Phase-locked loops

The phase-locked loop (**PLL**) is a type of linear IC which is now used to a considerable extent either as a stand-alone IC or incorporated into other ICs. The block diagram of the circuit is outlined in Figure 6.17 and consists of a voltage-controlled oscillator, a phase sensitive detector, and comparator units. The oscillator is controlled by external components, so the frequency of oscillation can be set by a suitable choice of these added components. An input signal to the PLL is compared in the phase-sensitive detector to the frequency generated in the internal oscillator, and a voltage

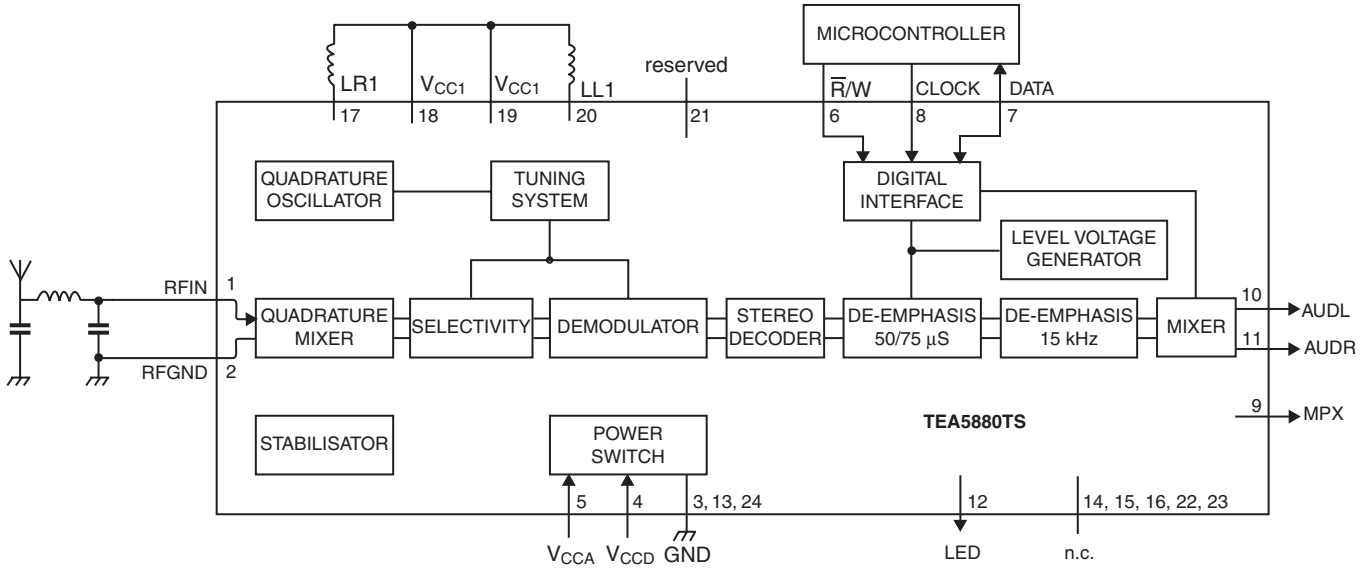


Figure 6.16

Block diagram of Philips TEA5880TS FM radio IC.

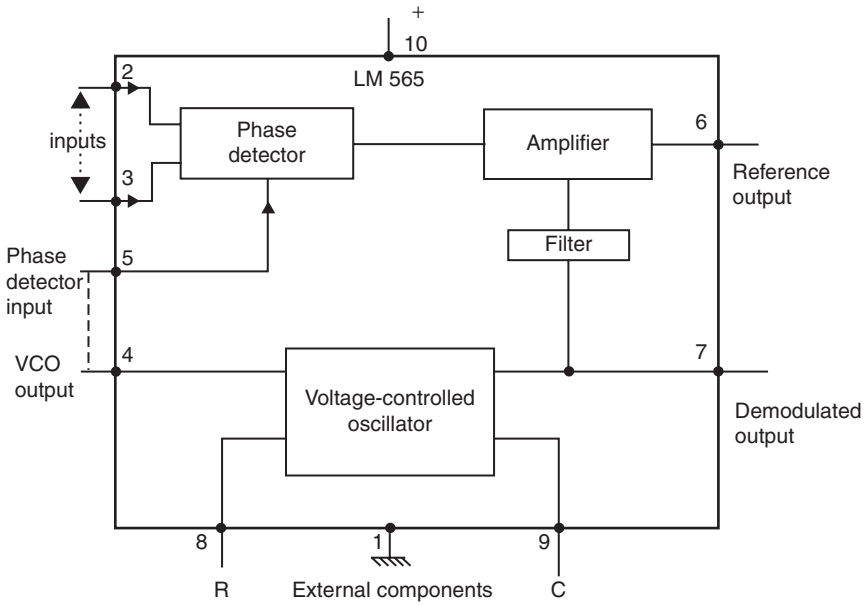


Figure 6.17

The PLL block diagram. The pin numbering is for the LM565. The signal input can be to pin 2 or 3 in this IC, and in normal use pins 4 and 5 are linked.

output is obtained from the phase-sensitive detector. Provided that the input frequency is not too different from the internally generated frequency (within the pull-in range), the voltage from the phase-sensitive detector can then be used to correct the oscillator frequency until the two signals are at the same frequency and in the same phase. Either the oscillator signal or the correcting voltage may be used as an output. The circuit can be used, for example, to remove any traces of amplitude modulation from an input signal, since the output (from the internal oscillator) is not affected by the amplitude of the input signal, but is locked to its frequency and phase.

The circuit may also be used as an FM demodulator, since the control voltage will follow the modulation of an FM input in its efforts to keep the oscillator locked in phase. PLL circuit examples are illustrated in Figure 6.18.

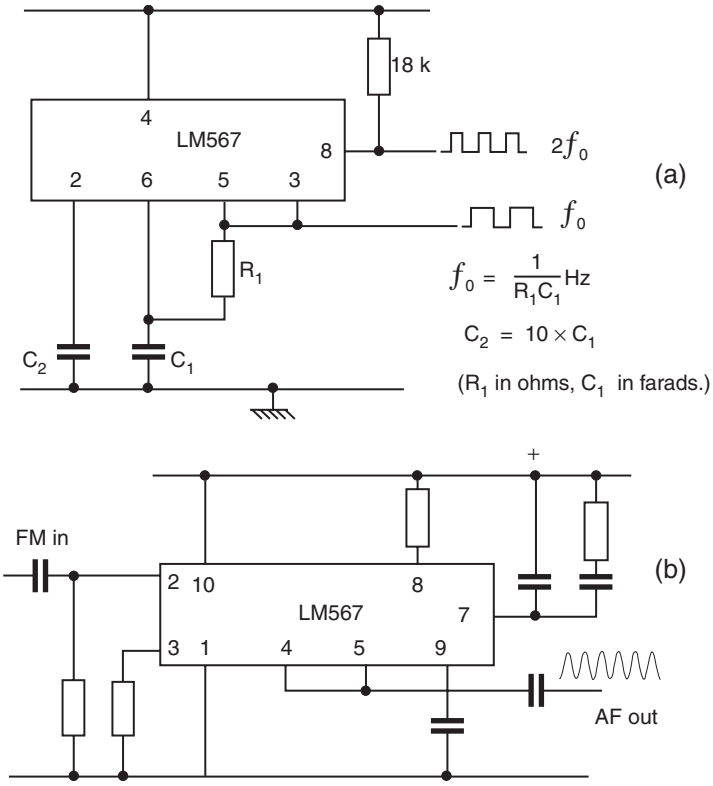


Figure 6.18

PLL circuits: **(a)** oscillator with fundamental and second harmonic outputs. **(b)** FM demodulator – the component values must be calculated with reference to the IF frequency which is used. In this example the IF cannot be as high as the normal 10.7 MHz because the operating frequency limit of this IC is 500 kHz.

Waveform generators

The requirement for a precise waveform generator is so common that it justifies the production of specialized ICs for that purpose. A typical example is the ICL8038 from Intersil. This is an IC that uses a triangular wave as a driver to generate other waveshapes. It can produce, with high accuracy, waveforms of sine, square, triangular, sawtooth and pulse shapes at frequencies ranging from 0.001 Hz to more than 300 kHz, using a minimum

of external components (resistors or capacitors). A frequency sweep can be obtained by using a sawtooth input voltage signal. Some important features include:

- low frequency drift with temperature, typically 250 ppm/°C;
- low distortion, typically 1% on a sine wave output;
- high linearity, typically 0.1% for a triangle wave output;
- wide frequency range, 0.001 Hz to 300 kHz;
- variable duty cycle, 2% to 98%;
- high level outputs, TTL to 28 V;
- simultaneous sine, square, and triangle wave outputs;
- only a few external components required.

Figure 6.19 shows a typical application circuit for an audio generator.

For higher frequencies, the Maxim MAX038 chip can be used; a circuit for a sine wave generator is shown in Figure 6.20 (other application circuits are noted on the application notes for the chip). The features of this IC include:

- operating frequency range 0.1 Hz to 20 MHz;
- choice of triangle, sawtooth, sine, square, and pulse waveforms;
- independent frequency and duty-cycle adjustments;
- frequency sweep range 350 to 1;
- duty cycle variable from 15% to 85%;
- low-impedance output buffer;
- low-temperature drift, 200 ppm/°C.

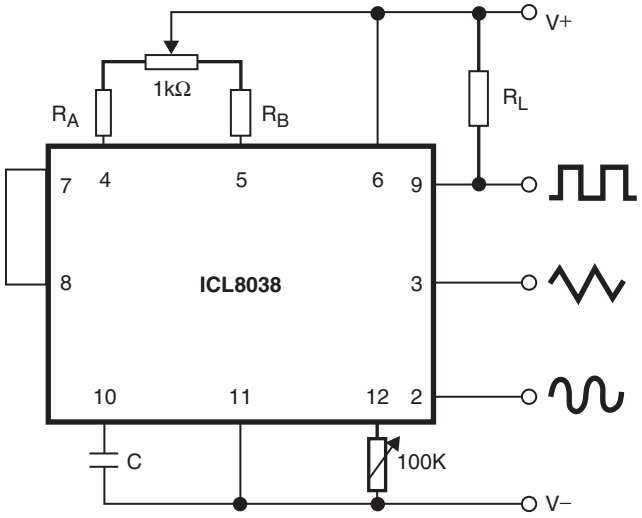


Figure 6.19

Audio generator using the ICL8038. (Courtesy of Intersil.)

Active and switched capacitor filters

Filters of various types are available as IC components, requiring only a minimum of additional components. These are classed as **active filters**, making use of operational amplifiers along with waveshaping components for their action, and they greatly ease the burden of designing filters from discrete components. The classic filter type of responses – Butterworth, Bessel, Cauer and Chebyshev – all require a mass of calculation to implement in passive components (inductors, capacitors and resistors). Active filters can be programmed simply by applying a clock input, by adding external resistors, or by adding external capacitors. One type, known confusingly as **switched capacitor filters**, uses a data switching technique, and a typical example, the National Semiconductor MF10C, is noted here. The MAX series of semiconductors is also a popular choice. The filters that include a switching action are not strictly speaking purely linear ICs, but are classed with the linear active filters because their outcome is that of a linear filter – often an outcome that would be impossible to achieve with practical passive components.

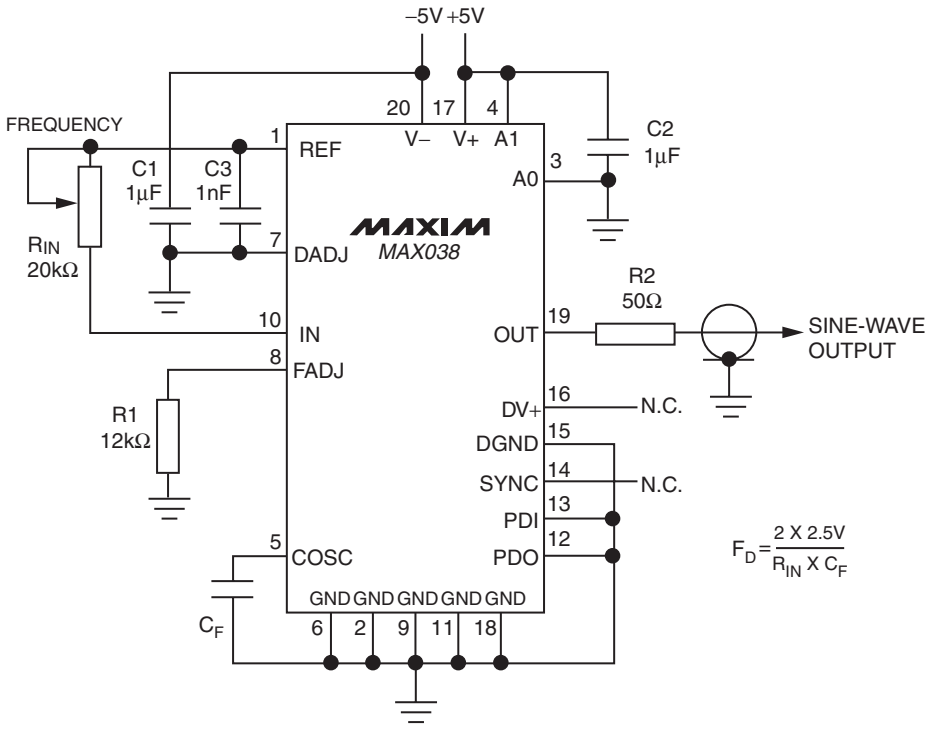


Figure 6.20

Using the MAX038 IC. (Circuit courtesy of Maxim Inc.)

The MF10 is a very versatile unit, consisting of two independent CMOS active filter units. To these a few resistors (usually three) can be connected so that the required filter action is obtained. In each block, one output can be configured as any one of all-pass, high-pass or notch filtering, and the other two outputs can be used for low-pass and band-pass actions. An external clock frequency can be used to set the centre frequency for the low-pass and band-pass actions; the high-pass centre frequency is determined both by the clock and by the external resistors.

Figure 6.21 shows a typical application circuit using three external resistors; Figure 6.22 shows normalized graphs for the band-pass, low-pass, high-pass, notch and all-pass filter actions.

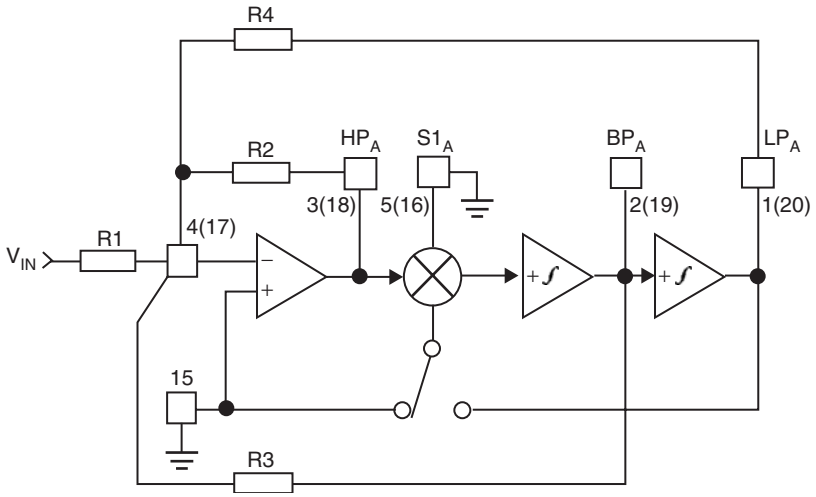


Figure 6.21

Filter circuit using the MF10. (Courtesy of National Semiconductor.)

The important features of the MF10 and other similar active filter units are:

- ease of use;
- clock to centre frequency ratio accuracy of $\pm 0.6\%$;
- filter cut-off frequency stability is directly dependent on external clock quality;
- low sensitivity to external component variation;
- separate high-pass (or notch or all-pass), band-pass, and low-pass outputs;
- $f_0 \times Q$ range up to 200 kHz;
- operation up to 30 kHz;
- 20-pin 0.3″-wide DIL package or 20-pin surface mount (SO) wide-body package.

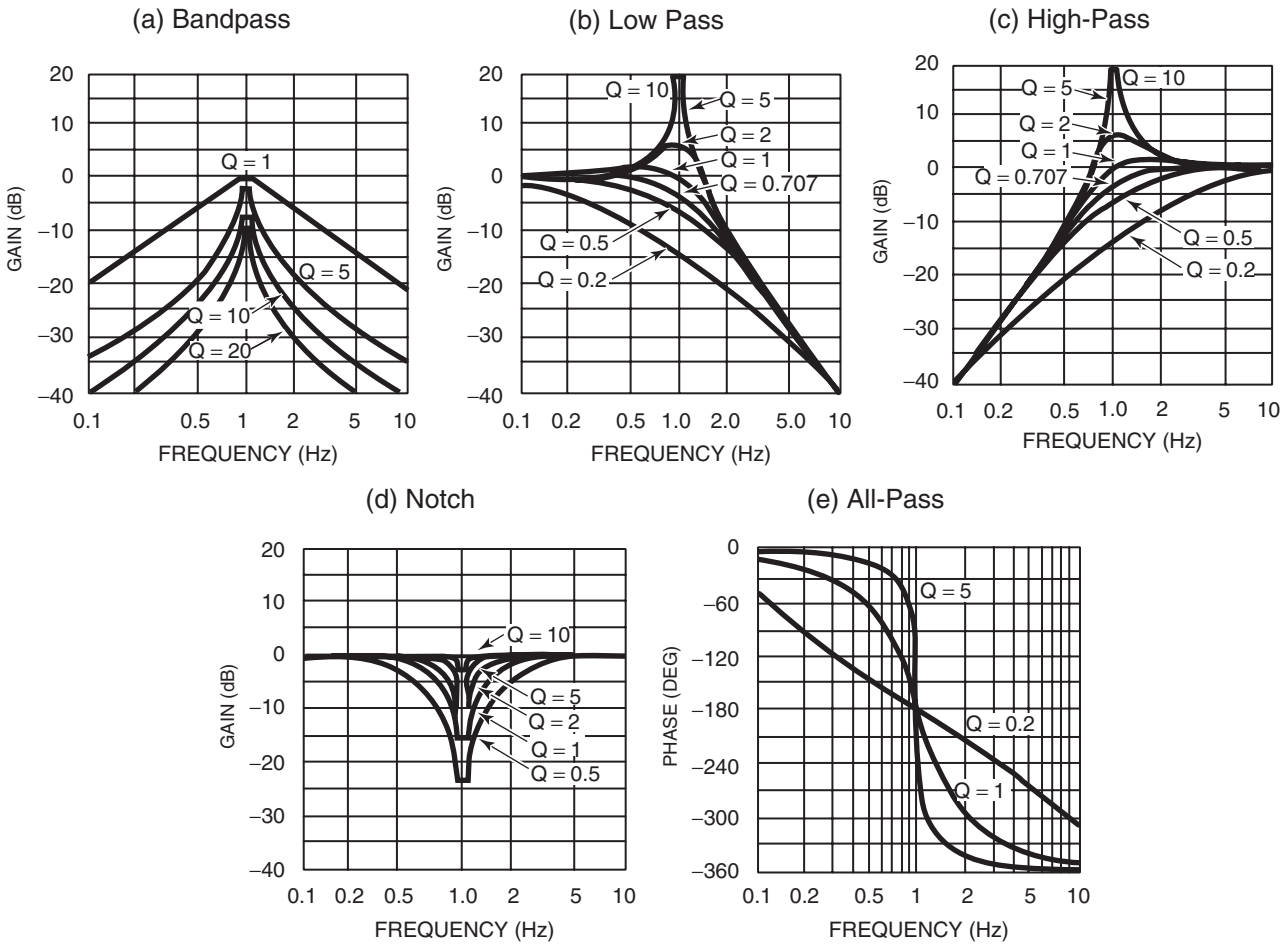


Figure 6.22

Normalized graphs for filter actions of the MF10C.

Voltage regulator ICs

The ease with which precision band-gap (regulated voltage) circuits and balanced amplifiers may be constructed in integrated form, together with the increasing demand for stabilized supplies and the steady increase in the power which can be dissipated from ICs due to improved heatsinking methods, has led to the extensive use of IC voltage stabilizer circuits, to the extent that discrete component stabilizers are almost extinct. The types considered here are the truly linear types; the switching types of regulators are dealt with in Chapter 7. The name 'regulator' is now replacing 'stabilizer' for this type of circuit, but the principle is the same – to provide a power supply whose output voltage (or current, for a current regulator) remains constant despite variations in load resistance and supply voltage. Most of the following refers to voltage regulators.

The older types of regulator, such as the well-known 78xx series, used much the same circuitry inside the IC as the discrete component counterpart. Voltage regulators can be of the fixed type, giving an output voltage fixed by the internal circuitry, or the variable type whose output voltage can be altered by connecting external resistors. Latterly, the performance expected of IC regulators has changed to reflect the extensive use of battery-powered supplies and the need to reduce the dissipation in the IC. Modern linear regulators fall into three categories, referred to as Standard (sometimes NPN Darlington, often un-named), low dropout (**LDO**), and **Quasi-LDO**. The feature that distinguishes these types is the **dropout voltage**, which is the minimum difference between input voltage and output voltage needed to maintain (voltage) regulation. A regulator which features low dropout voltage will dissipate less power than one with a higher dropout, and is therefore more efficient, with a lower earth current.

Figure 6.23 shows the basic internal circuitry, without details, of these three classes of regulators. The NPN type makes use of a power NPN transistor structure that is driven by a PNP-NPN Darlington circuit. This requires a minimum of dropout of about 1.5 V to 2.5 V to operate. The LDO type of structure uses a power PNP transistor, ensuring a dropout of less than 500 mV (as low as 10 mV on low loads); the Quasi-LDO provides a dropout voltage whose value lies between the other two.

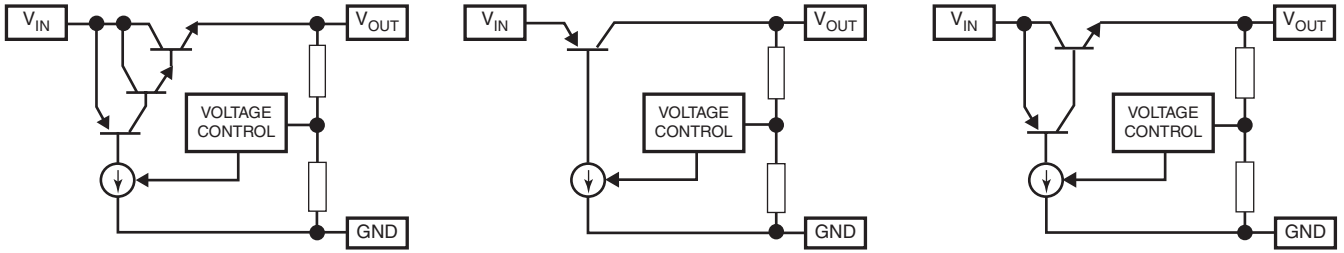


Figure 6.23

The basic circuits of the main voltage regulator types: **(a)** NPN, **(b)** PNP, **(c)** Quasi-LDO.

All types of linear IC regulators nowadays contain additional circuitry that is designed to prevent damage from either excess load current or excessive temperature. The design of regulators includes three separate control loops to deal with current limiting and thermal shutdown as well as the normal voltage error correction required in any regulator. Modern designs ensure that the thermal limiter can override all others, and that the current limitation (also called foldback protection) overrides voltage error. Anything that causes overheating or excess current will therefore lead to loss of regulation of voltage.

Figure 6.24 shows a typical circuit using a regulator, and the important point is that for the highest standards of regulation, the common earth pin of the regulator must be connected to the 'cold' side of the load rather than to a more local earth. The alternative is to ensure that the resistance between the earthy side of the load and the earth pin of the regulator is at a minimum. The regulator should also be placed as close to the load as possible.

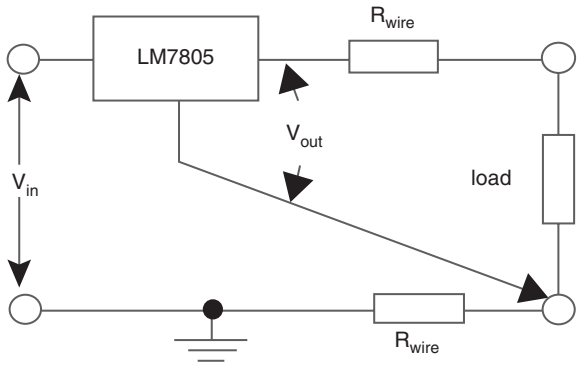


Figure 6.24

A typical regulator application circuit. (Courtesy of National Semiconductor.)

Adjustable regulator circuits

Any fixed voltage regulator can be converted to use as a variable voltage regulator by returning the earth (usually labelled **adjust**) pin of the regulator IC to a voltage divider circuit as indicated in Figure 6.25.

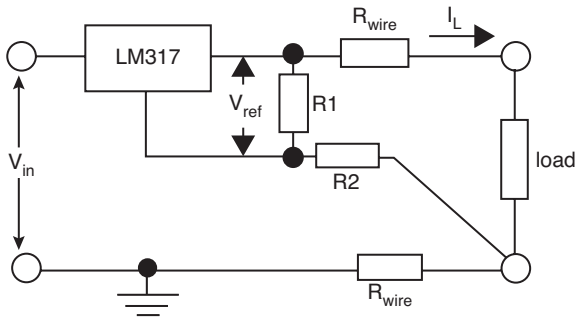


Figure 6.25

Extending the range of a fixed-voltage regulator. (Due to National Semiconductor.)

A later development has been the use of four-pin regulator ICs. This allows a separate earth pin to be used, connected, ideally, to the ground end of the load, along with a feedback pin fed from a resistive voltage divider (Figure 6.26). In this type of circuit, the lower resistor of the divider (R2) should be located as close to the regulator as possible. The circuit illustrated permits output voltage to be adjusted in the range 1.23 V to 29 V.

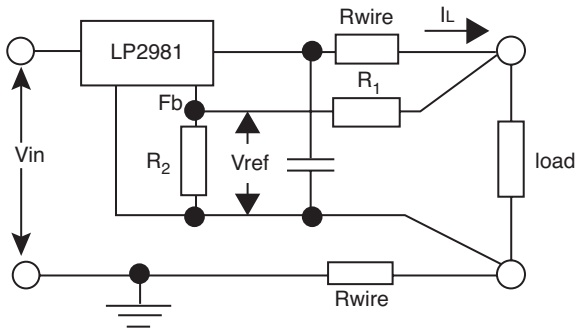


Figure 6.26

Using a multipin regulator (such as the National Semiconductor LP2951).

One very important class of linear ICs is concerned with television circuitry. The development of linear ICs has been such that virtually every part of an

analogue TV circuit with the exception of the tuner can now be obtained in IC form. Because of the specialized nature of such circuits, the reader is referred to the manufacturer's handbooks for further information. The coming of digital television along with digital displays (such as LCD and plasma) has made it possible to construct a completely digital TV receiver with no analogue circuitry, and this is steadily leading to making the dream of the 'one-chip TV' a reality.

By now, the chips that have been used for analogue TV circuits are beginning to be classed as maintenance items rather than design items, because the cathode-ray tube is rapidly being superseded as a TV display device, and all-digital TV is rapidly replacing the analogue system that has served us for almost 70 years. Several manufacturers of analogue PAL TV receivers have used the Panasonic AN5192K IC, which performs a remarkable amount of the processing requirements of a PAL TV receiver, using a 64-pin DIL package chip.

The AN5192K is a complete processor for analogue TV colour signals in either NTSC or PAL format. It includes video and sound IF stages, and the processing of chroma, RGB and synchronization signals. No manual adjustments are needed, nor are any inductors required. This chip is released only to manufacturers, and Panasonic do not wish the block diagram to be reproduced in this book, but you can see this (and more) information on the website:

www.ortodoxism.ro/datasheets/panasonic/AN5192.pdf

The 555 timer

This circuit is generally classed among linear circuits because it uses op-amp circuits as comparators. Though this is a very old device (in IC terms, at least) it is still in production and use in various forms because of its remarkable versatility. There is a CMOS version, coded as 7555, that operates with much lower currents. The purpose of the timer is to generate time delays or waveforms which are very well stabilized against voltage changes. A block diagram of the internal circuits is shown in Figure 6.27. A negative-going pulse at the trigger input, pin 2, makes the output of comparator (2) go high. The internal resistor chain holds the (+) input of comparator (2) at

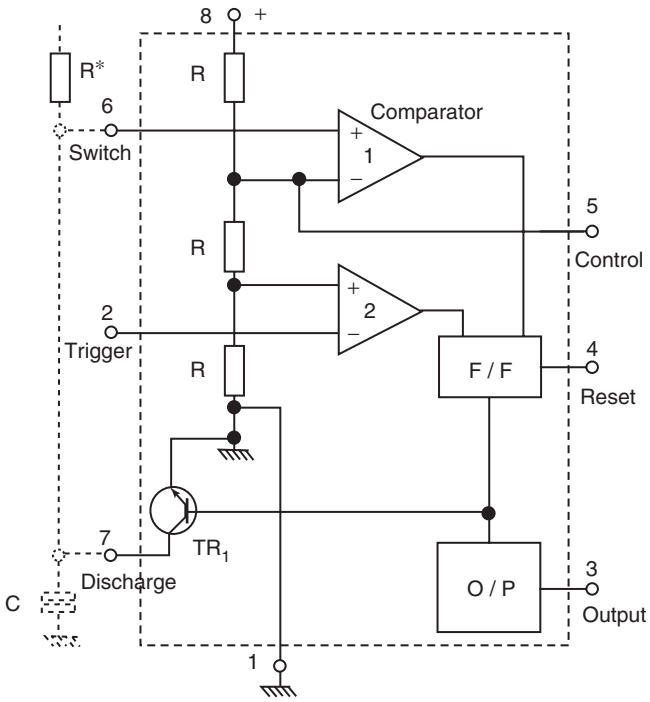


Figure 6.27

The 555 timer block diagram. R* and C are external components which are added in most applications of the timer. The delay provided by these components is found from 1.44R*C.

one third of the supply voltage, and the (−) input of comparator (1) at two thirds of supply voltage, unless pin 5 is connected to some different voltage level. The changeover of comparator (2) causes the flip-flop to cut off Tr_1, and also switch the output stage to its high-voltage output state. With Tr_1 cut off, the external capacitor C can charge through R (also external) until the voltage at pin 6 is high enough, equal to two thirds of the supply voltage, to operate comparator (1). This resets the flip-flop, allows Tr_1 to conduct again, so discharging C, and restores the output to its low-voltage state. Resetting is possible during the timing period by applying a negative pulse to the reset pin, number 4. The output of a 555 timer is rated to supply more than 100 mA, so transducers such as including loudspeakers, lamps, and even small motors can be connected directly to the output of the 555, pin 3.

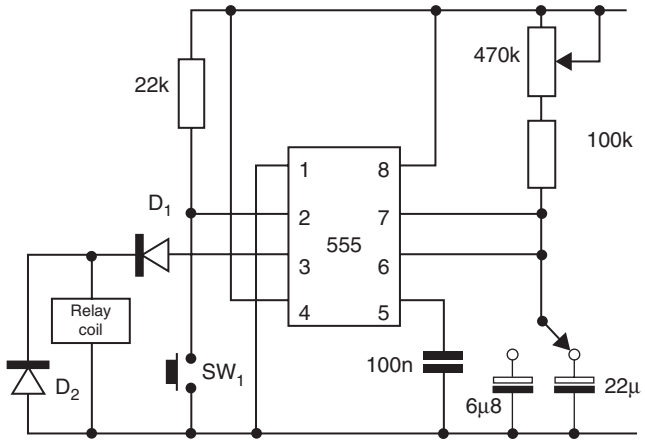


Figure 6.28

A relay timer circuit using the 555. On pressing the switch the relay is activated for a time determined by 1.1R*C, where R* equals 100k plus the setting of the 470 kΩ variable and C is the capacitor value that has been selected by the switch. Note the use of diodes to prevent latch-up and damage to the IC when the relay is switched off.

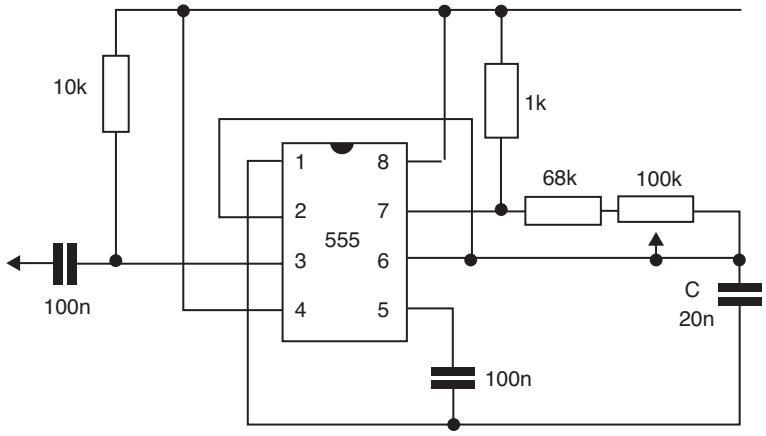


Figure 6.29

An astable pulse generator with variable frequency output controlled by the 100 kΩ potentiometer. The capacitor C can be a switched value if required. The frequency is given by the formula $\frac{1.4}{C(R_1+R_x)}$ with, in this example, $R_1 = 1k$, $R_x = 68\ k +$ variable setting, and $C = 20$ nF.

The triggering is very sensitive, and some care has to be taken to avoid unwanted triggering pulses, particularly when inductive loads are driven. Retriggering caused by the back-EMF pulse from an inductive load is termed 'latch-up', and can be prevented by the diode circuitry shown in Figure 6.28. Typical application circuits are shown in Figure 6.28 and 6.29. The timer is available from several manufacturers, all using the same 555 number though prefixed with different letter combinations which indicate the manufacturer.

CHAPTER 7

FAMILIAR LINEAR CIRCUITS

Overview

This chapter illustrates a selection of well-established circuits and data, and comments are reduced to a minimum so as to include the greatest number of useful circuits. The common-emitter and a few other basic amplifier circuits have already been dealt with in Chapter 5. Where several different types of circuits are shown, as for oscillators, practical considerations may dictate the choice of design. For example, a Hartley oscillator uses a tapped coil, but the arrangements for frequency variation may be more convenient than those for a Colpitts oscillator which uses a capacitive tapping. Note that crystal oscillator circuits may have to be modified to take account of the range of drive requirements for crystals of differing frequencies, mode and Q-values. As many variants on basic circuits have been shown as is feasible in the space. Discrete component circuit have been used in order to illustrate the action of each circuit, something that is usually hidden in the depths of the IC versions.

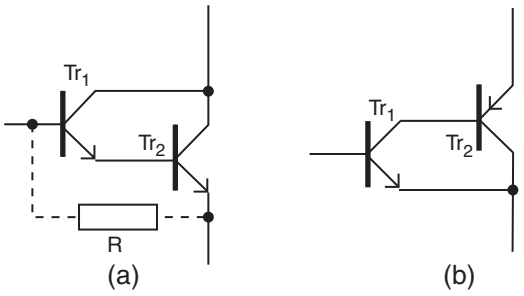
Discrete transistor circuits

We have looked earlier at the Darlington circuit as an example of a compound transistor circuit that gives an effective multiplication of h_{fe} value. There are some other two-transistor circuits that are still widely used, either in discrete form or incorporated in ICs. One example is the **complementary Darlington**, using both a NPN and a PNP transistor, sometimes called the **Sziklai pair**. In this circuit (Figure 7.1), the voltage between base and emitter is just that of a single transistor in contrast to that of a conventional Darlington. This type of circuit is commonly used in power output stages or in drivers for power transistors. The overall h_{fe} is, as for

a Darlington, the product of the h_{fe} values for the two transistors and so can have a very high value.

Figure 7.1

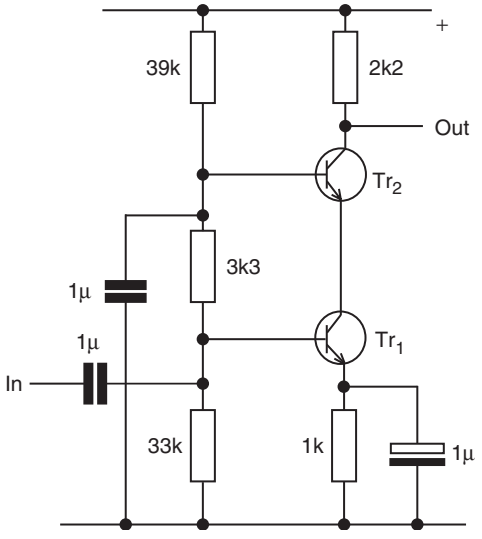
(a) NPN Darlington circuit, **(b)** complementary Darlington, or Sziklai pair, circuit.



The **cascode** circuit (Figure 7.2) is another way of connecting two transistors to obtain useful effects. The driver transistor operates in the common-emitter or common-source mode, and its load is another transistor operating in common-base or common-gate mode. The advantage of the cascode, as compared to a conventional two-transistor amplifier, is stability, because there is practically no feedback from output to input. The normal Miller effect is also greatly reduced because Tr_2 offers a low impedance collector load for Tr_1. This has led to the use of cascode circuits in both tuned and untuned amplifiers for high frequencies. The circuit has

Figure 7.2

Cascode connection of two bipolar transistors.



high gain over a large bandwidth. FET cascodes and combinations of FET and bipolar transistors can also be used; the FET types are now much more common. The combination of a JFET and NPN transistor has been used as a video driver stage for a CRT (see National Semiconductor application note no. 32).

The **long-tailed pair**, shown in both bipolar and in FET form in Figure 7.3, is the most versatile of all discrete transistor circuits, which is why it is so extensively used in the internal circuitry of linear ICs. A **common-mode signal** is a signal applied in the same phase to both bases or gates. Any amplification of such a common-mode signal can only be caused by a lack of balance between the transistors or FETs, so this value of gain is normally low, often very low. The difference signal is amplified with a considerably greater gain, and the ratio of the differential gain to the common-mode gain is an important feature of this type of circuit, called the **common-mode rejection ratio**, abbreviated to CMMR. The long-tailed pair is most effective when used as a balanced amplifier, with balanced inputs and outputs, but single-ended inputs or outputs can be provided for as shown in Figure 7.4a and b. The overall voltage gain of a long-tailed

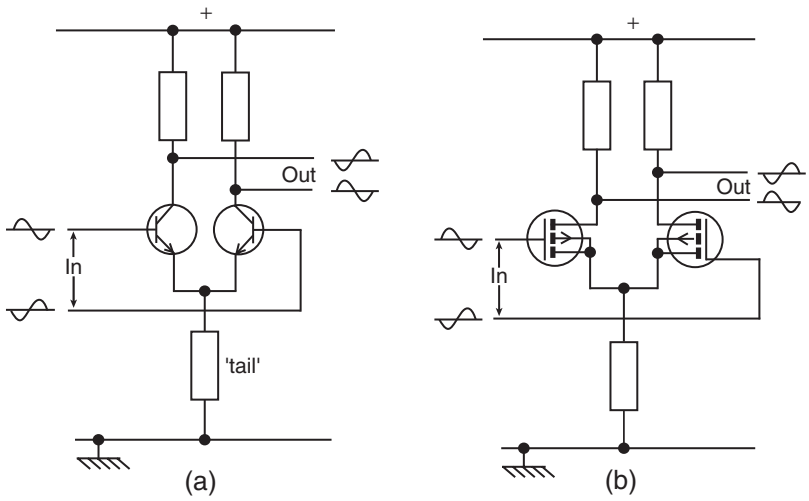


Figure 7.3

The long-tailed pair circuit using **(a)** transistors, **(b)** p-channel MOSFETs. Balanced input signals, as shown, are amplified, but unbalanced signals (in the same phase at each input) are attenuated.

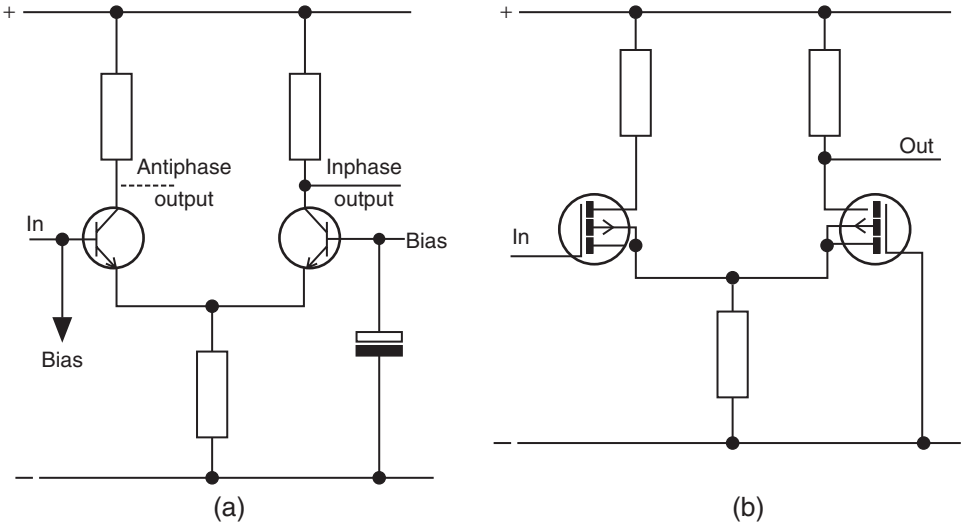


Figure 7.4

Single-ended inputs and outputs on a long-tailed pair circuit **(a)** using bipolar transistors, **(b)** using p-channel MOSFETs. The second input is earthed to signals. No bias arrangements are shown.

pair circuit is about half the gain that would be obtained from one of the transistors in a common-emitter circuit using the same load and bias conditions.

We have looked at the **current mirror** circuit earlier, and there are two other biasing applications that need to be noted, found mostly in IC form, but sometimes useful in discrete circuits. An ideal **current source** is one that will supply a fixed value of current irrespective of changes in load. This is another way of saying that the internal impedance of the source is infinitely high. For practical purposes, we can think of this as very high compared to the other impedances in the circuit.

Figure 7.5a shows a very simple type of current source, consisting of a PNP emitter follower. The V_{be} of the transistor is across the resistor R_2, and if R_2 is a small value then the current through R_2 will be much less than the i_{be} for the transistor. The same current, V_{be}/R_2, will also flow through R_3 and is the constant current that is required.

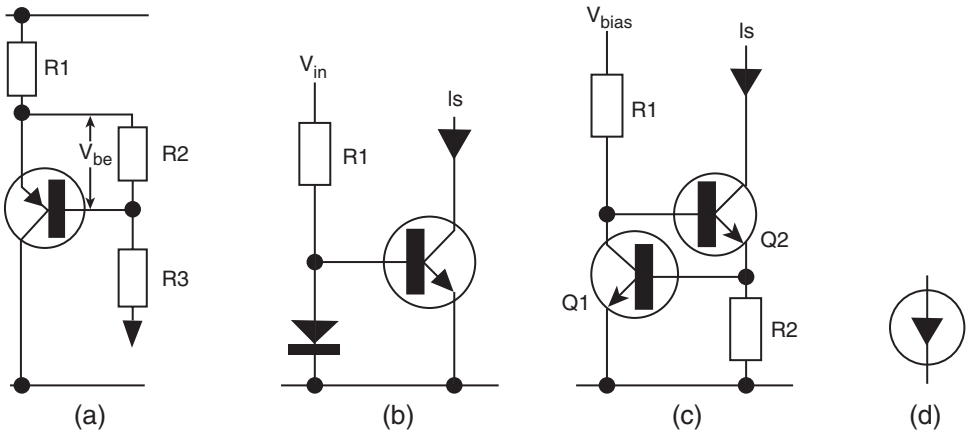


Figure 7.5

Current source circuits: **(a)** using a PNP transistor, **(b)** with diode biasing, **(c)** a circuit used extensively in ICs, **(d)** symbol for a current source.

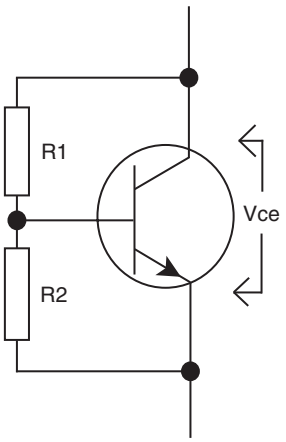
Figure 7.5b shows another current source circuit, using an NPN transistor. The bias voltage V_{in} passes current through R_1 and the diode, so the voltage across the diode is also the base-emitter bias for the transistor. Because the base-emitter junction of the transistor is constructed in the same way as that of a silicon diode, changes of temperature will affect both equally, so the bias current is fairly constant, as also will be the collector current, which is the steady current that is required. A variation on this circuit uses an LED in place of the diode, providing a higher bias voltage, and a resistor in the emitter circuit of the transistor to provide control over the collector current.

Figure 7.5c shows a type of current source circuit that is more elaborate and mainly used within integrated circuits. V_{bias} passes current through R_1 to bias the transistor Q_2. The collector current of Q_2 flows through the emitter resistor R_2, providing voltage feedback to the base of Q_1 and so controlling the collector current of Q_2 at $I_s = V_{be}/R_2$. The amount of controlled current can be altered by changing the value of R_2. Figure 7.5d shows the symbol for a current source, with the arrow indicating direction of current.

Another type of biasing circuit is the V_{be} **multiplier**. The simplest type is illustrated in Figure 7.6, and consists of a transistor biased by a potential divider between the collector and the emitter. The current through the potential dividing resistors is made high enough to swamp any changes in the base current of the transistor. The feedback ensures that the V_{be} for the transistor is almost constant, and because of that, the collector-emitter voltage is also constant, a constant multiple of the V_{be}, hence the use of the word **multiplier**.

Figure 7.6

A simple form of V_{be} multiplier circuit.



This type of circuit has for many years been used to set the bias of power output transistors in the conventional 'totem-pole' type of circuit used in audio power amplifiers. Figure 7.7 shows this type of output stage, using BJTs rather than the more usual MOSFETs, with the bias to the complementary Darlington power transistors obtained by using a V_{be} multiplier. The potential divider chain (R_1 and R_2) is often a preset potentiometer to allow the bias to be changed.

Audio circuits

Figure 7.8 shows a typical old-fashioned cassette recorder input circuit. This includes time constants that provide equalization to correct for the characteristics of tapeheads and tape. In addition to these 'standard' corrections, individual tape decks may need further corrections, a multiplex

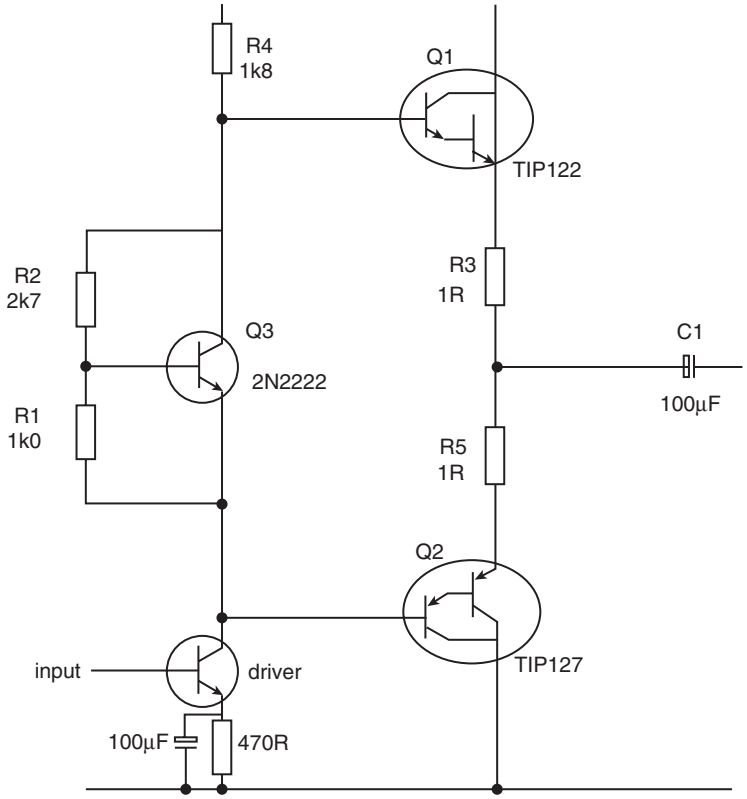


Figure 7.7

An output stage using a V_{be} multiplier to set bias. Note the use of power Darlington transistors.

filter may be included to remove FM stereo subcarrier signals, and noise-reduction circuits such as Dolby or dbx may be used. At the last count, equalization frequencies being used on replay were 3180 µs for all tapes, and either 70 µs or 120 µs for chrome and for ferric tapes respectively, with ferrochrome and pure iron tapes replayed at 70 µs. Equalization needed for recording amplifiers is too specialized to include here partly because recording equalization time constants depend much more on individual needs.

The use of discrete components in such circuits is vanishing except for specialized units intended to form part of a hifi system. Consumer cassette

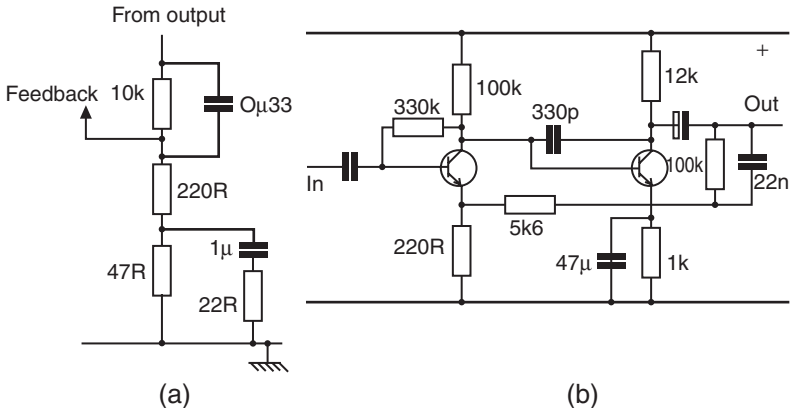


Figure 7.8

(a) and **(b)** A cassette recorder input stage using discrete components, showing time constants in the equalization networks.

recorders are much more likely to use ICs, and Figure 7.9 shows a one-chip solution using the Sanyo LA4160. This chip contains a preamplifier stage, automatic level control (ALC) and power amplifier with about 1 W output power using a supply voltage of 6 V and a loudspeaker of 4 Ω impedance. Only a few external passive components are needed.

Figure 7.10 shows the passive portion of the Baxandall tone control circuit, which is virtually the standard method of tone control used in audio systems. This was originally (about 1952) used in thermionic valve preamplifiers, but the principles have survived the transition, first to discrete transistors, and latterly to op-amps, proving the good design and durability of this circuit. There is very little interaction between the treble and the bass controls, low distortion, and a good range of control amounting to 20 dB or boost or cut.

Simple active filters

We looked at the elaborate type of programmed active filter in single-chip form in Chapter 6, but simpler circuits can be implemented either by discrete transistors or, more usually, op-amps.

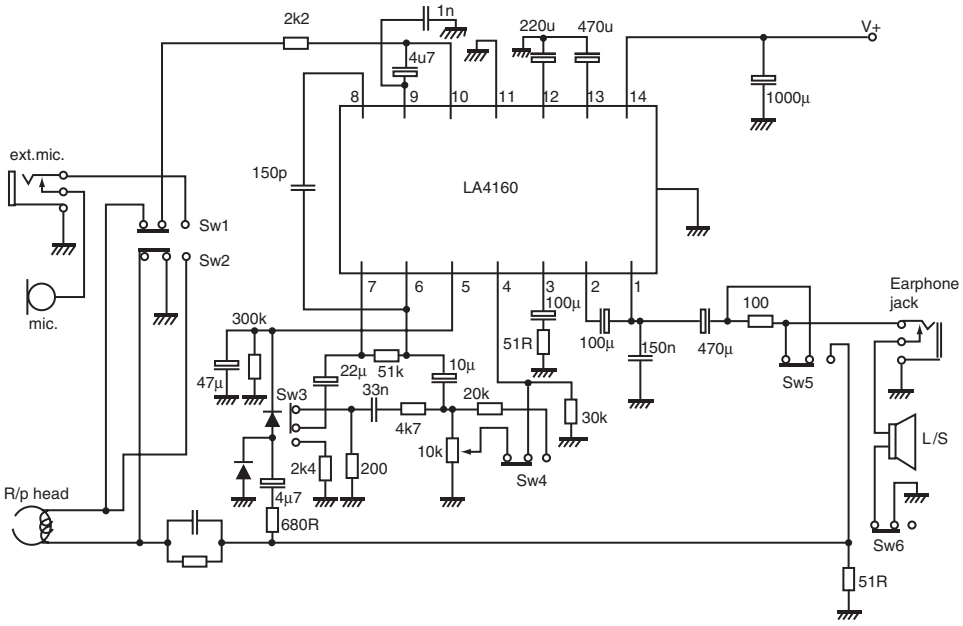


Figure 7.9

A complete cassette recorder circuit using the Sanyo LA4160 chip.

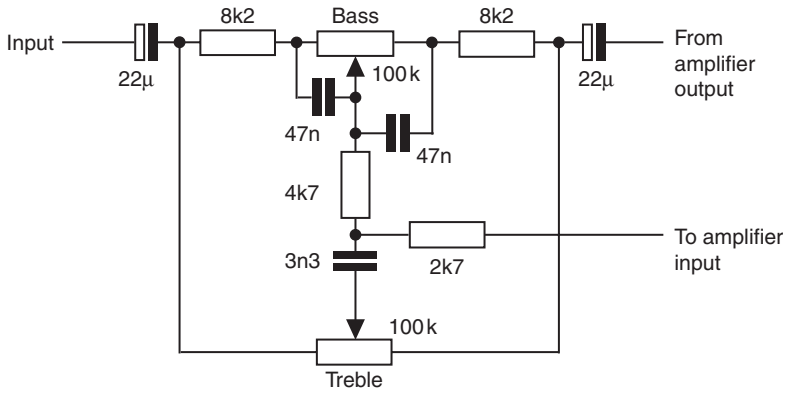


Figure 7.10

A Baxandall type of tone control circuit.

Figures 7.11 to Figure 7.13 deal with active filters using op-amps. These designs use only resistors and capacitors together with the ICs and are considerably easier to design than LC filters. Figure 7.11a shows a typical low-pass design. This is a first-order filter, meaning that only one reactive stage is used, so the slope is –6 dB per octave and turnover frequency given by $1/(2\pi RC)$. Figure 7.11b shows a Sallen & Key second-order type of circuit with higher slope value of –12 dB per octave. Figure 7.12 shows

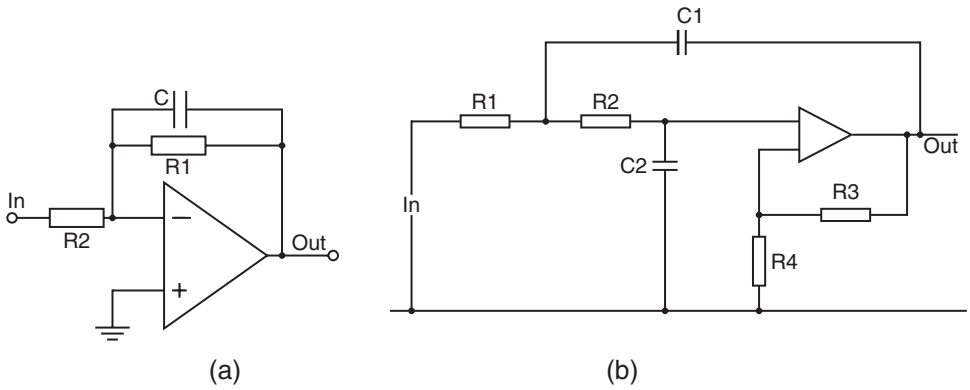


Figure 7.11

Active low-pass filters: **(a)** simple first order, **(b)** Sallen & Key second order.

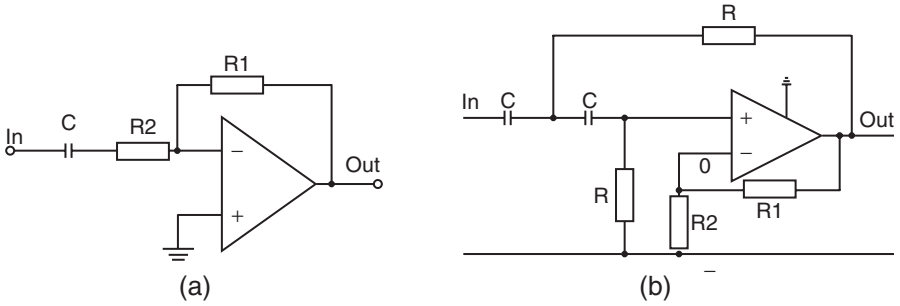


Figure 7.12

Active high-pass filters: **(a)** simple first order, **(b)** Sallen & Key second order.

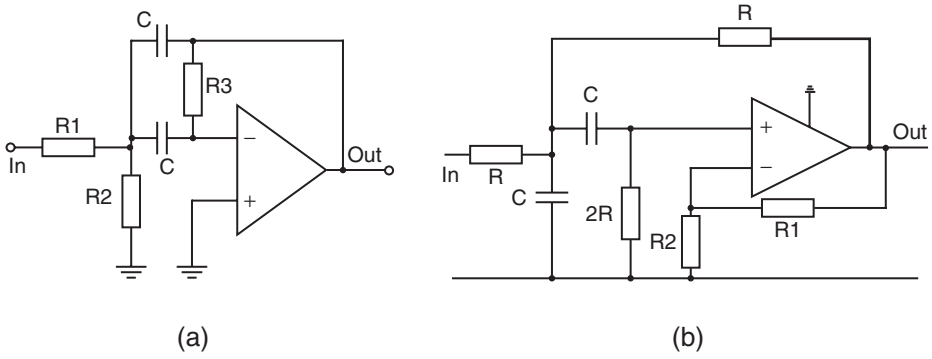


Figure 7.13

Active band-pass filters: **(a)** simple first order, **(b)** Sallen & Key second order.

the corresponding high-pass designs, and Figure 7.13 shows band-pass versions:

- Filter calculations can be difficult and time consuming, so you can either model the circuit using an aid such as SPICE (see Chapter 17) or try the websites devoted to such calculations such as:
  www.daycounter.com/Calculators/ Sallen-Key-Calculator.phtml

or

www.analog.com/Wizard/filter/filterUserEntry/

Circuits for audio output stages

The next three sets of circuits deal with audio output stages. **Class A** stages are those in which the transistor(s) are always biased on and never saturated (bottomed). A Class A stage may use a single transistor (a single-ended stage) or two transistors which share the current in some way (a push-pull stage), but the efficiency is low. Percentage efficiency is defined as:

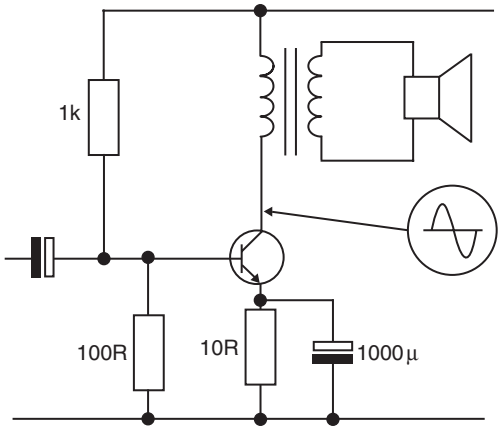
$$\frac{\text{power dissipated in the load}}{\text{total power dissipated in the output stage}} \times 100$$

and is always less than 50% for Class A operation.

A Class A stage should pass the same current when no signal is applied as when maximum signal is applied. Because of this, the dissipation is large, so large-area heatsinks are needed for the output transistors. Because of the high no-signal dissipation, Class A amplifiers are never found in IC form. Figure 7.14 shows a Class A, single-ended power output stage, once considered suitable for general-purpose use but hardly ever seen nowadays.

Figure 7.14

A Class A single-ended output stage, which needs a good heatsink.



Class B audio operation uses a pair of transistors biased so that one conducts on one half of the waveform and the other on the remaining half. Some bias must be applied to avoid 'crossover distortion' due to the range of base-emitter voltage for which neither transistor would conduct in the absence of bias. Class B audio stages can have efficiency Figures as high as 75%, though at the expense of rather higher distortion than with a Class A stage using the same layout. The higher efficiency enables greater output power to be obtained with smaller heatsinks, and the use of negative feedback can, with careful design, reduce distortion to negligible levels. Class B (or Class A–B, which uses higher no-signal current) is the favoured method of operation for IC amplifiers at power levels up to about 15 W output.

Figure 7.15 shows the **totem-pole** or **single-ended push-pull** circuit, which can be used for either Class A–B or Class B operation according to the bias level. This version uses complementary symmetry – the output

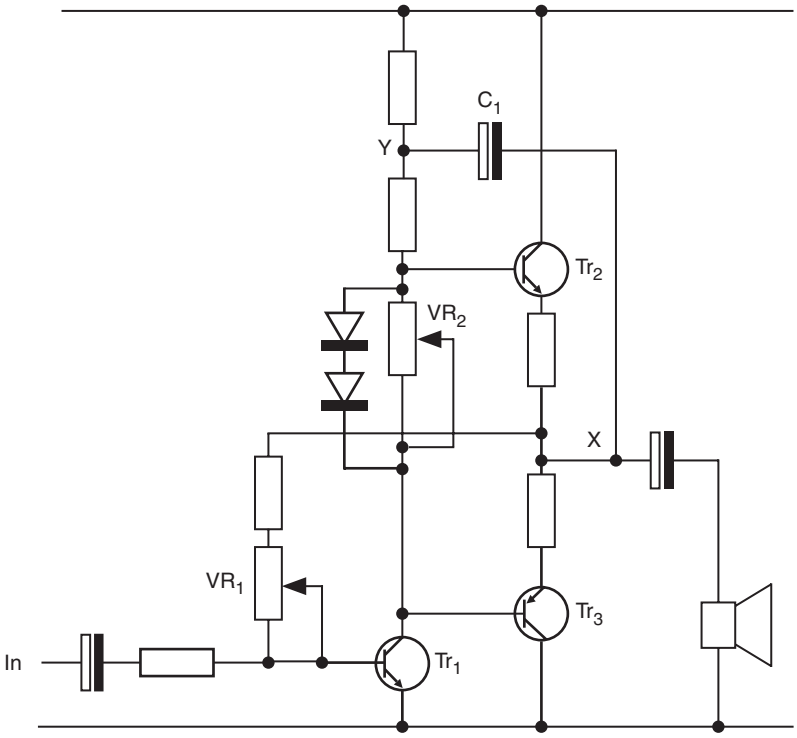


Figure 7.15

A single-ended push-pull (totem-pole) Class B output stage.

transistors are PNP and NPN types. In the circuit that is illustrated, VR_1 sets the voltage at point X to half of the supply voltage. VR_2 sets the quiescent (no signal) current through the output transistors. C_1 is a 'bootstrap' capacitor which feeds back in-phase signals to point Y, increasing input impedance. Oscillation is avoided because the gain of Tr_2 is less than unity. This type of circuit is usually the basis of IC power amplifiers.

When complementary output transistors cannot be obtained, a pseudo-complementary circuit, such as that of Figure 7.16, can be used, though this is not truly symmetrical. In fact, even a stage using complementary power transistors is not truly symmetrical because the characteristics of a PNP transistor can never be perfectly matched to those of an NPN transistor.

Figure 7.16

A quasi-complementary output stage. The low-power complementary transistors Tr_1 and Tr_2 drive the high-power output pair Tr_3, Tr_4, which are not complementary types. The circuit is not truly symmetrical and this can cause considerable distortion when overdriven.

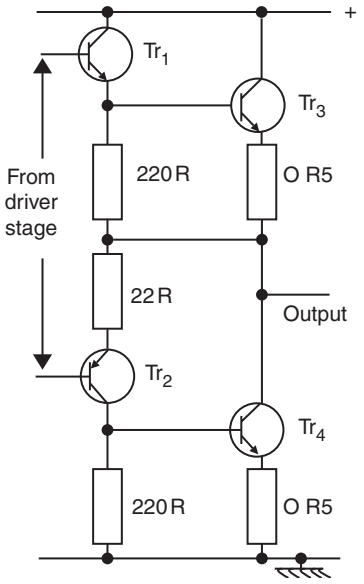
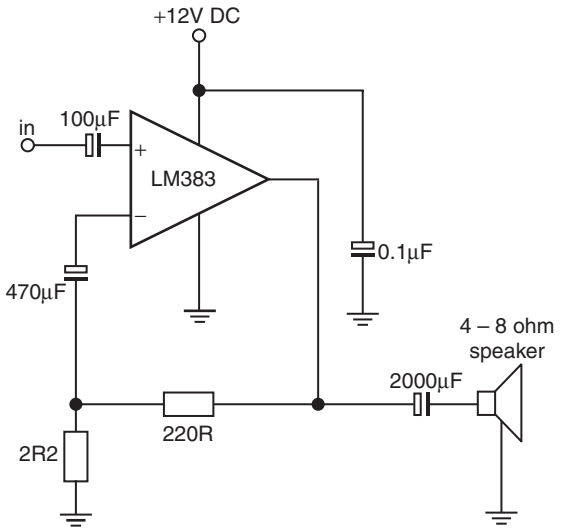


Figure 7.17 shows the circuitry for a Class B amplifier with 8 W output, based on the LM383 chip. This chip (by National Semiconductor) is a flat pack (TO-220) with five leads and a metal tab for bolting to a heatsink. This chip was designed with car radio applications in view, and has a 3.5 A

Figure 7.17

An 8 watt IC audio output amplifier stage.



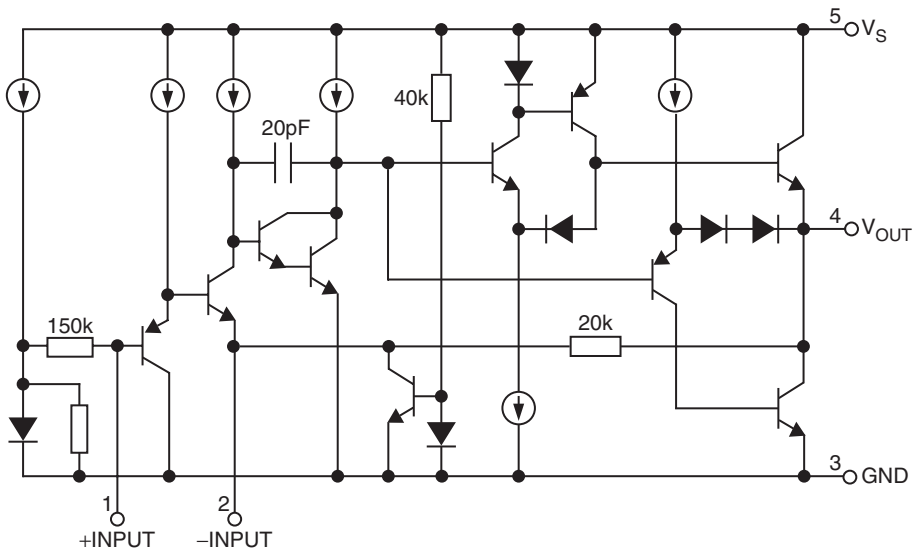


Figure 7.18

Internal circuitry of the National Semiconductor LM383. The symbol of the arrowhead in the circle represents a current source.

maximum driving current, with automatic current limitation and thermal protection (one version, the LM383A, also has over-voltage protection against transients). Figure 7.18 shows the internal circuitry for this chip which is typical of Class B IC power output stages.

Class D amplifiers

The idea of Class D amplification has been around for a long time (remember the kits that appeared briefly in the 1960s?) but only recently has been applied for use in amplifiers that can be taken seriously. The principle is to use fast-switching transistors with pulse waveforms that have been modulated with an audio signal, and one enormous advantage is that the dissipation in the transistors can be low even for outputs of several hundred watts. This allows higher output power for an IC chip to become a reality, particularly when fast-switching MOSFETs with very low forward resistance can be used.

The modulation system is usually a form of pulse time coding, in which the time of switching transitions (between minimum and maximum) is varied according to the input amplitude. The conversion to audio is carried out using a low-pass filter. A simple circuit that illustrates the principles is shown in Figure 7.19. The differential operational amplifier has inputs consisting of an audio wave and a fast pulse waveform, and the combined signal, the pulses modulated by the audio, is fed to high-power switching MOSFETs. At the output, a low-pass LC filter removes the high-frequency switching signals leaving the audio. Audio feedback can be applied from the switching output to an earlier audio stage. Practical IC implementations usually are of 'H' (or bridge) configuration, so the loudspeaker is fed from two switching circuits operated in antiphase. Hybrid circuits use an IC driver to feed the power FETs, as illustrated by the AudioMax LX1710 (Figure 7.20). Circuits of this type can be used for very large power outputs, as exemplified by the IR2011S assembly from International Rectifier, providing a maximum stereo output of 500 W + 500 W.

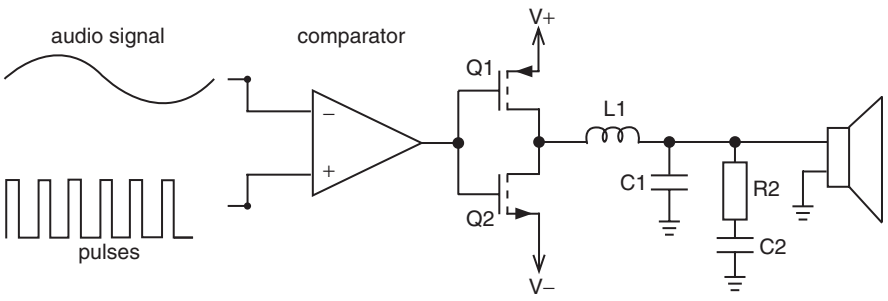


Figure 7.19

Principle of Class D operation.

The use of Class D can also allow the construction of single-chip power amplifiers of fairly high power requiring the minimum of heatsinking. The Texas Instruments TPA3100D2 is a modern example, delivering 20 W per channel into an 8 ohm load (with a supply voltage of 18 V) using bridge connected speakers of 20 W/ch into an 8 ohm load from a 18 V supply. Power output into 8 ohms for a 12 V supply is 10 W per channel, and for a 4 ohm load using a 12 V supply is 15 W per channel. The permitted supply

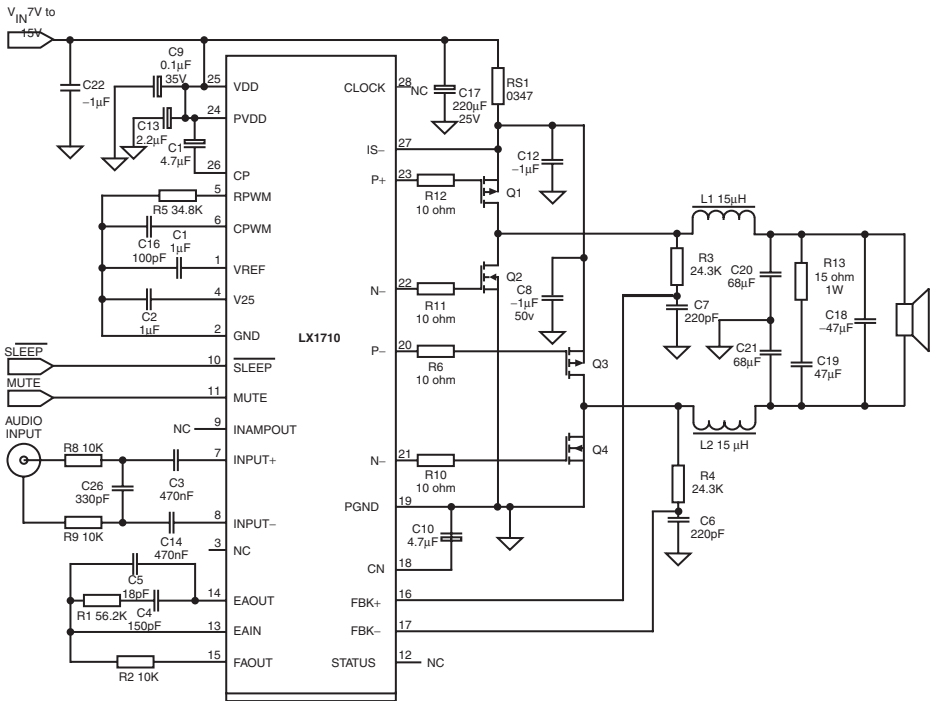


Figure 7.20

A circuit using the AudioMax LX1710 Class D driver chip.

voltage range is 10 V to 26 V, and the efficiency is 92%, emphasizing the advantages of Class D operation so that the chip does not require a heatsink. The circuitry incorporates thermal and short-circuit protection with auto-recovery, and two pins can be used to provide four gain settings of 20, 26, 32 and 36 dB. Differential inputs are used and the chip is packaged in SM format with 48 pins. A typical application circuit is shown in Figure 7.21. For full details of this chip, see the website http://focus.ti.com/lit/ds/symlink/tpa3100d2.pdf.

Another single-chip solution is the STA5150 from STMicroelectronics with a maximum mono output of 200 W. Other chips are in course of development, particularly with digital rather than analogue inputs.

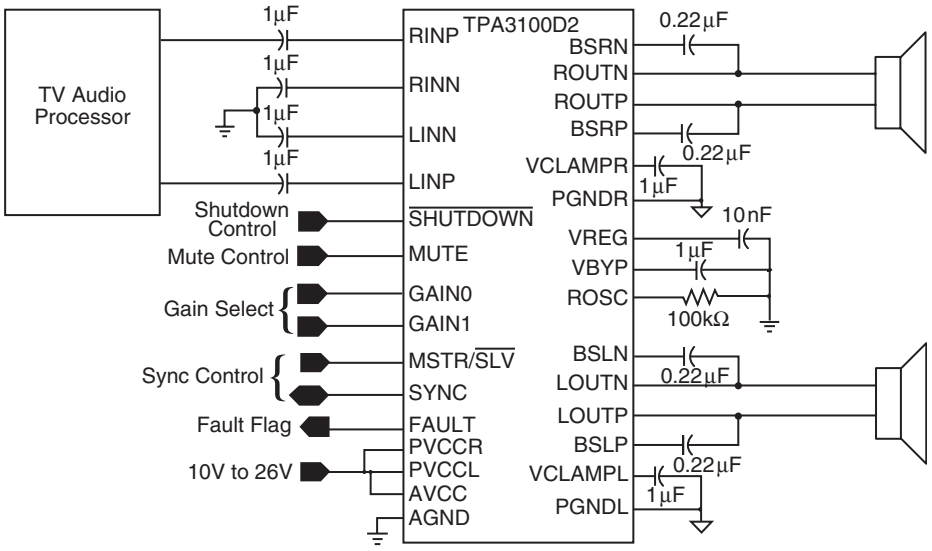


Figure 7.21

Typical circuitry around the T.I. TPA3100D2 chip as used in TV applications.

Wideband voltage amplification circuits

Figures 7.22 and 7.23 illustrate some of the circuits traditionally used for wideband voltage amplification with BJTs. Figure 7.22 deals with methods of frequency compensation using inductors or capacitors to compensate for the shunting effect of stray capacitances. Capacitive compensation can simply be applied with a capacitor across the emitter resistor (a), whose value is chosen so that the emitter resistor is progressively decoupled at high frequencies. As a rough guide, using the components illustrated, $C_S \times R_3$ should equal $C_2 \times R_4$. The other option is inductive shunt compensation, in which the value of L is chosen so as to resonate with the input capacitance of the transistor at a frequency above that of the uncompensated 3 dB point. These compensation methods are useful, but cannot compensate for low gain caused by an unsuitable transistor type. Transistors capable of amplification at high frequencies must be used in these circuits. Several single-chip wideband amplifiers are available.

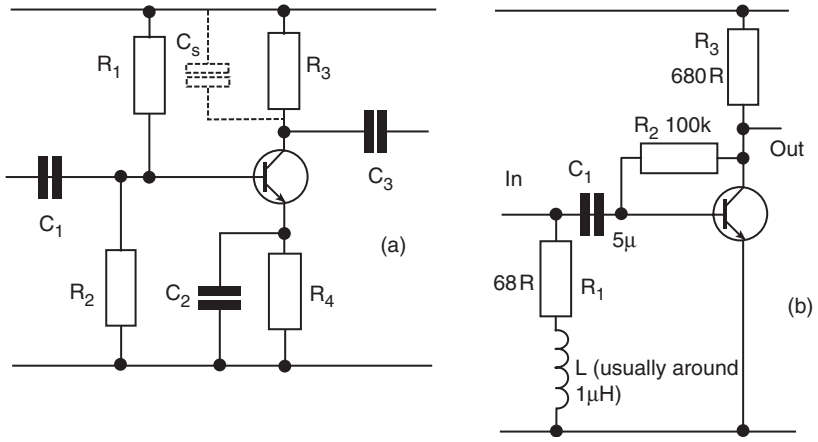
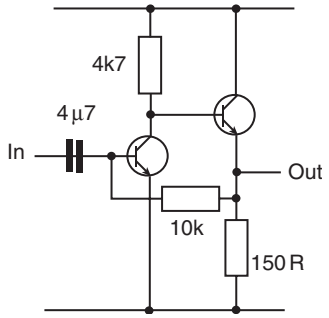


Figure 7.22

Frequency compensation for wideband amplifiers. **(a)** Capacitive compensation. **(b)** Inductive shunt compensation.

Figure 7.23

A circuit that uses feedback to reduce the gain and so extend the flat portion of the frequency range – this is a very useful basic circuit for video frequencies.



For truly remarkable wideband amplification, however, ICs can provide spectacular bandwidths from DC to microwave frequencies. The **HiMark DA1300** uses a GaAs process for fabricating heterojunction bipolar transistors, and can provide 20 dB of gain over a range of DC to 3 GHz. Packaging is in the standard SOT89 surface mount. On a more usual scale of wideband operation, the **NTE726** (NTE Electronics) has a typical gain of 75 dB at 4.5 MHz, and typical bandwidth of 100 Hz to more than 20 MHz. The same manufacturer also makes the **NTE7081**, a triple video amplifier for RGB monitors with 70 MHz bandwidth.

Yet another example comes from NEC, whose **mPC1663** is a differential amplifier using bipolar devices and intended for video amplification in high resolution equipment. A voltage gain of 300 is attainable with a bandwidth of 120 MHz, and, at a gain of 10, the bandwidth is 700 MHz. An external resistor can be used to control gain and no external frequency compensation components are needed.

Sine wave and other oscillator circuits

An **oscillator** is a circuit that can generate a frequency source such as a sine wave, square wave or pulse train. An oscillator is fundamentally a combination of a frequency-sensitive circuit (such as an LC circuit or a crystal) and a negative resistance (usually obtained by using an amplifier with positive feedback).

The circuits shown in Figures 7.24 to 7.27 are of BJT sine wave oscillators that operate at radio frequencies. The **Hartley** type of oscillator (Figure 7.24) uses a tapped coil and in the diagram, the resonant circuit is L_1C_1; the value of C_2 should be chosen so that the amount of positive feedback is not excessive, since, otherwise, a distorted waveform is created. R_1 should be chosen so that the transistor is just drawing current when C_1 is short-circuited.

The **Colpitts** type (Figure 7.25) uses a capacitor tap. Though these are not the only RF oscillator circuits, they are the circuits most commonly used for variable frequency oscillators. The Colpitts circuit is one that is often used for crystal-controlled oscillators. In the circuit illustrated the tapping is provided by the series combination of the two capacitors which are in parallel with the crystal.

- Note that oscillators of the same basic circuit can be drawn in a variety of different ways, and you often need to look closely to recognize an oscillator type, particularly if you have always seen it used in common-emitter form and you are confronted with circuits using common-base or common-collector mode.

The Colpitts oscillator can be found in common-base and common-collector formats as well as the common-base type illustrated in Figure 7.25,

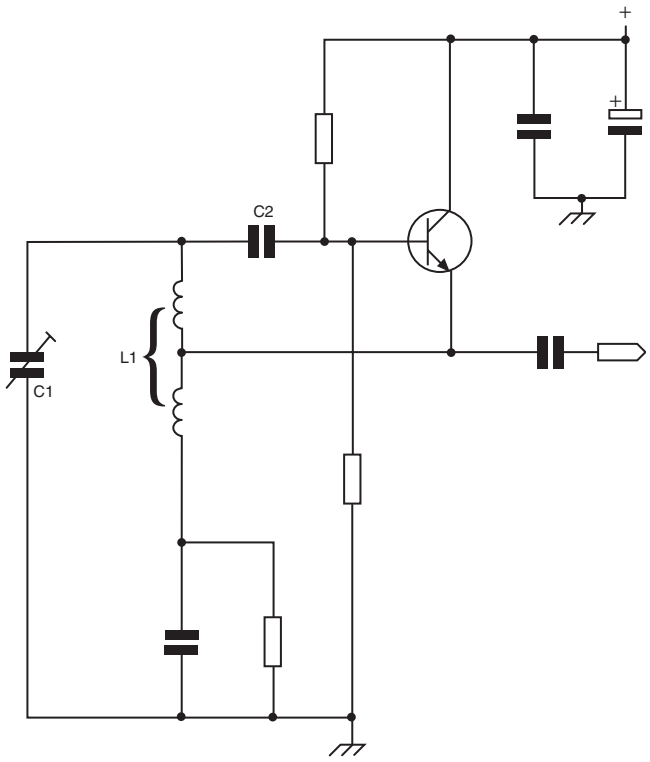


Figure 7.24

The basic Hartley oscillator.

but the circuit can always be recognized by the use of capacitors as a signal potential divider. One variation on the Colpitts design is sometimes referred to as a Clapp oscillator, and this type is illustrated in Figure 7.26.

Other crystal oscillators

Crystal oscillators generally are grouped into five classes, referred to by the abbreviations XO, VCXO, TCXO, OCXO and DTCXO. The **XO** type of oscillator is the simplest, using a straightforward circuit with no control circuits or any method of correcting frequency drift caused by changes

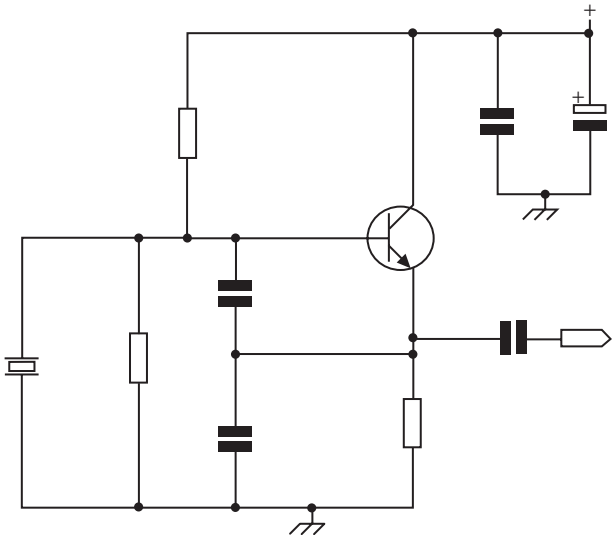
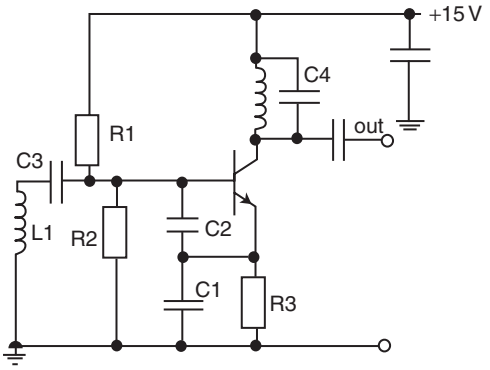


Figure 7.25

The Colpitts oscillator, in this example using common-emitter format and crystal-controlled.

Figure 7.26

A Clapp oscillator design.



of temperature. The **VCXO** circuits are arranged so that the frequency can be altered by applying a voltage applied to a circuit input. Preferably, the frequency change is directly proportional to the controlling voltage. **TCXO** (temperature compensated crystal oscillators) use a circuit in which a network of thermistors is used to sense ambient temperature and create

a correction voltage that reduces the change in frequency caused by changes in temperature. The **OCXO** (oven-controlled crystal oscillator) uses circuitry to maintain the crystal and any other temperature-sensitive circuits at a constant temperature, using a heated container (the oven). **DTCXO** is a more recent development, the digitally temperature controlled crystal oscillator.

The VCXO type of circuit alters the crystal frequency by means of a varactor diode fed with a tuning voltage. This type of circuit is used where frequency stability and low-phase noise are important; typical applications are spread spectrum systems. The VCXO type of circuit is valued for its ability to maintain a constant output frequency against changes in temperature or voltage supply even if the control signal is absent at intervals. Figure 7.27 illustrates two VCXO crystal-controlled oscillator circuits.

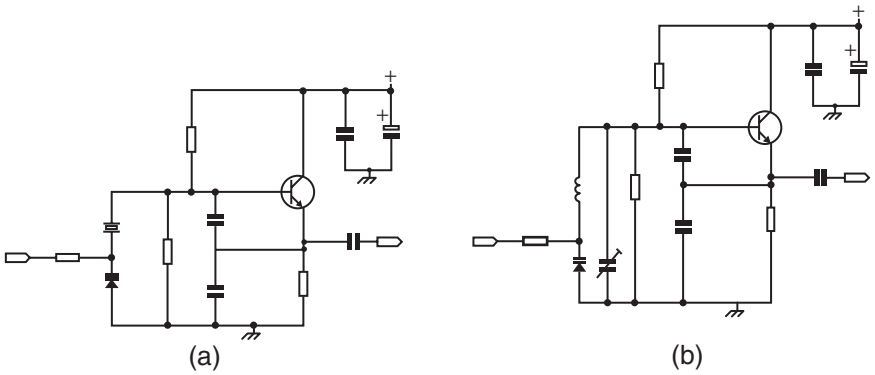


Figure 7.27

VCXO controlled oscillator circuits: **(a)** using the Colpitts configuration, **(b)** an alternative sometimes called a Clapp oscillator.

The frequency of the output need not be the fundamental crystal frequency, since most crystals will oscillate at higher harmonics (overtones) and harmonics can be selected at the output. Figure 7.28 shows a common-base Colpitts type of circuit in which the AT-cut crystal is used in overtone mode with a parallel inductor L2 resonating with the crystal shunt capacitance. Frequency multiplier stages can then be used to obtain still higher frequencies.

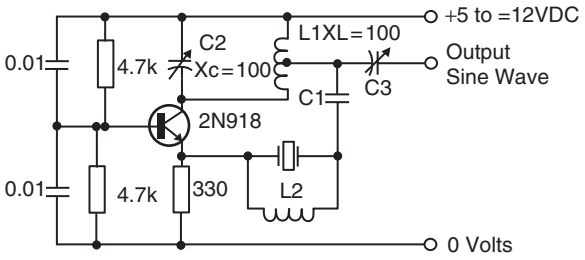


Figure 7.28

A form of Colpitts oscillator working in overtone mode.

The **Pierce** oscillator is a variant on the Colpitts design, and two crystal-controlled versions are shown in Figure 7.29, with a common-base format. The Pierce circuit is noted for excellent short-term stability and for its ability to operate over a large frequency range.

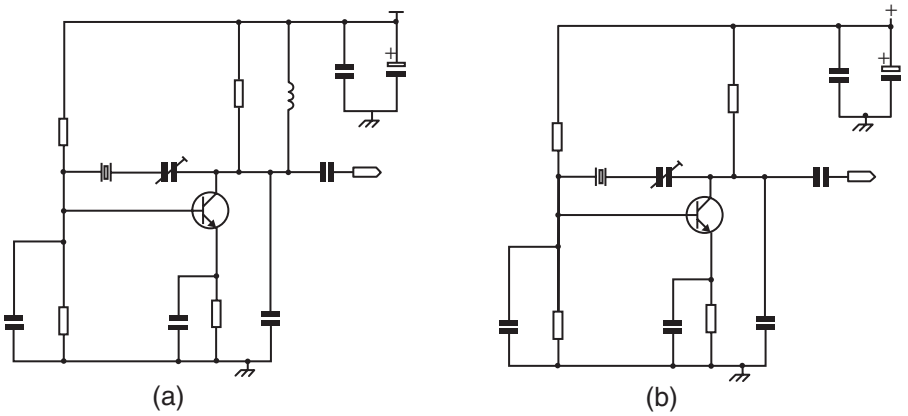


Figure 7.29

Two crystal-controlled Pierce oscillator circuits. The version shown in **(b)** is sometimes referred to as the Pierce Fund oscillator.

Another well-established variety of oscillator is the Butler, illustrated in Figure 7.30 in its VCO form. The Butler circuit allows higher-frequency response with good waveform shape and a lack of unwanted resonances

(*parasitics*) as compared to other designs. It is particularly favoured for crystal control using high-order overtones.

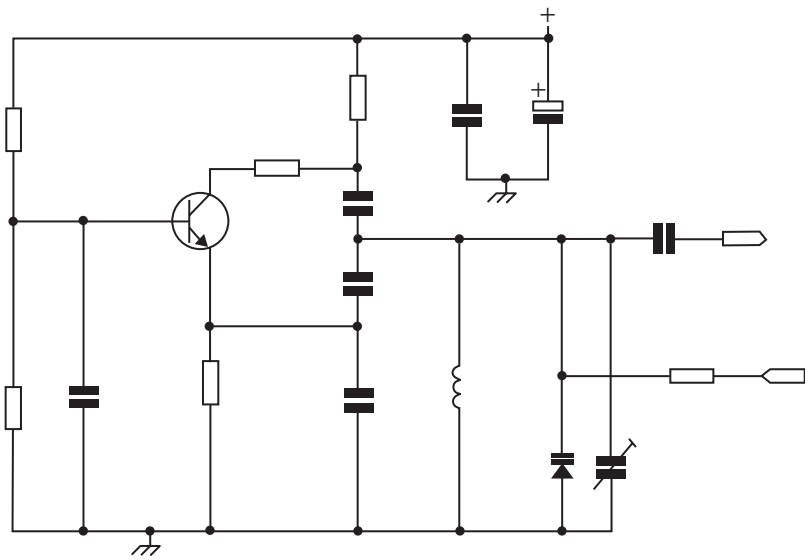


Figure 7.30

A Butler VCO circuit.

For low frequencies, oscillators such as the **Wien bridge** or **twin-T** types are extensively used. Circuits using discrete transistors are still used for these audio sine wave circuits, but the illustrations here indicate op-amps. Usable frequency ranges for these types of circuits are from 1 Hz, or lower, to around 1 MHz.

The Wien (sometimes spelled Wein) bridge, (Figure 7.31), is a frequency-selective circuit, and the corresponding oscillator uses this bridge circuit in a feedback loop. The amplitude of oscillation must be stabilized, and methods employed include the use of a light bulb, thermistor or, as illustrated, antiparallel diodes. The gain of the amplifying stage must be carefully controlled to ensure oscillation with a good sine wave form.

The twin-T oscillator (Figure 7.32) uses a twin-T notch filter in the feedback loop, and like the Wien bridge oscillator needs to be stabilized. In the illustration this also is done using a diode along with a potentiometer.

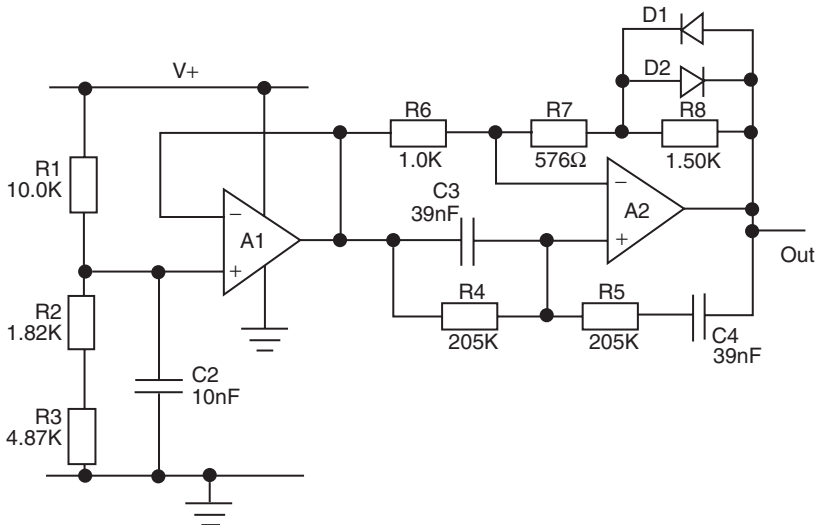


Figure 7.31

Typical modern Wien bridge oscillator circuit.

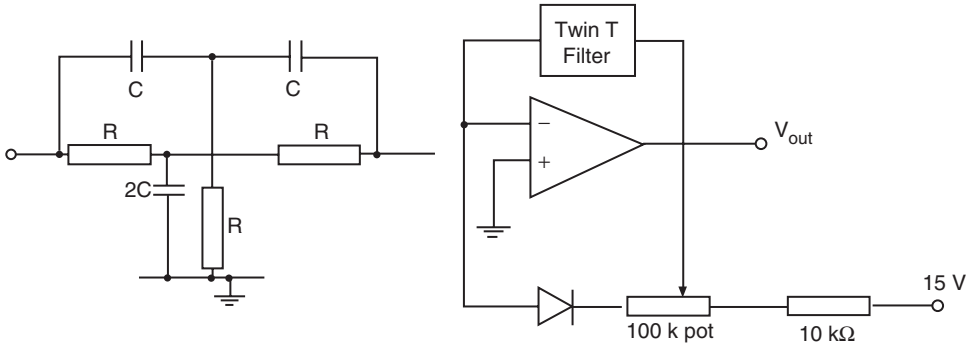


Figure 7.32

(a) Twin-T filter; **(b)** Typical Twin-T oscillator circuit.

Astable, monostable and bistable circuits

Untuned or aperiodic oscillators are important as generators of square and pulse waveforms. Figure 7.33a shows the familiar astable multivibrator with antiparallel diodes connected between base and emitter of each transistor. These additions have two important functions; they speed up switching by ensuring that neither transistor is driven into saturation, and they prevent the reverse biasing of the base-emitter junction that can lead to Zener-type conduction. This discrete circuit is still used because it allows the time constants to be different so that the square wave has a DC component, but it is more common to use the 555 timer for pulse generation (only one time constant needed) or cross-coupled NAND gates for high-speed operation.

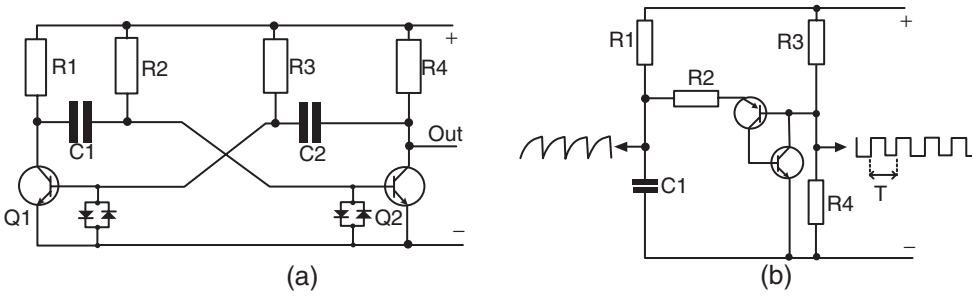


Figure 7.33

(a) A discrete transistor astable multivibrator, **(b)** serial multivibrator.

The less familiar serial multivibrator is shown in Figure 7.33b; this circuit uses only a single time constant and is a useful source of narrow pulses. It is seldom used nowadays because of the easy availability of ICs (such as the 555 timer and digital gates) that can obtain superior performance with little of no design complications.

When a pulse of a determined, or variable, width is required from any input (trigger) pulse, a monostable (also called one-shot) circuit must be used (Figure 7.34). The width of the monostable output pulse is determined by the time constant CR. The block diagram illustrates how a combination

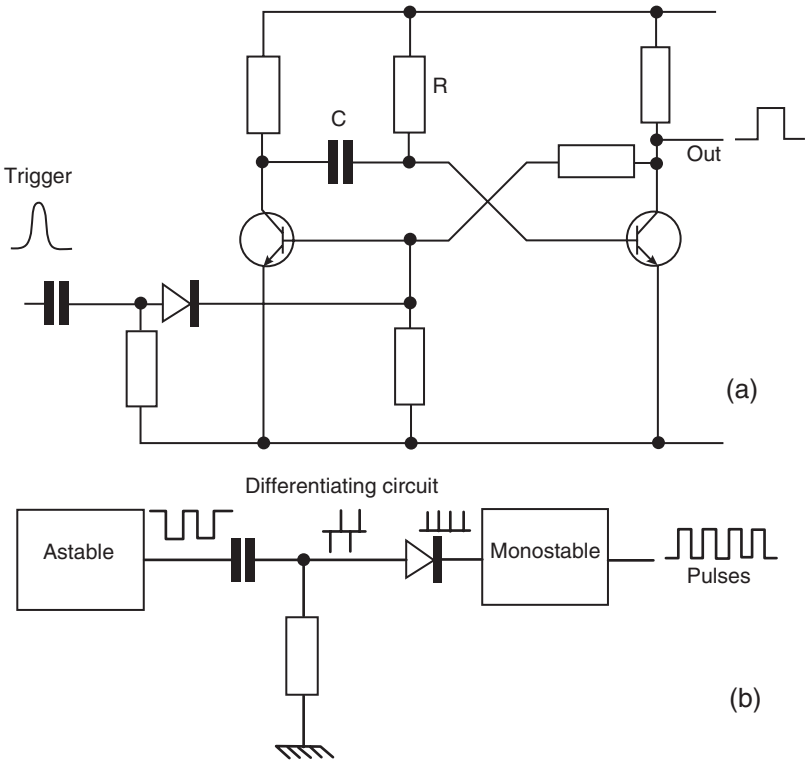


Figure 7.34

The monostable **(a)** and a block diagram **(b)** for a precise pulse generator.

of astable and monostable can form a useful pulse generator with control over both frequency and pulse width.

As ever, the familiar 555 timer can be used as a monostable, as can dedicated monostable ICs such as the 74S123 (also called 2602 or 26S02), 4528 and 4538.

Figure 7.35 shows the basic bistable circuit, now a rarity in the discrete form thanks to the low price of IC versions. The diodes are 'steering diodes' and their function is to guide the trigger pulse to the transistor that will cause the bistable to change over. The Schmitt trigger is illustrated in Figure 7.36; its utility is as a comparator and trigger stage which gives

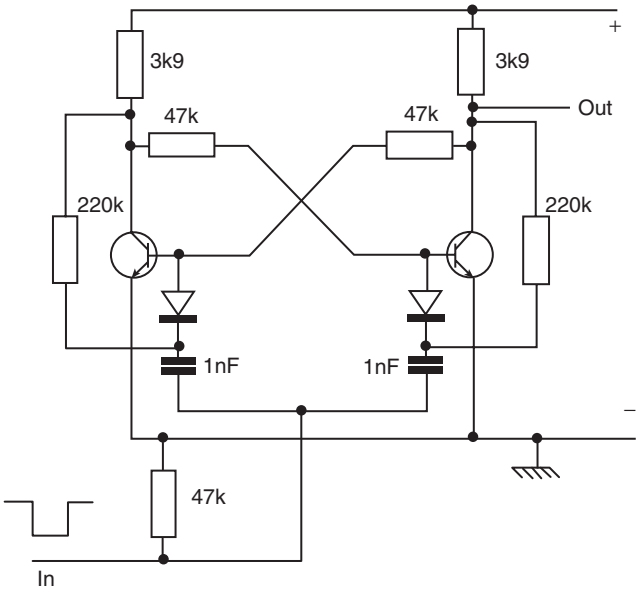


Figure 7.35

The bistable, or flip-flop. The output changes state (high to low or low to high) at each complete input pulse.

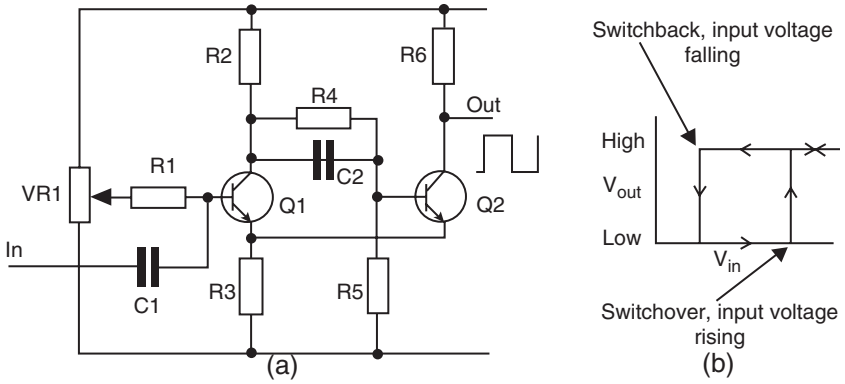


Figure 7.36

The Schmitt trigger circuit **(a)** and its characteristic **(b)**.

a sharply changing output from a slowly changing input. The hysteresis (voltage difference between the switching points) is a particularly valuable feature of this circuit. A circuit with hysteresis will switch positively in each direction with no tendency to 'flutter' or oscillation, so Schmitt trigger circuits are used extensively where electronic sensors have replaced purely mechanical devices such as thermostats.

Radio-frequency circuits

Radio-frequency circuits are represented here by only a few general examples, because the circuits and design methods that have to be used are fairly specialized, particularly for transmission; the reader who wishes more information on purely RF circuits is referred to the excellent amateur radio publications. At one time, a reference book would have shown discrete circuits for RF and IF receiver stages, but for conventional analogue radio reception these functions are now invariably carried out by ICs.

The Philips TEA5711 is an IC, now quite old (1992) and established, that integrates all the functions of an AM/FM radio from front end to AM detector and FM stereo output in a 32-pin DIL package. Figure 7.37 shows

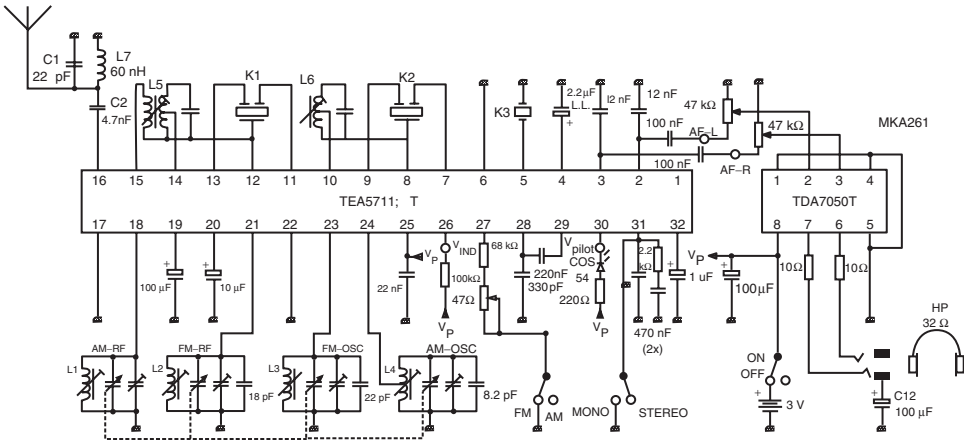


Figure 7.37

Philips TEA5711 application circuit.

a suggested application from the datasheet, using a separate TDA5070 output chip. The TEA5711 chip allows a wide range of supply voltage, from 1.8 V to 12 V, and has a low current consumption of 15 mA on AM and 16 mA on FM. The input sensitivity for FM is 2.0 μV, with high selectivity, and the FM input uses a high impedance MOSFET. The main applications are in portable radios.

The Chorus FS1010 from Frontier Silicon is a 179-pin BGA package that implements the most difficult sections of a DAB digital radio receiver, needing only an external RF stage, audio D to A, flash memory, keypad and display for a complete radio. The chip incorporates 16 K of ROM, 384 K of RAM, and two 8 K cache memories. It is likely that some day we shall have all of these functions on one chip, but until DAB radios sell in more significant numbers and until gaps in transmitting areas are filled in this is not likely to happen rapidly. One significant difference from radio as we used to know it is that there is no chance of using discrete components! Figure 7.38 shows the suggested block diagram for a DAB radio using the Chorus FS1010.

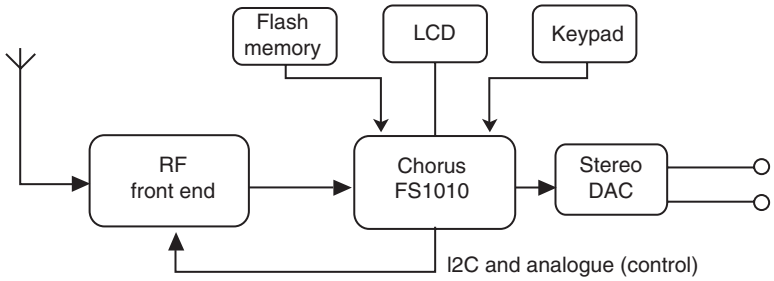


Figure 7.38

Outline of DAB receiver using the FS1010.

Figure 7.39 shows discrete component frequency multiplier and intermediate stages for transmitters; these are typical circuits for use in the amateur bands. In each example, the output is tuned to a frequency that is a multiple of the input frequency (usually from a crystal-controlled oscillator. Other common arrangements include the push-push multiplier circuit for even multiples, and the use of varactor diodes for harmonic generation at low power levels. Figure 7.40 illustrates a push-push multiplier for GHz frequencies.

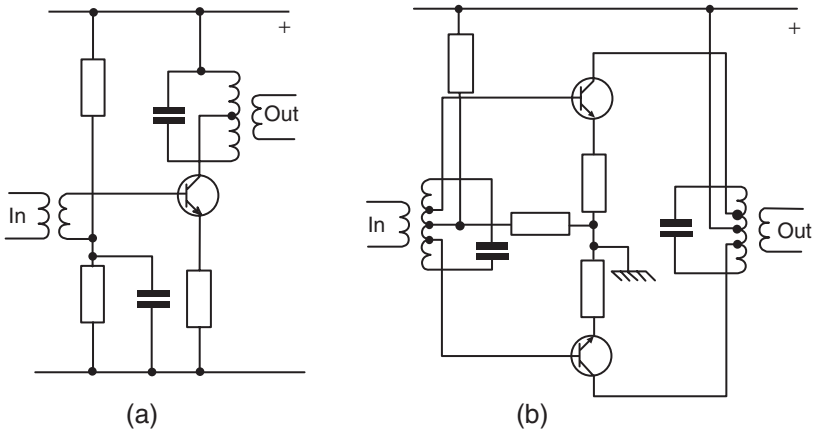


Figure 7.39

(a) Multiplier for even or odd multiples, **(b)** push-pull multiplier for odd multiples.

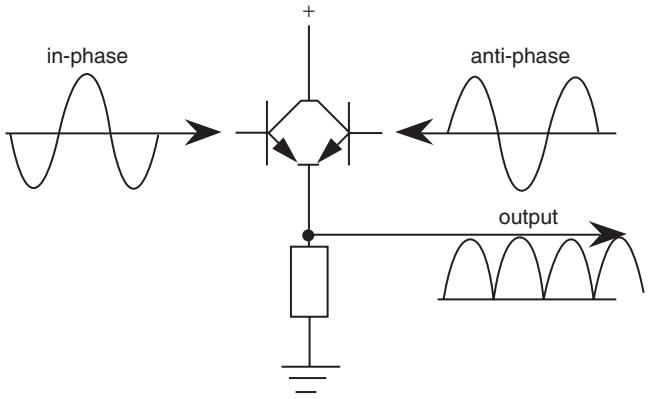


Figure 7.40

A push-push multiplier for GHz frequencies.

Much more specialized devices are used for microwave frequencies, and a specialist in semiconductors for these ranges is Tquint Semiconductors. As example, the Tquint TGC1430G multiplier is intended as a ×3 multiplier with an output in the range of 20–40 GHz using stripline architecture with GaAs semiconductors.

Figure 7.41 shows a selection of low-power output (power amplifier or PA) stages; again particularly for the amateur bands. Tuning inductors have been omitted for clarity. In both circuits some decoupling capacitors have not been shown – complete decoupling is essential. At the higher frequencies, circuit layout is critical, and the circuit diagram becomes less important than the physical layout. Transmitters which use variable frequency oscillators (VFO) will require broadband output stages as distinct from sharply tuned stages, and this precludes the use of Class C amplifiers (in which the transistor conducts only on signal peaks). Without a sharply tuned, high Q load, Class C operation introduces too much distortion (causing unwanted harmonics) and so Class B is preferable.

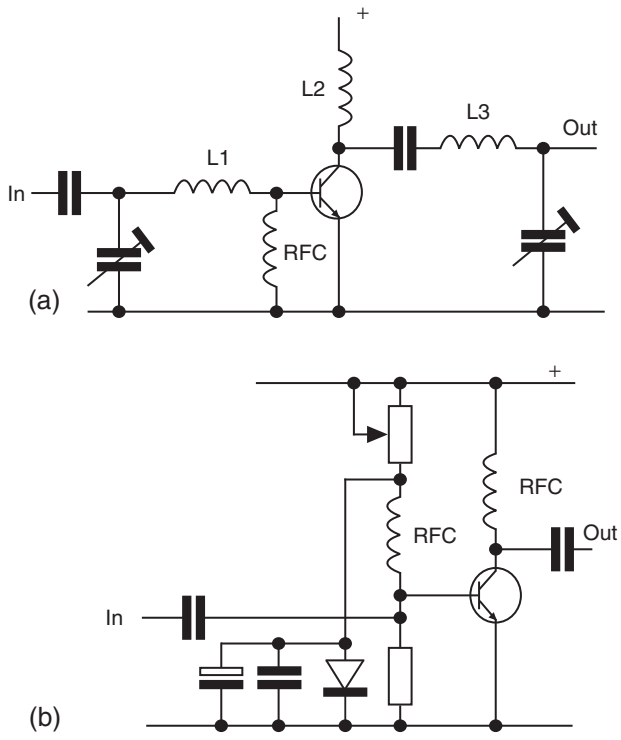


Figure 7.41

Power amplifiers for transistor transmitters: **(a)** a Class C single transistor PA stage, **(b)** a Class B design, necessary for single-sideband transmitters.

Modulation circuits

To carry information by radio or by digital signals requires some form of modulation and demodulation. For analogue radio use, amplitude modulation and frequency modulation are the most common techniques. Straightforward amplitude modulation produces two sidebands, with only one third of the total power in the sidebands, so double sideband AM is used virtually only for medium- and long-wave broadcasting. Short-wave communications use various forms of single sideband or suppressed carrier AM systems; VHF radio broadcasting uses wideband FM and other VHF communications use narrow-band FM.

Figure 7.42 shows two simple modulator circuits, excluding specialized types. Carrier suppression can be achieved by balanced modulators in which the bridge circuit enables the carrier frequency to be balanced out while leaving sideband frequencies unaffected. Sideband removal can be achieved by using crystal filters, a fairly straightforward technique which is applicable only if the transmitting frequency is fixed, or by using a phase-shift modulator which makes use of the phase shift that occurs during modulation.

Frequency modulation, unlike amplitude modulation, is carried out on the oscillator itself, so requiring reasonably linear operation of the stages following the oscillator. Figure 7.43 illustrates some types of discrete component demodulators. The AM demodulator (Figure 7.43a) uses a single diode. The time constant of C_1 with $R_1 + R_2$ must be long compared with the time of a carrier wave, but short compared with the time of the highest-frequency audio wave. The FM demodulator, or **discriminator**, (Figure 7.43b) is a ratio detector. An alternative to the older forms of FM discriminator is the pulse-counter type, a design that has been around for a long time but which was once too impracticable because of the limitations of early counting circuits. A pulse-counting discriminator operates by using a circuit that produces a narrow pulse on positive-going slopes of the input waveform (at the time of crossing the zero voltage point). The number of these pulses depends on the frequency of the input, so passing the pulses into a low-pass filter will produce the audio signal. The usual implementation of this type of discriminator is a PLL chip, and this is often a component of single-chip FM receivers.

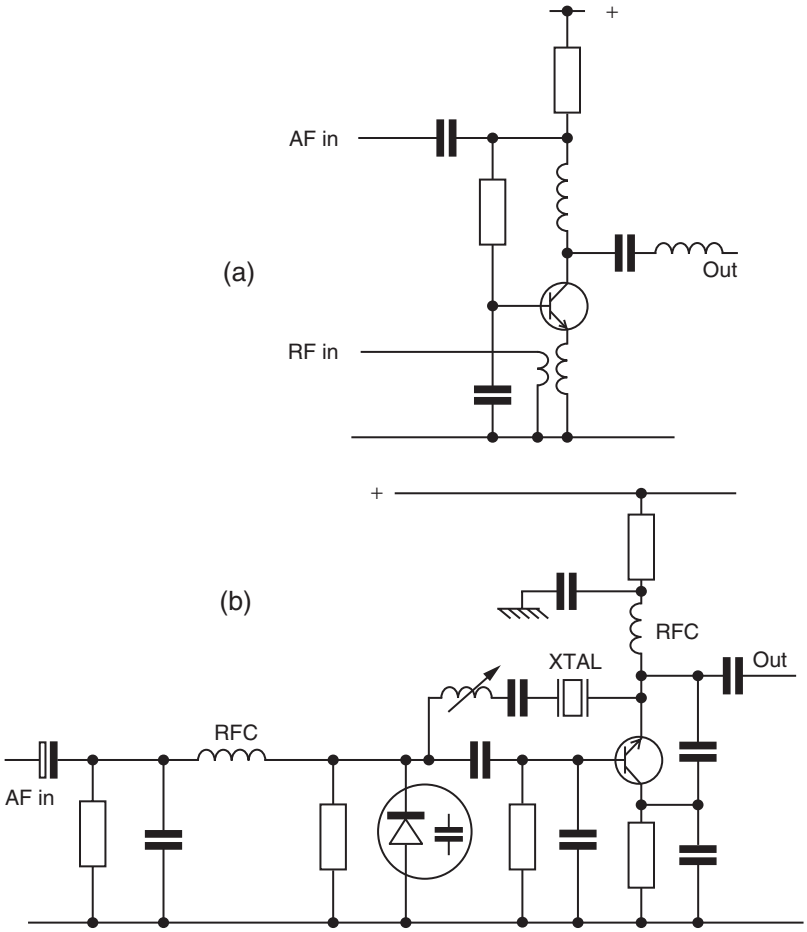


Figure 7.42

Two simple modulator circuits. **(a)** Collector-modulated stages for an AM transmitter, **(b)** a varactor diode FM modulator.

Pulse modulation systems are used extensively in applications ranging from data processing to radar. Pulse amplitude modulation and frequency modulation is essentially similar in nature to AM and FM of sine waves, and will not be considered here.

Forms of modulation peculiar to pulse operation are pulse-width modulation (**PWM**), pulse-position modulation (**PPM**) and pulse-code

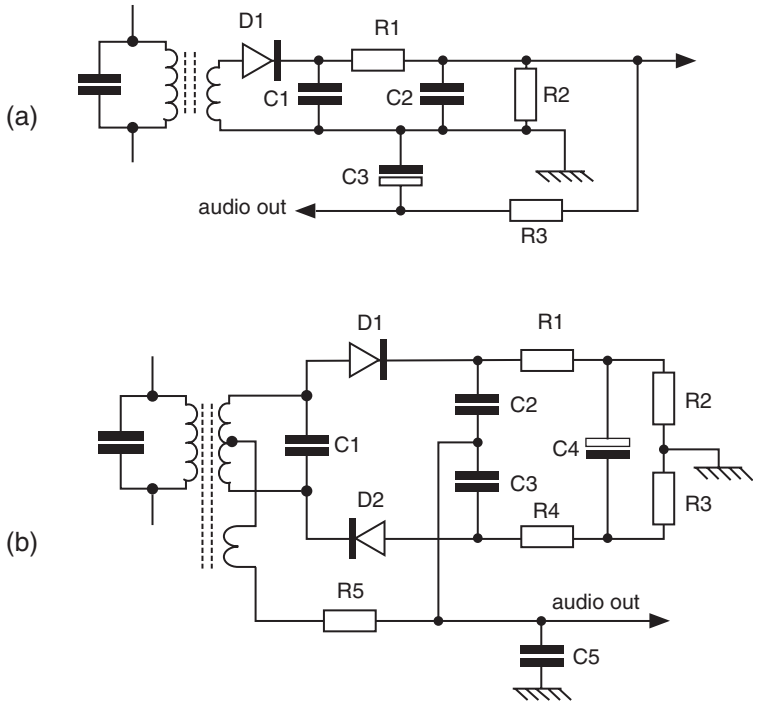


Figure 7.43

Demodulators: **(a)** AM, **(b)** FM.

modulation (**PCM**). A technique which is not a pulse modulation system but which is extensively used for coding slow pulse information is frequency shift keying (**FSK**) in which the high (logic 1) and low (logic 0) voltages of a pulse are represented by different audio frequencies. Pulse-code modulation is the system used for 'digital' coding systems.

Optical circuits

Figure 7.44 is concerned with optical circuits, including LED devices and light detectors. A current-limiting resistor must be used when driving a LED, and when the LED is operated from AC, a diode must always used to protect the LED from reverse voltages. The optocoupler is used to couple

signals at very different DC levels. This is useful for Triac firing, or for modulating the grid of a CRT, since DC signals can be transferred, which is not possible using a transformer.

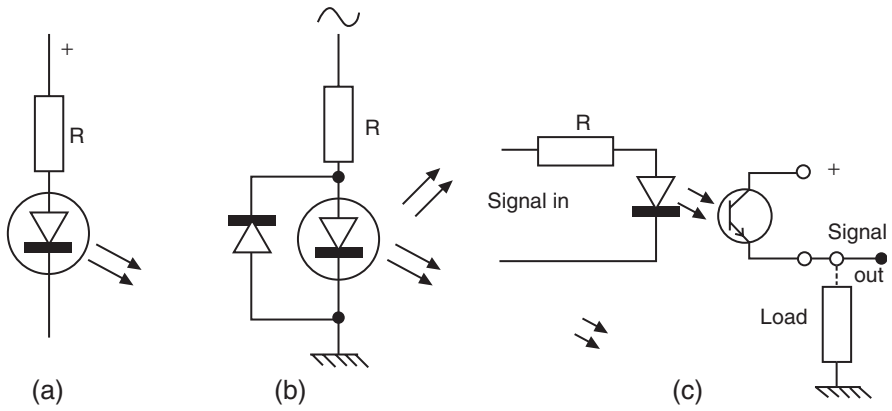


Figure 7.44

Optoelectronic circuits: **(a)** single LED, **(b)** AC operation, **(c)** optocoupler.

In addition there are many specialized optoelectronic assemblies available from suppliers such as Silonex.

Linear power supply circuits

The circuits shown in Figure 7.45 deal with the rectifier portion of linear power supply units. The Figure shows the no-load voltage output, and the relationship between DC load voltage, minimum voltage, and AC voltage at the transformer. Only capacitive input circuits have been shown, since choke-input filters are by now rather rare. All of these circuits will give a DC output on which is superimposed a fluctuating **ripple** voltage. For the half-wave rectifier, the ripple is at line frequency (50 Hz in the UK), but for the other two circuits the ripple frequency is at twice the line rate (100 Hz in the UK). Table 7.1 is a summary of the performance of these rectifier circuits.

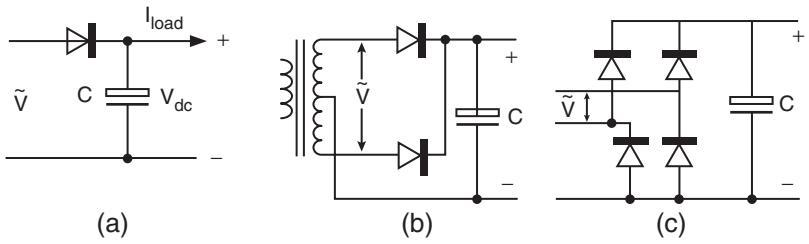


Figure 7.45

Rectifier circuits in detail: **(a)** half-wave, **(b)** biphase half wave, **(c)** full-wave bridge.

Table 7.1 Rectifier circuit performance

Parameter	Half-wave	Biphase half-wave	Full wave bridge
No-load V_{DC}	$1.4 \times V_{AC}$	$0.7 \times V_{AC}$	$1.4 \times V_{AC}$
Peak reverse voltage, diode	$2.8 \times V_{AC}$	$1.4 \times V_{AC}$	$1.4 \times V_{AC}$
Output V_{min}, full load	$0.44 \times V_{AC}$	$0.44 \times V_{AC}$	$0.44 \times V_{AC}$

V_{AC} is the AC input voltage

The relationship between the size of the reservoir capacitor and the peak-to-peak ripple voltage is given approximately by:

$$V = \frac{I_{DC} \times t}{C}$$

with I_{DC} equal to load current (in amperes), t in seconds (the time between voltage peaks) and C the reservoir capacitance in farads. V is then the peak-to-peak ripple voltage in volts. A more convenient set of units is I_{DC} in mA, t in ms, and C in μF, using the formula unchanged. Figure 7.46 shows a typical ripple waveform.

All power supplies that use the simple transformer-rectifier-capacitor circuit will provide an unregulated output, meaning that the output voltage will be affected by fluctuations in the mains voltage level and also by changes in the current drawn by the load. The internal resistance of the power

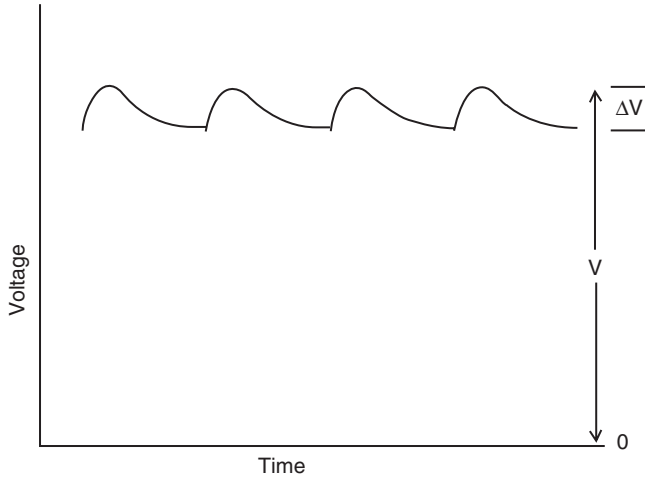


Figure 7.46

A typical ripple waveform, approximately a sawtooth.

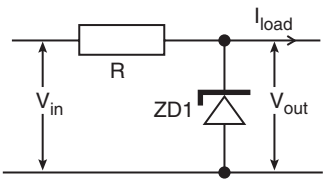
supply unit causes the second effect and can be the reason for instability in amplifier circuits, or of misfiring of pulse circuits. A regulator (stabilizer) circuit provides an output which, ideally, remains constant despite any reasonable fluctuation in the mains voltage and has zero internal resistance so that the output voltage is unaffected by the load current. IC regulators have been noted in Chapter 6, so what follows is partly a summary and partly a guide to simple regulator circuits.

Linear regulation is achieved by feeding into the regulator circuit a voltage that is higher than the planned output voltage even at the worst combination of circumstances – low mains voltage and maximum load current. The regulator then controls the voltage difference between input and output so that the output voltage is steady.

Figure 7.47 shows a simple Zener-diode regulator suitable for small scale circuits taking only a few milliamps. This is a **shunt regulating** circuit, so called because the regulator (the Zener diode) is in parallel (shunt) with the load. The value of the resistor R is such that there will be a 'holding' current of 2 mA flowing into the Zener diode even at the lowest input voltage and maximum signal current.

Figure 7.47

A simple Zener-diode shunt regulator.



The circuit shown in Figure 7.48 (sometimes known as the 'amplified-Zener') is a shunt regulator which does not depend on dissipating power in the Zener when the load current drops.

Figure 7.48

An 'amplified-Zener' or shunt-regulator circuit. The transistor dissipation is greatest when the load current is least.

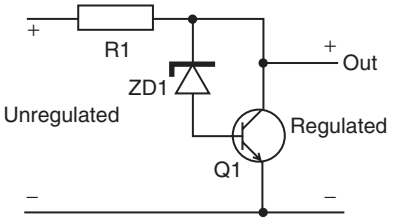
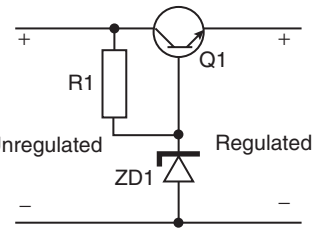


Figure 7.49 shows a simple series regulator, using a Zener diode to set the voltage at the base of an emitter follower.

Figure 7.49

A simple series regulator.



Switch-mode power supplies

Linear regulators are widely used, but they all suffer from the same set of drawbacks:

- They are most inefficient. It is unusual to find that more than 35% of the input energy reaches the load. The remainder is dissipated as heat. The inefficiency is greater for low-voltage high-current supplies.

- The mains transformer is invariably large. Its size tends to be inversely proportional to the operating frequency.

- The reservoir and smoothing capacitors need to be large to keep the ripple amplitude within acceptable bounds. This is particularly difficult for low-voltage supplies.

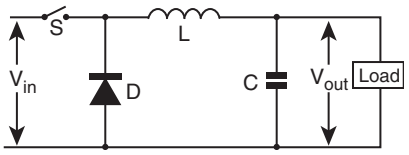
- Because the series transistor (or transistors) is operated in the linear mode it/they must be mounted on large heatsinks.

If the operating frequency can be increased significantly, both the transformer and the filter capacitors can be reduced in size. If the series transistor can be operated either cut-off or saturated, its dissipation will be greatly reduced. The power supply can then be made more efficient. Such operation can be achieved using a switch-mode power supply (**SMPS**). These circuits can operate with efficiencies as high as 85%.

The basic switching principle of the most common type of SMPS (sometimes called a Buck converter) is shown in Figure 7.50. When the switch is closed, current flows through the inductor or choke L to power the load and charge the capacitor C. When the switch is opened, the magnetic field that has been built up around L now collapses and induces an EMF into itself to keep the current flowing, but now through the flywheel or freewheel diode, D. The voltage across C now starts to fall as the load continues to draw current. If the switch is closed again the capacitor recharges. This switching cycle produces a high-frequency supply voltage.

Figure 7.50

The basic Buck converter circuit.



The duty cycle or switching sequence is shown in Figure 7.51 together with the output voltage V_{out} that it produces. Increasing the on-period will increase V_{out} whose average level is given by $V_{in} \times t_{on}/T$. V_{in} can be regulated by varying the mark-to-space ratio of the switching period. Any unwanted ripple can be filtered off in the usual way.

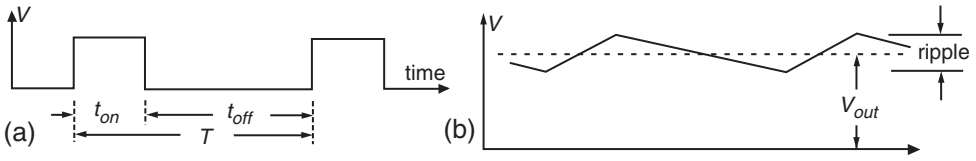


Figure 7.51

(a) Switching action and **(b)** waveform.

The typical SMPS whose block diagram is shown in Figure 7.52, consists of a mains rectifier with simple smoothing whose DC output is *chopped* or switched at a high frequency, using a transistor as the switch. For TV applications this switch is commonly driven at the line frequency of 15.625 kHz. The circuit generally needs some start-up arrangement that will ensure drive to the PWM switch when no DC output exists.

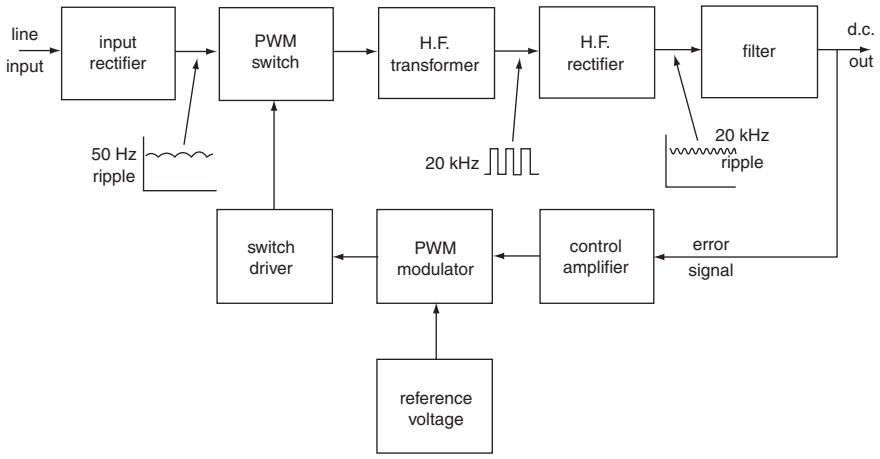


Figure 7.52

Block diagram of switched mode power supply.

For industrial applications or computer power supplies the switching frequency is usually in the order of 20–25 kHz. The chopped waveform is applied to the primary circuit of a high-frequency transformer that uses a

ferrite core for high efficiency. The signal voltage at the secondary is rectified and filtered to give the required DC output. This output is sensed by a control section that compares it with a reference voltage to produce a correction signal that is used in turn to change the mark-to-space ratio of the switching circuit to compensate for any variation of output voltage. This action is effectively pulse-width modulation.

The ripple frequency of 50 Hz at the input has been changed to a frequency of 20 kHz at the output so that the smoothing and filter capacitors can be reduced in value by the ratio 20 000:50, equal to 400 times. No electrolytic capacitors need be used, providing another bonus for reliability.

The oscillator/rectifier part of the circuit can be operated from a battery or any other DC input, so it becomes a device for converting DC from a high voltage to one at a lower level. Another option is to use a transformer whose input is the chopped high-frequency voltage, with several outputs that are rectified to produce DC at different voltage levels. Only one of these levels can be sampled to provide control, so only one output is stabilized against load fluctuations, though all are stabilized against input fluctuations.

SMPS circuits are universal in small computers, because of the need to regulate a low-voltage supply at a high current output. The usual circuitry rectifies the mains voltage directly (using no input transformer) so that the early stages operate at high voltage and low current, and a conventionally regulated supply is used to operate the control stages, ensuring that these are working at start-up. A transformer for the high-frequency voltage provides for isolation from the mains and for voltage output of +5 V (main output at high current) along with –5 V, +12 V and –12 V. A complete SMPS circuit can be obtained in IC form, and for higher outputs an IC can be used to control a high-power switching transistor. The low price of the PC computer units discourages home construction of switch-mode PSUs other than for unusual voltage outputs or applications.

The SMPS generates more radiated and line conducted noise than does a linear supply. This can be reduced to acceptable levels by using:

- mains input filters balanced to earth to give rejection of the switching frequency;

- suitable design of output filter;

- electrostatic screen between primary and secondary of the mains transformer;
- efficient screening of the complete unit.

Note that SMPS units for CRT-based analogue TV or monitor use generally make use of the horizontal sweep waveform for switching, and are closely integrated into the TV scan circuits, making them more specialized than the units used for computers. Switch-mode technology is also used for low-power supplies, particularly voltage converters (such as obtaining a 1.6 V regulated supply from a +5 V supply).

As an illustration of a lower-power type of circuit, Figure 7.53 shows a circuit published by ST Electronics for a 5 V, 6 W supply operating from a DC input that can be in the range 120 V to 375 V. Mains isolation

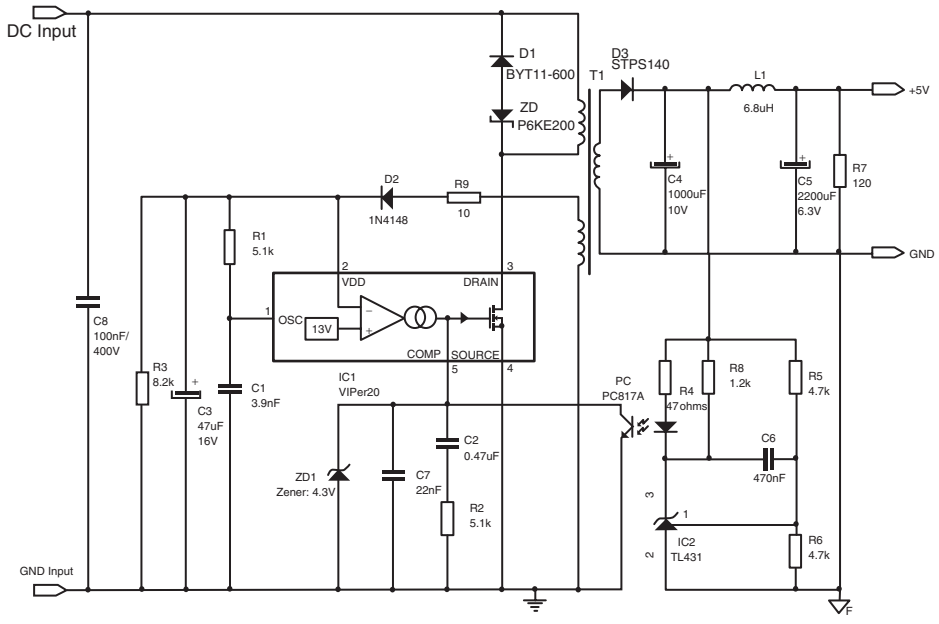


Figure 7.53

A miniature SMPS circuit. (Due to ST electronics.)

is achieved by using an opto link for feedback. The switching chip is the ST VIPer20, operating at 100 kHz. Details of this circuit, including components list and PCB trace, are available at the website address:

http://www.st.com/stonline/books/pdf/docs/6082.pdf

This page intentionally left blank

CHAPTER 8

SENSORS AND TRANSDUCERS

Introduction

Energy conversion components, as the name suggests, convert one form of energy into another, and those of interest for the purposes of this book convert other forms of energy either *into* electrical form or *from* that form. The importance of these components is that they allow electronic circuits to be used for detecting (sensing) and measuring other quantities such as acceleration, light flux and temperature, and they allow electronic circuits to form part of the control system for such quantities.

Conversion components are often classed as sensors or as transducers. The difference is often blurred, but in essence a sensor converts one form of energy into another with no regard to efficiency and is used for measurement purposes, and a transducer is used where the efficiency of transfer is more important, as in control systems or power generation. For example, a small anemometer propeller can sense wind speed, but a giant turbine with blades each weighing more than a ton each is needed to generate any useful power (at a cost in money and disruption that is quite disproportionate). For measurement purposes, the *resolution* of a sensor means the smallest change that can be measured for the detected quantity.

For any conversion component we can measure quantities that are termed **responsivity** and **detectivity**. The responsivity is a measure of the efficiency of the conversion and is defined as:

$$\frac{\text{output signal}}{\text{input signal}}$$

using whatever units are required for each form of energy. If the input signal and the output signal are both measured in watts, the responsivity

is equal to the efficiency and can be expressed as a percentage. If the units are different they must be quoted.

The detectivity measures the ability to detect the quantity that is being measured, and is defined as:

$$\frac{\text{S/N of output signal}}{\text{amplitude of input signal}}$$

where S/N means the signal-to-noise ratio for the output signal. This definition can be reworked into the more convenient form:

$$\frac{\text{responsivity}}{\text{output noise signal}}$$

There is a very large range of conversion components, and in this chapter we shall look only at some of the more common types that are used when one energy form is electrical. In addition, this chapter is concerned mainly with sensors, because high-efficiency energy conversion for power generation is outside our remit.

Strain and pressure

Stress is the force applied to a material per unit area, and is equal to pressure when the force is distributed evenly over an area. **Strain** means the fractional change in the dimensions of a material caused by stress, and is, up to a limit (the elastic limit), proportional to the stress on that component (Hooke's law). The strain on a material can be sensed by fastening a **strain gauge** to the material (the *host* for the strain gauge). The resistive strain gauge consists of a piece of thin wire whose change of length is measured by sensing its change of resistance using a bridge circuit. Metal wire strain gauges are insensitive, and semiconductor strain gauges are used wherever the operating temperature permits. The semiconductor strip (Figure 8.1a), is laid on an insulator such as mica and is passivated to prevent atmospheric contamination. Either type of strain gauge is fastened to its host using epoxy resin. The bridge circuit that is used for measurement must be temperature compensated, because the changes in resistance caused by temperature changes will be as large as, or larger than, those caused by strain.

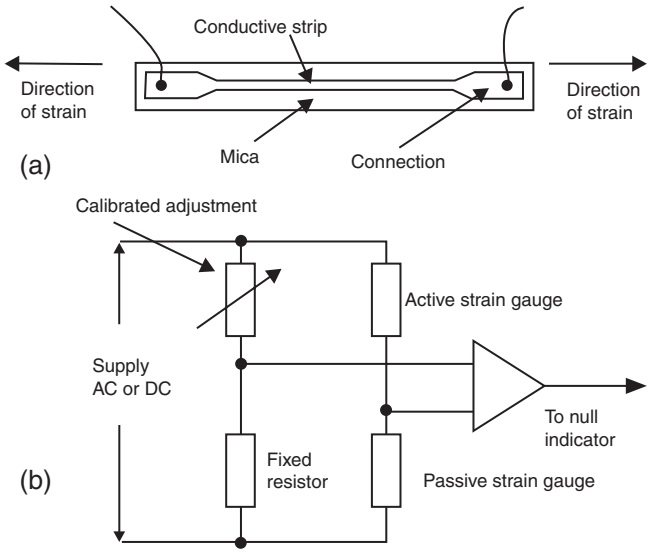


Figure 8.1

The physical appearance of a strain gauge **(a)**, and a typical bridge circuit that compensates for temperature effects **(b)**.

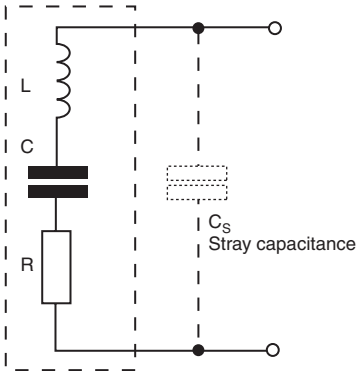
This is achieved by using a circuit (Figure 8.1b) which uses two identical strain gauges, only one of which is subjected to strain.

For rapidly changing strains, the **piezoelectric strain gauge** provides larger signal outputs. The piezoelectric crystal, using a material such as barium titanate, is metallized on opposite sides, and the signal output is a voltage between these sides that is generated when the crystal is strained (because of the displacement of the ions in the crystal). The voltage can be comparatively high, even into the kV region, but the output impedance is very large and is also capacitive, with an equivalent circuit as illustrated in Figure 8.2. This makes the sensor less useful for slowly changing strains but ideal for vibrational strains.

Pressure changes in gases and liquids can be measured by monitoring their direct effects on a piezoelectric crystal (as in a crystal microphone) or by way of a diaphragm. The use of a diaphragm separates the sensor from the liquid or gas whose pressure is to be measured, and also allows a greater choice of sensing methods. For example, the diaphragm can form one plate

Figure 8.2

The equivalent circuit of a piezoelectric crystal. In this equivalent the inductance is very high, the capacitance low and the series resistance almost negligible.



of a capacitor in a resonant circuit so that a change of pressure will cause a change in the resonant frequency, or the diaphragm can carry a coil which is within the field of a permanent magnet, so that changes of pressure will induce currents in the coil. The capacitor form can be used to detect slow pressure changes, but the electromagnetic type will detect only rapid changes.

The absolute measurement of low gas pressures is carried out using heat transfer measurements (the **Pirani gauge**) or ion current readings (the **ionization gauge**). The Pirani gauge is useful for pressures in the region of 1 mm of mercury to 10^{-3} mm of mercury, and uses the principle that the rate of heat conducted from a hot wire to a cold one, in a gas atmosphere, will drop as the gas pressure decreases. The heat energy reaching the cold wire is detected by measuring its resistance using a bridge circuit.

The ionization gauge, in various forms, is used for pressures below 10^{-3} mm of mercury down to the lowest pressures that are obtainable. Its operating principle consists of a beam of electrons ionizing the gas that remains in a vacuum, and the ions of gas can be attracted to a plate and the ion current measured. The lower the pressure, the lower the ion current. Both Pirani and ionization gauges require calibration if they are to be used for precise measurement.

Direction and motion

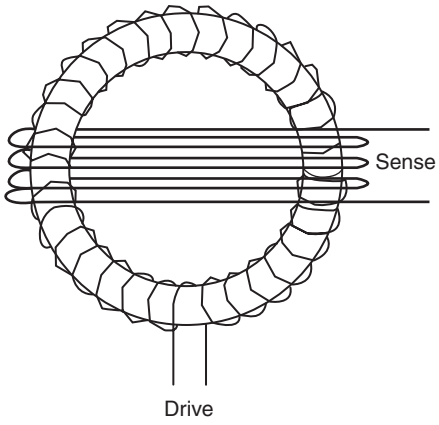
On a large scale, direction on the surface of the Earth can be sensed by the strength and direction of the Earth's magnetic field using a compass needle.

This can be adapted to provide an electrical output, but much more sensitive devices are available and are used for other applications that require sensing magnetic field.

The **fluxgate magnetometer** is an older type of device (dating back to the 1940s) which is still in use because it is easy to construct and yet has remarkably high sensitivity, particularly when combined with modern digital control circuitry. Its operating principle is illustrated in Figure 8.3, showing a toroid with one winding, the drive coil, around the core, and another, the sense winding, over the **outside** of the toroid, not threading through the toroid. The controlling circuitry will increase the current in the drive coil in one direction until the sense winding indicates non-linearity, and this is repeated with the current reversed in the drive coil. With no external field applied, saturation would occur at the same value of drive current in either direction, but when an external field is present, the values are different. This difference is proportional to the external field strength, and the sensitivity is greatest along the axis of the toroid, so the direction of the field can be sensed as well as the field strength.

Figure 8.3

Principle of the fluxgate magnetometer.



A more recent device is the Hall effect sensor (Figure 8.4). Constant current is passed through the semiconductor crystal to which a magnetic field is also applied. The force caused by a magnetic field affecting particles in the conductor creates an electric field which can be measured as a voltage between the faces of the crystal. The effect exists in all carriers, but is much greater in semiconductors. This voltage, the **Hall voltage**, is proportional to the size of the magnetic field.

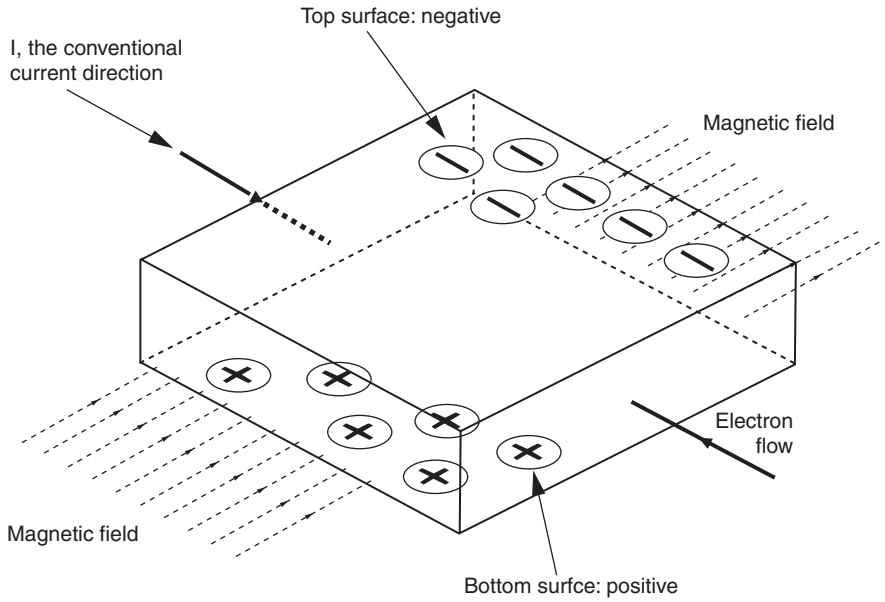


Figure 8.4

The Hall effect on a semiconductor crystal.

The most recent class of magnetic field measurements make use of **giant magnetoresistance**, an effect discovered in 1988. Magnetoresistance is the change in resistance of a material in the presence of a magnetic field, and the 'giant' part of the name comes from the discovery of a way of constructing magnetoresistive devices with much greater sensitivity, using many thin layers of magnetic materials. The detection of magnetic field with such devices is simply achieved by measuring the resistance, and the devices are used in magnetometers and also in computer hard-disk drives, in land-mine detection and many other applications.

Distance sensing on a large scale can be carried out using a radar system, sending out a pulse of waves (in the millimetre wavelength range) and measuring the time needed for reflected waves to return. The same principle can be applied (in the form of **sonar**) using sound waves in water, but the differences in wave speed require the time measurement methods to be very different.

For small-scale measurements, such as distances along a drawing board, much simpler methods can be used involving resistive, capacitive or inductive sensors. The most precise measurement can use laser interferometry, but this is outside the scope of this chapter. The most common method that is used involves the component referred to as the linear variable differential transformer, or LVDT (Figure 8.5). This device consists of three fixed coils in a moveable core, of which one coil is energized with AC. As the core is moved, the difference between the amounts of AC picked up in the other two coils will change, and this can be sensed by a phase-sensitive detector.

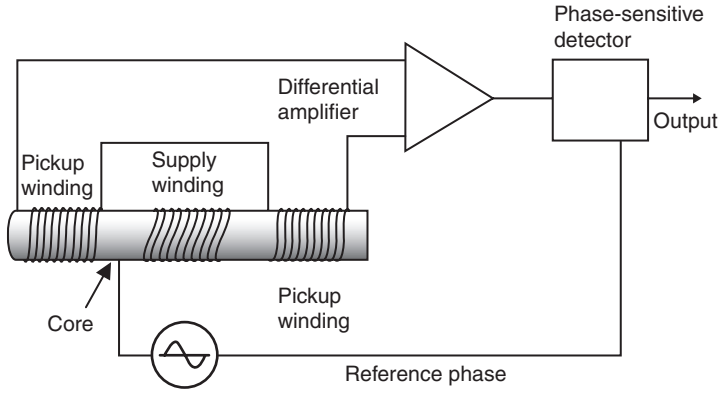


Figure 8.5

The principle of the linear variable differential transformer or LVDT.

For a reasonable range of movement, the output is linearly proportional to the distance moved, and because the core is not in contact with the coils the amount of friction can be very low (compared with that of a potentiometer, for example). The resolution is high and the output signal can be large. The device is rugged and is not readily damaged by excessive movement. Commercially available LVDTs will sense motions in ranges from ±l mm to ±65 mm, using an AC supply of typically 5 kHz. Some types contain an integral oscillator so that a DC supply can be used.

Optical encoders, which give a digital output, can be used for linear or rotary motion. The operating principle, illustrated in Figure 8.6, uses a transparent slide which has a printed pattern. A photocell is placed behind each track so that as the slide moves the outputs from the photocells will

represent a number in binary code. The code can be 8-4-2-1 or the more useful Grey code (Table 8.1) in which only one digit changes for each increment in the number.

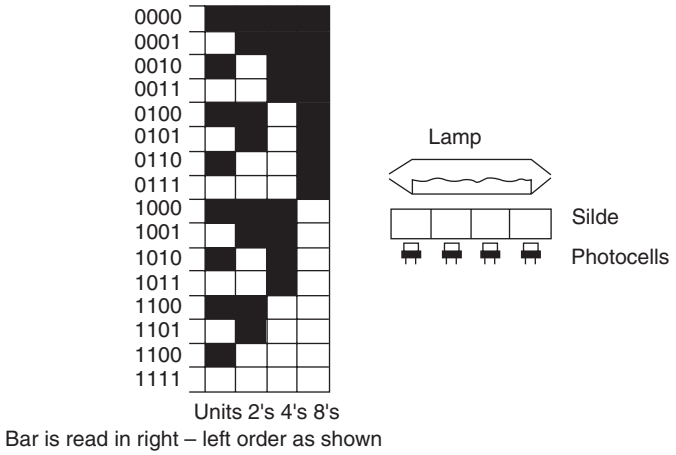


Figure 8.6

Optical encoder operating principle for four bits. Each position of the encoder will provide a binary number output from the photocells.

Table 8.1 The Grey scale and 8-4-2-1 binary compared

Denary	8-4-2-1 binary	Grey code	Denary	8-4-2-1 binary	Grey code
0	0000	0000	8	1000	1100
1	0001	0001	9	1001	1101
2	0010	0011	10	1010	1111
3	0011	0010	11	1011	1110
4	0100	0110	12	1100	1010
5	0101	0111	13	1101	1011
6	0110	0101	14	1110	1001
7	0111	0100	15	1111	1000

Note: Temperature (°C)/EMF (mV) data assume that the cold junction will be at a temperature of 0°C. Only the useful range is shown.

A familiar application of optical encoding is the computer mouse, which uses rotary encoders to provide positional information in two directions at

right angles. Figure 8.7 shows the interior view of a mouse, with the ball in contact with the spindle and toothed optical disc units. Movement of the ball will cause rotation of the spindles, so rotating the optical discs and altering the light passed between an LED and a photodiode. The photodiode outputs from the two encoders are combined into a serial position code that is sent to the host computer.

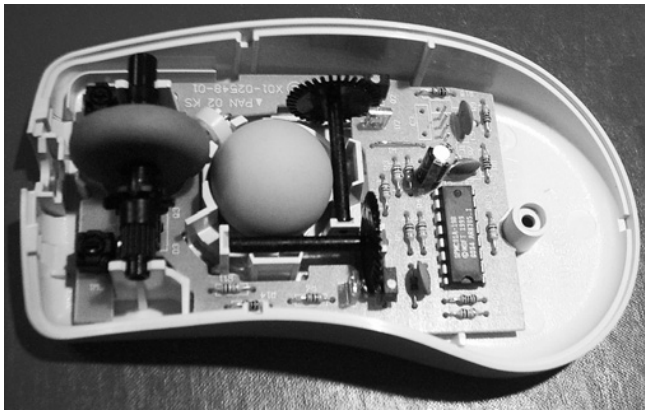


Figure 8.7

Interior of a computer mouse, showing encoders. This is a wheel-mouse, using a separate third encoder to allow wheel movement to scroll the computer screen. (Photo courtesy of John Dunton.)

An optical grating is another useful method of measuring small amplitudes of movement. The principle involved (Figure 8.8) is to use two identical grating patterns on transparent material. When one strip moves relative to the other, the transmitted light intensity will vary in a sine wave pattern, and the peaks can be counted. The number of peaks counted is directly proportional to the amount of movement, and can be calculated from the number of lines per centimetre in the grating and the colour of light being transmitted.

Light, UV and IR radiation

Light is an electromagnetic wave of the same type as radio waves but of much shorter wavelength, corresponding to a much higher frequency,

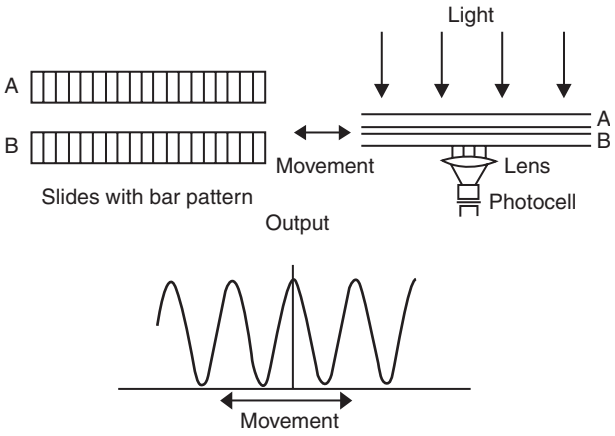


Figure 8.8

Using optical gratings to sense movement. Movement equivalent to the width of one spacing will produce a complete sine wave of output, making this very sensitive when gratings are ruled with several hundred lines per millimetre.

which also means that the energy content is higher. Wavelengths are generally measured in nanometres (10^{-9} m). The devices that are used to generate and to detect light are therefore very different from those used for radio waves, even for the shortest wavelengths of radio waves that we can use. Light detectors are collectively known as photosensors, and of these the older devices such as selenium cells and photoemissive cells are seldom used now. Photoresistors or light-dependent resistors (LDRs) are made from materials whose resistance value changes when light strikes the material.

The most familiar type is the cadmium sulphide cell, named after the light-dependent resistive material that is used. The cadmium sulphide is deposited as a zig-zag thread on an insulator, with a connector at each end (Figure 8.9) and is encapsulated in transparent resin to protect the material. Unlike most semiconductor devices, this cell can withstand a considerable range of temperatures and also of voltages. The cell is most sensitive to colours in the orange–red range, and is extensively used for controllers in oil-fired boilers. Table 8.2 shows the characteristics of the ORP12 type, and Figure 8.10 shows a typical application circuit. As illustrated, this will switch the relay on when the light level reaching the LDR is rising, but by

Figure 8.9

A typical LDR or photoconductive cell using cadmium sulphide.

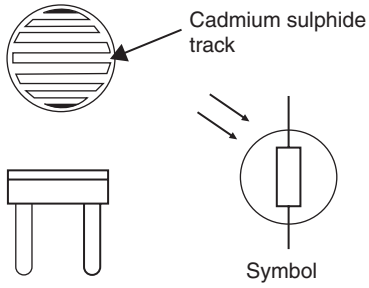


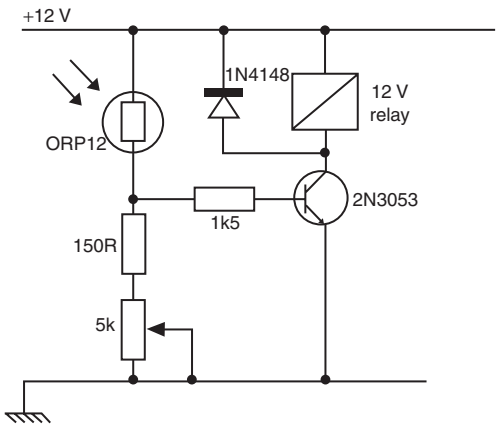
Table 8.2 Characteristics of the ORP12 photoconductive cell

Peak spectral response	610 nm
Cell resistance at 50 lux	2.4 kΩ
Cell resistance at 1000 lux	130 Ω
Dark resistance	10 MΩ
Max. voltage (DC or peak AC)	10 V
Max. dissipation at 25°C	200 mW
Typical resistance rise time	5 ms
Typical resistance fall time	350 ms

Data courtesy of RS Components Ltd.

Figure 8.10

A circuit in which light falling on the cell operates a relay. This can be easily altered to operate the relay when the light decreases or is switched off.



reversing the positions of the LDR and the resistive arm, the circuit can be made to switch as the light intensity is falling, making this a dusk-detector.

Photodiodes and **phototransistors** make use of semiconductor junctions rather than the action of a bulk material. A silicon photodiode has no opaque covering, so that light can affect the junction, and it is used with reverse bias. The effect of light is to cause the reverse current to increase, but the sensitivity is low, of the order of a fraction of a μA of current for each mW/cm^{-2} of illumination power. For a normal range of illumination (darkened room to sunlit room) this corresponds to currents that range from 2 nA to 100 μA. The dark current for a photodiode is the minimum figure, corresponding to the reverse leakage current. Figure 8.11 shows a typical application circuit, in which the resistor R is set at a value that will determine the sensitivity – a typical value is 1 MQ. In such a circuit, a graph of output plotted against input illumination is reasonably linear, and the response time is around 250 ns, so the diode can be used for detecting beams that are modulated with frequencies up to the video region.

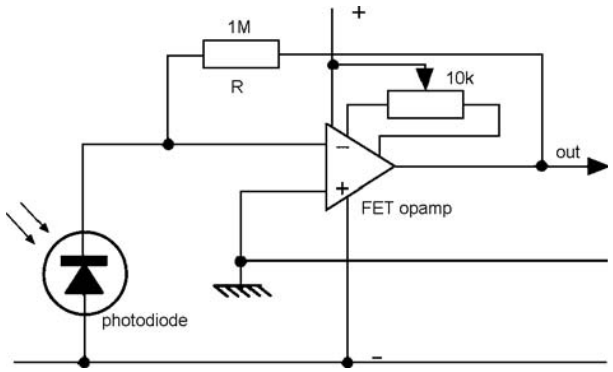


Figure 8.11

A typical circuit that makes use of a photodiode. The peak response for this photodiode is in the near infra-red.

A phototransistor is very closely allied to the photodiode, and is constructed so that light can reach the collector–base junction. This reverse-biased junction will have a low current in darkness, but the effect of light will be to increase the current, and this current will in turn be amplified by the normal transistor action. This makes the phototransistor much more sensitive than a photodiode, often by a factor as large as 1000. The response time is,

however, correspondingly poorer, and is measured in microseconds rather than in nanoseconds. This makes phototransistors unsuitable for detecting modulated light beams unless the modulation is at a comparatively low frequency.

The opposite conversion, electrical input into light output, has for many years been represented by filament lamps. These have been replaced for all but a few applications by light-emitting diodes (LEDs). The LED is formed from a semiconductor material for which the forward voltage of a junction is large, and the energy radiated when an electron meets a hole is in a suitable range (which can be visible or infra-red). The most usual materials are gallium arsenide or gallium phosphide and the most common colours of visible emitted light are red and green, with yellow obtained by mixing the other two.

Electrically, the forward voltage for conduction is around 2.0 V and the maximum permitted reverse voltage is low, often as low as 3.0 V. This makes it important to avoid reversed connections and also to avoid the possibility of AC reaching the LED. The light intensity depends on the amount of forward current, usually in the range of 2 mA to 30 mA depending on the size of the LED junction. A full table of characteristics is included in Chapter 5.

Opto-isolators (see also Chapter 5) are components that contain, typically, both LED and phototransistor in one package, so that an electrical input to the LED will provide an electrical output from the phototransistor, but with complete electrical isolation between the circuits. This isolation is used, for example, to allow the cathode of an instrument CRT to be modulated from a low-voltage circuit when the DC level of the cathode is –7 kV or more. Opto-isolators using triacs along with LEDs are also obtainable, but you should not use such devices to isolate mains voltages unless this is permitted by the local electricity supply company. For some applications, only a mechanical relay is permitted as a method of isolation.

Temperature

Heat is a form of energy, and temperature is the level of heat; the relation of temperature to heat is similar to that of voltage to electrical energy.

The heat content of an object cannot be measured in any simple way, but changes in the heat content are proportional to temperature change. Temperature sensors operate by using materials whose characteristics are affected by temperature. In this chapter we shall ignore simple mechanical devices such as bimetallic strips and concentrate on non-mechanical sensors.

The **thermocouple** (Figure 8.12) is one of the traditional methods of measuring temperature electrically, and its action depends on the contact potential that always exists when two dissimilar metals are joined. This contact potential cannot be directly measured for one junction, because in any circuit at least two junctions must exist. When these two junctions are at different temperatures, however, a potential difference (voltage) can be measured, and its value depends on the size of the temperature difference. The relationship between temperature difference and output voltage is not linear (Figure 8.12b) though a small part of the curve can be assumed to be linear.

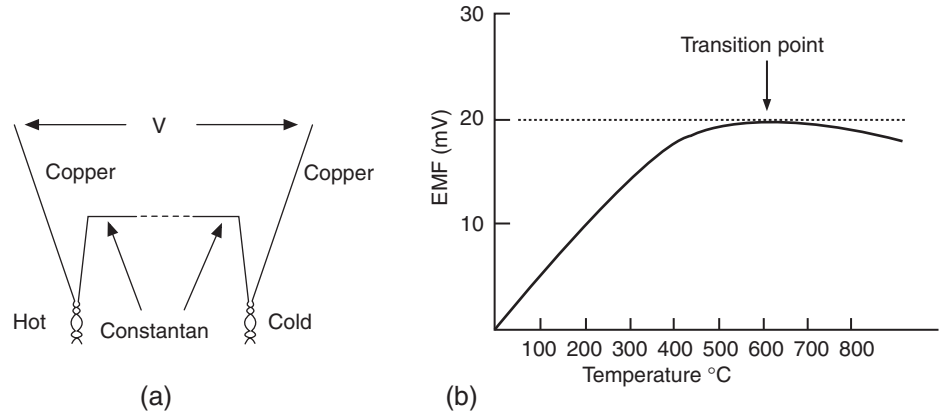


Figure 8.12

(a) The construction of a thermocouple and **(b)** a typical graph of output plotted against temperature.

Most combinations of metals show the type of characteristic that is illustrated, in which the output voltage peaks at a point called the **transition temperature**, and such thermocouples are normally used below this turnover point. The output of any thermocouple is of the order of a few millivolts, and a suitable DC amplifier must be used – either an operational amplifier or (preferably because of its better stability) a chopper type.

The outstanding advantages of thermocouples are that the sensing element can be very small, and that the temperature range can extend to high levels. Temperature/EMF data for three traditional thermocouple types are noted in Table 8.3, assuming that the cold junction will be at a temperature of 0°C.

Table 8.3 The thermoelectric behaviour of metals

Temperature (°C)	Copper/Constantan	Iron/Constantan	Platinum/ Plat.Rhodium
−20	−0.75	−1.03	
−10	−0.38	−0.52	
0	0.00	0.00	0.00
10	0.39	0.52	0.05
20	0.79	1.05	0.11
30	1.19	1.58	0.17
40	1.61	2.12	0.23
50	2.04	2.66	0.30
60	2.47	3.20	0.36
70	2.91	3.75	0.43
80	3.36	4.30	0.50
90	3.81	4.85	0.57
100	4.28	5.40	0.64
200	9.28	10.99	1.46
300	14.86	16.57	2.39
400	20.87	22.08	3.40
500		27.59	4.46
600		33.28	5.57
700		39.30	6.74
800		45.71	7.95
900		52.28	9.21
1000		58.23	10.51
1200			13.22
1500			17.46

Note: Temperature (°C)/EMF (mV) data assume that the cold junction will be at a temperature of 0°C. Only the useful range is shown.

Commercially obtainable thermocouples are normally used with **cold junction compensation** circuits which correct the measured voltage to allow for the temperature of the cold junction being at air temperature. When this system is used you must not alter the thermocouple connections in any way

(adding more cable, for example) except as instructed by the manufacturer, because such alterations would make the compensation invalid.

Metal resistance thermometers make use of the change of resistivity of a metal as its temperature changes. For most metals, the temperature coefficient of resistivity is positive, so the resistance increases as the temperature increases, and values of temperature coefficient are around 4×10^{-3}. The standard form of resistance thermometer uses platinum as the sensing metal in the temperature range of $-270°C$ to $+660°C$. The resistance change is measured using a bridge circuit, and for precise work a set of dummy leads is used in series with the balance resistor (Figure 8.13) to compensate for the effect of temperature on the leads to the platinum element – this set of dummy leads runs parallel to the leads to the platinum element and is subject to the same temperature changes. A typical sensitivity figure is 0.4 Ω change of resistance per Celsius degree of temperature.

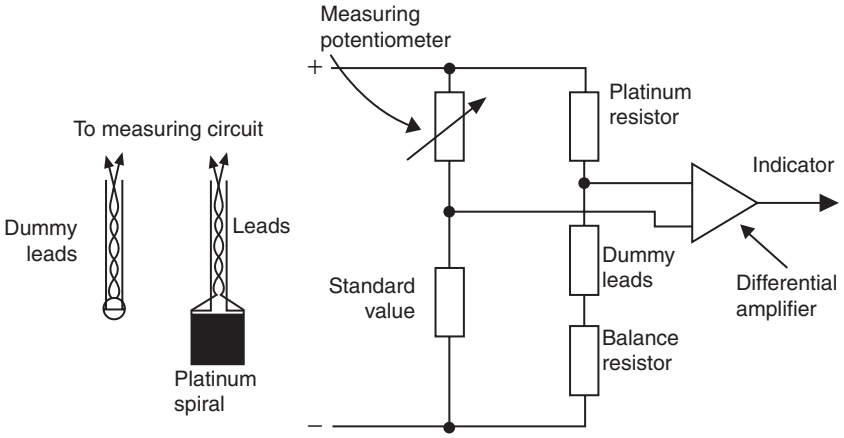


Figure 8.13

The form of bridge measuring circuit used for a platinum resistance thermometer, indicating how the dummy leads are connected.

Semiconductors have much larger temperature coefficients of resistivity, and materials termed **rare-earth oxides** have characteristics that are particularly useful. These materials are used to form thermistors, now the most common method of measuring temperatures by electrical means. A typical resistance/temperature characteristic is illustrated in Figure 8.14, showing

the non-linear shape and the negative characteristic (resistance decreases as temperature increases). Thermistors are normally used in the circuit such as that of Figure 8.15 – if part of the series resistor is made variable it can be used as a range setting.

Figure 8.14

A typical thermistor characteristic with negative temperature coefficient and a non-linear shape.

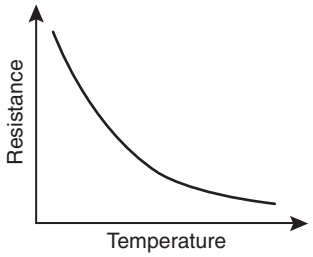
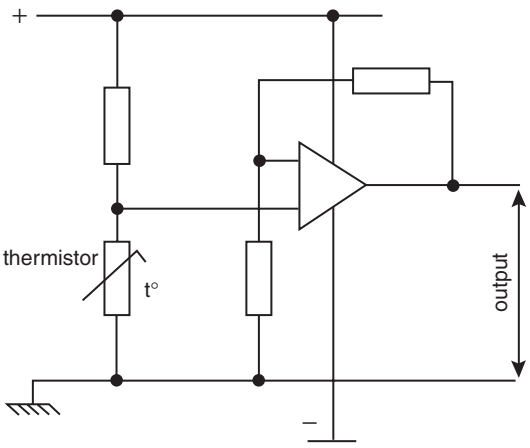


Figure 8.15

Using an operational amplifier in a thermistor temperature sensing circuit. The sensitivity can be adjusted by altering the feedback ratio.



Pyroelectric films are not so well known as temperature sensors, but are now widely available because of their sensitivity to radiated heat (infra-red), with the upshot that they are widely used in PIR (passive infra-red) alarm systems. The most favoured material at the time of writing is lithium tantalate, though several types of plastics will also provide pyroelectric effect.

A typical pyroelectric detector is constructed like a capacitor with one metal plate and one plate of the pyroelectric material that has been metallized on one side. The DC voltage between the plates will alter according to the amount of infra-red radiation striking the pyroelectric material. Because the

source impedance is very high, the output from a pyroelectric capacitor must be to a MOSFET, and most commercially available pyroelectric cells incorporate a MOSFet along with the cell.

Figure 8.16 shows a typical circuit in which the internal MOSFET is used in a source-follower circuit. This is followed by two stages of amplification and the output is used in a threshold circuit (IC_3, IC_4), which in turn will trigger a transistor. The LED is normally used to show that the unit is operating correctly, and the output will be used in an alarm circuit which can be turned on or off as required.

Sound

Sound is a wave motion in air or other materials, and sound sensors (microphones) depend on the vibration of materials (such as diaphragms) caused by the sound wave. Sound in this sense includes the type of waves that are called **ultrasonic**, meaning that they are in a frequency range that is above 20 kHz. Such waves cannot be detected by the human ear, but they are identical in nature to the sound waves that can be heard, allowing for the effects caused by the high frequency (such as being more directional).

Microphones can be designed to be pressure operated, velocity operated, or a mixture of both, and the differences are important. A microphone that is purely pressure operated will be **omnidirectional**; it will pick up sound equally well no matter from what direction the sound arrives. This is because air pressure is a non-directional (scalar) quantity. Velocity-operated microphones, by contrast, are directional, and have a maximum response when pointed in the direction from which the wave arrives. Whether a microphone is pressure or velocity operated depends much more on the constructional methods than the method that is used to sense sound. For example, if a diaphragm is open on both sides, it will be affected mainly by air velocity, but if it is open on one side only it is affected mainly by pressure. In microphones, any of a number of sensing systems can be utilized along with a diaphragm.

One very common system is the moving-iron (or variable-reluctance) type (Figure 8.17), in which the diaphragm carries a soft-iron armature and can move this armature between the poles of a magnet. This movement alters

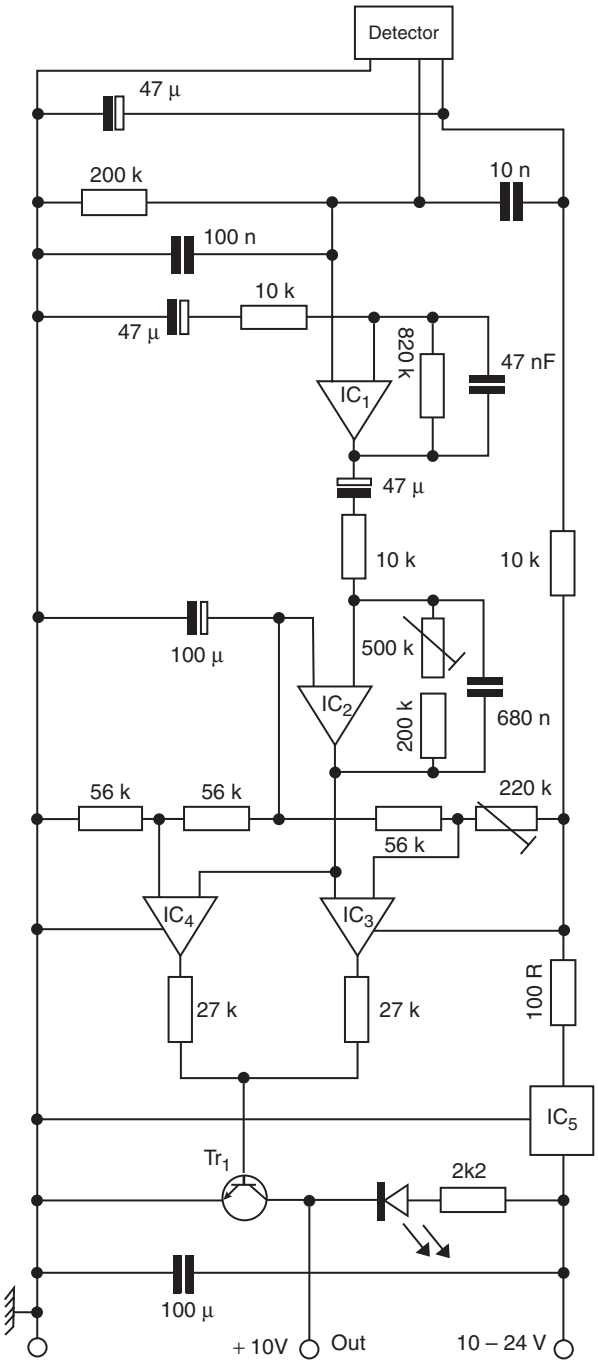
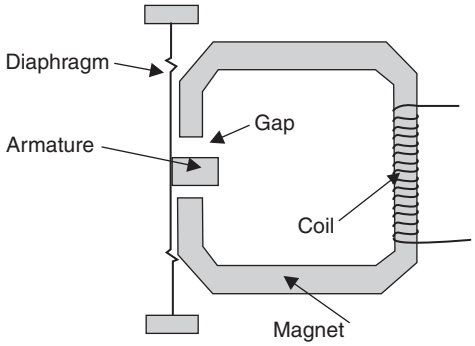


Figure 8.16

Typical circuitry for a pyroelectric burglar alarm.

Figure 8.17

The variable reluctance (or moving iron) microphone principle. The same construction can be used for an earphone or loudspeaker.



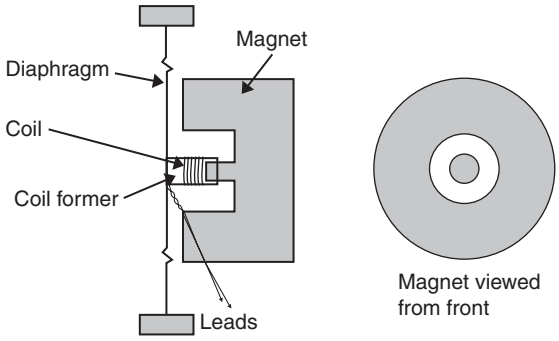
the magnetic flux and this in turn causes a voltage to be induced in a coil wound over the magnet. The output can be reasonably linear, but only if the shape of the gap and the armature are carefully designed. The voltage output level can be of the order of 50 mV, with an impedance of several hundred ohms. This in microphone terms is a fairly high impedance.

The moving-coil microphone uses a powerful magnet surrounding a small coil that is attached to the diaphragm (Figure 8.18). The impedance is very low, and this type of microphone is much less liable to pick up hum than is the variable-reluctance type. The output voltage is, however, much lower, of the order of a few millivolts. A more specialized type, the **ribbon microphone**, combines diaphragm and coil into one thin metal strip held between the poles of a long magnet. The output level and impedance figures are so low that such microphones often use a built-in transformer or a preamplifier. The ribbon microphone is very directional and is used extensively in broadcasting from noisy locations. Piezoelectric microphones can make use of a diaphragm connected to a piezoelectric crystal, or can be constructed so that the sound waves affect the crystal directly. The impedance level is very high, and the output is also high. Piezoelectric microphones are useful detectors, but are not favoured for sound recording or broadcasting because of poor linearity and distortion.

Capacitor microphones have always been highly regarded for high-quality sound recording. The principle employed is that one plate of a capacitor is also a diaphragm that is vibrated. This in turn will alter the capacitance between the plates, and if the capacitor is polarized by connecting one plate to a voltage (via a large-value resistor) the plate voltage will vary as the sound wave amplitude varies, providing an output. The impedance is very high

Figure 8.18

Principle of the moving-coil microphone, which is used also for earphones and loudspeakers.



and the output is low. The older form of capacitor microphone was always highly regarded, but the problems associated with the high impedance and the need for a polarizing voltage caused most manufacturers to use other systems.

The use of electrets has revived the capacitor microphone. An **electret** is the capacitor equivalent of a permanent magnet, a material that is permanently electrostatically charged. This eliminates the need for a polarizing voltage, and therefore a capacitor microphone can be constructed with a slab of electret material (metallized on one side for a connection) and a separate vibrating diaphragm. Using a MOSFET preamplifier in conjunction with an electret microphone element allows microphones of very good quality to be constructed at modest cost. Pyroelectric films (see earlier) can also be used in microphones of the capacitor type.

The conversion of electrical waves to sound or ultrasound involves the use of loudspeakers, earphone or crystal transducers and is outside the scope of this book. For a very full treatment of loudspeaker types and theory, see *Newnes Audio and Hi-Fi Engineer's Pocket Book* (Vivian Capel), Butterworth-Heinemann 1994. For further reading on conversion components and methods, see *Sensors and Transducers* (Ian Sinclair), Butterworth-Heinemann 1992.

This page intentionally left blank

CHAPTER 9

DIGITAL LOGIC

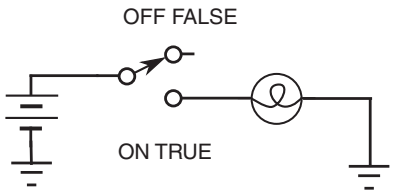
Introduction

Systems that use two or more discrete levels of voltage or current to represent states are referred to as **digital**. The vast majority of such systems use two levels only, so they are **binary** in nature. In a binary system the states are usually named TRUE and FALSE; by convention TRUE is equated with 1 (one) and FALSE with 0 (zero).

In order to represent these states electrically we could use a switch. When the switch is open no current flows, the zero (0) state, when the switch is closed current flows, representing one (1). Current flow can be indicated by a lamp or a meter (Figure 9.1).

Figure 9.1

A battery, switch and lamp.

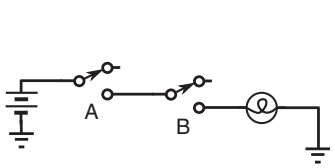


Given that the states of the system can be set, represented and indicated by these simple means we can extend the concept to include decisions based on reason, that is *deterministic* logic systems.

The basic decision-making logic operations or gates are AND, OR and NOT. These were defined in the 19th century by the mathematician/philosopher George Boole, hence the name *Boolean Algebra* given to the system of writing logic equations. The three elementary logic gates are simple but from these even the most complex systems can be built.

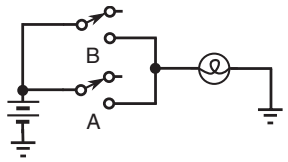
Boolean algebra provides a compact representation of logic functions. The notation of Boolean algebra is similar to that of arithmetic, OR is represented as +, AND is represented as ×. For example A + B × C is A OR B AND C. The NOT or inverse of a variable is indicated by a bar above the variable, for example $\overline{A}$. In a fashion similar to arithmetic there are rules for the use of brackets (parentheses) and the order of evaluation of expressions. AND, like multiplication, is distributive and so we can write A × B + A × C as A(B + C). It is usual to write A × B as AB, leaving out the dot as we do in normal algebra.

Figure 9.2 shows two lamp circuits, and it should be clear that the lamp will light only when **both** the series connected switches are closed, therefore A AND B, but will light when **either** of the parallel connected switches is closed, that is A OR B.



AND

A	B	LAMP
OFF	OFF	OFF
OFF	ON	OFF
ON	OFF	OFF
ON	ON	ON



OR

A	B	LAMP
OFF	OFF	OFF
OFF	ON	ON
ON	OFF	ON
ON	ON	ON

Figure 9.2

Switch implementation of AND and OR gates and their truth tables.

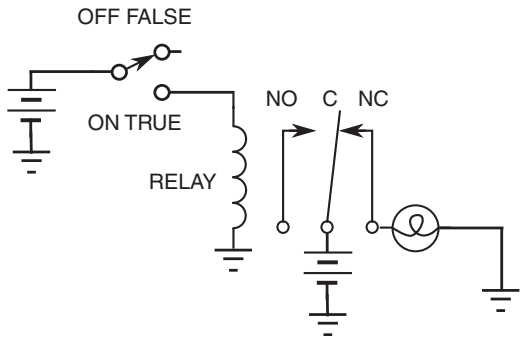
Tables of input and output states are called **truth tables**, since they indicate the relationships in the system that give TRUE or FALSE outputs.

The third of the basic logic gates is the NOT gate or inverter. This gate provides an output, which is the inverse of its input; Figure 9.3 shows an inverter implemented with a relay.

It is possible to build all higher-level logic functions from combinations of these three basic gates (Figure 9.4). So far we have looked at switches and

Figure 9.3

A relay inverter.



indicators as examples of logic states and basic gate functions but in order to build real systems we need electrically controlled switches so that the logic output of one stage can be the input of another following stage.

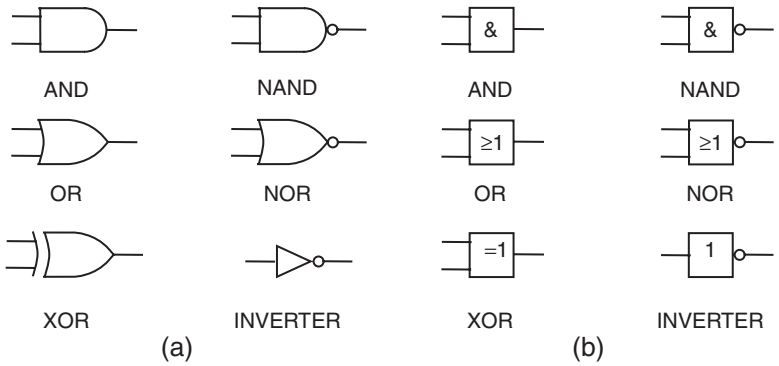


Figure 9.4

(a) IEEE logic symbols for gates, **(b)** IEC symbol for gates.

Relay logic was developed from the telegraph and railway signal technology of the early 20th century. The first general purpose programmable computer built to solve numerical problems was constructed in the early 1940s by Konrad Zuse in Germany, using thousands of relays; the program was stored on punched tape.

The relay circuit shown here uses a relay with change-over contacts; this allows either the inverted or the non-inverted output to be selected for each gate (Figure 9.3).

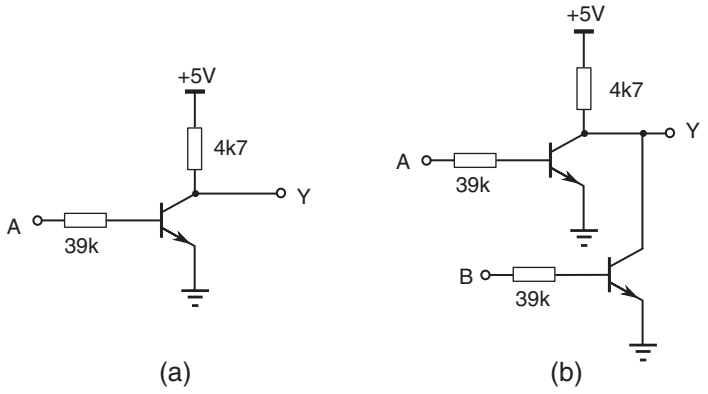


Figure 9.5

(a) Transistor inverter, **(b)** transistor NOR gate.

Relay logic still has some uses, but these are mostly safety applications like machine guards on machine tools and interlocks on dangerous systems of one sort or another. The majority of logic systems use semiconductor switches, either FET or bipolar transistors.

Konrad Zuse's electromechanical computer has been largely overshadowed by the almost simultaneous development in the UK and USA of electronic computers based on thermionic valves used as switches; these were in turn overtaken by the invention of the transistor in 1948 at AT&T Bell labs.

Using transistors as switches made it possible to build much more complex logic circuits, which used less power and which were faster than previous systems. Integrated circuits further increased the complexity that could be achieved, and single chip transistor counts passed 100 million transistors on a single chip by the year 2000.

We have introduced the three fundamental logic gates, AND, OR and NOT. The combination of AND or OR functions followed by NOT gates are named NAND, NOT AND and NOR, NOT OR. NAND and NOR functions implemented with fewer transistors in most logic systems, in fact in CMOS and TTL an AND gate would be implemented as a NAND gate followed by an inverter.

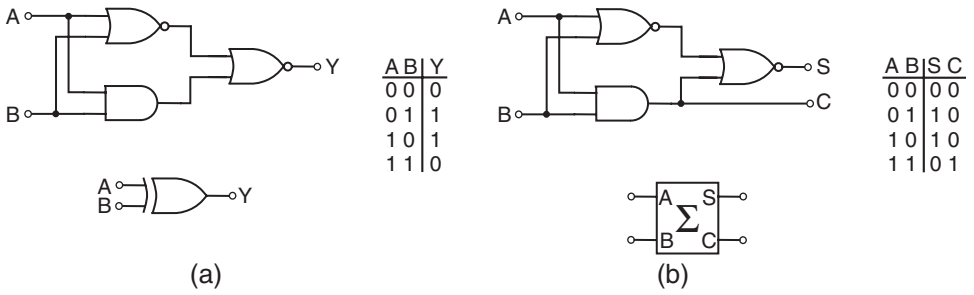


Figure 9.6

(a) XOR gate and **(b)** half adder.

There is one more logic function that is referred to as a gate and which is fundamentally arithmetic in nature. This is the exclusive-or gate (XOR) which is a one bit adder. Figure 9.6 shows the circuit of a half adder, which, without the carry output, is a XOR gate; the output of the XOR gate is only 1 when the inputs are different, hence the name exclusive. As the truth table shows, the XOR gate output is the binary sum of its inputs. The XNOR gate is an XOR followed by an inverter.

The XOR function is written as $\oplus$, thus $\overline{AB + \overline{(A + B)}} = A \oplus B$.

The half adder is so called because there is no carry input; full adders require two half adders per bit to provide a carry input and a carry output. Note that XOR and half adder functions can have only two inputs and are not like OR or AND gates which can have arbitrary numbers or inputs.

Logic families

Since the introduction of the first integrated logic circuits in the 1960s there has been an evolution in logic *families*, with ever increasing speed and decreasing size and power consumption. Resistor–transistor logic (RTL) and diode–transistor logic (DTL) were the predecessors of transistor–transistor logic (TTL) and low-power Schottky TTL (LSTTL).

Figure 9.7 shows a TTL 2-input NAND gate and an inverter. The inputs are to emitters of a transistor (in the NAND gate a transistor with two emitters formed onto one base). The output is from a series transistor circuit so that rise and fall times are short.

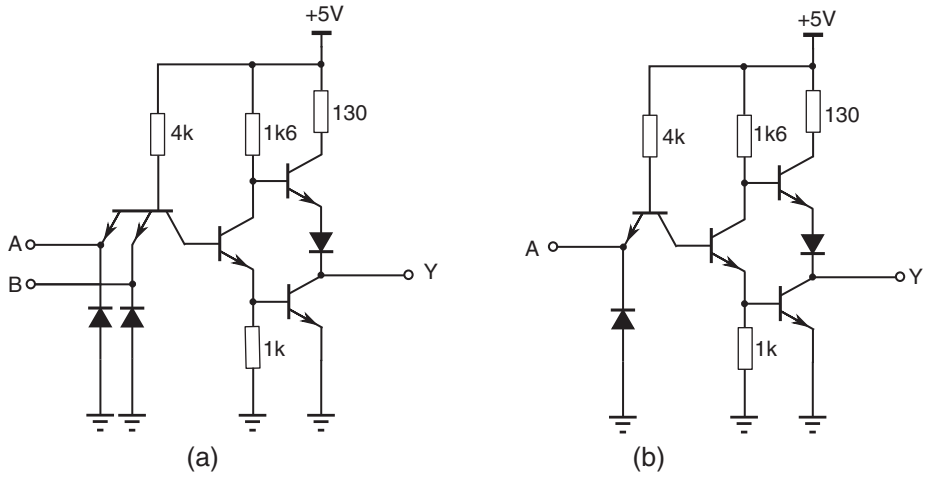


Figure 9.7

TTL gates: **(a)** NAND gate, **(b)** TTL Inverter.

Since the input is always to an emitter, the input resistance is low, and because the base of the input transistor is connected to the +5 V line, the input passes a current of about 1.6 mA when the input voltage is earth, logic 0. If an input is left unconnected, it will 'float' to logic 1, but it can be affected by signals coupled by stray capacitance, so such an input would normally be connected to +5 V through a 1k resistor. At the output, a totem-pole type of circuit is used. This can supply a current to a load which is connected between the output and earth (current **sourcing**), or can absorb a current from a load connected between the output and the +5 V line (current **sinking**). The normal TTL output stage can source 0.4 mA or sink 16 mA.

TTL ICs which use these output stages must *never* be connected with the outputs of different units in parallel, since with one output stage at logic 1 and another at logic 0, large currents could pass, destroying

the output stages. Modified output stages which have open collector outputs, are available for connecting in parallel – an application which is called a **wired OR**, since the parallel connections create an OR gate at the output.

The first family of CMOS logic was the 4000 series developed in the early 1970s. High-speed CMOS, HCMOS was introduced in the early 1980s and is generally pin function compatible with 74LSXX TTL devices although not necessarily interoperable.

High-speed CMOS 74HCXX devices and 74HCTXX devices which have TTL level compatible inputs represent the commonest logic family in use at the time of writing. TTL gates are now becoming hard to obtain and are not recommended for use in new designs.

Figure 9.8a shows a CMOS inverter circuit. This is simply composed of two MOSFETs, one P-channel and one N-channel. The FETs are connected so that, if the input is near zero, the top P-channel device is enhanced and provides a low impedance between the supply rail and the output pin. At the same time the bottom N-channel device is switched off. When the input is +5 V the situation is reversed. In between there is a region around 2.5 V when both transistors are partly enhanced and current can flow between the 5 V supply and ground; it is for this reason that CMOS inputs should *never* be allowed to float and inputs should change (transit)

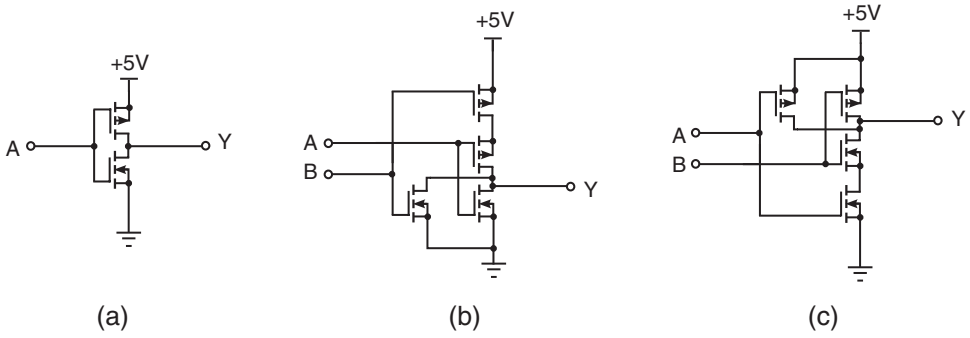


Figure 9.8

CMOS gates: **(a)** inverter, **(b)** NOR gate, **(c)** NAND gate.

between high and low states as fast as possible. If this is not done the IC will draw excessive current from the supply and may not work at all.

Battery-powered microcontroller circuits are particularly susceptible to problems with floating pins because the function of a pin, either input or output, is often controllable in software; unused pins should be set to output if possible or tied to one or other supply rail with a resistor. The NAND and NOR gates operate in a similar fashion to the inverter. An advantage of the CMOS circuit topology is that gates can be connected in parallel to increase the available output drive current.

In practice, protection diodes are built-in at the gate inputs and outputs to prevent excessive voltages from damaging the gates in circuit. CMOS devices are, however, very sensitive to electrostatic damage when not connected in circuit, and they should therefore be handled appropriately and stored on antistatic foam or in dissipative tubes.

Figure 9.9 shows the relationship between input threshold, output level and supply voltage for 74HC, 74HCT and 74LS gates. From this you can see

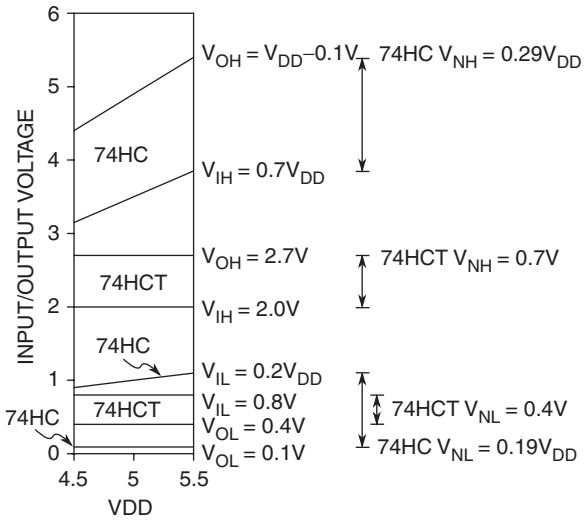


Figure 9.9

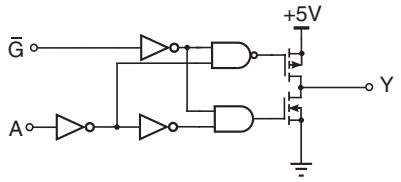
Input and output thresholds and noise margins for 5 V CMOS and TTL gates.

that the 74HCT input thresholds are fixed at 0.8 V and 2.0 V, whereas the input thresholds for 74HC devices are related to the power supply voltage 0.2 VDD and 0.7 VDD. 74HC devices can be operated at supply voltages between about 2 V and 6 V. 74HCT and 74LS devices need to have fixed supply voltages, usually specified as 5 V ± 0.5 V.

The **fanout** of a gate is the number of gate inputs that a gate output can drive while still satisfying its output level and rise and fall time specifications. This is more of a problem for logic families like TTL and LSTTL which have relatively large input currents. Typically a TTL output can source 0.4 mA and sink 16 mA (LSTTL 0.4 mA and 8 mA). The input currents are high-level 40 μA and low-level 1.6 mA (20 μA and 0.4 mA respectively for LSTTL). This means that a TTL output can drive 10 inputs and a LSTTL output can drive 20 LSTTL inputs. CMOS gates on the other hand have typical output drives of ±5 mA and typical input currents of ±1 μA. The practical limit of CMOS fanout is often determined by the input capacitance of the gates being driven; so, for example, an input capacitance of 5 pF per pin gives approximately 50 pF for 10 inputs, which for 74HCXX gates is the maximum capacitance for data sheet rise and fall times to be met (Figure 9.17).

Figure 9.10

CMOS tri-state inverter.



Tri-state outputs (Figure 9.10) provide a third, high-impedance, output state to a logic gate; therefore one, zero and high-impedance states are possible. In effect the output is disconnected from the rest of the circuit in the high-impedance state because both the top and bottom FETs in the output stage are turned off. This has useful applications in microprocessor buses where multiple devices can drive the bus but only one device is enabled at any given time.

Other logic families

There are several other logic families, but one that is used most frequently is emitter-coupled logic, ECL. This is utilized where very high speed or

differential signalling is required. ECL devices are available that work at frequencies up to 5 GHz and above and they find application in frequency synthesis and fast optical communications systems as well as in some areas relating to specialized fast computers.

Combinational logic

Circuits whose outputs are entirely determined by the combination of their inputs are referred to as **combinational** and they find application in logic functions, addition, encoding, decoding and pattern detection. As we noted earlier, any logic function can be built from a combination of AND, OR and NOT gates.

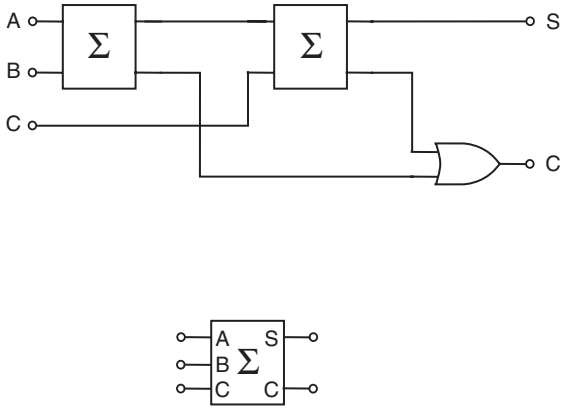
De Morgan's theorem is very useful for minimizing and implementing combinational logic equations; it states that AND and OR functions can be equated by appropriate inversions of the input and output variables. That is:

$$\overline{A \cdot B \cdot C} = \overline{A} + \overline{B} + \overline{C}$$

$$\overline{A + B + C} = \overline{A} \cdot \overline{B} \cdot \overline{C}$$

The half adder was discussed earlier in the chapter; Figure 9.11 shows a full adder made from two half adders and an OR gate. Cascades or full adders of this type can be built to add binary numbers of arbitrary width; usually the least significant bits require only a half adder because there will be no carry-in required.

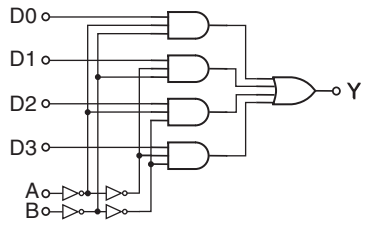
Combinational logic finds good application for pattern detection and data selection. Figure 9.12 shows a multiplexer or data selector circuit and a de-multiplexer or decoder circuit. These are the duals of each other; i.e. the multiplexer sets its output Y to the value of the input D0 to D3 that is selected by the binary code applied to pins A and B. The de-multiplexer reverses the operation by setting one of the outputs Y0 to Y3 to the value of the G input based on the binary code applied to A and B. The de-multiplexer can also be used as a decoder; by setting the G input high the outputs Y0 to Y3 represent the value of the binary bits applied to inputs A and B.



INPUT			OUTPUT	
A	B	C	S	C
0	0	0	0	0
0	1	0	1	0
1	0	0	1	0
1	1	0	0	1
0	0	1	1	0
0	1	1	0	1
1	0	1	0	1
1	1	1	1	1

Figure 9.11

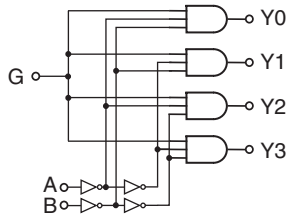
Adder circuit with symbol and truth table.



(a)

SEL		INPUTS				OUTPUT
B	A	D0	D1	D2	D3	Y
L	L	L	X	X	X	L
L	L	H	X	X	X	H
L	H	X	L	X	X	L
L	H	X	H	X	X	H
H	L	X	X	L	X	L
H	L	X	X	H	X	H
H	H	X	X	X	L	L
H	H	X	X	X	H	H

(b)



(c)

INPUTS			OUTPUTS			
G	B	A	Y0	Y1	Y2	Y3
L	X	X	L	L	L	L
H	L	L	H	L	L	L
H	L	H	L	H	L	L
H	H	L	L	L	H	L
H	H	H	L	L	L	H

(d)

Figure 9.12

(a) Multiplexer (selector) and **(c)** demultiplexer (decoder). The truth tables describe the relationship between the inputs and outputs of the circuits. An X in an input column indicates that the input does not affect the state of the outputs.

There are many applications, such as display drivers and memory addressing, that are based on selectors and decoders.

Number bases

Large binary numbers are awkward to handle, and can be difficult to copy without error and denary (decimal) numbers are unsuitable because they are not directly relatable. For these reasons, octal, base 8, or hexadecimal (hex), base 16, number representations are used for many applications, particularly in microprocessor machine code (see later). Hex coding is used when binary numbers occur in groups of four (called a **nibble**), eight (called a **byte**) or multiples of eight. The conversions are shown in Table 9.1. The use of hex coding makes the tabulation of binary numbers considerably simpler.

Table 9.1 Binary, octal and hexadecimal numbers and their decimal equivalents

Denary	Binary	Octal	Hexadecimal
0	0000	0	0
1	0001	1	1
2	0010	2	2
3	0011	3	3
4	0100	4	4
5	0101	5	5
6	0110	6	6
7	0111	7	7
8	1000	10	8
9	1001	11	9
10	1010	12	A
11	1011	13	B
12	1100	14	C
13	1101	15	D
14	1110	16	E
15	1111	17	F

As an illustration of how hexadecimal notation can help in manipulating binary numbers, the binary number 11001100 is 102 decimal, 66 hex.

Because each group of 4 binary bits is directly related to the hexadecimal code it is easy to convert between the two.

Sequential logic

The RS latch is an *asynchronous sequential circuit*, sequential meaning that the state of its outputs depends not just on the state of its present inputs but on the previous output that is fed back to its inputs. The circuits for RS latches made from NOR and NAND gates are shown in Figure 9.13; the difference between the two types is that the NOR based circuit has active high inputs, and the NAND active low inputs, that is the NAND based RS latch changes state when one of its inputs is connected to ground. In operation the circuit is very simple; the cross-coupled feedback between the two gates means that an input that causes the output to change is reinforced by the change in output, and so when the input is removed the output that it caused remains, that is the circuit has memory.

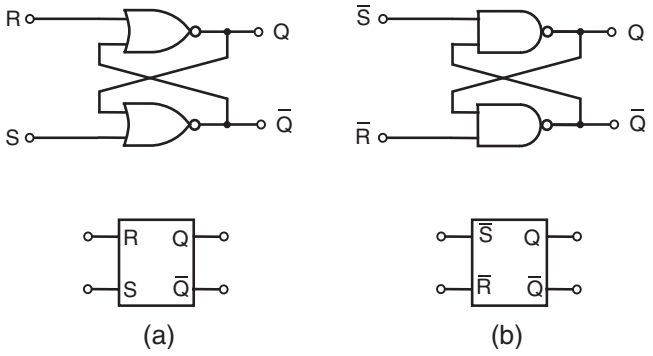


Figure 9.13

RS Latch **(a)** from NOR gates **(b)** from NAND gates.

While this simple cross-coupled circuit is useful for switch de-bouncing (see Figure 13.12), it is more usually used as a building block for other more complex sequential circuits. The RS latch also has a problem in that it must not have both its R and S input active at the same time since this would lead to the outputs being the same, no longer complementary, also making the next output state unpredictable.

The asynchronous nature of the RS latch is also a problem because the inputs are immediately effective, which can lead to problems particularly where feedback is envolved. The solution is to make the circuit synchronous by adding a clock input to control when the inputs are effective.

Figure 9.14 shows the evolution of the RS flip-flop into the more generally useful J-K flip-flop, with the addition of clock input and asynchronous reset and clear inputs greatly enhancing the usefulness of the circuit. The additional feedback cross-coupling from the output to input removes the possibility of both inputs to the inner RS flip-flop being active at the same time and guarantees complementary outputs.

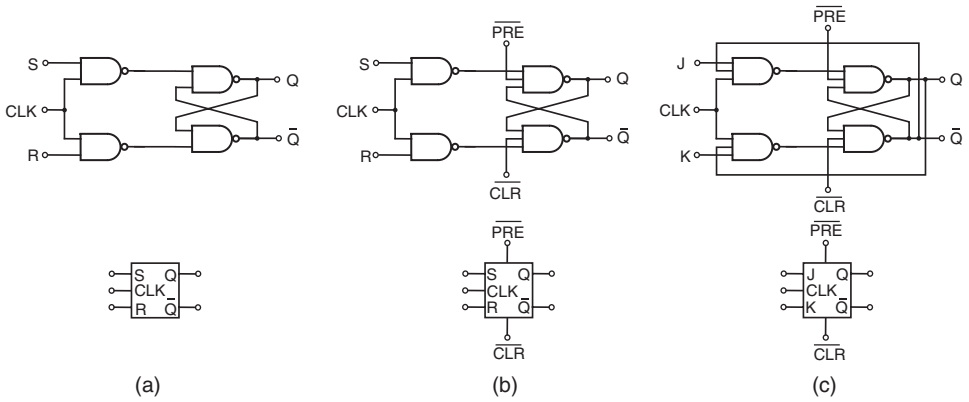


Figure 9.14

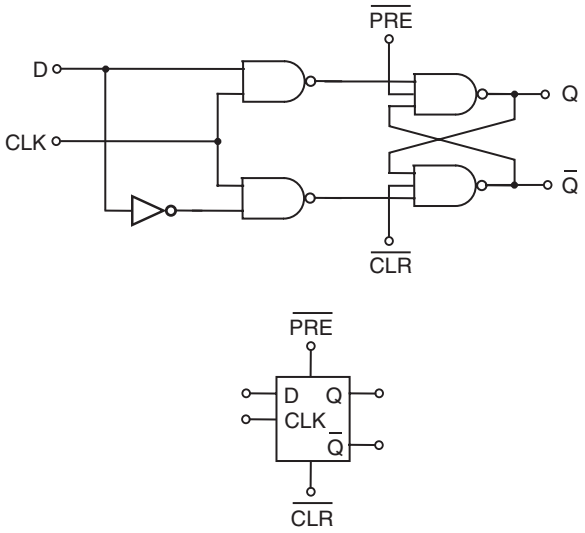
(a) Clocked RS flip-flop, **(b)** with asynchronous preset and clear, **(c)** J-K flip-flop.

The D flip-flop is probably the most widely used of all flip-flop circuits. It is usually implemented by using a J-K flip-flop with an inverter driving the K input from the J input signal as shown in Figure 9.15. The advantage of the D flip-flop is that its output Q copies the D input when the clock is active. If the D flip-flop is level triggered it is referred to as a *transparent latch* meaning that its output follows the input while the clock input is high and is latched to the last input while the clock is low.

The transparent latch is of limited use because race hazards can occur in feedback from stages, which can lead to unpredictable results. In order to

Figure 9.15

D flip-flop.



solve this problem edge-triggered rather than level-triggered flip-flops have been developed.

The master–slave flip-flop (Figure 9.16) is designed to provide isolation between the input and output stages to prevent race hazards in feedback circuits like counters. This is achieved by clocking the slave stage from the inverted master clock, so that the slave can change its outputs only when the master's inputs are disabled. This prevents the slave output transitions from affecting the state of the master's inputs. An alternative way of avoiding problems with feedback from outputs changing while the inputs are enabled, is to make the clock pulse very short; this is usually achieved by using edge-triggered inputs rather than level-triggered ones, as we shall see next. In very fast logic systems it may be impractical to implement edge-triggering and in these circumstances the master–slave approach is preferred. The J-K flip-flop is the building block from which most integrated up/down counters and non-modulo 2 counters are constructed

The rise and fall times and the propagation time for CMOS gates are characterized between the 10% and 90% points on the waveform as shown in Figure 9.17. Typical data sheet times for a 74HCXX gate at VDD = 5 V load 10 pF are $t_r = t_f = 10$ ns and $t_{plh} = t_{phl} = 15$ ns. The propagation

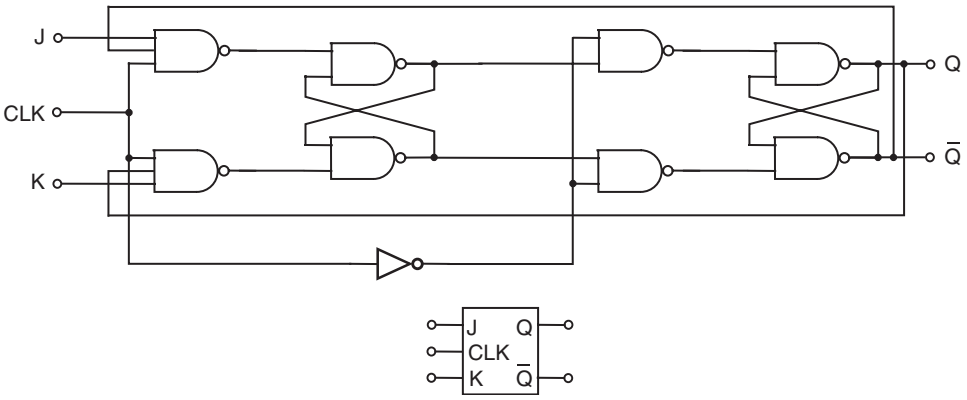
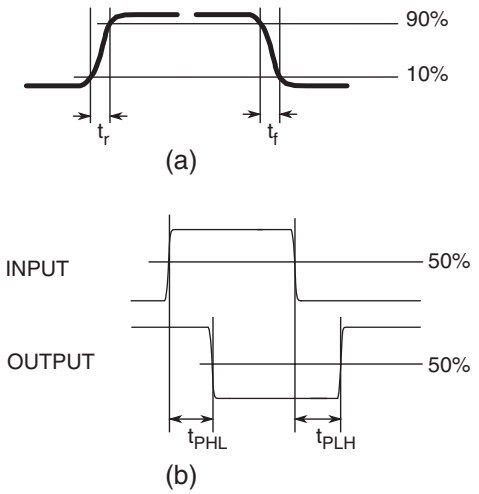


Figure 9.16

Master–slave J-K flip-flop.

Figure 9.17

(a) Output rise and fall measurement points and **(b)** gate propagation delays.



delay and rise and fall times are functions of both supply voltage and temperature, supply voltage having the greater effect.

A circuit termed a transition detector that will detect the rising edge of a waveform and provide a narrow output pulse, is shown in Figure 9.18, based on the delay through three inverters. This circuit gives a pulse

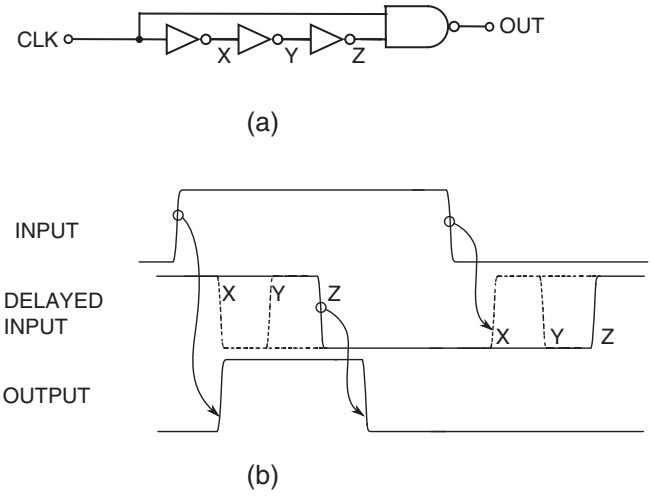


Figure 9.18

(a) Edge detector circuit using gate delays and **(b)** timings.

of about 45 ns wide when implemented in 74HCXX un-buffered logic. Un-buffered logic uses a single stage per inverter gate, whereas buffered logic uses three stages with the last using large area transistors to provide larger output drive capability.

An edge-triggered flip-flop using a propagation delay generator might be implemented as shown in Figure 9.19; the symbol for edge-triggered clock input is a triangle pointing inwards and the direction of transition to which the gate is sensitive may be shown by a rising or falling edge symbol or by inversion marking, that is rising edge-triggered CLK and falling edge-triggered $\overline{CLK}$.

If the $\overline{\mathrm{Q}}$ output of an edge-triggered D flip-flop, made from either a master–slave J-K or a propagation-delay version of the J-K flip-flop, is fed back to the D input the flip-flop will change state at every clock pulse; this means that the output changes state once for every two transitions of the input clock, dividing the frequency of the input clock by two. Apart from the obvious use as a counter, the D flip-flop used in this way guarantees that its output mark space ratio is 50%, that is the output spends the same amount of time high as it does low.

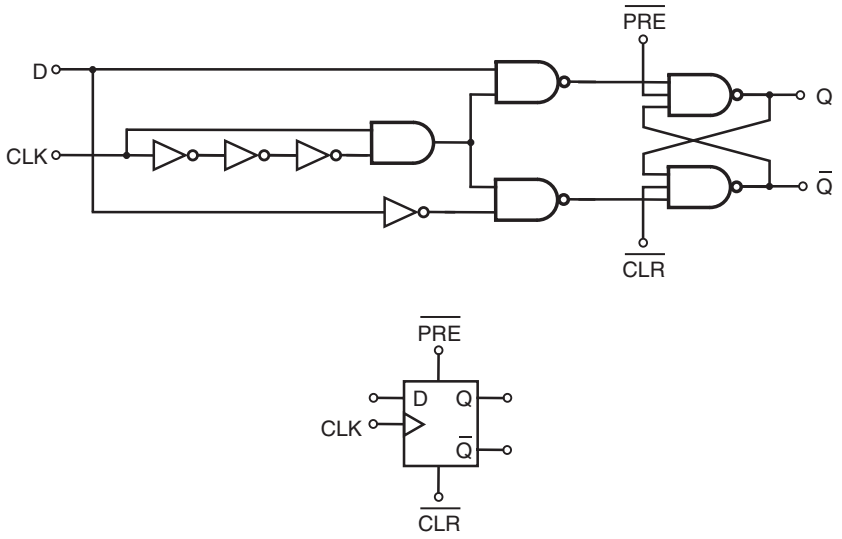


Figure 9.19

Edge-triggered D flip-flop.

Shift registers can be formed from J-K flip-flops or D-type flip-flops connected as shown in Figure 9.20. The action of a shift register is to pass a logic signal (1 or 0) from one flip-flop to the next in line at each clock pulse. The input signals can be serial, so that one bit is shifted in at each clock pulse, or **parallel**, loaded into each flip-flop at the same time, using the *preset* (sometimes called *set*) and *clear* (sometimes called *reset*) inputs.

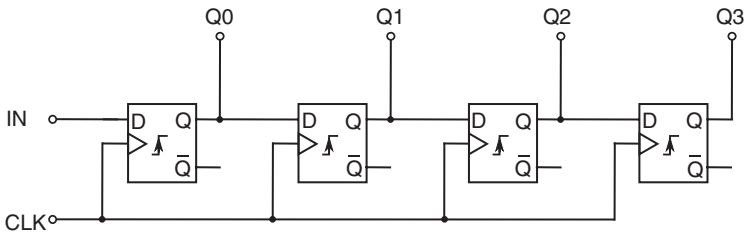


Figure 9.20

Shift register.

The output can similarly be serial, taken from one terminal at each clock pulse, or parallel at each flip-flop output. The shift can be designed to shift left, right or be selectable in either direction.

Serial communications interfaces use shift registers to perform serial to parallel and parallel to serial conversion; usually all the functions necessary for such an interface are built into a single block called a universal asynchronous receiver transmitter (UART) or universal synchronous/asynchronous receiver transmitter (USART).

Shift registers also find application in binary multiplication and division, shifting left n places to multiply by 2^n, shift right to divide. Shift registers with gated feedback can be used to produce binary sequences, referred to as pseudo random sequence generator (PRSG). The feedback gates are usually designed for maximum length without repeat; these systems find application in devices like mobile phones and noise generators.

Counters and dividers

The **ripple counter**, a chain of divide-by-two D flip-flops, is probably the simplest counter to construct. It is useful for frequency division but should be used with care in other applications because the outputs are not synchronous; the clock for each stage is generated by the output of the previous stage with the result that a race hazard exists – that is, edges take more time to propagate through to the later stages of the counter, earlier stages winning the race to set their outputs. This means that decoding the count from a ripple counter can result in very short pulses between the changing of the first and later stages. These are referred to as **runt pulses** because since they are very short they may not reach full logic swing. Figure 9.21(a) shows a four-stage ripple counter and the outputs over a number of clock cycles. The expanded section shows a close-up of how the outputs change as the count increases from 7 to 8. The delays can be clearly seen, and the result is that for a period the output is undefined and decoding circuits attached to the outputs of the counter could be falsely triggered.

The solution to the ripple counter problem is to build synchronous counters. Figure 9.21(c) shows a synchronous up-counter; because all the

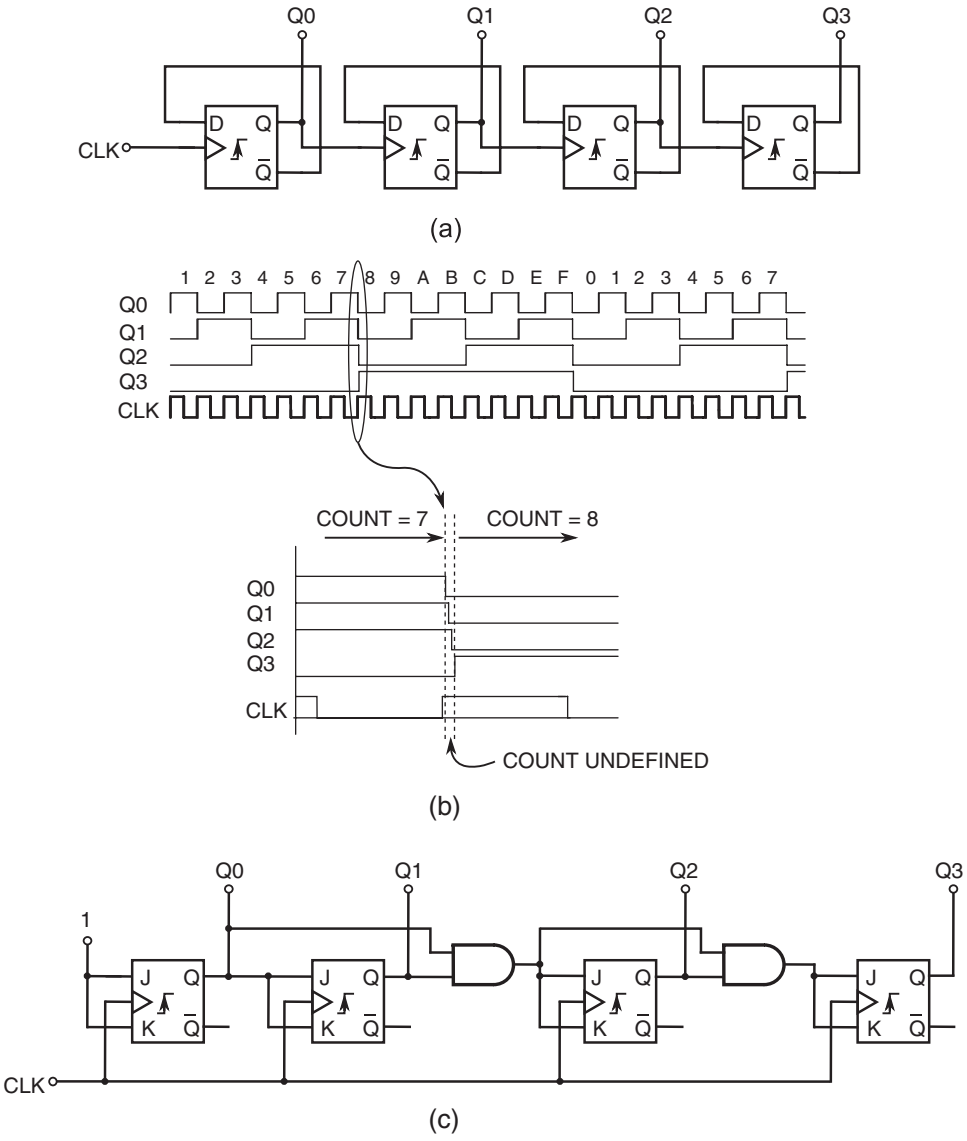


Figure 9.21

(a) Ripple counter, **(b)** race hazard decoding counter output, **(c)** sequential counter.

flip-flops are clocked from the same clock the outputs all change together eliminating the undefined outputs between counts. This technique can be used to produce both up and down counters, including selectable up/down counters and counters of arbitrary modulo, for example 5 or 10.

Complex logic circuits should usually be designed using synchronous logic; this is necessary to avoid the possibility or race hazards existing in the circuit causing unexpected results, as with increasing circuit complexity the delays get longer and the potential paths through the circuit get more difficult to analyse.

Simple clock sources are often required in logic circuits and RC oscillators of the type shown in Figure 9.22 are often provided as single pin oscillators on microcontroller chips and other complex ICs. By suitable choice of the resistor and capacitor, frequencies in the range a few hundred Hz to several MHz may be generated. The frequency may be varied by making the resistor

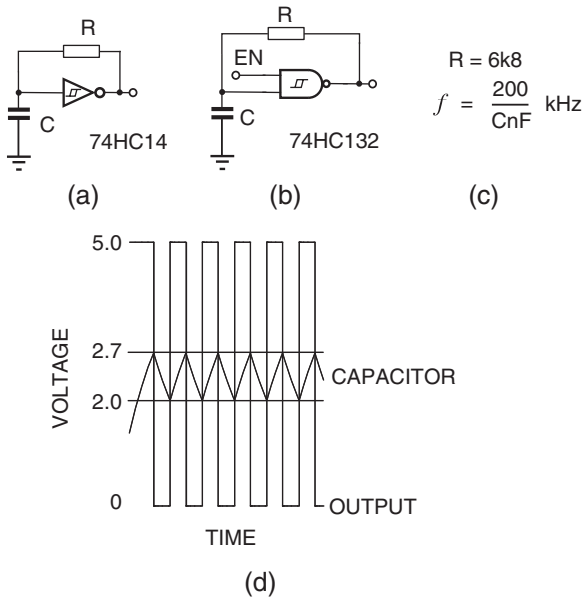


Figure 9.22

RC oscillator based on an inverter **(a)**, using a NAND gate **(b)**, formula relating frequency to R and C values **(c)** and waveforms **(d)**.

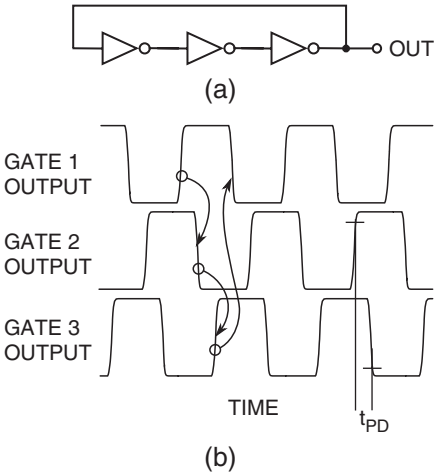


Figure 9.23

(a) Ring oscillator made from three inverters and **(b)** waveforms.

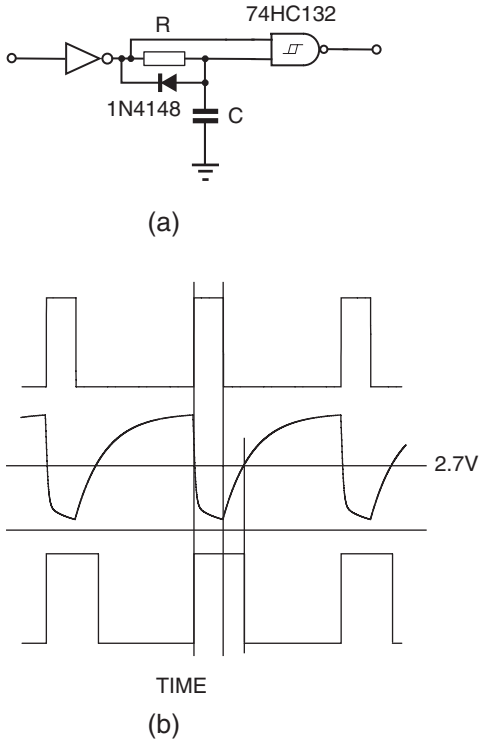


Figure 9.24

(a) Pulse stretcher circuit and **(b)** waveforms.

variable. Using a thermistor or light-dependent resistor, or other resistance that is dependent on an environmental variable, can be useful in measuring circuits; microcontrollers without an analogue-to-digital converter can usually measure frequency using a counter.

Earlier in the chapter we saw how propagation delays due to several gates could be used to detect transitions of waveforms and it is also possible to build oscillators based on propagation delay of gates. A three-gate circuit, as shown in Figure 9.23, is the simplest of this type of oscillator that can be built, termed a ring oscillator. Circuits of this type are often used where low cost, relatively low accuracy oscillators are required. The oscillation frequency is a strong function of supply voltage and also affected by temperature. Depending on the gate used, the manufacturer and the supply voltage a frequency between 10 MHz and 30 MHz is likely for this three gate circuit; any odd number of gates can be used.

Figure 9.24 shows a pulse stretching circuit that can be useful when trying to get an analogue (or low cost digital) oscilloscope to trigger on narrow or glitch pulses in digital circuits, for example runt pulses on the output of a ripple counter. The use of such circuits in actual logic designs is not recommended except in the simplest applications!

This page intentionally left blank

CHAPTER 10

PROGRAMMABLE DEVICES

Memory

Devices that can store information or settings, either permanently (non-volatile memory) or while the power supply remains on (volatile memory), form an essential part of almost every modern electronic system. Even equipment that has no apparent programmable functions may contain devices that are configured at or after assembly, reducing the inventory that the manufacturer has to keep and making designs more flexible by allowing modifications during production.

Solid state or integrated circuit memory devices for microprocessors and other computer applications fall into two categories. One type is memory that is not changed in normal operation and whose contents are not lost if power is turned off (non-volatile), typically containing the program commands and data that determine how a system operates. This type of memory tends to be called read-only memory (**ROM**), and historically ROMs were produced by manufacturing chips with the data defined during manufacture of the silicon, by configuring the connections of one or more layers of poly-silicon or metal. Even in very high volume production equipment, true ROM is rare; today, most systems use a form of programmable read-only memory (PROM) – these are often reprogrammable, although not necessarily in the system in which they are used.

The other type of memory is volatile, and is referred to as random access memory (RAM). Most modern memory devices support random access; that is, data can be accessed or written to any location independent of the location of the previous read or write – however, this was not always the case and the name has stuck.

Volatile memory with battery back-up can be used in place of non-volatile memory and the CMOS configuration memory of personal computers is used in this way. This can have advantages over volatile memory in the case where the user can make changes to the configuration that prevents the computer from operating. The cure is to disconnect the battery and let the volatile memory lose its stored data.

Another key difference between volatile and non-volatile memories is access speed. Non-volatile memory usually takes significantly longer to write than volatile memory, hundreds of times slower on average, and often uses considerably more power during writing because of the need for high voltages or current. Volatile memory is usually quicker to read than non-volatile although the difference in speed is much smaller.

Read-only memory (ROM)

A 64-bit memory formed from 8 locations each 8 bits wide is shown in Figure 10.1. The operation is straightforward, the 3- to 8-line decoder has active low outputs and one output is active at a time, pulling the row wires low. The column wires are pulled high by pull-up resistors. The diodes conduct, pulling a column wire low when the row wire is pulled low. The pattern of diodes along each row represents the bits that are zero; where the diode is not fitted the column line will remain high so the bit is a one. The hexadecimal numbers down the right-hand-side of the array show the data represented by the pattern of diodes for each row. Until quite recently circuits like this were used to provide small amounts of set-up information to embedded microprocessor systems like burglar alarms; such systems now use electrically programmable non-volatile memories.

The principle of operation of the circuit in Figure 10.1a is similar to that of integrated ROM devices; Figure 10.1b shows a diode connected between a row wire and a column wire. To make an integrated ROM the connections have to be made when the silicon circuit is fabricated. Figure 10.1c shows a diode with a series-connected fuse for programming. The first programmable read-only memories used fusible links. The fuse is blown to disconnect the diode and extra circuitry is used to direct high currents to the appropriate fuses to blow them.

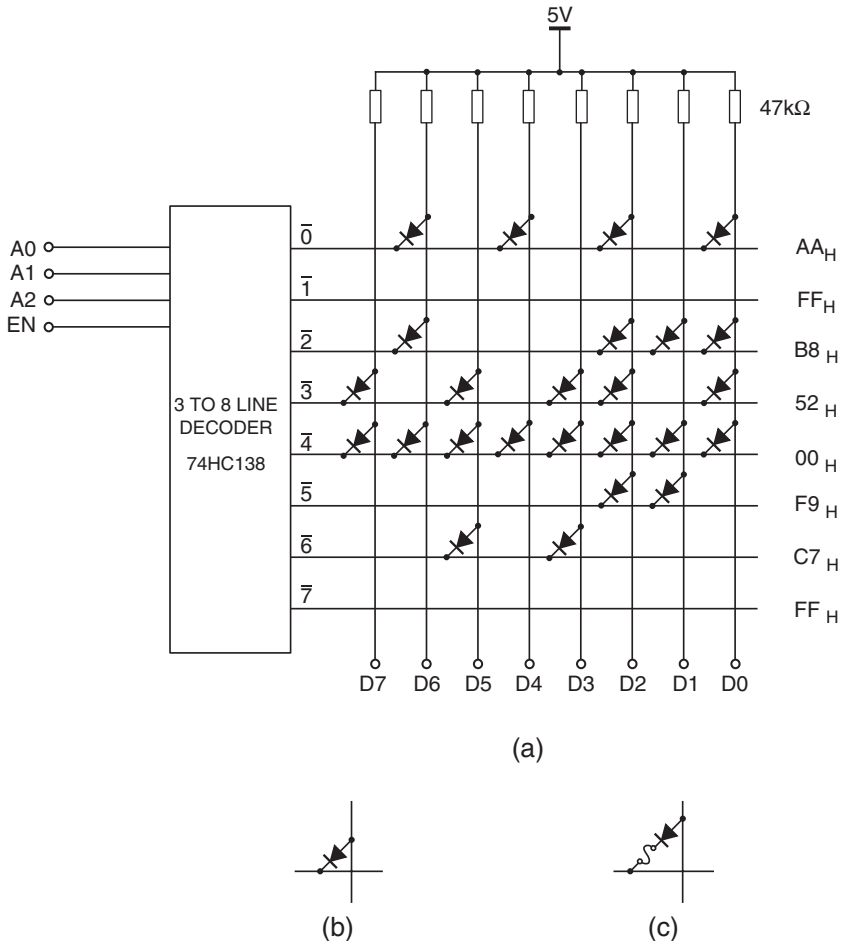


Figure 10.1

(a) An example of a memory array, **(b)** diode connecting row and column, **(c)** diode and fusible link.

Programmable read-only memory (PROM)

Fusible link PROM has now largely been superseded by ultraviolet erasable programmable read-only memory (UVEPROM) and electrically erasable programmable read-only memory (EEPROM).

While fusible link devices are effectively permanent, UVEPROM and EEPROM have expected data retention times of 10 to 40 years at room temperature; this has implications for system reliability so they may not be suitable for some systems like those that are exposed to very high temperatures or radiation, such as satellites.

UVPROM and EEPROM use floating gate FETs as the programmable elements. These operate like a normal FET except the gate structure contains an extra isolated conducting layer, the *floating gate*, which forms a capacitor that can be charged by application of a much higher voltage than used for normal operation. The effect of charging the capacitor is to change the threshold voltage of the FET. In the uncharged state, the floating gate prevents the FET from turning on when the row line is pulled high, and does not pull the column line low. Once the floating gate is charged the FET can be turned on, pulling the column line low. FLASH memory is based on similar physical effects but the logical architecture is different.

The charge will remain on the capacitor until it leaks away over time, taking 10 to 40 years at room temperature; this leakage can be accelerated by exposure to ultraviolet (UV) light or a high voltage. UVEPROMs are designed to be erased by exposure to short-wavelength UV radiation for about 20 minutes. It should be noted that the device will be erased by leaving it in direct sunlight for a few days, or under bright fluorescent light for a few months to a year. The package has a quartz window (Figure 10.2) to allow the light in, and this should be covered with a lightproof label if the device is likely to be exposed.

UVEPROMs are available without the window in the package, and these devices are referred to as one time programmable (OTP) devices. The silicon die is identical to that used in the windowed part but the cost of the package is lower. Microcontrollers are often provided in UVEPROM for development work and in OTP for production. EEPROM do not need the window because they have additional circuitry to erase/re-write the bits Figure 10.3 shows simplified schematics of UVEPROM and EEPROM elements.

Fusible link memories are permanent and they can not be reprogrammed, although it is sometimes possible to design a program arrangement so that sections of program can be bypassed by blowing more fuses.

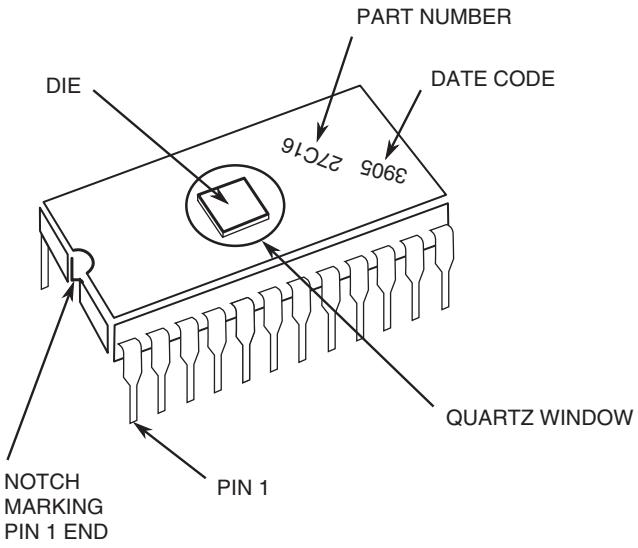


Figure 10.2

27C16, a 2k × 8 UVEPROM in a dual inline package.

The reason that the no-operation (NOP) instruction of some older microprocessors is FFH is to allow changes to programmable devices that cannot be erased. An instruction can be changed to NOP by blowing all the unblown fuses of a byte. Modern microcontrollers often use 00H as the NOP instruction for the same reason, since OTP versions of UVEPROMs allow program code to be deleted by programming all the bits of a byte.

Small-memory devices of up to about 256 bytes could be made in a similar way to the 8-byte example shown in Figure 10.1, however, as memory devices get larger the address decoding overhead becomes an issue. Square arrays of memory cells are more efficient in their use of silicon. Using 8 square arrays, one for each bit of the byte, reduces the decoding requirement from 4096 row drivers to 512 row drivers and 512 column lines, making the whole device smaller and nearer a square in shape which makes layout of the row and column interconnect easier. Figure 10.4 shows a simplified example of the structure of a 4096 byte memory consisting of 8 arrays each 64 × 64 in size.

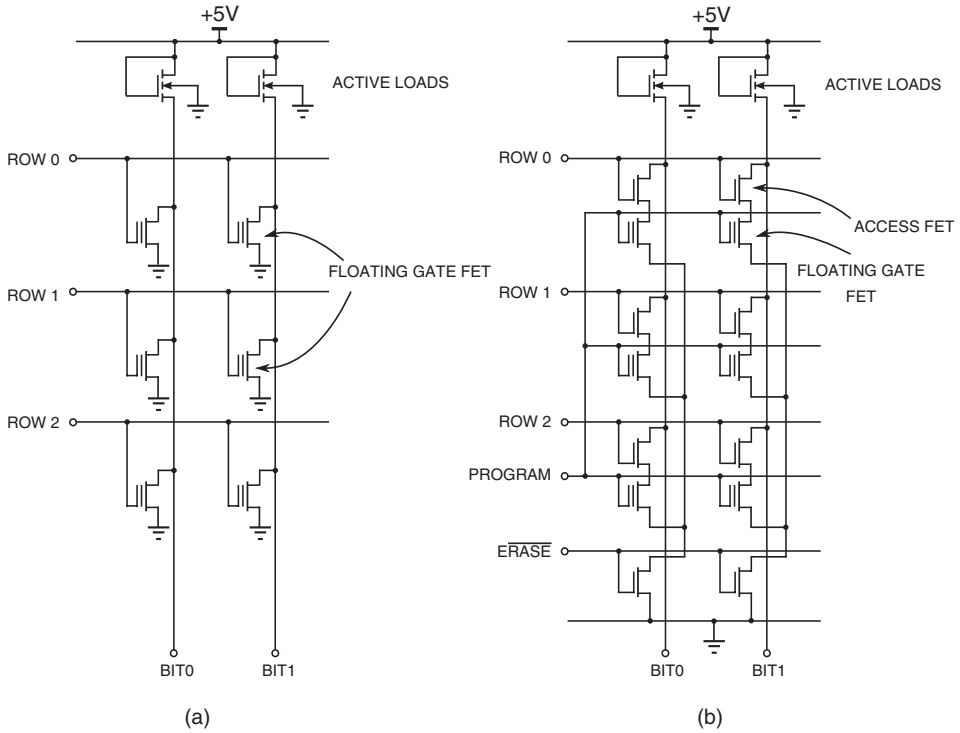


Figure 10.3

Erasable memory based on floating gate FETs: **(a)** UVEPROM, programming circuitry not shown, and **(b)** EEPROM.

Volatile memory (RAM)

Volatile memory may use flip-flops as storage elements, so-called **static memory**, or be based on charging capacitors, a system called **dynamic memory**. The terms static and dynamic derive from the fact that while a flip-flop stays in the state in which it was left unless the power supply is removed, a capacitor will discharge slowly over time and so needs to be refreshed regularly if the memory is not to be lost.

Dynamic memory can be fabricated with much higher density than static memory because each bit in memory requires fewer transistors. Dynamic memory chips can have built in refresh circuitry that takes care of

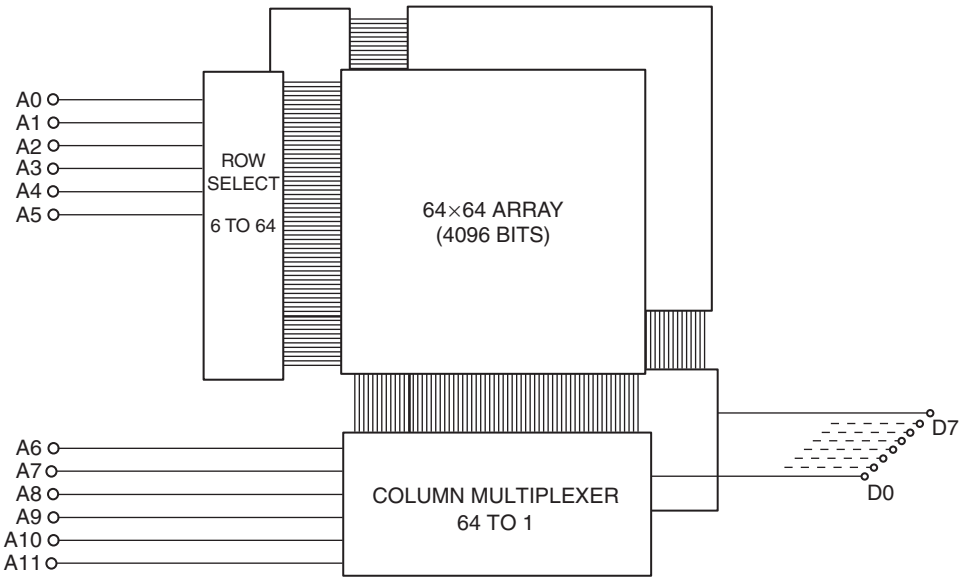


Figure 10.4

Row and column arrangement of integrated memory devices.

the recharging necessary to keep the data stored on the capacitors – these are sometimes refereed to as *pseudo-static* memories because the circuit designer does not need to provide external refresh circuitry.

The basic static memory cell consists of a pair of cross-coupled inverters, much like an RS flip-flop. The inverters are built with FETs whose resistance, when turned on, is relatively high. This allows them to be forced into the required state by pulling their outputs up or down with external drive circuitry. Figure 10.5 shows a simplified schematic of part of a static memory. This is a conventional row and column array, with the row driver selecting a row of cells and the column multiplexer selecting the specific cell from the row. The column multiplexer is differential, unlike the single-ended design used in EPROMs. The column multiplexer also serves as a driver, and when selected to write to a bit the C and /C outputs override the outputs of the flip-flop to drive it into the desired state.

Dynamic RAM requires a different access arrangement to allow read, write and refresh of the memory capacitors. In Figure 10.6 separate row-read and

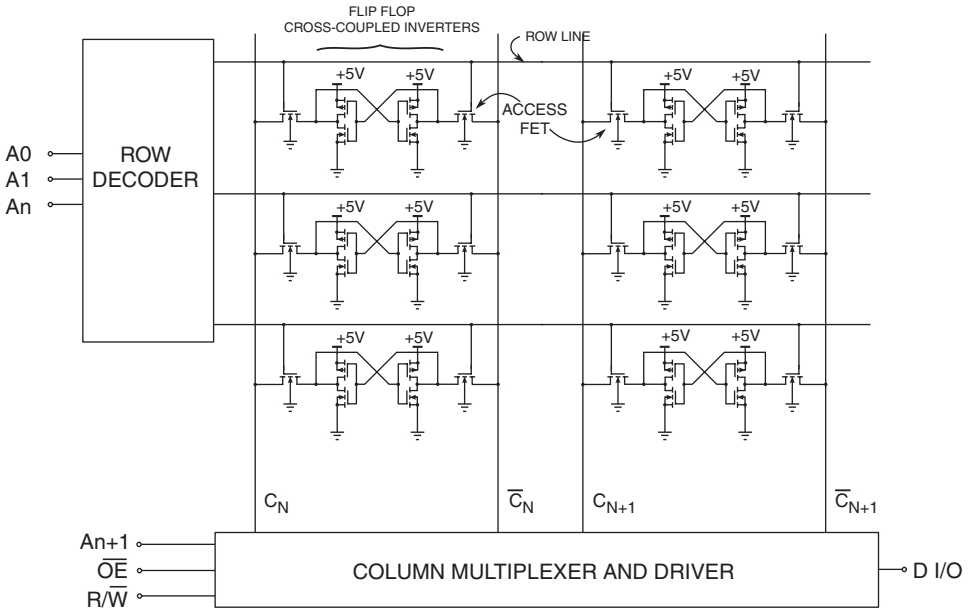


Figure 10.5

Part of a static memory device simplified to show the main features.

row-write lines turn on access FETs for each element, and separate bit-read and bit-write lines carry the data to and from the element. In order to refresh the memory capacitors, each bit-read has the bit-value rewritten, at shorter intervals than the capacitor discharge time.

Dynamic RAM often includes error detection and error checking logic. In its simplest form this is a parity bit that is calculated when the data are written and checked when read. There are also more elaborate error-correcting systems which allow correction of single-bit errors and detection of multi-bit errors.

Programmable logic

Building one-off systems or small-volume production from large numbers of standard logic integrated circuits is possible but not very efficient in

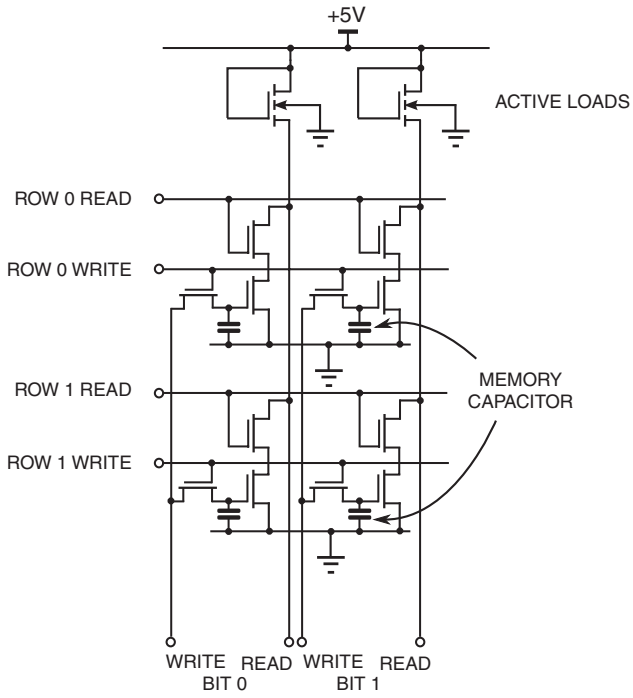


Figure 10.6

Simplified D-RAM cells; four bits, 2 × 2 array shown.

terms of finished equipment size or development cost. In mass production there are significant size and cost savings to be made in designing a custom integrated circuit to do the job, and improvements in reliability are also possible. The development cost of custom integrated circuits is high, but easily justified for equipment with high production volumes.

Programmable logic devices have advantages similar to those of custom chips, but with additional advantages of reduced inventory of standard parts and reduced development times, plus the ability to modify the design without necessarily redesigning the printed circuit board.

Programmable logic was first developed in the 1970s. The original devices were based on fusible link memory. Programmable logic devices based on a sum of products structure, like that shown in Figure 10.7, have

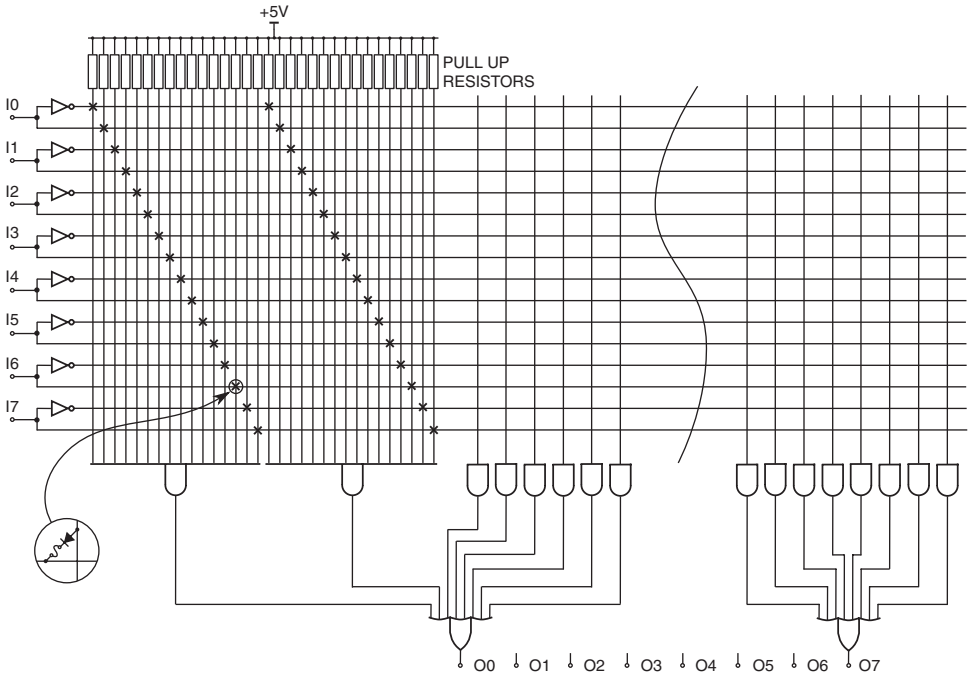


Figure 10.7

Simplified schematic of a PLD (programmable logic device), showing the location of the programmable links.

complementary inputs driving a programmable AND array whose outputs are ORed to drive the device outputs. They are configured by blowing fuses to disconnect inputs from AND gates. These are the simplest programmable logic devices and fuse patterns can be designed by hand and programmed directly with the appropriate programmer. The PAL16L8 is a device of this type.

As an example of this, if we want output 0 of the PLD to be high when inputs I0 AND I1 AND NOT I3 are high we would blow all the fuses in the leftmost AND array except for I0, I1 and NOT I3. This means that the unused inputs have no effect, that is the unused AND inputs are pulled high by the pull up resistors. The inputs I0, I1 and NOT I3 are now the only ones that can pull an AND input low, and the unused AND arrays all have low outputs because they receive all inputs and their

complements. The OR gates' output is then solely derived from the first AND array.

The development of programmable logic devices has consistently produced larger and more complex devices, in a similar way to the development of microprocessors and memory chips, since many design aspects of the silicon chip are common to these devices. Figure 10.8 shows a sum of products array with a register output; the D flip-flop and feedback term allows this type of device, the PAL16R8, to be used in synchronous logic circuits, which greatly enhances the complexity of the designs that can be developed.

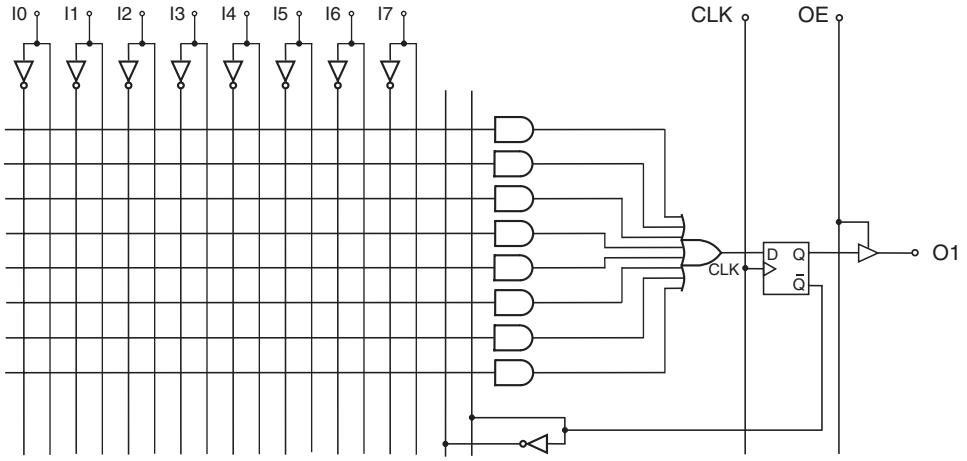


Figure 10.8

One section of a register PAL.

Complex programmable logic devices (CPLD)

Complex programmable logic devices use EEPROM memory to program the elements and have more complex internal structures that the

PAL described above; 1600 gate devices with 72 I/O pins are typical of the larger devices. Such devices can be clocked at speeds in excess of 100 MHz.

Field programmable gate array (FPGA)

Field programmable gate arrays use volatile memory rather than fuses to control the settings, and can be configured using a serial interface. FPGAs can be set up by a microprocessor if there is one in the system but usually this is done by a companion configuration device. Configuration devices have to be programmed; they typically use FLASH memory with a state machine base serial loader to load the configuration data into the FPGA at power-up.

FPGAs are very effective tools for development of complex systems because they can be reprogrammed quickly, directly from the development software running on a PC. The standard JTAG interface is used to program FPGAs; this is a serial interface developed to support boundary scan testing of memory and other complex chips and it also allows programming of various microcontroller, EEPROM and FLASH memory devices. Most FPGA and CPLD development tools have drivers for either generic or vendor-specific JTAG interfaces.

FPGAs are much larger and more complex devices than CPLDs or PALs; 50000 gate devices are easily capable of implementing entire microprocessor systems, and the vendors of these large devices provide so-called soft microprocessor cores. A soft microprocessor core is simply the hardware description language code used to implement the microprocessor, without peripherals; the user can then develop their own specific peripherals. There are several open source soft microcontroller and microprocessor projects in place; these implement microprocessors such as the 6502 and Z80 and microcontrollers like the 8051 and PIC16C84. Typically the FPGA implementation of an older microprocessor like the 6502 is significantly faster than for the original processor. FPGAs, like CPLDs, can be clocked at 100 MHz or more. There are also projects in place to implement entire computers like the ZX Spectrum and ATARI 600 in FPGAs.

Hardware description language (HDL)

The complexity of the logic that can be implemented with programmable logic devices makes the use of software tools to develop the design essential. Design tools are available from programmable logic device vendors in much the same way that assemblers and compilers are provided by microcontroller vendors. The description of logic is similar to that of a computer program and is written in hardware description language. There are two main dialects favoured by device vendors and electronic design automation (EDA) tool providers: VHDL and Verilog. VHDL stands for Very high speed integrated circuit Hardware Description Language, and is standardized as IEEE 1076 and IEEE 1164. The advantage of using VHDL or Verilog is that the logic design is independent of the vendor or technology of the target device, which renders code portable and reusable, rather like using the C programming language for microcontroller programming.

The full details of using VHDL are beyond the scope of this book, but a simple example of VHDL that implements an OR gate using three I/O pins – a, b and y – is given below.

```
LIBRARY ieee;
USE ieee.std_logic_1164.all;

ENTITY or_gate IS
PORT (a,b: IN BIT;
        y: OUT BIT);
END or_gate
ARCHITECTURE simple_or_gate of or_gate IS
BEGIN
       y <= a OR b;
END simple_or_gate;
```

The code above tells the compiler to use the standard logic library, then defines an entity called or_gate which has associated with it three ports – a, b and y. The final part defines a function to associate with or_gate; in this case y is defined as equal to a OR b. To program the OR gate into hardware the compiler would need to be told what the target hardware was and the physical pin assignments for the port a, b and y. If the desired

target hardware were changed, only the physical pin assignments would need to be altered before recompiling.

Other programmable devices

There is a growing number of programmable devices that use non-volatile memory to store configuration information or settings. These range from digital potentiometers to configurable analogue arrays and programmable system on chip devices (mixed signal array) which combine microcontroller, data converters and analogue devices like programmable gain amplifiers and programmable bandwidth filters. Several vendors produce programmable mixed signal arrays, among them ATMEL, Analog Devices and Cypress Semiconductor. Figure 10.9 shows a simplified diagram of the CY8C27643, which is typical of this type of IC.

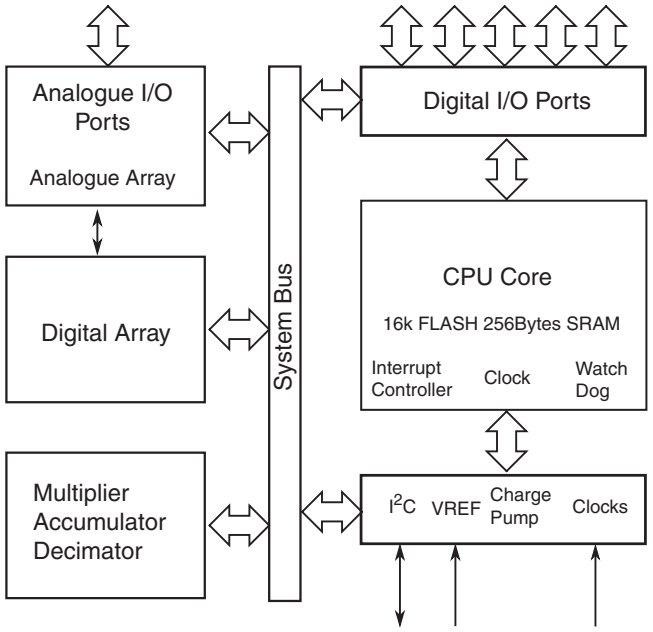


Figure 10.9

Simplified block diagram of Cypress PSoC™ CY8C27643 device.

Other applications of memory devices

Memory devices like EPROMs can be used in applications other than storing microprocessor programs. Hardware look-up tables use memory devices to store data values in a way that obviates the need for a microprocessor. Examples of such tables are digital temperature compensation systems which use ADC data from a temperature sensor to drive the address pins of a memory, and arbitrary waveform generation which uses a counter to drive the address pins and cycles through the locations of the memory continuously.

In Figure 10.10 a 64k × 8 EPROM is driven by a 16-bit synchronous counter composed of four 74HC191 up/down counters. The data lines drive a R2R ladder DAC (see Chapter 13) via a latch (74HC574); this latch is required because the EPROM outputs are always enabled and do not all change at exactly the same time, when the address is changed, which can give rise to glitches in the output. The latch is clocked from the opposite edge of the clock to the counter, which allows the EPROM and counter almost half a cycle for the data to settle before the latch is triggered; this sets the maximum frequency of the clock.

If the EPROM used has an access time of 200 ns, the set-up time for the latch is 10 ns and the propagation delay for the counter is 30 ns, the maximum clock frequency is given by:

$$f_{max} = \frac{1}{2(200\text{ ns} + 10\text{ ns} + 30\text{ ns})} = \frac{1}{480\text{ ns}} = 2.08\text{ MHz}$$

The clock could be run at a very low frequency if slowly changing data was required; a 0.76 Hz clock would take a day to scan the EPROM once. If the system was clocked at about 8 kHz, just over 8 seconds of speech could be stored and played back.

The clock source is a variable-frequency RC oscillator based on a CMOS inverter (see Chapter 9) but a crystal oscillator could be used if frequency stability was important. The R2R ladder output would need to be low-pass filtered to remove the switching steps and buffered to prevent the load affecting the output amplitude.

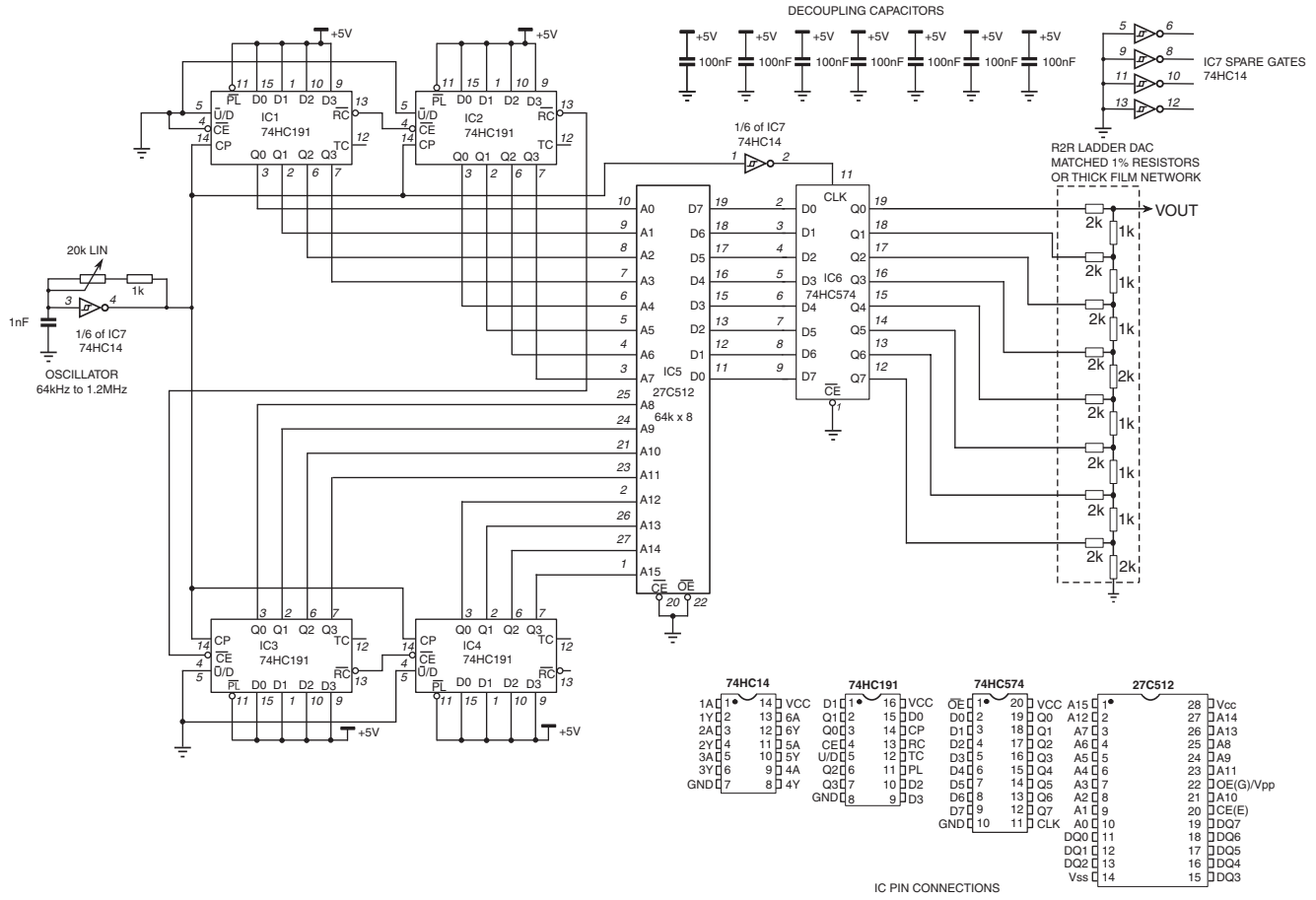


Figure 10.10

An arbitrary waveform generator using an EPROM and DAC.

The data stored in the EPROM can be generated with a computer program or spreadsheet or recorded using a PC sound card, or even captured with a digital oscilloscope depending on the application. In order to program the data into the EPROM they need to be converted into a format that an EPROM programmer recognizes; there are two common formats, Intel HEX and Motorola S record. Some EPROM programmers also support raw binary files.

The memory does not have to be an EPROM; a static RAM chip could be used but would require extra logic to allow for writing the data to it, probably using a microcontroller. The advantage is that static RAM with access times as low as 15 ns is available, allowing clock speeds of up to about 30 MHz with the use of 74AHC logic for the counter and latch circuitry.

Useful websites

Memory devices

Advanced Micro Devices	**www.amd.com**
ST Microsystems	**www.st.com**

Programmable logic devices

Altera	**www.altera.com**
Xilinx	**www.xilinx.com**

Programmable system on chip

Cypress Micro Systems	**www.cypress.com**
Analog Devices	**www.analog.com**

This page intentionally left blank

CHAPTER 11

MICROPROCESSORS AND MICROCONTROLLERS

Introduction

Until the late 1940s the accepted meaning of the word computer was 'a person who carries out a series of mathematical calculations', usually according to a set of rules – an *algorithm* – but sometimes as part of a 'production line' where each person (computer) took the result from the previous one, performed their allotted calculation and passed the result on to the next person.

An algorithm is a rule or process for solving a mathematical problem in a finite number of steps; this rule or process would be determined by a mathematician and given as instructions to one or more computers. The computers would then perform the appropriate calculations and return the finished answer, with any mistakes that they had made!

To speed up the process and help reduce the errors that the human calculators made, mechanical calculators that could add, subtract, multiply, divide and later also sort in order of size were introduced. To reduce the transcribing errors at each stage the data was punched into cards rather than being written down. The punch card idea was borrowed from a method of setting up weaving looms to repeatedly produce the same complex pattern, named after its inventor Joseph Marie Jacquard.

The scene was therefore set for the evolution of the computer as we currently know it. The computer – the person who followed the algorithm – using their calculator to perform the steps of calculation, was replaced by

a controller and the algorithm became the program. In a computer the calculator is called the arithmetic logic unit (ALU).

Consider an algorithm for converting temperature from Celsius to Fahrenheit; the input data, e.g. 27°C × 9/5 + 32, gives 80.6°F and the algorithm would be as shown in Table 11.1.

Table 11.1 Steps for Celsius to Fahrenheit algorithm

Step	Operation	Result	Output
1	Input the temperature in Celsius	27	
2	Multiply by 9	243	
3	Divide by 5	48.6	
4	Add 32	80.6	
5	Output temperature in Fahrenheit		80.6

Binary stored program computers

A special-purpose computer designed to perform one task (such as Colossus, designed at Bletchley Park to decode German radio transmissions) can be built specifically to do that task; therefore the algorithm is part of its design and it cannot be easily changed. Stored program computers are general purpose and do not perform any function until a program is provided; however, changing the program is simple.

The stored program computer requires a means of data input and output, a controller to follow the program, a calculator to perform operations and a store for intermediate data. With a few early exceptions, computers use binary numbers that are composed only of ones and zeros, each binary digit (**bit)** of a number representing a power of 2. The right-most binary digit is 2^0 (1). Eight bits are termed a **byte** and four bits a **nibble**. The number of bits used by a computer to represent numbers is called a **word**, and the word width in bits is commonly used as a classification of processor type. Most microcontrollers use 8- or 16-bit words; microprocessors use 8-, 16-, 32- and 64-bit words. Table 11.2 shows the largest numbers represented by common word widths.

Table 11.2 Largest unsigned integers represented by binary words

Word width in bits	Largest unsigned integer
4	15
8	255
16	65535
32	4294967295
64	18446744073709551615

The controller reads each instruction in turn, sets up the input, output, intermediate store and calculator to perform the desired operation, and then executes the operation. The controller then advances to the next step in the program. This arrangement would replace a room full of people with mechanical calculators dumbly following one instruction after another, removing errors and speeding up calculations. The real breakthrough was the introduction of *conditional branching*: the controller can be instructed to test whether one number is less than, equal to or greater than another and based on the result follow different branches in the program. Conditional branching allows computers to 'make decisions', rendering them much more than just machines that speed up the calculations of arithmetic.

Early machines used reels of paper tape onto which data and instructions were punched. The controller would index the tape forward one step each time it read from the tape. In modern solid-state systems the program is stored in memory, either ROM or RAM, and the controller contains a register that points to the address of the next instruction to read from memory. This register is called the **instruction pointer** (IP) or **program counter** (PC) and is incremented every time an address in memory is read. The system requires a clock signal to trigger the controller to perform the load, decode and execute sequence. The clock is typically a square wave signal derived from a quartz-crystal controlled oscillator. The controller, including the program counter, and arithmetic logic unit are usually combined together and called the **central processing unit** (CPU).

The clock frequency of a computer determines the rate at which instructions can be executed; the instruction rate is measured in instructions per second (ips) and usually this is thousands or millions per second. CPUs for

desktop computers operate at rates of around 1000 Mips or more, while the average microcontroller operates at around 1 Mip. Some CPUs perform one operation per clock cycle; the clock frequency needed to operate a microcontroller is usually a multiple of the operation rate. The Atmel AVR requires 1 clock cycle, and the Microchip PIC10/12/16 series 4 clock cycles per instruction, so achieving 1 Mip takes a 4 MHz clock. Some older machines used many more clock cycles per instruction, for example the Intel 8051 required 12.

In order to implement conditional branching the controller needs to be able to observe the status of the calculations being performed. This is done by providing a single-bit output from the arithmetic logic unit, called a **flag**, which indicates if a calculation resulted in a zero or a carry. The controller also needs to be able to add and subtract numbers from the instruction pointer to follow branches in the program.

An example of conditional branching that requires subtraction from the instruction pointer (program counter) is **looping**, that is executing the same set of instructions repeatedly. Loops can be used to wait for some event to occur or to do something a predefined number of times. Unconditional branching is a special case, in that no test is required: a number is simply added to or subtracted from the instruction pointer. Unconditional branches are often called **jumps**, the terms branch to and jump to being synonymous.

There are advantages in being able to jump to a different place in a program and then return to execute the instruction that should have been next had the jump not happened. In order to do this, the address in the instruction pointer must be stored before the offset is added or subtracted to cause the jump. In this way the controller can jump to a section of program whose last instruction tells the controller to reload the previously stored value of the instruction pointer and therefore return to the original place in the program. Sections of program used in this way are referred to as **subroutines**.

A mechanism for storing the program position and returning to it, allows another important feature of computers, the **interrupt**. This means interrupting one program to perform another by sending the controller a signal from outside. Interrupts allow computers to interface with users and the real world. For example, interrupts can be used to tell a computer that

a counter has reached zero or that the user has pressed a key of the keyboard. Using interrupts to get the controller's attention when irregular or infrequent events occur allows it to process other tasks more efficiently. Interrupts make time sharing and real-time systems possible.

If only one store is made available for the return address, only one subroutine may be active at a time, so it is convenient to allow subroutines to be **nested**, one inside another, so a means of storing return addresses in the order that they will be used needs to be provided. This is a last-in first-out (LIFO) memory, usually known as a **stack**. Each time a subroutine is called, the address in the instruction pointer is stored on top of the stack, and the stack grows one location bigger. When a subroutine returns execution to the calling routine the instruction pointer is recovered from the stack and the stack becomes one location smaller. Since the stack must be of finite size the result of trying to put too many return addresses onto the stack, or load too many back from it can be unpredictable and usually causes the processor to crash.

A microprocessor is a single chip containing all the circuitry for the CPU, that is the controller, arithmetic logic unit and memory access arrangements. In some cases microprocessors include memory too. Usually a complete system is made up of microprocessor, memory, ROM containing program and RAM for data and programs, as well as interfaces for devices such as keyboards, displays and communications (Figure 11.2).

A microcontroller is a single chip microprocessor with data memory and input/output ports on chip. Program memory (ROM) is usually on chip as well but some vendors produce devices that can access external ROM instead of, or as well as, the internal ROM. An example of this is the Intel 8051 with internal PROM; the 8031 is identical with the exception of the PROM. Most microcontrollers do not have external address and data buses so they cannot access external memory directly.

Von Neumann and Harvard architecture

There are two architectures defining how a computer accesses memory. The Intel x86 family of microprocessors used in IBM PC compatible computers, and most other general purpose microprocessors, are based on *Von Neumann architecture*. The Von Neumann architecture (Figure 11.1)

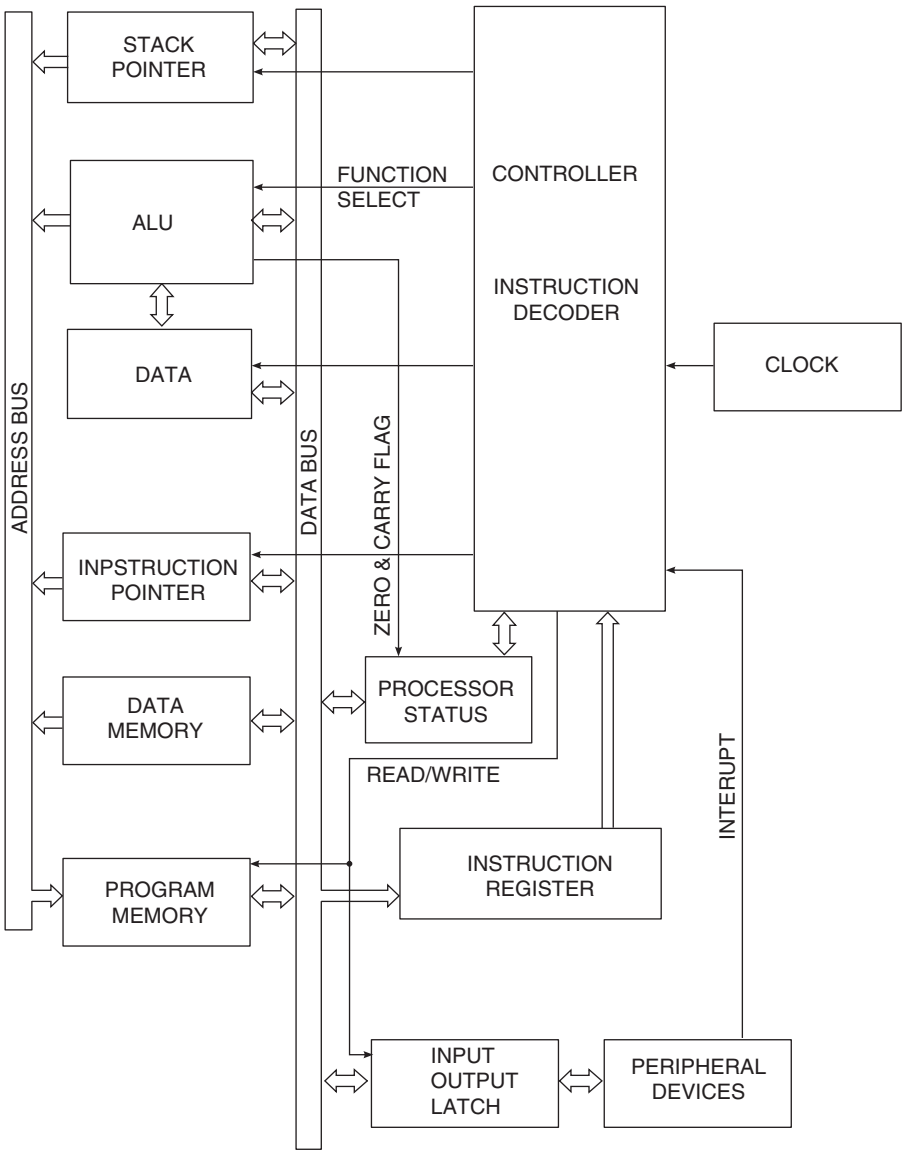


Figure 11.1

Simplified architecture of a stored program computer, of the Von Neumann type.

uses a common memory space for programs and data, this meaning that a program can write to the memory locations that stores the program, or that data can be run as a program.

The other architecture is *Harvard architecture* and this tends to be used in microcontrollers. The Harvard architecture uses separate data and program memories, and this can have advantages for control applications because data and program memories can have different characteristics such as width and access time. Harvard architecture machines often use a hardware stack built into the controller to store index pointer return values. Von Neumann machines tend to allocate space at a fixed position in main memory for the stack.

When an instruction is read by the controller it has to be decoded to provide the settings for the various blocks of the processor (Figure 11.1). An instruction can include data; the part of the instruction that determines the operation to be carried out is referred to as an operation code (**op-code**) and data is termed an **operand**. The advantage for Harvard architecture in having different width data and instruction memory is that the instruction can be wide enough to include a data word with the op-code. The von Neumann architecture might require two reads from memory to perform the same function.

Micro code is the name given to the table of instructions that the controller uses to determine how to configure the arithmetic logic unit and other internal functions of the processor. In effect it is a look-up table, the address value being the op-code and the output being the control lines to the processor functions.

There are also two schools of thought about the design of instruction sets for computers, complex instruction set computer (**CISC**) and reduced instruction set computer (**RISC**). The Intel Pentium processor is typical of a CISC processor with in excess of five hundred instructions. Microcontrollers using Harvard architecture tend to be RISC processors and use typically 30 to 60 instructions. The advantage for microcontrollers is smaller silicon area for micro code and simpler control logic; there is also an advantage for the programmer learning the instruction set. CISC processors may make software development more efficient in certain types of application.

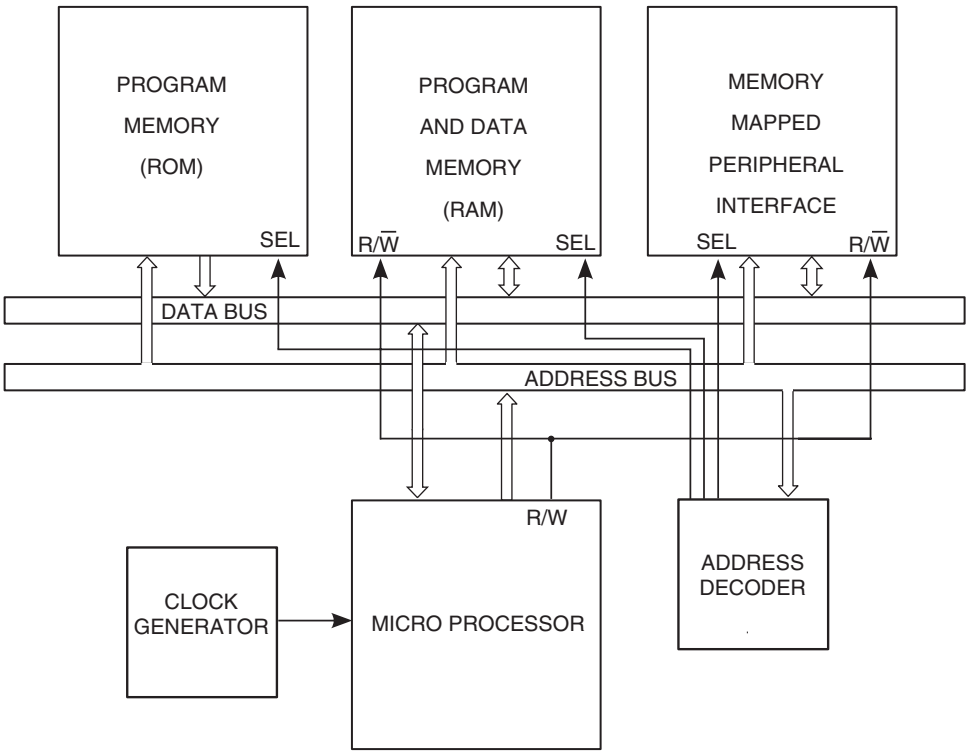


Figure 11.2

Block diagram of a microprocessor system.

Microprocessor systems

Microprocessor systems are assembled from the microprocessor chip, ROM, RAM, address decoding logic, peripheral interfaces and clock. In addition to these devices, voltage regulator, power-up reset circuit and various passive devices like power supply decoupling capacitors and pull-up resistors for input lines are required.

Memory is connected to the microprocessor via buses; a bus is a group of signals that link two or more devices. Microprocessors usually have

external address, data and control buses. The control bus carries the read/write, chip select, reset and interrupt signals that are required to interface memory and peripherals to the address and data buses.

The address bus is usually driven by the address output port of the microprocessor. The microprocessor writes the address from the program counter register to the bus so as to select the desired location in memory. The address bus is usually controlled by only one microprocessor; however, in some circumstances more than one microprocessor may access the same bus, for example in systems with shared memory or with special purpose co-processors, like floating point arithmetic units. Direct memory access (**DMA**) controllers also need to write to the address bus. A DMA chip performs functions like fast block copying of memory without the microprocessor needing to spend time doing it and DMA controllers are often used to interface to disk storage.

The address bus output port of a microprocessor will usually be a tri-state output if the processor is designed to share the address bus, and an address bus enable (**ABE**) or similar signal will be provided to control it. When more than one device needs to access the address bus, control signals enable the address bus outputs of individual controllers and handshaking arrangements so that devices can pass control of the bus to each other cleanly, typically BusRequest and BusAck.

The data bus carries data to and from the microprocessor. Multiple memory and peripheral devices attached to the data bus may write to it but only one at a time. In systems with a single microprocessor and no DMA controller a read/write signal is usually sufficient to control the transactions. The CPU sets the address, to read or write, on the address bus, and the address decoding logic decodes the address and outputs the chip select signal for the desired device. The read/write line indicates to the memory device whether it is to output data onto the bus or wait for data from the CPU. Figure 11.3 shows the sequence of events involved. When data are written to a memory the data are latched as the read/write line returns to the read state. To ensure that the data will be correctly written, the data set up and hold time, that is the time during which the data is stable on the data bus before the read/write line, is raised and the data hold time, the time after the read/write line is raised until the data bus can change value, must meet the memory specification for the chip used.

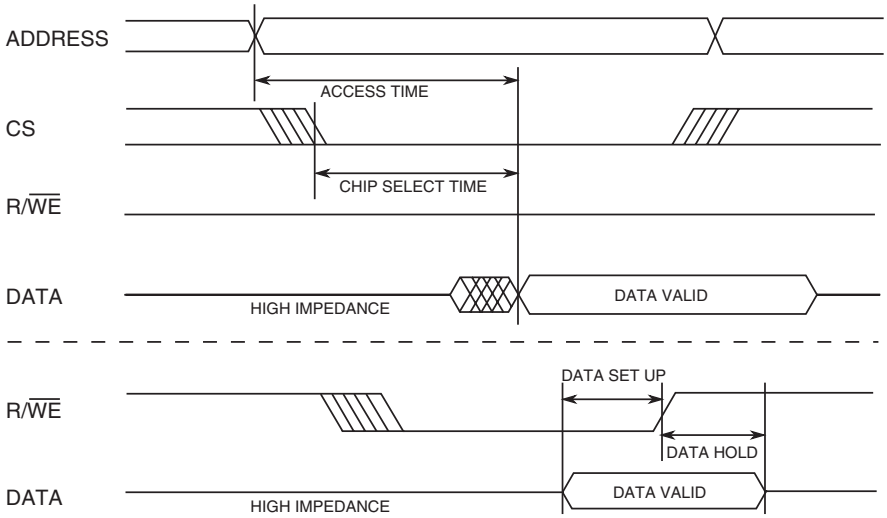


Figure 11.3

Read (top) and write (bottom) timing for a static RAM, EPROM read timing is similar.

If memories with very different access times are to be used on the same bus some CPUs allow an external circuit to report back to the CPU the length of time a transaction will take; this can be done with combinational logic using the outputs of the address decoding logic and dip switches or links on the board. A similar arrangement can be used to inform the CPU of attempts to write to ROM memory; this can use an interrupt input to the CPU, if the write signal is asserted while ROM memory is selected (Figure 11.4) an interrupt or reset event can be caused. If multiple interrupts are needed either the CPU requires multiple interrupt inputs or an external input port can read a latch, the bits of which represent various events that can cause an interrupt. If writing to ROM memory illegally caused an interrupt and set a bit of this interrupt status latch, the interrupt service routine can read the latch and determine how to respond, for instance printing a message like 'user program tried to write ROM area'. Further information could be provided for the user; if the stack containing the return address for the interrupt service routine can be read by programs (this is not possible on some processors), the address of the offending instruction that attempted to write ROM memory can be calculated by subtracting one from the address stored on the stack.

Power-up reset and program execution

When power is first applied to a microprocessor system, or it is reset, it needs to start executing program instructions. There are two common approaches to determining what program should be run. Fixed reset address systems usually start executing at the bottom of the memory; vectored systems load the start address from a set place in memory (the reset vector) which can

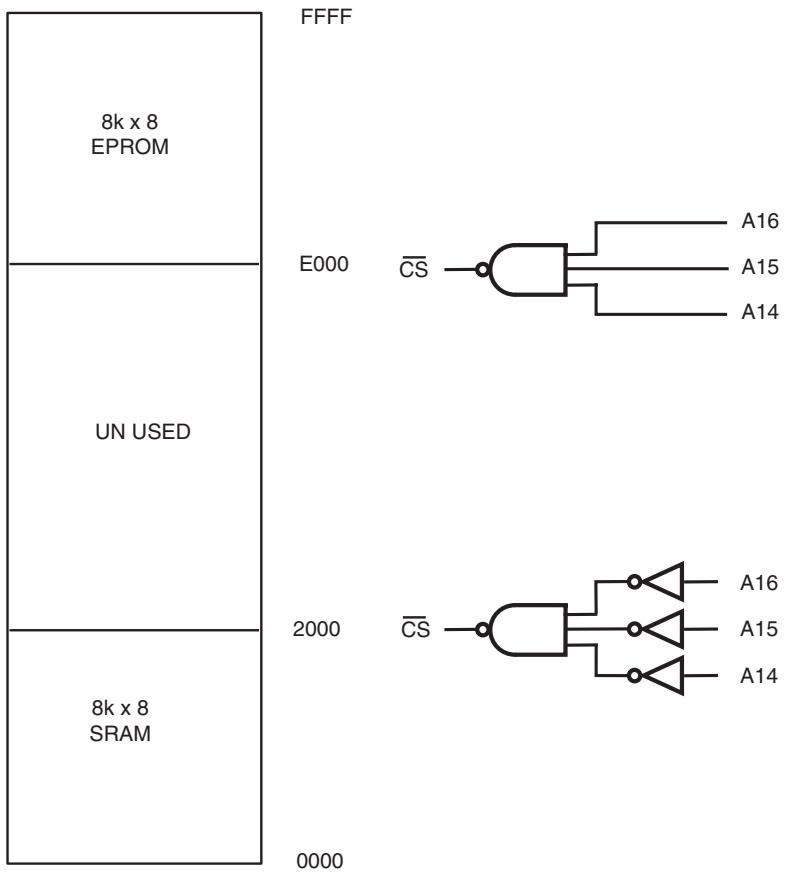


Figure 11.4

Memory map of a microprocessor system: two memory devices mapped at the top and bottom of memory with unused space in between; suitable address decoding logic is shown.

be at the bottom or the top of memory. Interrupts are handled in a similar way, an interrupt vector, or group of vectors being located at a fixed place in memory. This can be in system ROM, or software may be required to write the address of the interrupt service routine to the vector before an interrupt can call the routine. This latter method is the approach used in IBM PCs, and under certain circumstances the interrupt table can be modified by user programs.

Usually the first program or sequence of instructions to be run is required to set up input and output devices and other configurable peripherals. In general purpose microprocessor systems this first program usually prepares the system to run an operating system, and then loads the operating system program from disk or other storage media. This process is called **bootstrapping**, or simply **booting**. PCs have a ROM on the mother board, which contains software called the 'built-in input output system' (**BIOS**) – this sets up the essential functions of the peripherals, like disk, keyboard and video interfaces, performs the system power-up self-tests and attempts to load an operating system from a mass storage device like disk drive.

Programming

It is possible to program a computer in its native machine language, but this is not very understandable to humans and not normally necessary – although the first computers were programmed this way, and before the common availability of desktop computers microprocessors and microcontroller programs were frequently hand coded.

The following set of hexadecimal numbers represent 3 instructions for the AVR microcontroller made by Atmel. The AVR uses 16-bit instructions so each instruction is represented as a 4-digit hexadecimal number.

```
B396
3094
F0F1
```

In this form it is not very easy to understand. The program does the following set of actions:

1. loads register R25 with the value on the input pins of port B, whose address is $16;

2. then compares the value, now in register R25, with literal value 04;

3. if the values are equal, branches forward 30 words in the program.

In order to make things easier **assembly language** uses mnemonics (short words that are easy to remember) to represent op-codes in a more human friendly form; this time the program is a bit easier to understand.

```
IN   R25,$16
CPI  R25,4
BREQ  30
```

The format of assembly language instructions is often very strict because a computer program needs to convert the lines of assembly language into the binary numbers required by the target CPU. Symbolic names may be assigned to program sections, registers and ports to improve human readability (and write ability) of a program so the example could become:

```
DEFINE PORTBDATA $16
DEFINE TARGET 04

    IN   R25,PORTBDATA
    CPI  R25,TARGET
    BREQ  MATCH

MATCH:
```

When the assembler is run it substitutes the value $16 wherever it finds PORTBDATA and 04 where it finds TARGET; the assembler will also calculate the value to substitute for MATCH by working out how many memory locations separate the branch instruction and the label for the destination of the branch – this reduces errors because otherwise you would have to recalculate the value every time instructions were added or removed in between the branch and the label. Assemblers usually need several passes over the program to perform all the substitutions and address calculations.

Programming microcontrollers in assembly language has the advantage of giving direct insight into what the hardware is doing. Assemblers for microcontroller programming, provided by the microcontroller manufacturers, usually allow the use of **macros**, that is user-defined functions written in

assembler which can be represented by a key word. Macros are like subroutines in that they are reusable, so once a macro is defined it can be used repeatedly in a program wherever appropriate. The difference is that a subroutine is stored in memory outside the main program and requires an instruction to call it and on completion of its task must return execution to the address stored on the stack, and a macro name is expanded, inserting the macro definition in the program at assembly time; thus if the macro is used ten times, ten copies of it will be inserted in the code, taking up ten times the code space of a subroutine to do the same thing. Macros do not require stack space since they are not called, and execute faster for the same reason.

The alternative to assembly language is to use a language more abstracted from the hardware, for example the C programming language. There are good reasons for using C: if the application needs to perform complex mathematical calculations for example, or if the finished program is larger than a few hundred lines, C code will generally be easier to write and maintain. Languages like C are compiled rather than assembled, and in fact the output of the compiler may be assembly code that then needs to be assembled for the program to be used. The main differences between assemblers and compilers are that assembly language mnemonics have a one-to-one relationship with the machine instructions that they represent, a C program statement may be compiled as many machine code instructions, and the output generated by compiling the same C statement may vary depending on variable types, etc. Very few PC applications are written in assembly language; it is used, however, for writing hardware device drivers, or the libraries that the C and C++ compilers use for hardware access – these are the same sorts of programs as used in most microcontroller applications.

The ARM processor

The ARM processor family has been developed as cores that can be embedded in larger integrated circuits rather than as chips themselves. Advanced Risc Machines (ARM) license the design for use in applications from mobile phones to laser printers. A number of companies like Philips and Atmel manufacture microcontrollers based on an ARM core (Figure 11.5) with

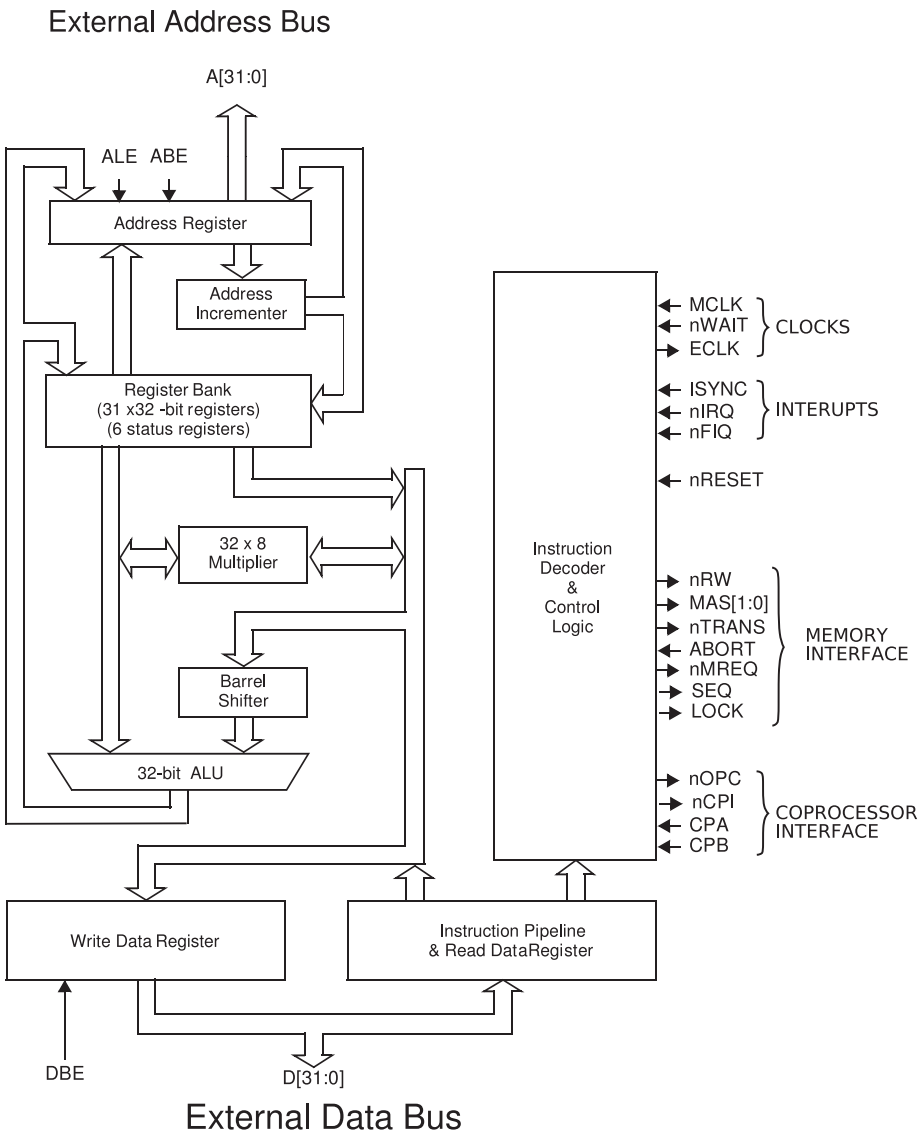


Figure 11.5

Simplified block diagram of the ARM 7 processor core.

peripherals and memory tailored for specific applications like set-top boxes and GPS receivers. The ARM 7 is a 32-bit microprocessor; microcontrollers using the ARM core are typically capable of addressing up to 64 Mbytes of external memory and operating at speeds of 20 MHz to 40 MHz, yielding up to 36 Mips. The ARM can access external data buses of 8-, 16- or 32-bit width, which means that different memory can be provided to meet cost or performance needs.

The ARM 7 is a RISC processor as the name ARM implies; the instruction set mnemonics are shown in Table 11.3. There are only 32 mnemonics to remember and even allowing for addressing modes and register names, it is not difficult to remember the ARM assembly language instructions and syntax. All of the ARM family are well supported by commercial assembler and compiler tools. The wide use of ARM microprocessors in consumer goods like phones and set-top boxes as well as the original Acorn Archimedes computer and later the Acorn RISC PC has driven the availability of ARM versions of Linux and the GNU Compiler Collection, and as a result there is a large amount of information available on the Internet regarding use of the ARM processor.

Developing microprocessor hardware

Developing applications that use microprocessors or microcontrollers is greatly assisted by the use of manufacturer development boards, both as software development targets and for understanding the hardware design required to support the chosen device.

Developing high-speed digital hardware like PC motherboards and graphic cards is generally not to be approached without significant engineering resources. At frequencies above a few tens of megahertz the interconnect between chips on a printed circuit board begins to exhibit the characteristics of transmission lines, that is inductance and capacitance conspire to delay signals and reflections occur if the track is not terminated in its characteristic impedance or has sudden changes in width. The physical design of a motherboard for a 3 GHz Pentium-class processor may take several months for a team of experienced engineers with access to sophisticated

Table 11.3 ARM 7 microprocessor assembly language mnemonics

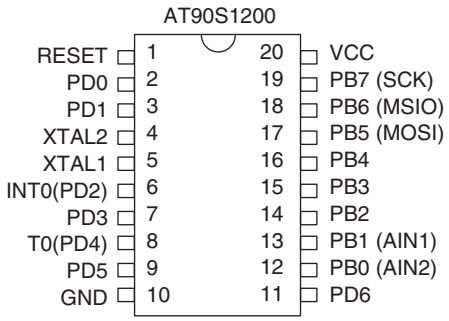
Mnemonic	Instruction action
ADC	Add with Carry
ADD	Add Rd
AND	AND Rd
B	Branch
BIC	Bit Clear
BL	Branch with Link
BX	Branch and Exchange
CDP	Coprocessor Data Processing (Coprocessor-specific)
CMN	Compare Negative CPSR flags
CMP	Compare CPSR flags
EOR	Exclusive OR Rd
LDC	Load coprocessor from memory
LDM	Load multiple registers
LDR	Load register from memory
MCR	Move CPU register to coprocessor register
MLA	Multiply Accumulate
MOV	Move register or constant
MRC	Move from coprocessor register to CPU register
MRS	Move PSR status/flags to register
MSR	Move register to PSR status/flags
MUL	Multiply
MVN	Move negative register
ORR	OR register
RSB	Reverse Subtract
RSC	Reverse Subtract with Carry
SBC	Subtract with Carry
STC	Store coprocessor register to memory
STM	Store Multiple
STR	Store register to memory
SUB	Subtract register
SWI	Software Interrupt
SWP	Swap register with memory
TEQ	Test bit wise equality
TST	Test bits

CAD tools and test equipment, but luckily the typical microcontroller clocked at 32.768 kHz to 20 MHz is easier to handle.

The majority of microcontrollers used are embedded in equipment performing simple automatic tasks, such as TV infra-red remote control, washing machine controllers, burglar alarms and answering machines. The Atmel chip shown in Figure 11.6 is typical of the microcontrollers used in these applications. The major microcontroller manufacturers between them have shipped in excess of 3 billion microcontrollers since the year 2000. One of the largest markets for microcontrollers is intelligent battery protection in mobile phones.

Figure 11.6

Pin connections of the Atmel AT90S1200.



A few simple rules should help towards achieving success in hardware design; digital logic, generally CMOS-based, requires well decoupled power supplies. Manufacturers' data sheets usually give recommendations for power and ground connections and these should be followed closely. The general rule should be to keep connections as short as possible, to minimize loops in current paths and to decouple all power pins. Signal routes should also be as short as possible and signal returns should form the smallest area loops possible. Double-sided PCB with a ground plane is essential if clock frequencies above about 10 MHz are to be used in microprocessor systems with external address and data buses; the advantage of single chip microcontrollers is that in many applications very few high-speed signals need be routed around a PCB.

Crystal oscillators for clocks are among the most problematic areas in microcontroller hardware design. Crystals should be as close as possible to the oscillator pins of the microcontroller, the ground connections of the oscillator capacitors should go directly to the nearest ground of the

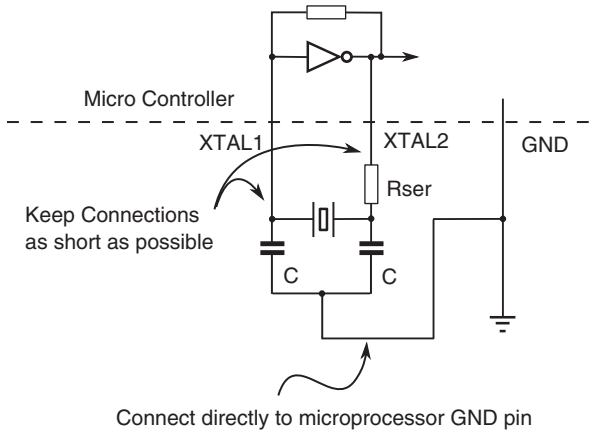


Figure 11.7

Crystal oscillator connections must be as short as possible.

microcontroller and the track should not be shared with any other ground connection (Figure 11.7). If a solid ground plane is used, the capacitors should be connected to the ground plane as near to the microcontroller ground pin as possible.

Electromagnetic compatibility

Microprocessor systems with fast clocks are often sources of radio frequency noise because it is not possible completely to eliminate radio frequency radiation short of putting the system in a sealed metal box. If all else fails, sealed metal boxes can be used in the form of metal screening cans soldered on to the PCB with low-pass filters on signals entering and leaving them. Good design practice and early testing make the process of meeting the levels required by the EMC directive in Europe and the FCC in the USA much easier.

Microcontroller manufacturers

Microchip
www.microchip.com
PIC 10F, 12F, 16F, 18F and 24F microcontrollers and support ICs

Atmel
www.atmel.com
AVR, ARM and 8051 based microcontrollers

Intel
www.intel.com
The original 8051 microcontroller, MCS51

ARM
www.arm.com
ARM microprocessor core

ST Microelectronics
www.st.com
ST6, ST7, ST8 microcontrollers and support ICs

Philips Semiconductor
www.philips.com
ARM and 8051/8751 based microcontrollers

Zilog
www.zilog.com
The original Z80 microprocessor and the Z8 family of microcontrollers

The GNU compiler collection
www.gcc.org www.gnu.org

The small device C compiler, microcontroller C compiler, based on GNU/GPL tools
www.sdcc.org

CHAPTER 12

MICROPROCESSOR INTERFACING

Even the simplest microcontroller system has to interface to the outside world in some way. There are many ways by which this is achieved and this chapter is a survey of some useful techniques. The reader will note that the methods mentioned here are aimed at the sort of microcontroller system, usually described as **embedded applications**, that might be built into systems to provide user interface or specific automation functions. It is not the purpose of the chapter to cover general purpose computers or personal computers.

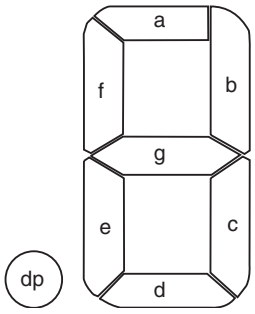
Output circuits

Display devices

The seven-segment numeric display is probably the single most recognizable feature of the digital age. The segments of the display are named by convention 'a' to 'g' with an eighth segment providing a decimal point. Figure 12.1 shows the arrangement and the state of the segments for displaying the numbers 0 to 9. The hex codes assume that the most significant bit (MSB) drives the 'a' segment; the decimal point is then the least significant bit and can be turned on by adding 1 to the hex code. Driving a display in this way uses one 8-bit-wide microprocessor port per digit.

Light-emitting diode (LED) displays

Seven-segment displays have been made with light-emitting diodes, liquid crystal displays, vacuum fluorescent displays, filament lamps and even solenoid-operated flags. We will concentrate on the LED (Figure 12.2)



Digit	Segments	HEX Code
0	a b c d e f	FC
1	b c	60
2	a b d e g	DA
3	a b c d g	F2
4	b c f g	66
5	a c d f g	B6
6	a c d e f g	BE
7	a b c	E0
8	a b c d e f g	FE
9	a b c d f g	F6

Figure 12.1

Seven-segment display and segment drive information, suitable for microcontroller applications.

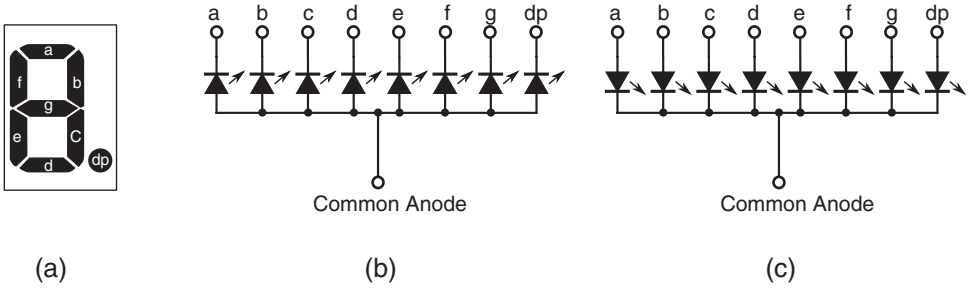


Figure 12.2

Seven-segment LED displays: **(a)** segment arrangement, **(b)** common anode and **(c)** common cathode connections.

and LCD types here since these are the commonest and have the best ranges of power consumption, cost and size.

Common anode and common cathode seven-segment LED displays are available in all the usual colours – red, green, yellow and even blue; they are commonly made in heights of 7.62 mm, 9 mm, 14.2 mm and 25.4 mm, and even up to 100 mm. The larger the display, the more power is required to light it. A typical 7.62 mm high display requires 7 mA for reasonable brightness, which conveniently is about the same as the drive capacity of CMOS microprocessors; however, driving more than a couple of digits

could easily approach the maximum supply pin or ground pin currents for the microcontroller, when several digits display eights for instance. As with other LEDs it is important to limit the current flow through the LED and so typically for a 5 V supply, resistors of between 330 Ω and 470 Ω are used, based on the forward voltage drop for a red LED being about 2 V.

While a microcontroller can drive displays directly it is not always an efficient use of I/O pins. Display driver ICs designed to drive displays at higher currents and decode binary coded decimal inputs, thus using fewer I/O pins, are available. Figure 12.3 shows the 74HC4511, a common cathode, binary coded decimal (BCD) to seven-segment decoder driver. The 4-bit BCD input has a latch, controlled by the active low latch enable input ($\overline{LE}$), allowing multiplexing of microprocessor outputs or being driven directly from the outputs of a counter. The IC also has two other inputs: lamp test ($\overline{LT}$), which can be used to light all the segments of the display as part of built-in test, and blanking input ($\overline{BI}$), which can be used to blank leading zeros in displays; it is common practice to use the lamp test input to flash the display to indicate over-range errors.

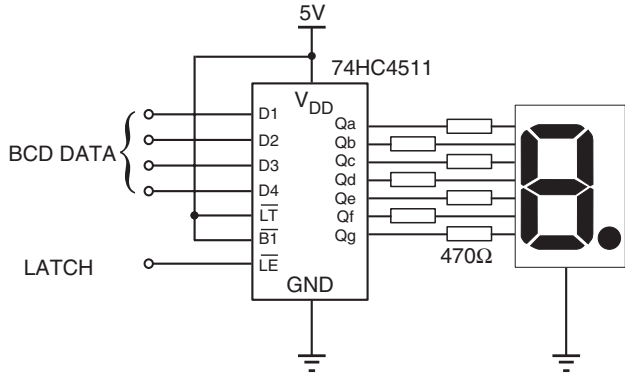


Figure 12.3

A common cathode LED display and driver IC.

When more than a few digits are required in a display it is usually convenient and cost effective to multiplex the display drive electronics.

The display in Figure 12.4 has 8 digits but only a couple of ICs, 8 resistors and 16 FETs that would probably be in the form of a packaged array. The whole circuit requires only an 8-bit port to drive it. In operation the digits are each driven in turn, but they are switched on and off sufficiently rapidly that the viewer is fooled into thinking that they are all on at once. The human eye is not able to discriminate on and off periods for light flashing at more than about 40 flashes per second, so if the display is driven faster than this it appears to be on all the time. In effect the eye integrates the light reaching it over time and so sees the average brightness rather than the flashing.

In operation the microprocessor controlling the display must update each digit in turn at least once every 20 ms or so. Each digit must be on for about one eighth of the time, 2.5 ms in 20 ms. This limits the brightness of such displays and 8 digits is probably close to the limit of practicality; however, because the display is not on continuously it can be driven with more current and thus give brighter light when it is on.

The circuit is designed so that the microcontroller outputs 3 bits to the 74HC138; a three- to eight-line decoder with active low outputs turns on one p-channel FET at a time, connecting the common anode terminal of the selected digit to the supply rail. At the same time another 4 data bits of the microcontroller output drive the 7HC4511 decoder IC, which turns on n-channel FETs to drive the individual segments via 120 Ω resistors giving a maximum current of 25 mA per segment. The decimal points are driven by the last remaining bit of the microcontroller port.

Usually such a system would be driven under interrupt control by the microcontroller; 20 or 30 instructions once every 2.5 ms is not a huge overhead, probably about 1% of the processing power of a simple microcontroller.

Seven-segment displays can display numbers and a few other characters; more complex segmented displays are available that allow a wider range of alpha-numeric characters to be displayed but dot matrix displays are more versatile, giving the option of graphical symbols as well as user-defined patterns. Figure 12.5 shows the arrangement of a 7 × 5 array that can be used to build up larger displays, for example the moving message displays used in airports and railway stations.

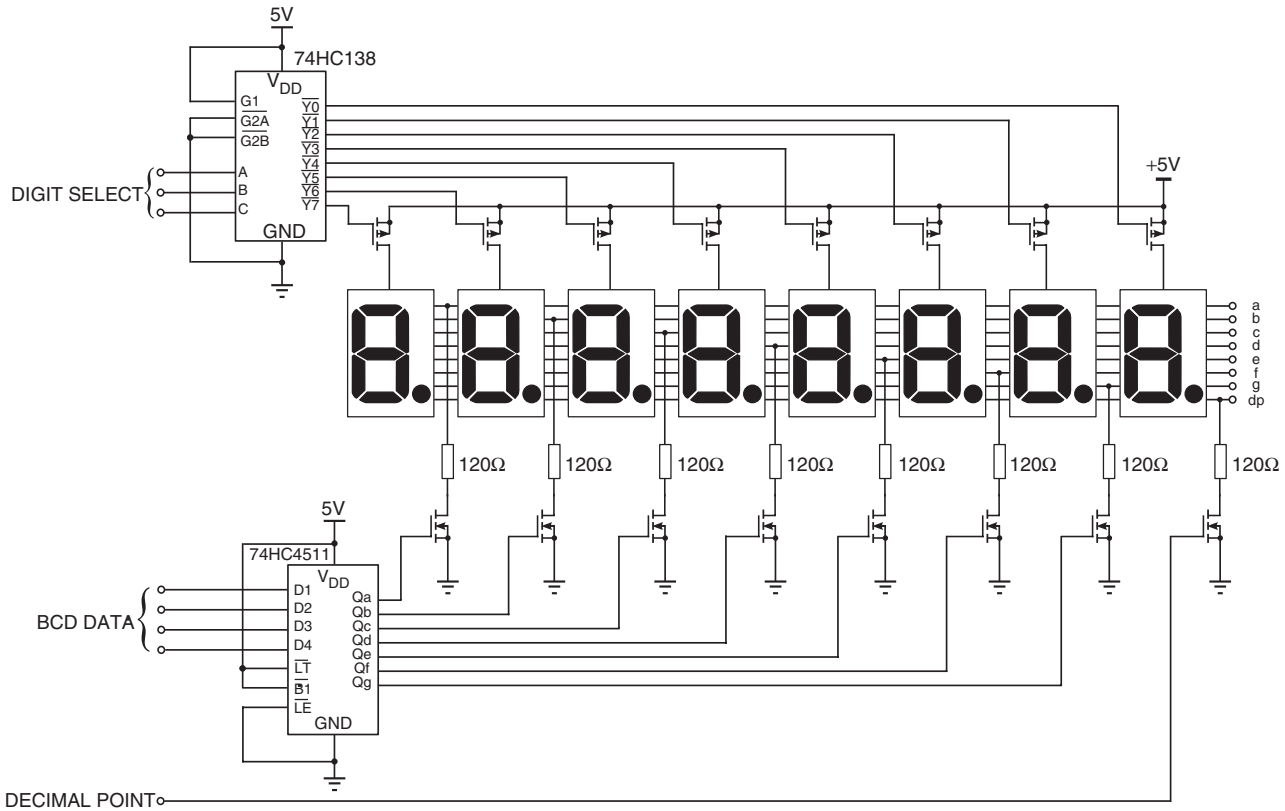


Figure 12.4

A multiplexed display that has eight digits and requires only one 8-bit microprocessor port to drive it.

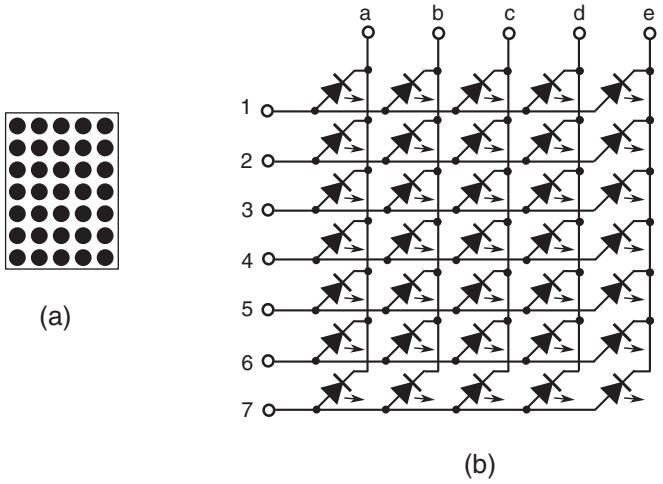


Figure 12.5

(a) Typical form of a 7 × 5 dot matrix LED display and **(b)** schematic of LED matrix.

Liquid crystal displays (LCDs)

Liquid crystal displays can be used in most of the applications where LED displays are found. They have some advantages as compared with LEDs; they use much less power and so are ideal for battery-operated equipment, and they can be made thinner and even transparent where required. LCDs have two important disadvantages however, these being viewing angle and temperature sensitivity.

The display relies on the fact that the molecules of particular organic compounds are cylinder shaped and electrically polarized. In the absence of any applied electric field the molecules align with each other, but if an electric field is applied across a sample of the liquid the molecules will rotate to align along the applied field gradient. The display is formed by sealing the liquid between two transparent sheets of glass with metal contacts deposited on them. There are various means of operation but generally the method relies on the molecules either being aligned to the surface of the glass when no field is applied allowing light to pass, or being aligned to the field to stop light passing. The display may be lit from behind or can have a mirror behind it and relay onwards the ambient light being reflected.

The temperature sensitivity effects result from the properties of the liquid, meaning that very low and very high temperatures reduce the contrast of the display. The field drive to the display must not contain a DC component because electrolytic disassociation of the liquid will occur if so.

The metallization patterns can be of any shape; examples of the common plate and segment patterns for a seven-segment display are shown in Figure 12.6. The metal film contacts are transparent, and in fact the resistance of the segment material can be many thousands of ohms.

Liquid crystal displays present a capacitive load to the drive circuit so that the drive current requirement is very low, typically less than 1 μA per segment at 50 Hz, and a drive voltage between 4 V and 15 V peak-to-peak may be used. These characteristics make CMOS ideal to drive LCD displays and there are many custom LCD drive chips and microcontroller ICs with LCD drivers. It is possible to drive an LCD display from a standard seven-segment decoder using XNOR gates and a clock signal to produce the differential drive signal required (Figure 12.6b).

Dot matrix LCD character display modules are available in standard sizes of 16 or 40 characters by one, two or four lines. There is a de facto standard for the interface based on the Hitachi HD44780 chip. The display has an 8-bit data interface, which can optionally be driven in 4-bit mode, using just the upper 4 bit inputs and 3 control lines making seven connections in all. The control signals are Enable, Register Select and Read/Write. There is also a display bias voltage input pin, which can be used to set the display contrast; this can be connected to the wiper of a 10 kΩ potentiometer connected between the supply and ground.

In operation the register select (RS) signal sets whether the input is data (low) or an instruction to the display driver IC (high). The read/write signal sets the controller IC to read or write. Data is written on the falling edge of the E signal.

The module contains a character ROM containing ASCII characters and some symbols. Some displays also offer user programmable character cells. Figure 12.7d shows typical character patterns and cursor position.

Figure 12.7c shows the relationship between the interface signals. The HD44780 and its derivates have a relatively complex start-up procedure,

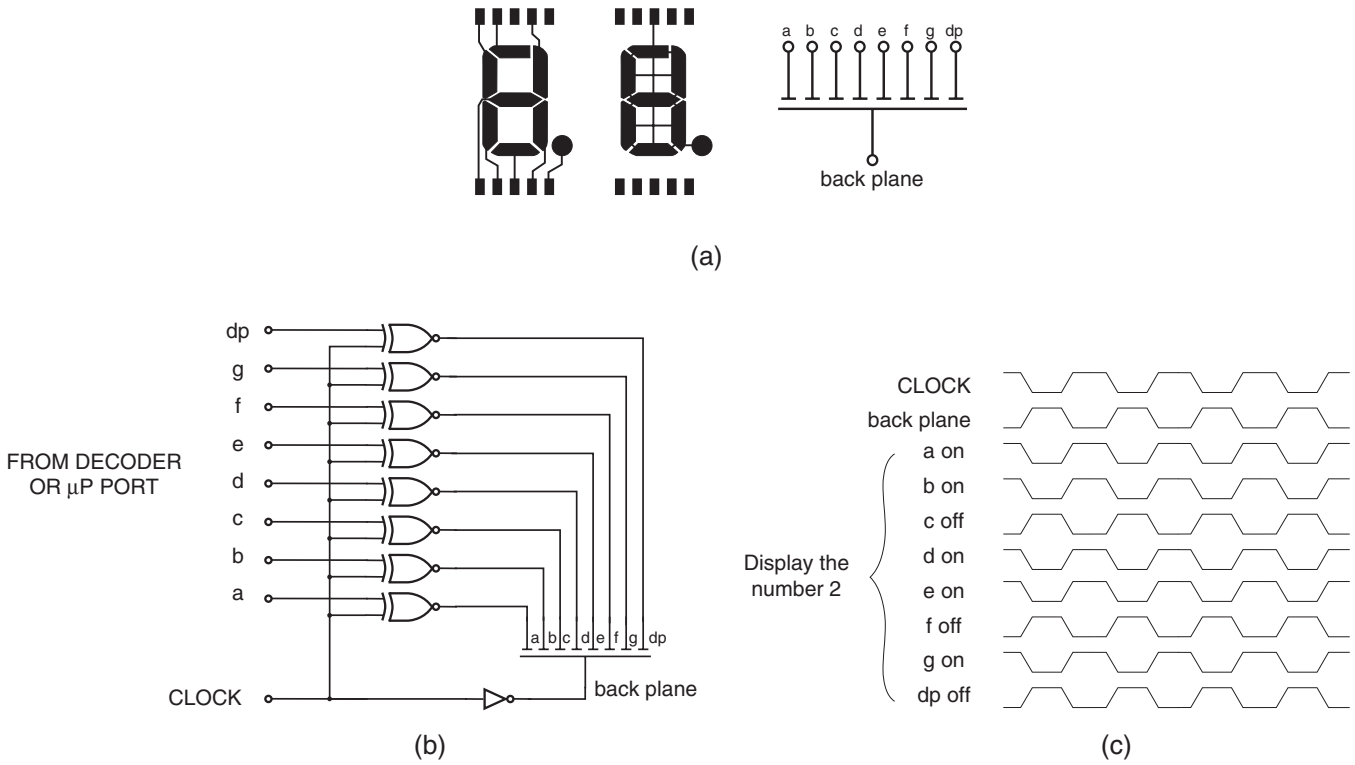
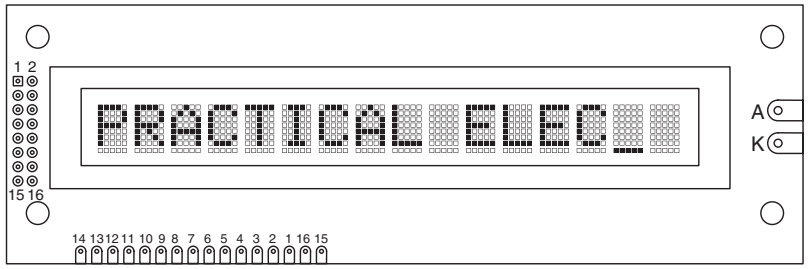


Figure 12.6

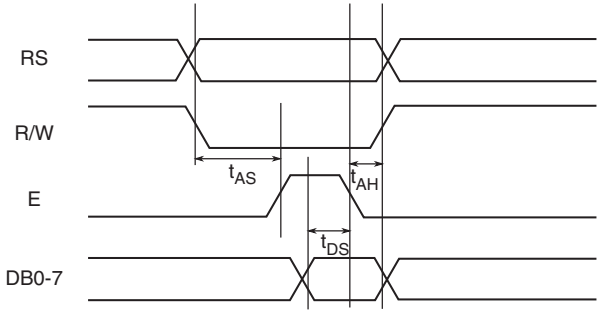
(a) Segment metallization patterns and equivalent circuit symbol, **(b)** LCD display drive circuit and **(c)** drive signals.



(a)

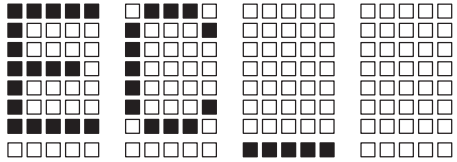
PIN	FUNCTION
1	VSS
2	VDD
3	Vo
4	RS
5	R/W
6	E
7	DB0
8	DB1
9	DB2
10	DB3
11	DB4
12	DB5
13	DB6
14	DB7
15	NC
16	NC
A	BACK LIGHT LED ANODE
K	BACK LIGHT LED CATHODE

(b)



t_{AS}	40ns	Address/Register Setup time
t_{DS}	60ns	Data Setup Time
t_{AH}	10ns	Address/Data Hold Time

(c)



(d)

Figure 12.7

(a) LCD character display module with **(b)** pin connection information for the Hitachi HD44780 and compatible driver ICs, **(c)** drive signals and **(d)** display elements, showing cursor.

because the chip requires a sequence of bit patterns and delays to set up the interface in either 4-bit or 8-bit mode before it can be communicated with. Most microcontroller vendors provide example code optimized for their ICs to drive these displays.

Dot matrix graphical displays are also available; these are largely similar to character ones, but the parallel interface is often simpler and I2C and SPI bus versions are available as well.

Microcontrollers are often required to drive devices other than displays. Infra-red LEDs, filament lamps and heaters can usually be switched on with simple transistor switching circuits like those shown in Figures 12.8a and b.

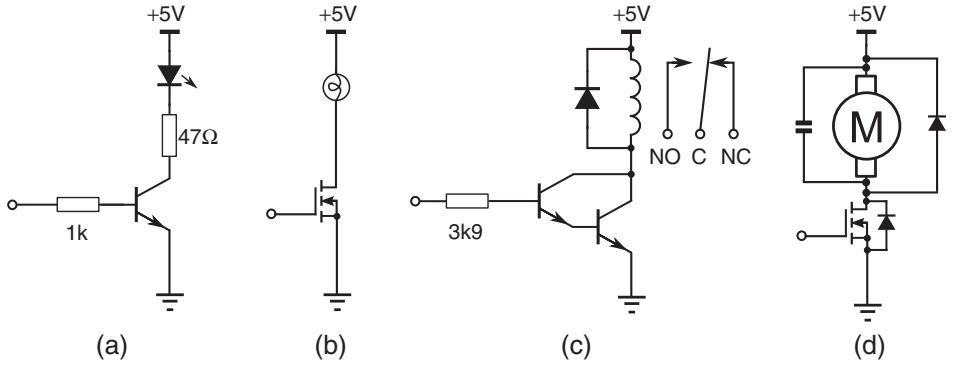


Figure 12.8

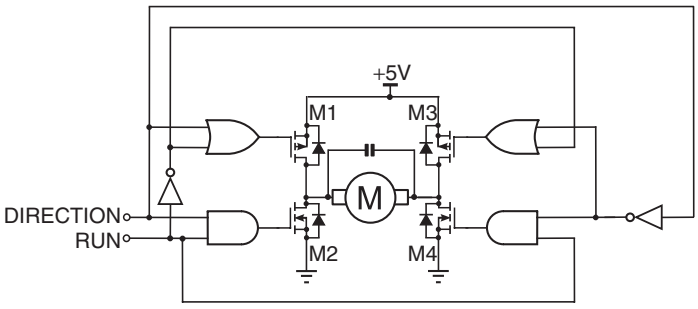
Using transistors to switch loads larger than microprocessor or logic outputs can drive **(a)** infra-red LED, **(b)** filament lamp, **(c)** power relay and **(d)** small permanent magnet DC motor.

Relays can be used to turn on larger loads and provide isolation from the load supply. This is particularly useful for mains voltages; because of the inductive nature of the relay coil it is important to protect the driving transistor. When the transistor turns off, the voltage across the coil will rise rapidly as the energy stored in the magnetic field is returned to the circuit. A diode, referred to as a flywheel diode or catch diode, connected across the coil as shown in Figure 12.8c, clamps the EMF due to the collapsing magnetic field at 0.6 V above the supply rail, protecting the transistor from over-voltage breakdown and possible permanent damage.

Darlington driver ICs with 7 or 8 integrated Darlington pairs are made by a number of manufacturers; some even have integrated protection diodes.

DC electric motors can also exhibit back-EMF effects and so when driving them with a transistor it is wise to provide catch diodes. DC motors also cause electromagnetic interference problems if the commutator is not suppressed with a capacitor; typically a 100 nF 50 V or 100 V disc ceramic capacitor soldered directly across the contacts on the motor will be adequate.

DC motors often need to be run in both clockwise and anti-clockwise directions. The circuit shown in Figure 12.9 is a bridge driver; such drivers are available as integrated circuits from several manufacturers but for small motors it is often cost effective to build this bridge with small discrete power FETs and conventional 74HC series logic. Each half of the bridge is very much like the tri-state CMOS gate output stage (see Figure 9.10). The motor spins one way when the current flows from left to right, i.e. M1, M4 turned on and M2, M3 turned off, and spins in the opposite direction when the current flows from right to left, M2 and M3 turned on.



(a)

RUN	DIR	M1	M2	M3	M4
0	0	OFF	OFF	OFF	OFF
1	0	ON	OFF	OFF	ON
1	1	OFF	ON	ON	OFF

(b)

Figure 12.9

Reversible DC motor drive: **(a)** schematic and **(b)** FET states for each combination of control input signals.

The motor circuits shown in Figures 12.8 and 12.9 are on–off drives. If speed control needs to be implemented, using a pulse width modulated output from a microcontroller can provide the solution. PWM can be

implemented in software at low frequencies using interrupts but many microcontrollers have peripheral hardware PWM modules.

Input circuits

Switches

The switch should be the simplest device to interface to a microprocessor or logic circuit, but, being mechanical, it exhibits an effect known as **contact bounce**: when one switch contact is moved to contact another there is often a mechanical oscillation, a series of make–break cycles (Figure 12.10).

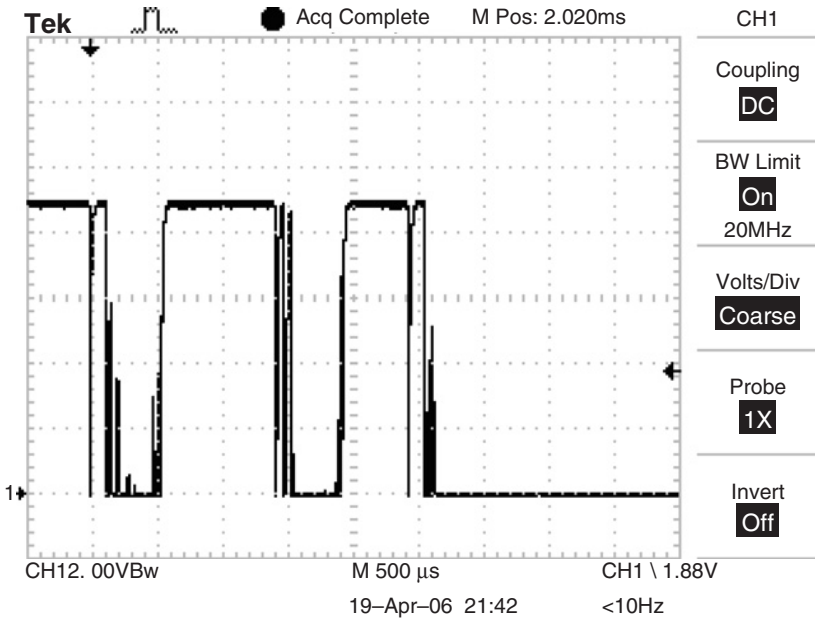


Figure 12.10

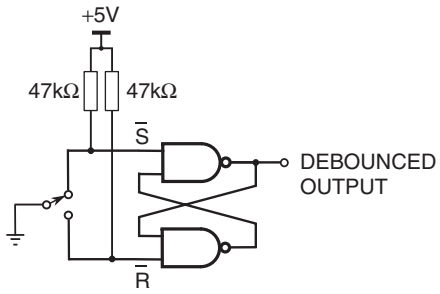
Contact bounce captured with an oscilloscope. The sequence lasts for less than 2.5 ms. A fast digital input would detect at least six transitions.

Contact bounce is too fast to be noticed when switching on a light but when electronic circuits that can respond in 10 nanoseconds (1 ns = 10^{-9} s) are being driven, many individual make–break cycles can be detected. The amplitude of the bounce is tiny; the contact movement is in the order of a few hundredths of a millimetre.

The RS flip-flop (Figure 12.11) can be used with a change-over switch and a couple of resistors to provide clean logic inputs to a system. The RS flip-flop 'remembers' the previous state and therefore is only triggered once per changeover; this works because the moving contact cannot bounce back as far as to the other contact.

Figure 12.11

RS flip-flop used to de-bounce a switch input.



Single-pole switch inputs can be debounced in software; a software loop reads the input until it stops changing, and has remained stable for a predetermined length of time, before registering the input status.

Keypads, made up from arrays of push switches, provide a convenient way for humans to interact with microcontrollers. In operation the controller scans the array by pulling each of the row outputs to ground in turn; if one or more of the switches on that row is pressed the column to which it is connected is pulled low also. The column state (bits 4, 5, 6 and 7) is then read for each row and de-bouncing is performed by software. Figure 12.12a shows a keypad circuit suitable for use with standard CMOS inputs and outputs. The diodes isolate the unselected row and the pull-up resistors are required to bias the CMOS inputs. Figure 12.12b shows how much simpler it can be if the microcontroller has tri-state outputs and internal pull-up resistors for the inputs, as many microcontrollers do have. Using a single 8-bit port allows a 4 × 4 array; using a 3- to 8-line decoder with

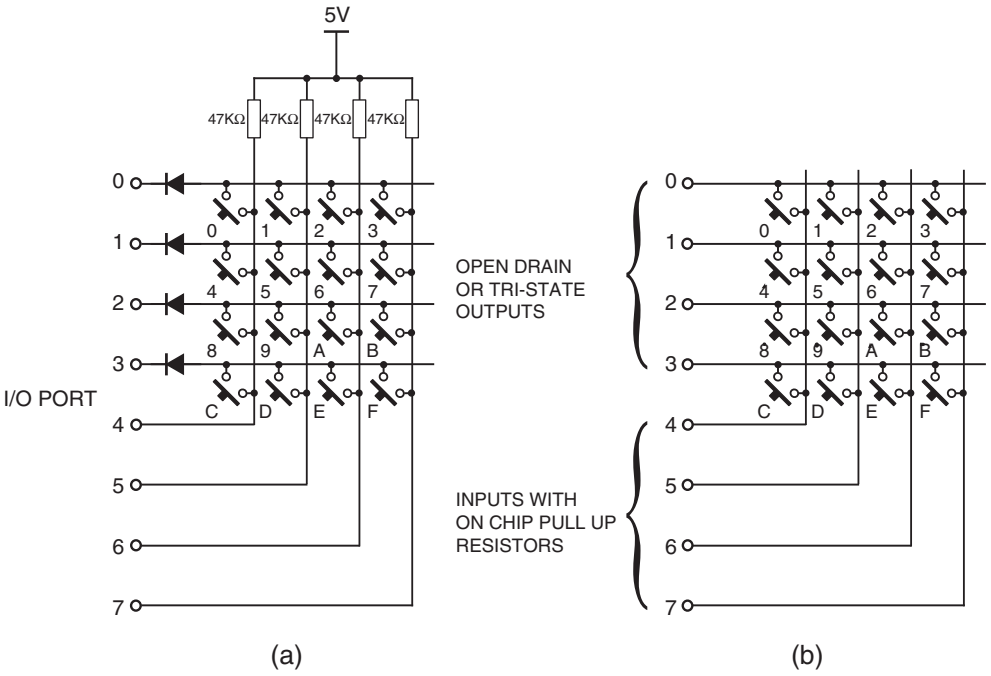


Figure 12.12

Keypads: **(a)** microprocessor logic I/O and **(b)** microcontroller version, a 4-input NAND gate with inputs connected to pins 4, 5, 6 and 7, can be used to generate an interrupt when any key is pressed.

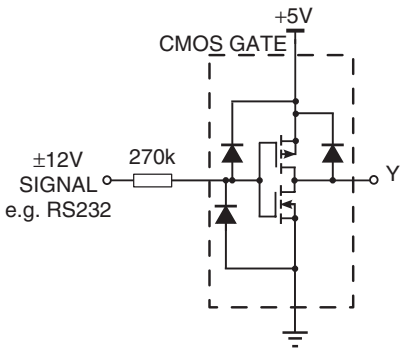
active low outputs like the 74HC138 allows a single 8-bit port to address an 8 × 5 matrix.

The speed at which the keypad is scanned can be quite slow in microcontroller terms, typically 120 ms between polling each row under interrupt control. This does not impose a heavy overhead on processing time even for a PIC processor which needs 4 clock cycles per instruction running from a watch crystal; at 32.768 kHz this is only about 2.5% of the available processor cycle.

Sometimes it is necessary to use an input that exceeds the supply voltage range of the power supply. Level shifters and resistive dividers can be used but for some applications there is a convenient feature of the CMOS input that can be employed. CMOS devices are very static sensitive and have input

Figure 12.13

Using CMOS inputs outside the IC supply voltage range.



and output protection structures which include diodes clamping the input line to the supply and ground rails. These diodes typically have an absolute maximum current handling capacity of ±20 mA, so a current-limiting resistor is necessary to protect these diodes. Limiting the input current to between ±20 μA and ±200 μA ensures safe operation. Figure 12.13 shows a CMOS gate with a 270k input resistor which allows it to safely convert RS232 input levels ±12 V; this is very much simpler and cheaper than using an RS232 level converter. This circuit can also be used to detect the zero crossing of the mains and other AC signals, if a suitable power supply transformer is used. It is important to ensure that the correct series resistor is used to guarantee that the input current is limited to a safe value.

This page intentionally left blank

CHAPTER 13
DATA CONVERTERS

Introduction

The physical world is a place that is characterized by processes that are continuous in time and amplitude. Parameters such as temperature, speed, pressure and length in the physical world are all continuous functions of time and value; that is, for the temperature of an object to change from one value at one time to a different higher or lower value at a later time, it will have to pass through every arbitrarily chosen intervening temperature. No parameter can change value instantly.

When we measure these parameters it is normal to use a sensor to produce a voltage or current output that varies in sympathy with the parameter that we are measuring, so we might, for example, represent temperature between 0°C and 100°C as a voltage between 0 V and 10 V. The voltage then is varying in ratio 100 mV/1°C with the parameter that we are measuring. The word analogue is derived from the Greek *analogia* (*ana*, according to; and *logos*, ratio) meaning equality of ratios and is commonly used to describe electrical signals that represent measured parameters and the systems that process these signals directly.

The advent of digital electronic systems at the end of the second world war, followed by the invention of the transistor and then the integrated circuit changed the way we process, display and store data. Digital systems are now dominant, replacing devices such as moving coil meters and chart recorders in almost all applications. In order for digital systems to process the parameter data it first has to be converted into digital form so that analogue to digital converters are required. There are many applications where a digital system has to *control* parameters as well as measure them so digital to analogue converters are also required, in fact the successive approximation

analogue to digital converter uses a digital to analogue converter internally along with a comparator.

There are some applications that use sensors that have effectively digital outputs, the thermostat for example closes or opens its contacts depending of the setting of a predefined temperature trip point, in effect we can view this as a one bit analogue to digital converter.

There is a growing trend to integrate silicon sensors and analogue to digital converters with data communications functions. The development of nanotechnology and integrated semiconductor/thin film sensors has lead to many automotive and medical electronics applications. It is convenient and often essential to have integrated calibrated sensors and converters, the advantages are that there is no need to carry small analogue voltages or currents in noisy environments because digital signalling with much larger signals can be used, also the sensor and converter can be matched in the same environment which can help with accuracy.

Digital-to-analogue converters (DACs)

Most digital systems use binary numbers, so to produce an analogue output it is necessary to assign a scale to the digital values; therefore if we want an output of between 0 V and 5 V and have 8 bits of unsigned data we would need to give the bits the weights shown in Table 13.1.

Table 13.1 Bit weights for unsigned 8-bit DAC

Bit number	2^n	$(2^n/2^8)*5V$
bit 0	1	0.01953125 V
bit 1	2	0.0390625 V
bit 2	4	0.078125 V
bit 3	8	0.15625 V
bit 4	16	0.3125 V
bit 5	32	0.625 V
bit 6	64	1.25 V
bit 7	128	2.5 V

The sum of all bits would give 5 V output; the least significant bit (LSB) is 19.6 mV, so this is the smallest change that can appear at the output of the converter, i.e. the difference between successive binary values.

There are many ways to generate outputs in these ratios; this section covers some of the main techniques.

Digital potentiometer

Probably the simplest D to A converter (Figure 13.1a), is a chain of resistors of equal or log scaled ratio, with switches at the nodes as shown. One switch is turned on at a time, and the voltage at the chosen node is available at the output. This is in reality a digital potentiometer; it is commonly available with chains of 64 to 256 resistors.

Since digital potentiometers are designed to replace a conventional potentiometer in a circuit they are not usually capable of fast switching; however, they often have integrated non-volatile memory to store the switch setting. Typically these devices are set or programmed using serial data via the I2C or SPI bus. Digital potentiometers find uses in automatic calibration circuits, variable gain amplifiers and many other applications where a conventional potentiometer could have been used. One of the advantages of resistor chains is that since they can all be identical in value they can be well matched on an IC; this makes for good linearity and accurate steps. The converter output is inherently *monotonic*, meaning that the relationship between analogue output and digital input data always has the same sign; therefore the output value always changes in the same direction with increasing input data value.

A disadvantage is that since it requires one resistor per step there are 2^n resistors for an n bit converter. The 10-bit 1024 resistor converters are close to the practical limit for this kind of converter.

Binary weighted resistor converter

The binary weighted resistor converter is typically implemented as shown in Figure 13.1a. The converter uses only one resistor per bit in addition to the bias and gain setting resistors for the summing amplifier. The disadvantage

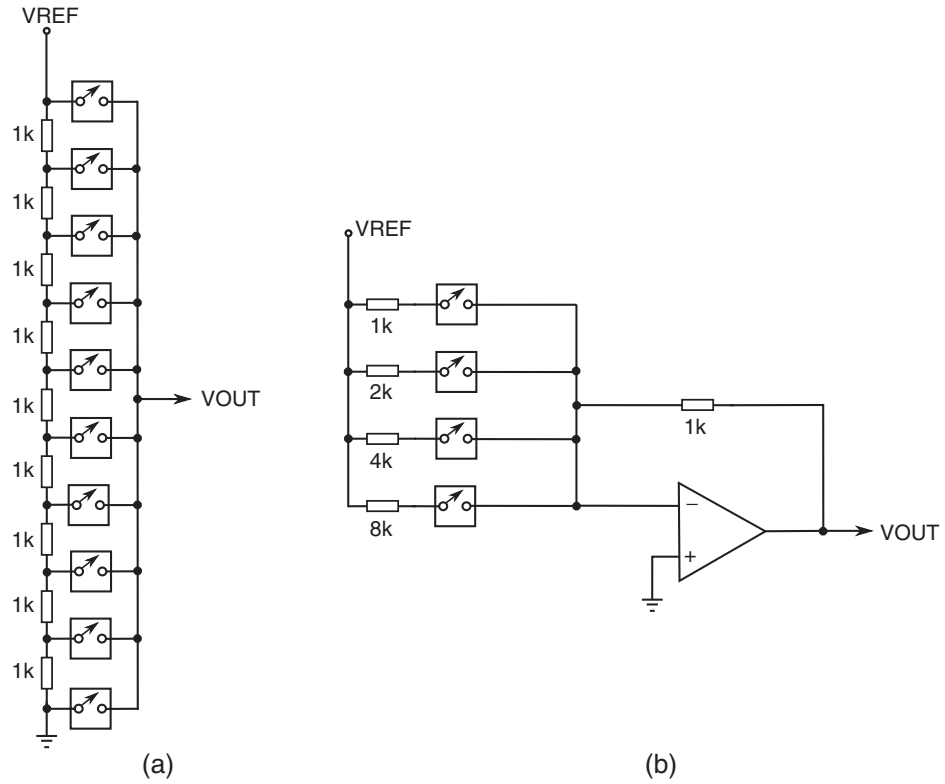


Figure 13.1

(a) Digital potentiometer, **(b)** weighted resistor DAC.

is that matching the resistors is critical to achieve linearity and monotonicity (see later), and in fact the range of resistor values required to make this type of converter are generally impractical as regards implementation. An 8-bit converter needs a 128:1 ratio between the LSB and MSB resistor, and therefore the tolerance of the MSB resistor needs to be better than 0.78% to avoid missing output values and better than 0.039% to match a 5% LSB resistor. In order to realize such ratios on an IC you typically use 2^n identical resistors and connect them into series chains to produce the weighted values, so the LSB resistor thus includes 128 resistors, while the MSB is a single resistor. It is faster and requires fewer switches than the digital potentiometer. This type of converter is most often used where only a few bits, say 4 or 6, are required and performance is not critical.

The R2R ladder

The difficulties associated with integrating weighted resistor converters led to the development of the R2R ladder. This solves problems of both the weighted resistor converter and the digital potentiometer. Because the converter only uses two resistor values the matching is considerably easier to achieve, this making it easy to integrate on an IC. The R2R ladder can be implemented as either a current or voltage mode converter.

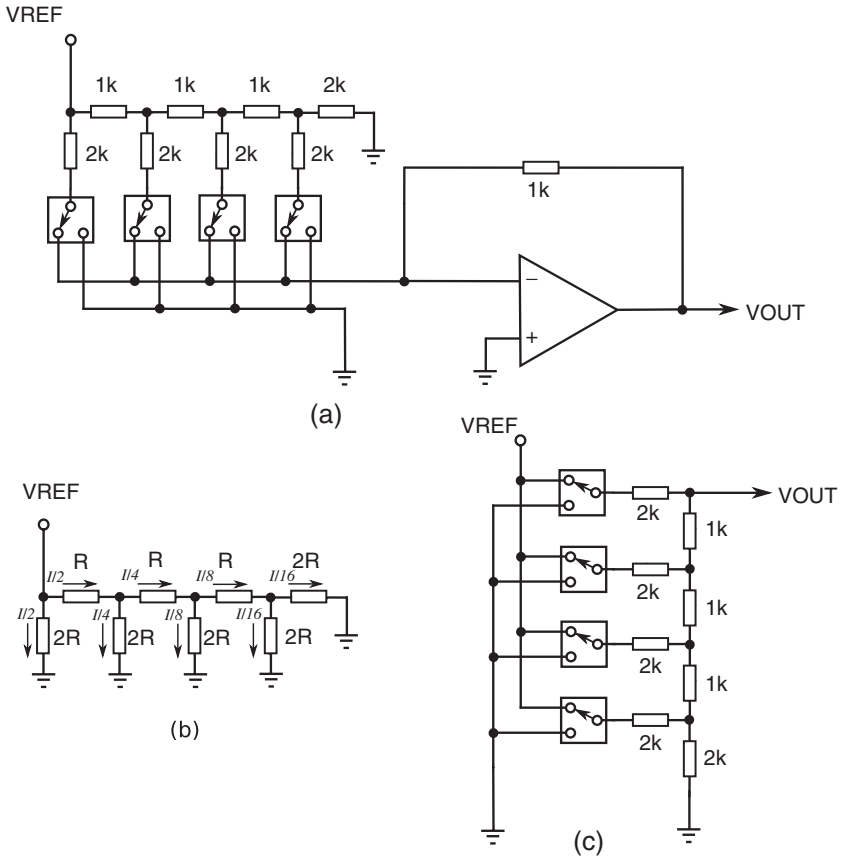


Figure 13.2

R2R ladder network: **(a)** current mode R2R ladder DAC, **(b)** current divider operation of R2R ladder, **(c)** voltage mode R2R ladder DAC.

In current mode it is similar to the weighted resistor converter. Rather than scaling the resistors so that applying the same constant reference voltage across each of them give scaled currents, the resistor network and switches are designed with two values resulting in successive halving of the voltage at each stage and, thus, binary weighted currents, as shown in Figure 13.2b. The network is arranged so that half the current from each node is either fed to the summing node (for a one) or dumped to ground (for a zero). The total current arriving at the output is thus directly dependent on the binary word that sets the switches.

In voltage mode the R2R ladder is driven in reverse; it is particularly convenient that the typical CMOS output stage looks quite like a changeover switch at reasonable currents and thus a resistor network can be driven directly from microprocessor pins. The resistors of the R2R ladder must be well matched, particularly if 8-bit or greater resolution is required. Resistor matching is less easy to achieve with discrete components; typically, 1% parts should be used but a better solution is to use an array of resistors manufactured as an R2R ladder. Resistor arrays in the form of R2R ladders are manufactured as either thin film or thick film parts in both pin-through-hole and surface-mount packages – these are designed to be used with CMOS microprocessor output ports.

Charge distribution DAC

The dominance of CMOS integrated circuits has led to the development of DAC topologies that are more suitable for integration in standard CMOS processing. It is easier to match values and takes less space on the die to use capacitors in place of resistors in these circuits, so the charge distribution DAC is frequently used with a successive approximation register and comparator to make an ADC in low-cost CMOS circuits like microprocessors.

The charge distribution DAC is inherently serial in its architecture; the schematic shows that, apart from four switches and a buffer amplifier, it consists of just two capacitors, which must be closely matched.

The operation of the charge distribution DAC is quite straightforward. Referring to Figure 13.3, initially C_1 and C_2 must be discharged by closing S_2 and S_4. The DAC is inherently serial in nature and receives the LSB

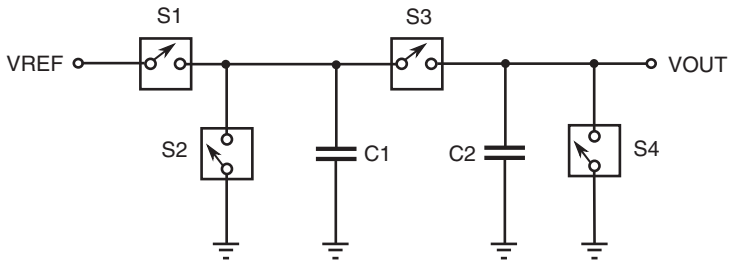


Figure 13.3

Charge distribution DAC.

of data first. If this is zero, S_2 is closed, keeping C_1 discharged; if it is a one, S_1 is closed instead and C_1 is charged to V_{ref}. Before the next bit is processed S_1 and S_2 are opened and S_3 is closed to equalize the voltage on C_1 and C_2. If the bit was a one, the resulting voltage is half V_{ref}.

The next bit is processed in the same way: if it is a zero, S_2 discharges C_1 and then S_3 equalizes the capacitor voltage at a quarter of V_{ref}; if it is a one, S_1 is closed and C_1 charges to V_{ref}; and when S_3 is closed it causes the capacitor voltage to equalize at three quarters of V_{ref}. This process is then repeated for the remainder of the bits, typically 8 to 12 bits. S_4 is used only before the first bit to guarantee that C_2 is discharged. Note that for accuracy to be maintained the match between C_1 and C_2 must be at least as good as the ratio of the LSB to full scale; so for an 8-bit converter this is a match of better than 0.4%, and for 12 bits the match has to be better than 0.02%.

Pulse width modulator

Probably the commonest form of DAC provided as a peripheral in single chip microprocessors is a pulse width modulator (PWM) output. This has the advantage of being entirely digital in implementation on the IC and requires only a single output pin. They are, however, relatively slow, with speed being directly related to the resolution of the converter. For example since it takes one clock cycle per resolution step, a 4 MHz clock will give 15.625 kS/s for an 8-bit resolution and 62.5 kS/s for a 6-bit converter.

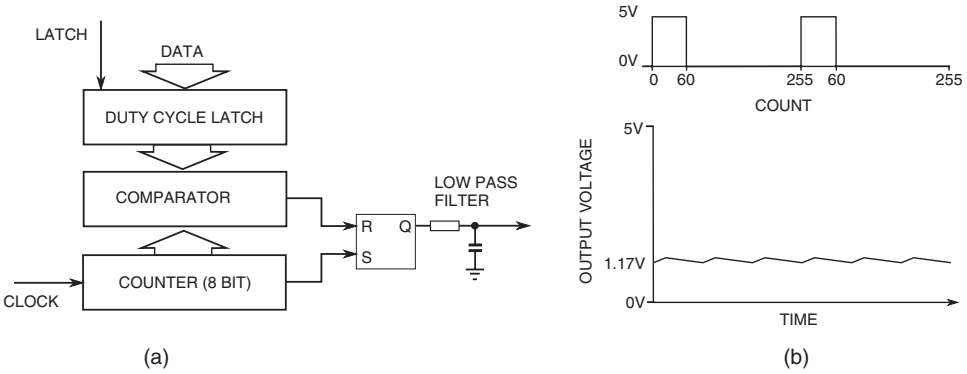


Figure 13.4

Simplified PWM DAC, typical of microprocessor **(a)** and output waveforms **(b)**.

Typically the PWM module, illustrated in Figure 13.4, consists of a free-running up-counter clocked by the microprocessor master clock – this sets the output high every time the counter reaches zero, and a magnitude comparator compares the counter value with the duty cycle value; when the counter matches this the output is cleared, so producing a pulse which is high for the number of clock cycles defined by the number in the duty cycle register.

A low-pass filter following the output gives an output that has an average DC value in ratio with the duty cycle.

Reconstruction filter

The output of a DAC contains switching and clock artefacts; that is the output can be seen to include instantaneous changes at the point of switching between codes as well as changes unrelated to the input data that can be directly attributed to the edges of the clock waveform and which distort the intended output waveform. In some applications these are not a problem but in most cases it is necessary to follow the DAC with a low-pass filter. The term **reconstruction filter** is often used for this low-pass filter, although strictly it applies to the filter used with a PWM modulator.

In most cases the low-pass filter should be designed to attenuate the clock and switching artefacts to a level below the LSB of the converter; often a simple RC filter is sufficient. One technique that can be used to improve the output signal is to oversample the data; that is, clock the DAC at a much higher frequency than the data was sampled at, e.g. 4 or 8 times, this allowing the simple RC filter to be more effective in removing the clock energy from the output signal. Where such techniques are not practical an active low-pass filter may be used.

Analogue-to-digital converters

In the previous section we saw that the digital potentiometer is the simplest form of DAC, the simplest form of analogue to digital converter, and also the fastest is the dual of the potentiometer. It consists of a chain of resistors just like the DAC, these provide scaled reference voltages to the reference inputs of an array of comparators, the other input of each comparator is connected to the input signal, thus each comparator compares the input signal with a different fixed reference voltage. It is this parallel operation that makes it capable of very high sampling rates.

A simple flash converter is shown in Figure 13.5, in this case for a simple thermometer. In this circuit the output voltage from the sensor is fed to the non-inverting inputs of all three comparators, and the inverting inputs are connected to nodes of the chain of resistors. So long as the supply voltage to the circuit is stable the LEDs will light, which can be depicted in the form of a bar graph of increasing temperature.

If the resistors in the chain are all of the same value, the steps between the LEDs lighting will be equal voltages, so the outputs of the group of three comparators provide a digital representation of the analogue input voltage. The digital output of this type of converter is often referred to as *thermometer code* because of the similarities between the output states and a mercury thermometer. While this type of output is very useful for driving bar graph displays, e.g. the LM3914 circuit shown in Figure 13.6, it is not so useful for providing a digital input to a microprocessor system since it has one signal or wire per step. If we feed the outputs of the comparators

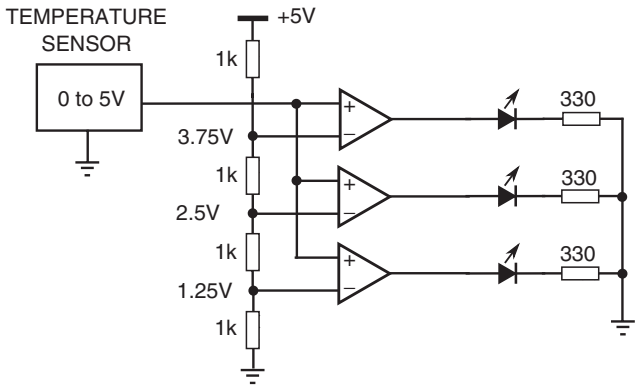


Figure 13.5

A simple ADC, with four output states, all LEDs off, and 1, 2 or 3 LEDs on.

to a logic circuit known as a **priority encoder** we can get a conventional binary output representing the analogue input.

Resolution and quantization

The three-comparator circuit shown (Figure 13.5) has four possible output states: 0 no comparator outputs, 1 the first comparator, 2 the second and first comparators and, finally, 3 all three comparators. We need to represent these four states with a binary number. To calculate how many binary bits the number needs to have, we take the base 2 log of the number of states, which is $\log_{10}(\text{states})/\log_{10}(2)$ rounded up to the next integer value, in this case 2 bits.

A 2-bit converter has limited use; if we represent a 0 to 5 V input signal with it we can only tell the difference between 0, 1.25, 2.5 and 3.75 V inputs – in fact if the output of our converter is 2 then the input could be any value from 2.5 V to 3.75 V. This is referred to as **quantization**, the inability to represent a level smaller than a given size. This error is directly related to the bit resolution of the converter: an 8-bit converter has 2^8, i.e. 256 full-scale range (FSR), and a 10-bit one has 2^{10}, 1024 steps. All analogue-to-digital converters quantize, and these errors are usually of the order of 1 least significant bit (LSB). It is possible to minimize the effects

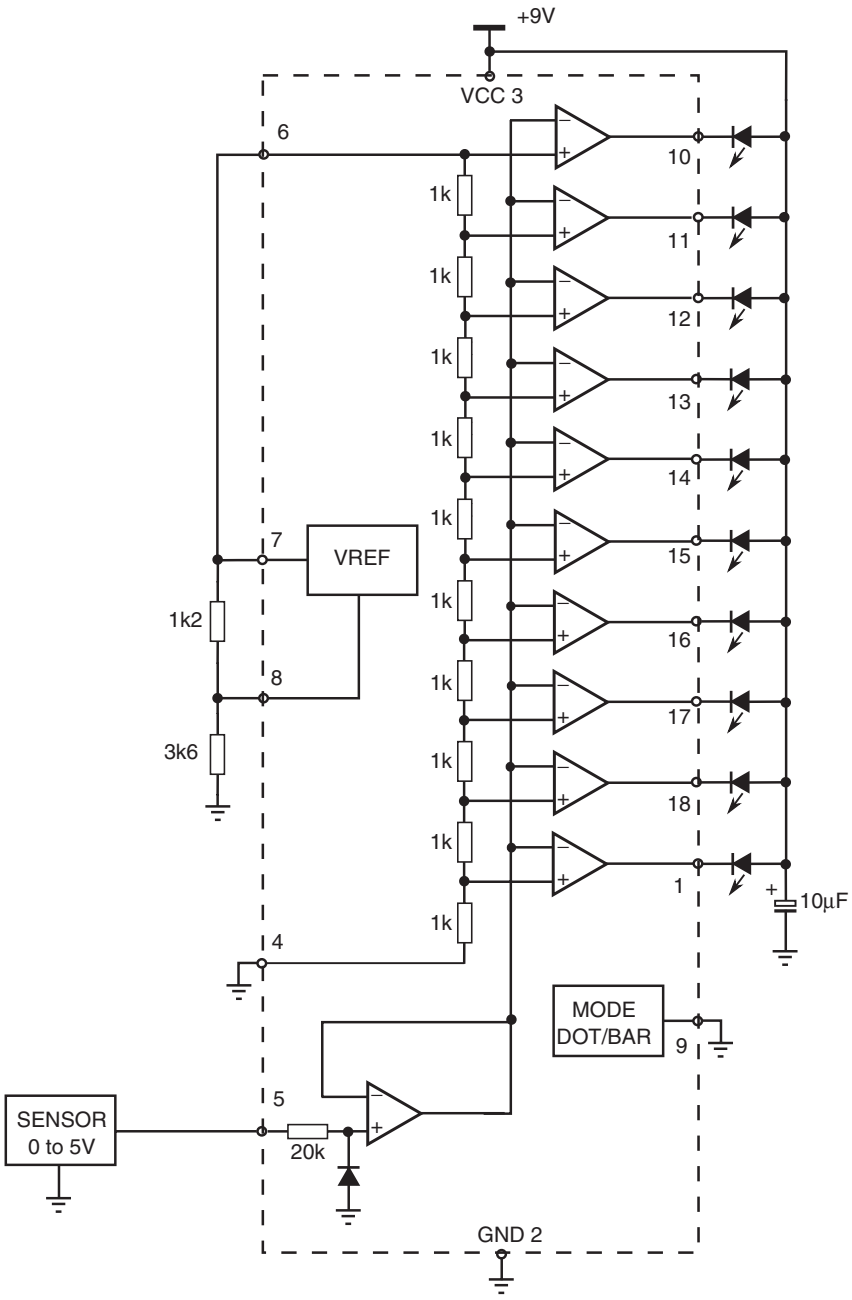


Figure 13.6

LM3914 bar graph display IC (resistors at pins 7 and 8 set the reference voltage).

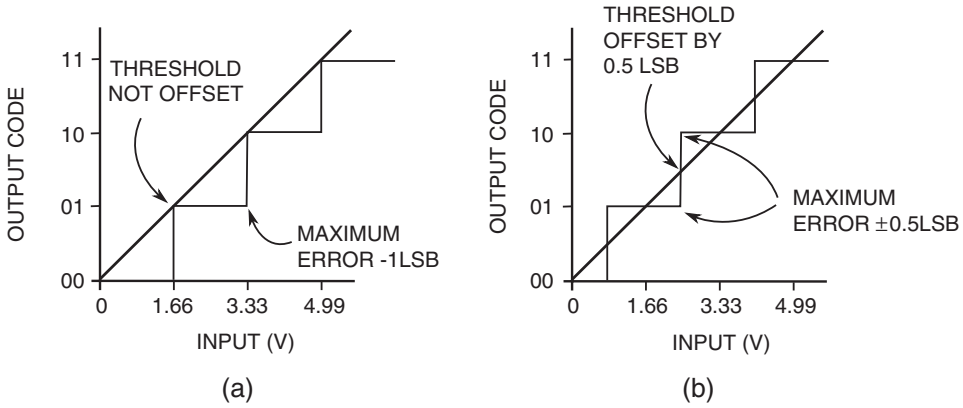


Figure 13.7

Quantization and offset errors in analogue-to-digital conversion.

of quantization errors by offsetting the decision threshold by 1/2 LSB – this makes the error symmetrical ±1/2 LSB rather than being asymmetric between –1 LSB and zero (Figure 13.7).

Quantization is an inevitable consequence of the resolution of analogue-to-digital conversion but there are other types of error which can be designed out depending on how much one is prepared to spend on the ADC. Apart from quantization and offset errors the main errors are gain and differential non-linearity (DNL). Differential non-linearity is a measure of how different successive quantization steps are; they should obviously be the same, however there are a range of possible effects including a worst case scenario of altogether missing codes, that is when there are no analogue

Table 13.2 ADC resolution and dynamic range

Number of bits	Resolution LSB/FSR	LSB/FSR	Voltage dynamic range
4	1:15	6.7%	24 dB
6	1:63	1.6%	36 dB
8	1:255	0.4%	48 dB
10	1:1023	0.1%	60 dB
12	1:4095	0.025%	72 dB

input values to a converter that can produce a particular digital output code. Figure 13.8 shows how DNL and gain errors affect the relationship between digital output code and analogue input level.

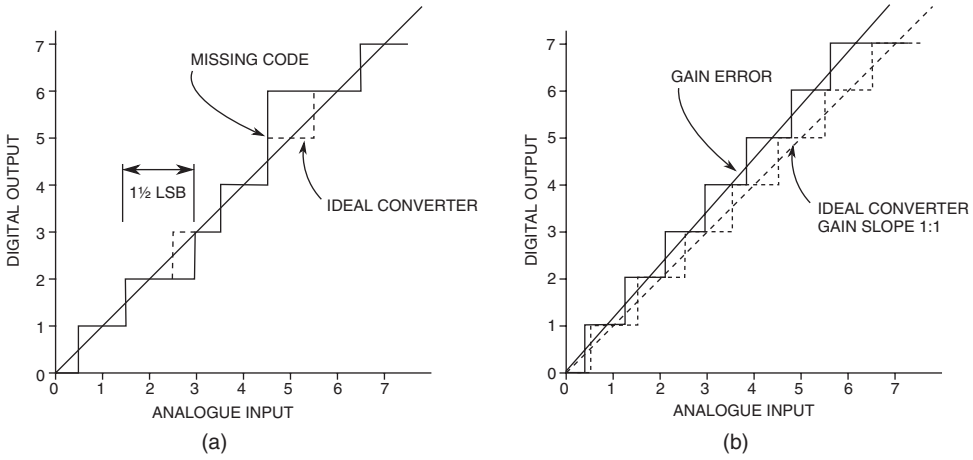


Figure 13.8

ADC errors: **(a)** differential non-linearity, **(b)** gain.

In analogue-to-digital converters of the type described so far these errors might typically be produced by matching errors between the resistors in the reference divider chain and input loading issues, etc.

Analogue-to-digital converters based on a chain of comparators as described are called **flash converters** because the conversion happens very rapidly, depending only on the speed of the comparators and the combinational logic that does the conversion from thermometer code to binary output. As we expand the number of inputs to improve the resolution of our digital representation of the analogue input signal, we start to need very large numbers of comparators if we use this type of converter; for most applications needing more than 8-bits resolution this is unnecessarily expensive in terms of silicon area and complexity. Flash converters and those based on a pipelined version of the flash architecture are often used in video and instrumentation applications where sample rates of 10 MS/s to in excess of 200 MS/s are required.

Sampling

When an analogue signal is converted to a digital representation it is very important to know how often the sample is converted; for instance if we were to sample the ambient temperature once per day at breakfast-time we might assume that the temperature was fairly constant over a period of a few weeks (Figure 13.9).

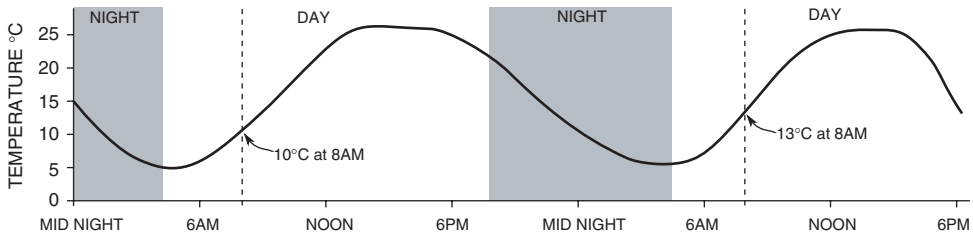


Figure 13.9

Day–night temperature variation.

Since we only have the digital representation of the analogue input signal at discrete times, i.e. at the instant of sampling, we would however have missed the 20°C range between minimum night-time temperature and maximum daytime temperature. The solution to this problem is to increase the number of samples that we take. This is simple when we are dealing with slowly changing quantities like daytime temperature; we can sample say every minute (1440 samples in 24 hours) without any problem, however, there are many other applications where, because of the input frequency spectrum, we also need to account for frequency aliasing in our sampled data.

Aliasing

If a sine wave, f_{in}, is 1.25 times the sampling frequency, f_s, the data would not be distinguishable from samples of a sine wave that has a frequency of 0.25 f_s. This is why the effect is called **aliasing** – one frequency appearing

as another (Figure 13.10). Aliasing is not always a bad thing; in some types of digital radio, aliasing is used deliberately in the system to under-sample the IF frequency, allowing the use of slower ADCs – this only works because there is no low-frequency signal at the input for the under-sampled high frequency to be confused with.

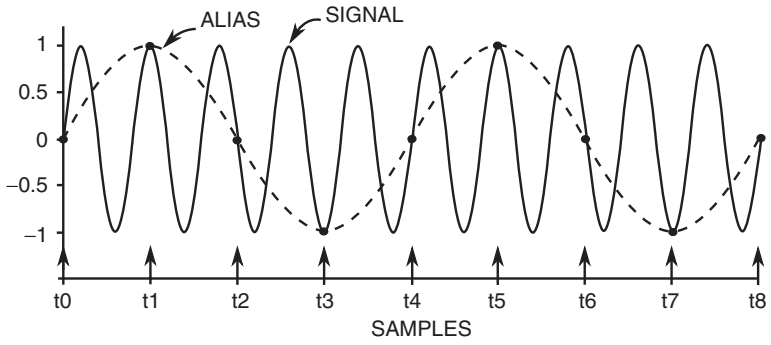


Figure 13.10

Aliasing, two sine waves which are indistinguishable when sampled.

The Nyquist sampling theorem tells us that we must sample at greater than twice the highest frequency present in the input signal in order to avoid aliasing. However, if we sample a pure 1 kHz sine wave at 2000 samples per second we will have only two data samples per cycle so we could not reproduce the sine wave; in fact to accurately represent a sine wave it is necessary to sample significantly quicker than the Nyquist rate.

To avoid aliasing, the analogue input signal must be filtered to remove the frequency components above the Nyquist limit. Anti-aliasing filters are generally required to have a steep transition from pass band to stop band because of the need to keep sampling rates as low as possible both for cost and data storage reasons.

If we use an 8-bit analogue-to-digital converter, this gives a dynamic range of 48 dB ($20\log_{10}(255)$) – the ratio of the amplitude of the biggest signal that can be converted to the smallest signal. In order to suppress

signals that might result in aliases within the sampled bandwidth, a filter needs to reduce the input amplitude for all frequencies above $1/2f_s$ to less than 1 LSB, or 48 dB below full scale. A sixth-order Butterworth filter can achieve 24 dB/octave transition between pass bands and stop bands, and this would allow an 8-bit ADC sampling at 750 kS/s to convert signals up to about 125 kHz faithfully, since signals of 375 kHz and above would be attenuated by more than 48 dB. In fact in most cases the situation is not as bad as this, because it is rare to have signals of significant amplitude above the Nyquist frequency, the exception being test instruments.

So far we have considered simple Flash ADCs as illustrative examples of the general characteristics of analogue-to-digital converters; there are, however, many different topologies of ADC to chose from, each tailored for specific types of application.

Successive approximation analogue-to-digital converter

In measurement and control applications successive approximation ADCs are very popular because they take relatively little space on silicon and have the inherent characteristic that they produce serial data output.

The successive approximation ADC works by making tests of the size of the input signal and refining the approximation of output at each test (Figure 13.11). This method uses a single comparator and a digital-to-analogue converter (DAC) to test the size of the input. The successive approximation register implements a binary search algorithm in hardware.

On the *start conversion* signal the DAC output is set to half full-scale range (FSR), i.e. only the MSB is set. If the input is greater than this comparator returns a 1, which is the MSB of the conversion. On the next clock the SAR sets the next highest bit of the DAC; if the input is less than 3/4 FSR, the comparator output is zero – this is the next output bit. The process continues until all the bits are converted; the conversion takes $n + 1$ clock cycles, where n is the number of bits resolution of the ADC.

When the ADC output is clocked out as serial data while the conversion is being done the data word is usually padded with leading zero bits,

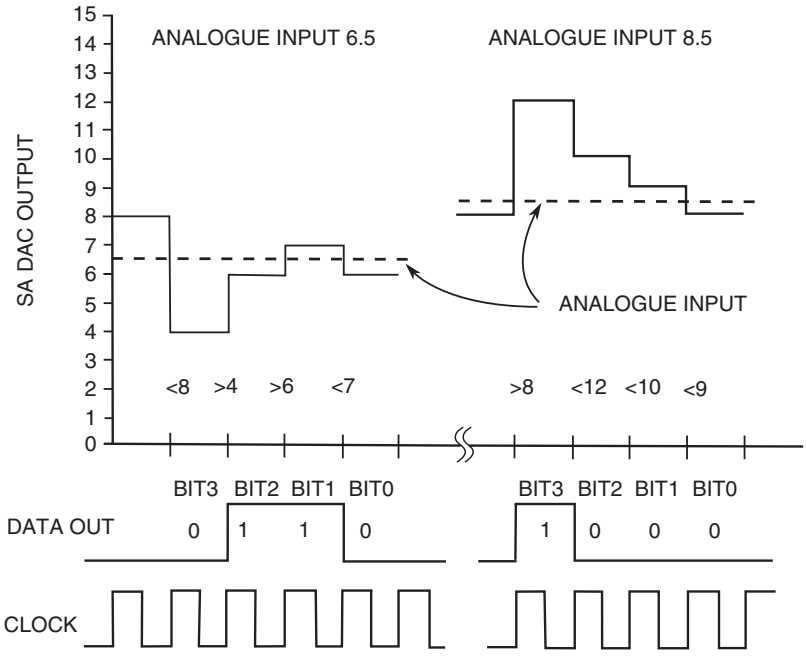


Figure 13.11

Successive approximation ADC binary search with serial output.

and hence an 8-bit conversion might require 10 or 11 clock cycles with the first two or three data bits being zero.

It is very important for the analogue input to remain stable during the whole conversion process otherwise the SAR sequence can fail, resulting in large errors in the data. To ensure that the input is stable most SA ADCs include a track-and-hold amplifier in front of the comparator stage. The track-and-hold circuit consists of a switch, a capacitor, and a buffer amplifier (Figure 13.12). Between conversions the switch is closed and the voltage on the capacitor follows the input voltage; when the switch is opened the voltage stored on the capacitor remains there until leakage currents drain it away – this usually takes quite a long time, say tens of milliseconds. The conversion time of the whole ADC is now slightly longer, as it needs to account for the settling time of the track-and-hold amplifier and opening the switch; as a result conversion times of $n + 2$ clocks plus sample and hold settling times are quoted.

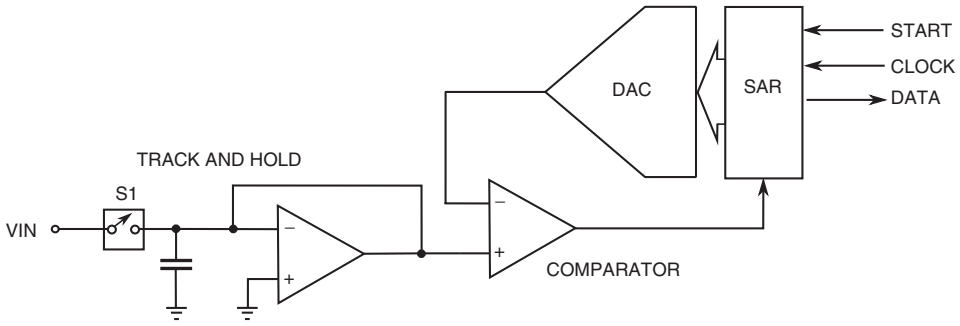


Figure 13.12

Successive approximation ADC with sample-and-hold amplifier.

An SA-ADC with built-in sample-and-hold circuits can achieve sample rates of about 2 Msps but are generally slower. The 10-bit ADC in the Microchip PIC16F microcontroller family is quoted as requiring 12 clocks of 1.6 μs, giving a conversion rate of 52 ksps.

Sigma–delta ADC (over sampling or bitstream converter)

Sigma–delta converters (Figure 13.13) are typically used in VLSI DSP applications, such as digital audio and high-resolution instrumentation. In fact without VLSI these converters would be impractical since they rely on decimating low-pass filters. Over-sampling is a significant advantage in wide dynamic range converters and high-resolution applications because it simplifies the design of anti-aliasing prefiltering since the Nyquist frequency of the converter is greatly in excess of the signal frequency. This also means that the tolerance issues usually associated with analogue filter and signal processing circuitry in volume production are avoided.

The bitstream generated by the closed loop consisting of the integrator, latched comparator and 1-bit DAC, has over a long enough interval, the same average value as the input signal. This bitstream data over a short interval is nearly random. The digital low-pass filter and decimator produce and output at the desired bit rate; e.g. a 256:1 decimator with

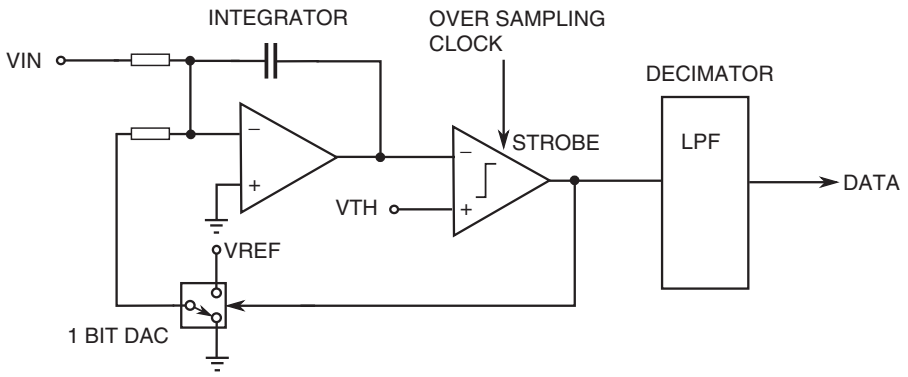


Figure 13.13

First-order sigma–delta ADC.

a 8 kS/s output rate is preceded by a sigma–delta modulator clocked at 2.048 MHz. This has several advantages including improving the signal-to-noise ratio of the converter (24 dB for a 256:1 decimator). Another advantage of 1-bit sigma–delta ADCs is that they have inherently linear transfer characteristics.

Dual-slope ADC

Dual-slope ADCs (Figure 13.14) are used in instrumentation applications such as digital voltmeters (DVMs). The advantage of the dual-slope architecture is that it is not dependent on the linearity of the slope and hence the integrating amplifier and capacitor.

Initially the integrator is zeroed, and then the input signal is connected to the integrator for a fixed time, controlled by the counter. At the end of this period the counter is zeroed, the reference is connected to discharge the integrator capacitor and the counter is started. The counter is stopped when the integrator output reaches zero. The discharge slope is constant and hence the counter output has a value proportional to the input voltage. The ratio of the counter output value to the fixed count represents the ratio of the input voltage to the reference voltage.

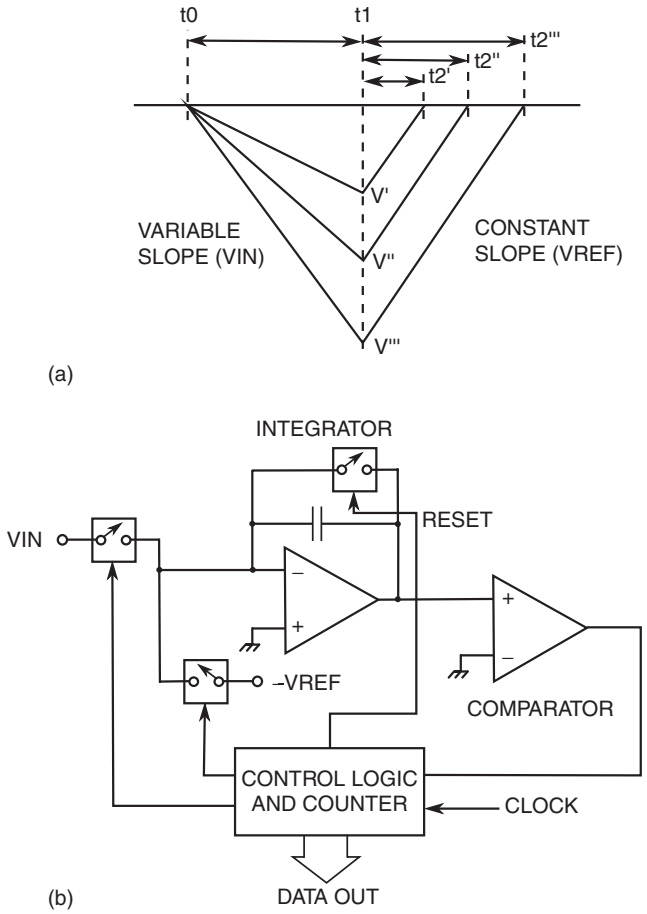


Figure 13.14

Dual-slope ADC: **(a)** slope timing and **(b)** schematic.

Voltage references for analogue-to-digital converters

Analogue to digital converters often require an external voltage reference; the critical thing here is to keep the reference stable and noise free. If an ADC with 10-bit resolution and 1 LSB accuracy is being used, it will need a reference that is stable over the temperature range of interest.

An accuracy of 1 LSB in ADCs is equal to an error of 0.1% in output of the reference. Typical three-terminal voltage regulators like the 78L05 have temperature coefficients of 0.1 mV/°C, meaning that an increase in temperature of about 48°C is required before the error introduced is the same size as the LSB of the converter. A bigger problem is the absolute accuracy of the voltage reference, since typical 2.5% regulators would result in 25 LSB of error between two otherwise identical units. In applications where 10-bit or higher ADC resolution is required, high-precision shunt references with laser trimmed nominal accuracy of 0.1% and 50 ppm/°C temperature coefficient can be used.

Noise on the reference supply is also a problem, as 5 mV of noise on the reference or the analogue ground of the ADC can reduce the overall accuracy of the converter by 1 LSB. Noise is a particular problem in circuits where there is a lot of fast digital logic or in circuits that control power devices like stepper motor drives, etc.

- **PCB LAYOUT**

PCB layout is often critical for successful design of ADC circuits; resolutions of more than 12 bits are less tolerant of noise simply because 1 LSB is of the order of 1 mV or less. Grounding is the key to success, as with all small-signal circuits; however, the proximity of digital switching waveforms, inherent in the design of DACs and ADCs, and the associated microprocessor circuitry makes for a noisy environment. Careful decoupling of supplies, segregation of analogue and digital grounds and avoiding ground loops are essential.

Connecting a serial ADC to a PC

The typical PC has many interfaces, among them the game port and sound card input, which are analogue-to-digital converters, but it is often inconvenient to use these ports for general purpose applications. The circuit shown in Figure 13.15 shows a simple application of the LTC1096 8-bit micro power serial ADC. Connected to the parallel printer port of a PC, the circuit allows the PC to make measurements on unipolar analogue voltages from DC up to a few kHz in frequency. The RC input filter is not

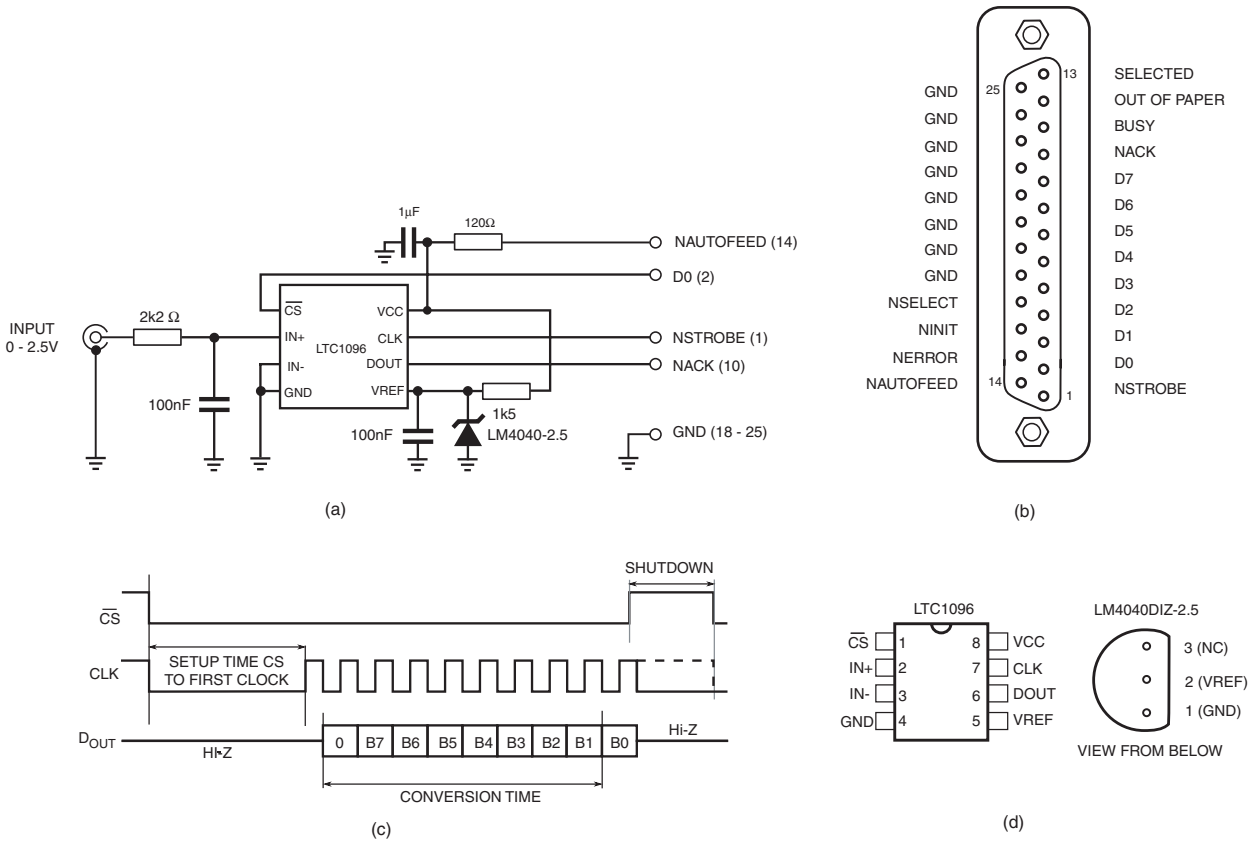


Figure 13.15

Using a serial micro power ADC: **(a)** LTC1096 schematic, **(b)** PC parallel port connector, **(c)** chip select, clock and data timing and **(d)** pin connections of LTC1096 and LM4040.

designed as an anti-aliasing filter but intended to keep the input of the successive approximation ADC stable during conversions.

The converter is powered from one of the parallel port outputs (AUTOFEED, pin 14). The LTC requires supply current of about 120 μA and LM4040 voltage reference is about 160 μA; the total current drain is well within the drive capability of the port. The circuit could be built on a small piece of Vero board and mounted inside a 25-pin D connector shell.

The chip-select line is driven from the D_0 (pin 2) output of the port and the STROBE (pin 1) is used to clock the ADC. The data from the ADC is read by the PC on the ACK input (pin 10).

The parallel port can be conveniently used in this way to interface many 3-wire serial bus interfaced integrated circuits, like ADCs, digital potentiometers and EPROMs; this is often useful for development prototyping purposes.

The parallel port or line printer port (LPT) as it is called under DOS and Windows, may be mapped at one of several addresses in the I/O address range; typically the default on recent machines is &H378, but &H278 and &H3BC are also possibilities. The parallel port interface uses three 8-bit registers, the output data register is located at the base address (&H378), the status input is at base +1 (&H379) and the control register is at base +2 (&H37A). The power supply and clock signal are controlled by bit 1 and bit 0 respectively in the control register, the chip-select signal is provided by bit 1 of the data register and data is read back using bit 6 of status register.

The exemplar program below uses Qbasic and runs under DOS and Windows 9X. The program enables the ADC, reads data back from the ADC and prints the raw ADC value on the screen before shutting down the ADC and exiting.

Listing 13.1 ADC interface program.

```
'adc control program for LTC1096
'port address registers definitions
portbase% = &H378
datareg% = portbase%
statusreg% = portbase% + 1
control% = portbase% + 2
```

```
‘bit pin map
‘power supply is provided by AUTOFEED pin, this is inverted
‘in hardware so clear bit 1 of the control register to turn the
‘ power on.
poweroff% = 255
poweron% = 253

‘clock is driven with the STROBE pin, bit 0 of the control register
‘this bit is inverted in hardware so set bit to drive clock line low
clocklow% = 255
clockhigh% = 254

‘chip select is driven from bit 0 of the data register, this allows
‘up to 8 devices to be used with out decoding hardware
chipselect% = 254
chipshutdown% = 255

‘data is read back from the ACK pin, bit 6 of the status register
datamask% = 64

‘define delay values to use for setup time, clock high and clock low
‘delay values allow scaling for different speed machines
‘and different sampling rates
delay0% = 1000
delay1% = 260
delay2% = 120
delay3% = 280

‘start
‘power up
OUT control%, poweron%
‘wait until LTC1096 and Voltage regulator has started up
FOR d% = 0 TO delay0% STEP 1: NEXT d%

‘clock low
OUT control%, poweron% AND clocklow%
OUT datareg%, chipselect%

‘wait for chip select to clock time
FOR d% = 0 TO delay1% STEP 1: NEXT d%

‘first clock pulse
OUT control%, poweron% AND clockhigh%
FOR d% = 0 TO delay1% STEP 1: NEXT d%
OUT control%, poweron% AND clocklow%
```

```
'clock data in from LTC1096
'first bit is always zero
'data will be returned in value%
value% = 0
FOR bit% = 8 TO 0 STEP-1
OUT control%, poweron% AND clocklow%
databit% = INP(statusreg% ) AND datamask%
'add weighted bit value to total
value% = value% + 2^ bit% * databit%
FOR d% = 0 TO delay2% STEP 1: NEXT d%
OUT control%, poweron% AND clockhigh%
FOR d% = 0 TO delay3% STEP 1: NEXT d%
NEXT bit%

'shutdown converter

OUT datareg%, chipshutdown%
OUT control%, poweroff% AND clockhigh%

PRINT value%
```

The LM4040DIZ-2.5 reference is a 1% part so an absolute error of 25 mV in readings may be observed. The LTC1096 is specified as having ±0.5 LSB offset and linearity errors and ±1 LSB full-scale error. The code-to-code error should be less than 5 mV. The maximum conversion rate depends on the speed of the PC used but clock frequencies of 33 kHz to 100 kHz giving conversion rates of about 3 kS/s to 9 kS/s should be achievable.

Useful websites

Data sheets and application notes for most types of data converters can be downloaded from the following companies' websites.

Analogue-to-digital converters and Digital-to-analogue converters

Analog Devices
www.analog.com
National Semiconductor
www.nsc.com

Maxim Integrated Circuits (Dallas)
www.maxim-ic.com
Linear Technology
www.linear.com

Digital potentiometers

Intersil (Harris) (Xicor)
www.intersil.com
Maxim Integrated Circuits (Dallas)
www.maxim-ic.com
Microchip Technology
www.microchip.com

Voltage references

Analog Devices
www.analog.com
Texas Instruments (Burr Brown)
www.ti.com
Zetex Semiconductor
www.zetex.com

CHAPTER 14

TRANSFERRING DIGITAL DATA

Introduction

It is very often necessary to transfer information from one piece of equipment to another, or one place to another – for example printing data from a computer or monitoring remote equipment. There are many applications that use high-speed links like Ethernet, WiFi, GPRS, USB or Firewire (IEE1394) to transfer data. These systems take considerable effort and resources to implement from scratch and are best addressed by using standard off-the-shelf systems. This chapter is aimed at providing information on methods that are applicable to microcontroller-based designs where specific needs or cost rules out the use of the previously mentioned systems.

Some applications of digital circuitry make use of the digital data at the time when they are obtained. A digital voltmeter, for example, displays the obtained data as soon as the data are collected, and there is not necessarily a requirement to transfer digital data from one piece of equipment to another. In many other applications, however, and particularly in control and computing, data have to be transferred over distances that range from a metre or less up to the maximum distance that a radio signal can reach. In this chapter, we shall look at data transfer methods.

The simplest method of transferring digital data is to connect to a microprocessor bus, usually by way of buffer or transceiver chip. The transfer of data from one board to another in a digital system makes use of the microprocessor bus either directly or by way of buffer circuits, the *bus drivers*. In such connections, all of the microprocessor signals are transferred, including data lines, address lines and all of the synchronizing and timing lines. These types of applications are covered in the chapter of this

book on microcontrollers (Chapter 15), including various serial interface buses like SPI and I2C.

This chapter focuses on the transfer of digital data between different pieces of equipment, of which probably the most familiar example in computing is the use of a printer. For industrial and instrumentation purposes, the requirements are much more varied but the basic methods are much the same. The choice is of either parallel or serial transmission and reception, and in many cases only serial transmission is possible. No matter which method is used, some form of synchronization will be needed because the rate at which data can be received by a device such as a printer is never as fast as the rate at which it can be transmitted from a microprocessor. Both serial and parallel data transfer systems must therefore ensure synchronization of transmission and reception, and the problem is more acute for serial links. The signals that are used for this purpose are called *handshaking* signals. Parallel transmission means that all the data lines of the microprocessor bus, or a set of data lines, will be used, along with a few synchronizing lines. For instrumentation purposes, a more complete set of signals will be needed than is the case for a computer printer. Many of the microprocessors used in industrial equipment are of the 8-bit variety, and for parallel transmission of data all eight data lines will be used.

Parallel transfer

The main problems of parallel data transfer are of line length and pulse frequency. These two problems are interconnected because they both arise from the stray capacitances between the leads of the cable. A parallel cable will be driven from a low-impedance source and will connect into a comparatively high impedance at the receiver end. The stray capacitance between signal wires, together with the very fast rise and fall times of the pulses, can therefore induce a 1 signal in a line which should be at level 0. The longer the line, the greater the induced signal, until the voltage becomes great enough to drive the receiver circuit, at which point a false signal will be received. The practical effect is to restrict parallel printer leads from computers to around 1–2 metres. Greater lengths can be obtained by using correctly matched 500 Ω lines, but these are rare in computing applications though fairly common for instrumentation applications.

IEEE 1284 Centronics printer interface

Table 14.1 shows the signals that are used in a Centronics printer output, in this case from a computer that is IBM compatible. Not all printers make use of all of these signals, nor do all computers, but the Centronics standard is sufficiently flexible to ensure that any computer that provides a parallel printer output to Centronics standard can be matched to any printer with a Centronics input. There are various minor deviations between printers and computers, all of which can be dealt with by omitting one or more links in the cable.

Plugs and sockets are standardized only at the printer end of the cable, which uses a 36-pin plug of the Amphenol 57-30360 type. IBM-compatible computers and most other equipment like oscilloscopes with printer outputs use a female 25-pin D-connector for the parallel output.

Table 14.1 Pin assignments of the D-type 25 pin parallel port connector

25-way D type plug (computer)	36-way Centronics (printer)	Signal name	Direction In/out	Register	Hardware inverted
1	1	NSTROBE	IN/OUT	CONTROL	YES
2	2	DATA 0	OUT	DATA	
3	3	DATA 1	OUT	DATA	
4	4	DATA 2	OUT	DATA	
5	5	DATA 3	OUT	DATA	
6	6	DATA 4	OUT	DATA	
7	7	DATA 5	OUT	DATA	
8	8	DATA 6	OUT	DATA	
9	9	DATA 7	OUT	DATA	
10	10	NACK	IN	STATUS	
11	11	BUSY	IN	STATUS	YES
12	12	OUT OF PAPER	IN	STATUS	
13	13	SELECTED	IN	STATUS	
14	14	NAUTOFEED	IN/OUT	CONTROL	YES
15	32	NERROR	IN	STATUS	
16	31	NINIT	IN/OUT	CONTROL	
17	36	NSELECT	IN/OUT	CONTROL	YES
18–25	19–30	GND	GND		

Only about 22 pins of the 36-pin connector are likely to be used, and on the 36-pin connector, pins 2 to 9 inclusive are used for the eight data lines of the data bus, D0 to D7. The Centronics standard provides for signal return lines, each at signal ground level, so that twisted pairs of signal/ground return wires can be used. Return pins 20 to 27 inclusive are used in this way for data, but this provision is not always used, particularly for short cables. Pin 1 (return pin 19) handles a strobe signal which is driven low by the computer in order to send data to the printer. The width of the strobe pulse must be at least 0.5 μs, and this pulse is one of three that forms the main handshaking provisions in this type of interface. The other two are BUSY (pin 11, return on 29) and ACKNOWLEDGE (pin 10, return on 28).

The BUSY signal is a steady-level signal that is set high by the printer to indicate that no more data can be received. The BUSY line is taken high when data starts to be entered, during printing and when the printer is off-line, or disabled because of a fault. Most printers contain buffer memory which can range from one line to several pages of printed characters, and the BUSY signal is taken high when this buffer is full. For such printers, transmission of signals is intermittent because of the time needed to fill the buffer at the normal rate of parallel transmission, which is as fast as the signals can be clocked (subject to the minimum pulse widths that can be used).

The use of a printer buffer allows the computer to be used during a printout, and some computers provide buffering in their own memory to assist this detachment of printing from other operations. The snag is that if you want to stop the printer you cannot do so immediately by stopping data being sent out from the computer because the printer will stop only when the buffer is empty. You can, of course, switch off or reset the printer, but this will empty the buffer and you will need to retransmit this data when the printer is ready for use again. The better option is to use the off-line switch. The ACKNOWLEDGE pulse signal is sent out by the printer to indicate that the printer has received data and is now ready to receive more data – usually when the buffer is empty. The relationship of these signals to each other is fairly flexible, as the timing diagrams in Figure 14.2 indicate.

The timing diagrams shown in Figure 14.2 are typical of the Centronics standard. The important point is the maintenance of the 0.5 μs minimum

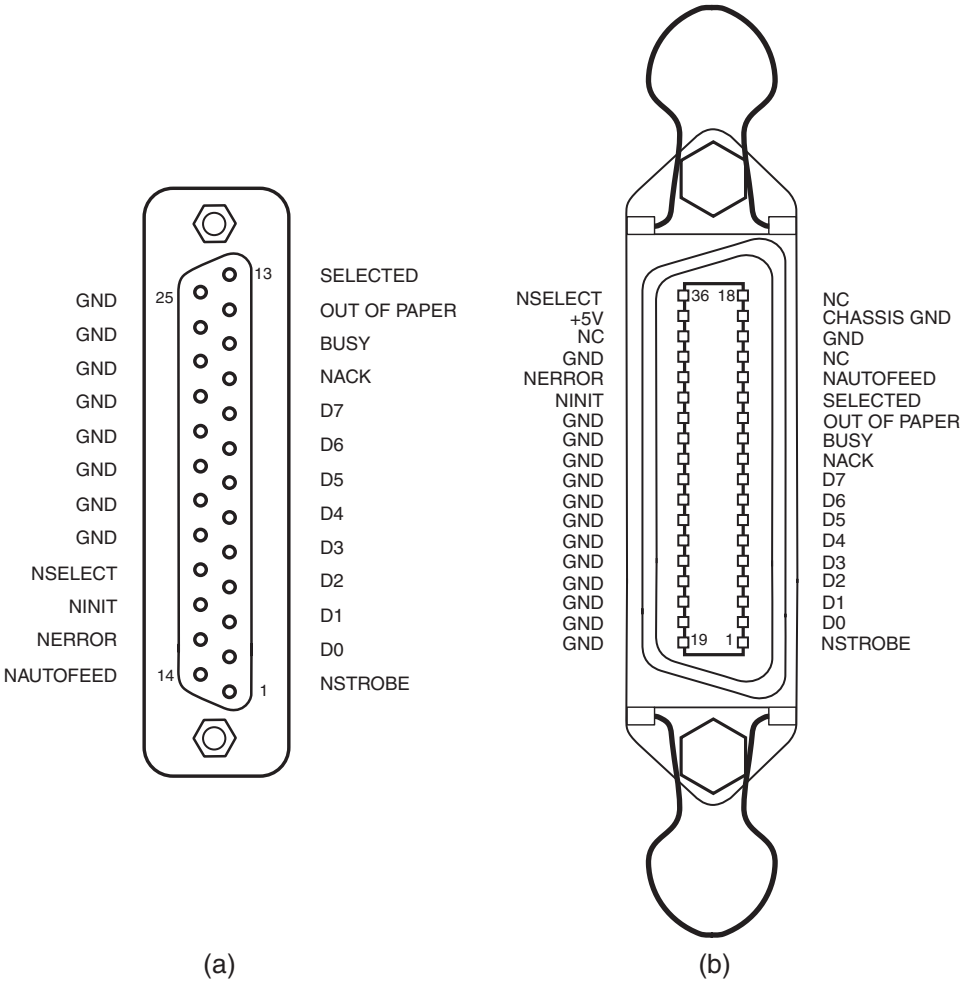


Figure 14.1

Parallel printer connector: **(a)** typical computer **(b)** at the printer.

pulse width and timing intervals, so the clock rate of output is usually considerably lower than the clock rate for the computer itself. The data, strobe, BUSY and ACKNOWLEDGE signals are the most important parts of the Centronics interfacing; the remaining signals and their uses are summarized in Table 14.1.

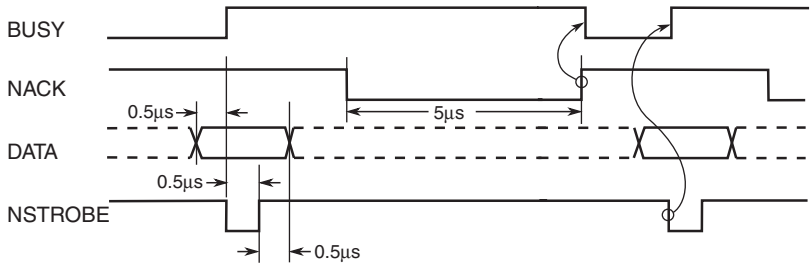


Figure 14.2

Typical timing specifications.

There is some minor variation between printers in the use of these signals. Not all printers, for example, make use of the Autofeed signal, but such variations are not generally important unless you are trying to use one printer with a cable that was intended for another. It is very unusual to find major problems of compatibility between computers and printers using the Centronics interface. One example in the past arose from the computer manufacturer earthing pin 14, with the result that printers which used the Autofeed signal were forced to take an additional line spacing. The register addresses and bit functions for printer interface of IBM PC compatible machines are shown in Table 14.2 and 14.3.

The SLCT IN output, also active low, inhibits all output to the printer unless this line is held low. Problems with 'dead' printers are often traced to this latter line being disconnected. For some purposes, notably low-cost network connections, a parallel connector can be used for two-way data communication. In this mode, which needs a controlling program, four of the data lines are used for the outward signals and four of the control pins are used for the input signals. This makes the use of a parallel port for two-way communication considerably faster than the use of a serial port, but slower than the use of a genuine 8-bit two-way port such as would be used for fast networks.

The IEEE-488 bus

The IEEE-488 bus is a parallel data transfer system for connecting complete systems rather than parts of systems and it is widely used in digital

Table 14.2 PC printer port registers

Register bit	Signal port	Direction	Register bit	Signal port	Direction inversion	Register bit	Signal port	Direction inversion
DATA 0	DATA 0	OUT	STATUS 0			CONTROL 0	NSTROBE	INVOUT
DATA 1	DATA 1	OUT	STATUS 1			CONTROL 1	NAUTOFEED	INVOUT
DATA 2	DATA 2	OUT	STATUS 2			CONTROL 2	NINIT	OUT
DATA 3	DATA 3	OUT	STATUS 3	NERROR	IN	CONTROL 3	NSELECT	INVOUT
DATA 4	DATA 4	OUT	STATUS 4	SELECT IN	IN	CONTROL 4		
DATA 5	DATA 5	OUT	STATUS 5	OUT OF PAPER	IN	CONTROL 5		
DATA 6	DATA 6	OUT	STATUS 6	NACK	IN	CONTROL 6		
DATA 7	DATA 7	OUT	STATUS 7	BUSY	INVIN	CONTROL 7		

Table 14.3 PC printer ports

PC port	Data	Status	Control
LPT1	0378H	0379H	037AH
LPT2	0278H	0279H	027AH
LPT3	03BCH	03BDH	03BEH

electronic instrumentation. Often called the Hewlett Packard Interface Bus (HPIB) or the General Purpose Interface Bus (GPIB), the standard dates back to 1974 and much of the detail of the standard derives from the Hewlett-Packard original.

The bus is used to connect devices that can carry out actions described as *controlling*, *listening* and *talking*. A device might carry out just one of the functions, any two of these functions, or all three. A controller device will control other devices and is almost always a microcomputer- or microprocessor-based controller system. A talker device will place data on to the bus, but does not receive data, and a listener will receive data from the bus but does not place any data on the bus. A counter might, for example, be connected as a talker, placing the data from its count onto the bus but not receiving any data (though it would obey command signals) from the bus. By contrast, a signal generator might be used as a listener, generating signals as commanded by data read from the bus, though not placing any digital signals onto the bus. Many devices will be used as talkers and listeners, receiving signals from the bus (for changing range or function) and placing signals onto the bus to indicate readings. Since the IEEE-488 is primarily intended for instrumentation, the prime example of a talker/listener is a digital multimeter.

The bus, like the Centronics parallel system, consists mainly of data lines with no address information and uses a total of 16 lines on a 24-pin connector. Of these, eight are data lines that are bidirectional, five are bus control lines, and three are handshaking lines. Figure 14.3a shows the standard pin layout, with data on lines 1–4 and 13–16 inclusive. The handshaking lines are on pins 6, 7 and 8, with the 'transfer-control' lines on 5, 9–11 and 17. The handshaking lines use open-collector outputs and are active low, so these lines can be connected to common output lines forming wired-OR connections. The handshaking lines are DAV (data valid) on pin 6, NRFD

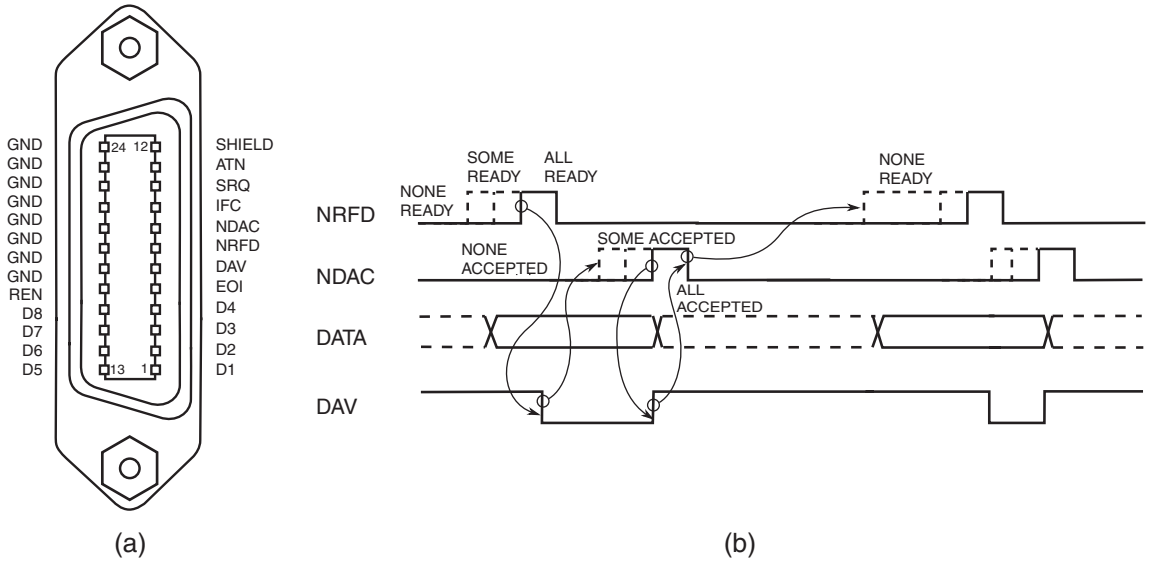


Figure 14.3

The IEEE-488 connector: **(a)** pin out and **(b)** the timing of a talker–listener IEEE-488 exchange. When more than one listener exists, DAV is not asserted until all listeners are ready.

(not ready for data) on pin 7 and NDAC (not data accepted) on pin 8. The DAV signal is sent out by a talker device to signal that data have been placed on the data lines and are valid for use. The other two lines are controlled by listeners, with NRFD signifying not ready for accepting the data, and NDAC signifying that data have not been read. When both NRFD and NDAC lines go high, the data are read. The action for a single talker and listener is as shown in Figure 14.3b.

The DAV line from the talker remains high even in the presence of data until the NRFD signal goes high. This is not such a simple action when several listeners are present because the NRFD line is ANDed; it cannot go high until all listening devices are ready. When the NRFD line from the listener(s) goes high, the talker activates the DAV line (low state) so that data can be transferred. The data transfer is complete when the NDAC line rises to the high level and the rate of transfer is controlled by the slowest listener. Typical maximum rates range between 50 and 250 kilobytes per second if the listeners are fast-acting devices.

The bus control lines are used to determine how devices interact with the controller. The simplest of these is the *interface clear* on pin 9, pulled low to reset the system in preparation for use. By contrast, the end or identify (EOI) signal on pin 5 is used to indicate that data transfer is complete. The attention (ATN) line, pin 11, decides the use of the eight data lines. When this pin voltage is low, all eight data lines are used for data, which need not necessarily use all eight data lines. When the ATN pin voltage is high, the lower lines of the data bus are used to hold an address number to which a specific device will respond. In the absence of such a specific address, all listening devices can receive signals sent over the data lines.

The other two bus control signals are service request (SRQ) and remote enable (REN). The SRQ, pin 10, is used by a device to indicate to the controller that the device needs attention. This is the equivalent of an interrupt signal to a microprocessor, and is normally used by a talker to indicate that it has data to transfer, or by a listener which needs data. The REN signal allows any device to be operated either from the IEEE-488 bus (remotely) or locally, as from its own front panel or from a test connector.

The IEEE-488 bus is used to a considerable extent in the computer control of automated electronic test systems. PCI bus IEEE-488 interface cards are

available from National Instruments and Agilent; however, it is often more convenient to use a USB to IEEE-488 converter which can be used with laptop computers as well as desktop machines. PCMCIA IEEE-488 cards are also available for laptop computers.

IEEE-488 is beginning to be replaced by other interfaces such as USB and Ethernet; it is likely that Ethernet will become the standard for connecting test equipment within a few years. major instrument companies like Agilent, Tektronix, LeCroy and Keithley already offer Ethernet interfaces for many of their instruments, and many oscilloscopes and spectrum analysers use embedded Intel x86 or ARM based control computers often running Linux or a version of Microsoft Windows like Windows CE.

Serial transfer

The serial transfer of data makes use of only one line (plus a ground return) for data, with the data being transmitted one bit at a time. Arrangements have to be made to convert the parallel data used by computers into serial form, and this is done by loading the data word in parallel into the flip-flops a shift register and clocking it out one bit at a time. At the other end of the link the reverse of this process clocks the data through a shift register until the data bits are in the correct places to be read in parallel, and in order to achieve this the same clock signal is required at each end of the link. Synchronous links transmit both the clock and data signals, whereas asynchronous links provide information to identify the beginning of the data and rely on the clocks at each end of the link being close enough to each other in frequency for the data to be correctly decoded. Typically, serial links like the RS-232 system, which is synchronized about once every 10-bit period, require the clock frequencies to be matched by better than 2.5% to ensure correct data reception; crystal oscillators are usually used to ensure this.

EIA/TIA 232E serial interface

The most widely used and best known standard system is formally known as EIA/TIA-232E, and has developed from the RS-232C standard that came

into use in 1969. RS stands for recommended standard; the designation RS has been dropped by the EIA and TIA (Electronic Industries Association and Telecommunications Industries Association respectively) – however, it is still widely used. The description RS-232 is generally used to cover both RS-232C and EIA/TIA-232E. Unfortunately, because the standard is so old and has gone through so many versions it is not always possible for two devices described as having RS-232 interfaces to communicate. The full implementation of RS-232C is seldom found nowadays, although many manufacturers have mislabelled EIA-232E interfaces as RS-232C. One obvious deviation concerns signal voltages. The original RS-232C system called for voltage levels of greater than +12 V and −12 V, up to a maximum of ±25 V, to be used for logic 0 and 1 levels respectively. Many modern systems, particularly laptop computers, make use of levels as low as −3 V and +3 V.

The second complication of RS-232 relates to its two different uses. When RS-232 was originally specified, two types of device were specified as data terminal equipment (DTE) and as data communications equipment (DCE). A DTE device can send out or receive serial signals, and is a terminal in the sense that the signals are not routed elsewhere. A DCE device is a translator for signals, like a modem, which converts serial data signals into tones for communication over telephone lines. The original conception of RS-232 was that a DTE device would always be connected to a DCE device, but with the development of microcomputers and their associated printers it is now more common to need to connect two DTE devices to each other. This requires the use of a crossover cable called a null modem, since it replaces two modems, each end of the cable appearing to be a DCE connection.

The original specification also stipulated that DTE equipment would use a male connector (plug) and the DCE equipment would use a female connector (socket), but you are likely to find either gender of connector on either type of device nowadays. The original specification was for a connecting cable of 25 leads, as shown in Figure 14.4c. Many of these reflect the use of old-fashioned telephone equipment and teleprinters, and very few applications of RS-232 now make use of more than eight lines. The standard connectors are the 9-way and 25-way D-type, but other connectors such as 8-way IDC and 5-, 6-, and 7-way DIN or XLR connectors are used. Some equipment makes use of the standard 25-pin D-connector but uses the ‘spare’ pins to carry other signals or even DC supply lines. The moral is

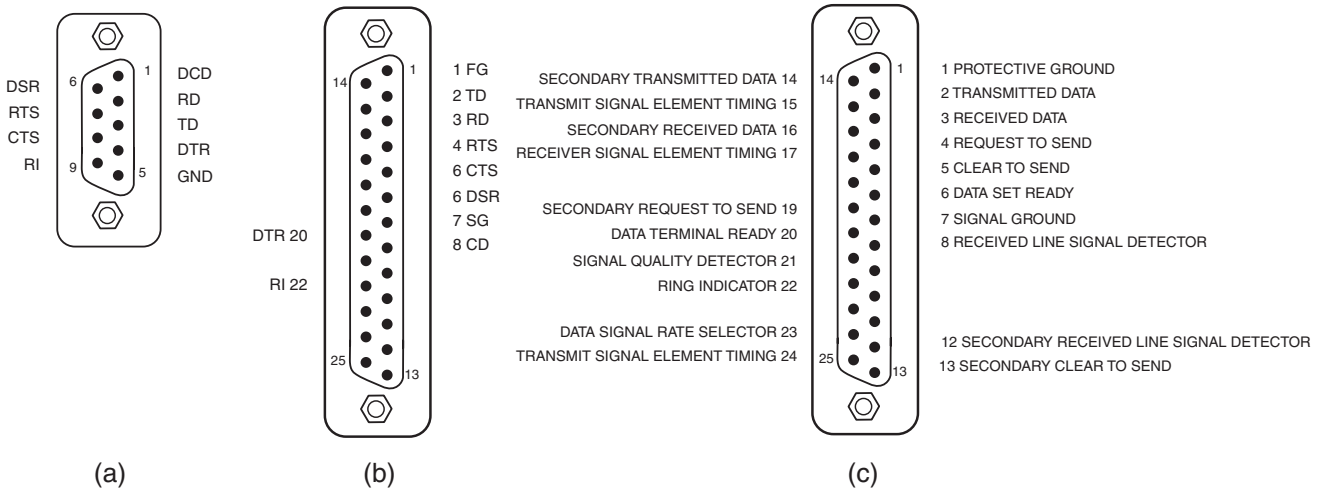


Figure 14.4

Typical computer RS-232 serial port pin assignments for **(a)** 9-pin D-connector, **(b)** 25-pin D-connector and **(c)** full RS-232 DTE-connector; many of these reflect the origins of the standard in teleprinter equipment.

that any link that is alleged to be RS-232 must be regarded with suspicion unless the wiring is known from a wiring diagram or from investigation of the connections.

The majority of RS-232 links can make use of eight pins of the 9-pin connector – these are pins 1 to 8; pin 9 is used to indicate when a modem detects that the line has become active, i.e. ring indicator. Figure 14.5 shows typical connections for 9-pin serial links. The 25-pin RS-232 connector has two ground pins, and a protective ground called *frame ground*, pin 1, is connected to the chassis of the equipment and signal ground, pin 7. The signal ground must be connected at both ends of the link but the frame ground should only be connected at the DTE end of the link. The 9-pin

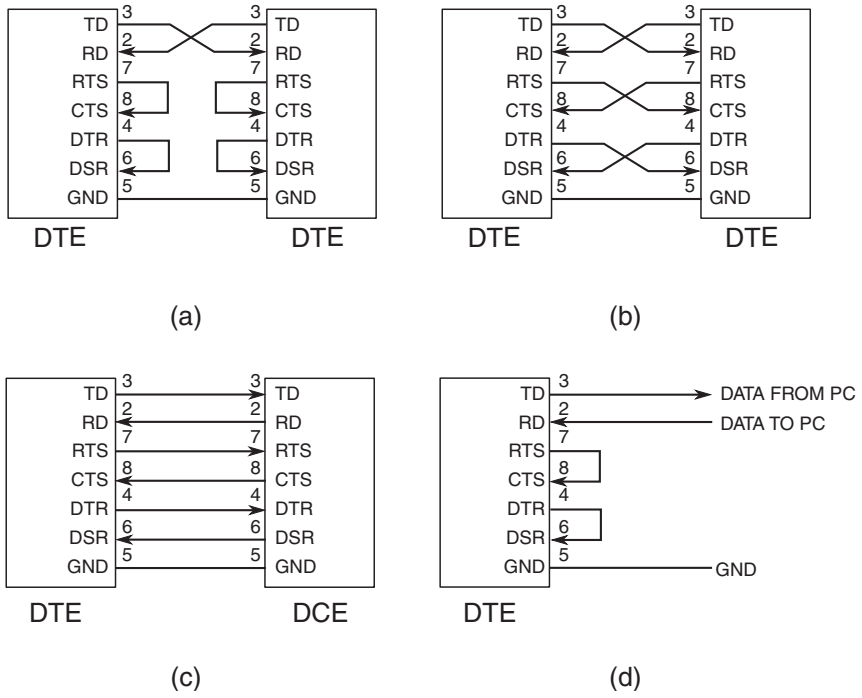


Figure 14.5

Common RS-232 connections using 9-pin connectors: **(a)** null modem without hardware handshaking, **(b)** null modem with handshaking, **(c)** straight connection DTE to DCE and **(d)** 3-wire connection typically used to connect devices like multimeters to a PC.

connection has only a signal ground pin. The main data pins are pin 3, the output pin for data transmitted from the DTE to the DCE, and pin 2, the input pin for data from the DCE to the DTE.

The use of separate transmit and receive lines means that the serial channel can be used in duplex, allowing transmission and reception of data simultaneously. A full 25-way RS-232C implementation allows for two sets of these transmit and receive lines (Figure 14.4c).

Note that a 25-pin serial connector on a PC machine is a male type (DTE). The 9-pin type of serial connector is used extensively on PC machines now, and it is fairly common to find that two serial ports are supported by the motherboard but only one using the 9-pin D-type connector fitted with an 8-way header usually provided on the motherboard for the other port. Either a 9- or a 25-way connector can be wired to the header. This arrangement is a convenient one because it avoids the need for adapters. Many laptop computers have no RS-232 or parallel printer connectors, but do have USB ports. USB serial and parallel port adapters are available, which are compatible with Windows 98 and later versions of Windows, and with Linux with USB support.

When a DTE is to be connected to a DCE, pin 2 of the DTE is connected to pin 2 of the DCE and pin 3 of the DTE is connected to pin 3 of the DCE. The other pins of the DTE are also connected to their corresponding numbers on the DCE. When two DTE devices are connected to each other, however, some links must be crossed. Pin 2 on one DTE must be connected to pin 3 on one other DTE, and similar cross-connection may be needed on handshaking lines. This difference in cabling is indicated by the naming of cables as modem (DTE to DCE) or non-modem (DTE to DTE), and failure to get a serial link working is very often the result of this very elementary difference.

An RS-232C serial link can be operated either *synchronously* (a data bit sent at each clock pulse) or *asynchronously* (data sent when ready), and since practically all modern applications of RS-232 only make use of asynchronous operation, the pin connections for synchronous use are frequently omitted. For asynchronous use, each transmitted byte has to be preceded by a start bit and ended by one or more stop bits. Ten or eleven bits must therefore be transmitted for each byte of data, and both transmitter and receiver must use the same number of parity and stop bits; the complete

group of bits, start, data, parity if used, and stop, are referred to as a frame. The way that the frame is made up (Figure 14.6) is often abbreviated to a three-character sequence; for example 8E1 stands for 8 data bits, even parity bit and 1 stop bit, whereas 7N2 is 7 data bits, no parity bit and two stop bits. Serial data are usually converted to and from parallel form used by microprocessors by a device known as a universal asynchronous receiver transmitter (UART). There are standard implementations of UARTs used in PCs, typically 16450/16550 chips. Microcontrollers often have hardware UARTs within their standard on chip peripherals.

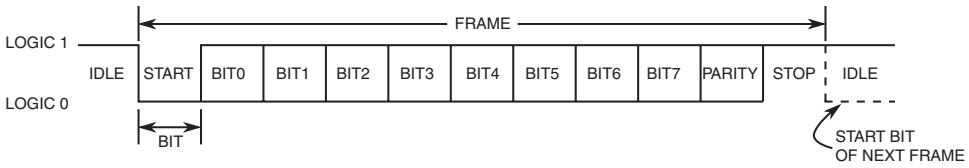


Figure 14.6

RS-232 asynchronous serial frame; each frame starts with a logic 0 start bit and finishes with a logic 1 stop bit.

The transmitter and receiver must use the same *baud rate*, the number of bit periods per second. Table 14.4 shows the RS-232 standard baud rates; very slow rates like 50, 75 and 110 tend to be used by audio frequency shift keyed radio links. Acoustically coupled modems used 300 baud, and modern modems connected to PCs via RS-232 use 9600 to 115200, although the transmission rate between the modems themselves may be different. Rates higher than 9600 are often used with equipment like set top boxes for configuration and diagnostics.

The serially transmitted data commonly uses ASCII (American Standard Code for Information Interchange) although other codes like IBMs EBDIC may be used, though these are almost obsolete. ASCII, which requires only 7 data bits to transmit control characters, numbers and un-accented characters of the Roman alphabet (Table 14.5), is being replaced in computer use by a 16-bit code, UNICODE, which allows representation of many accented characters and other alphabet systems including, among others, Arabic, Cyrillic, Hebrew, Greek, Roman, and Japanese kanji. Seven-bit data is thus becoming less common.

Table 14.4 Standard RS-232 baud rates

Baud rates supported by modern PC serial ports	Baud rates used by older equipment	Bit period	Frame rate	
			8N1	8E1
	50	20 ms	5	4.5
	75	13.3 ms	7.5	7
110	110	9.1 ms	11	10
	150	6.67 ms	15	14
300	300	3.33 ms	30	27
	600	1.67 ms	60	54
1200	1200	0.833 ms	120	109
2400	2400	416.7 µs	240	218
4800	4800	208.3 µs	480	436
9600	9600	104 µs	960	872
19200	19200	52.1 µs	1920	1744
38400		26 µs	3840	3497
57600		17.36 µs	5760	5237
115200		8.68 µs	11520	10473
230400		4.34 µs	23040	20947
460800		2.17 µs	46080	41894
921600		1.085 µs	92160	83787

ASCII data is still very useful for communicating engineering data and is likely to remain the standard for interfacing small microcontroller systems, owing to its code efficiency and wide acceptance as a standard.

The optional *parity* bit offers a check of the integrity of the data. The parity system can be even or odd. In the even parity system, the number of logic 1s in the data is counted, and the parity bit made either 1 or 0 so that the total number of 1s is even. In the odd parity system, the parity bit will be adjusted so as to make the number of 1s an odd number. At the receiver, the parity can be checked and an error flagged if the parity is found to be incorrect. This simple scheme will detect a single-bit error in a byte, but cannot necessarily detect multiple errors or correct errors. Methods such as cyclic redundancy checking (CRC) and Hamming or Reed–Solomon codes are needed to perform such correction; these are not implemented in the hardware of the RS-232 system but in the data itself.

The ASCII control codes include transmission control, display control and text formatting codes. STX (start of text) and ETX (end of text) along

Table 14.5 ASCII codes for control, and standard printing characters

	0	1	2	3	4	5	6	7
0	NUL	DLE	[space]	0	@	P	a	p
1	SOH	DC1	!	1	A	Q	b	q
2	STX	DC2	"	2	B	R	c	r
3	ETX	DC3	#	3	C	S	d	s
4	EOT	DC4	$	4	D	T	e	t
5	ENQ	NAK	%	5	E	U	f	u
6	ACK	SYN	&	6	F	V	g	v
7	BEL	ETB	’	7	G	W	h	w
8	BS	CAN	(	8	H	X	i	x
9	TAB	EM	)	9	I	Y	j	y
A	LF	UB	*	:	J	Z	k	z
B	VT	ESC	+	;	K	[	l	{
C	FF	FS	,	<	L	\	m	|
D	CR	GS	−	=	M	]	n	}
E	SO	RS	.	>	N	^	o	~
F	SI	US	/	?	O	_	a	DEL

Notes: the most significant hex digit of the code is shown at the top of each column, the least significant digit at the beginning of each row. For example G is represented by the ASCII code 47H.

with EOT (end of transmission) and SYN (synchronous idle) are typical of the transmission control codes and can generally be ignored. BS (back space), TAB, LF (line feed), VT (vertical tab), FF (form feed), CR (carriage return) and DEL (delete) are printer and display formatting codes which, along with letters, numbers and characters like !”#$%&’()* etc., form the bulk of the data transmitted.

As long as the transmitter and the receiver are set up to use the same protocols, meaning that the same baud rate, number of stop bits and parity method data can be transferred. A method of handshaking to ensure that signals are transferred only when both transmitter and receiver are ready is also required in most circumstances. The signals that are used for handshaking are referred to as RTS (ready to send), CTS (clear to send), DSR (data set ready) and DTR (data terminal ready). It is common for only RTS and CTS to be used.

When a DTE is connected to a DCE (computer to modem, for example), the RTS signal is sent out by the DTE to the DCE to indicate that the DTE has data to transmit. The DCE then responds with the CTS signal to indicate that data can be accepted, and data will be sent until the end of the data – note that handshaking is used only at the start and end of each block of data, not between bytes. The handshaking can be correctly implemented for connection of a DTE to DCE simply by using a cable that contains all the pins, a 9-way or 25-way cable. For the connection of two DTE devices to each other, a null-modem, or crossover cable, in which leads are crossed, can be used.

Software handshaking can be implemented in software by using the XON/XOFF system. This uses the ASCII codes 17 and 19 transmitted by the receiving device to stop and start the sending device. Data can be sent only once the transmitting device has received the ASCII 17 code, and disabled following the ASCII 19. Since these codes are sent over the normal data lines, only the data lines and earth need be connected. The rate of data transfer is slower because of the time that is needed to send the XON/XOFF signals.

RS-232 line levels need to be converted to logic levels so that microcontrollers and other logic ICs can be interfaced to systems using the bipolar signalling levels used by RS-232. Figure 14.7b shows a DCE schematic using the MAX232A chip from MAXIM Integrated Products. The chip, and similar ones from other manufacturers, has built-in charge-pump power supply circuits to generate the positive and negative signal levels used by the RS-232 system. The MAX232A needs 5 external 100 nF ceramic capacitors and can transmit and receive at up to 200 kbps (5 μs bit length).

RS-422/RS-485

The distance over which RS-232 links can be used – the cable length – is inversely proportional to the baud rate, as a general rule for baud rates above 19 200 cable length should be less than 10 m. Using the single-ended system, ground referenced voltage signals become susceptible to noise, voltage drop along the line and the stray capacitance of the line – all these effects become worse with shorter bit lengths. For transmission of data over longer distances differential signalling has advantages, using twisted

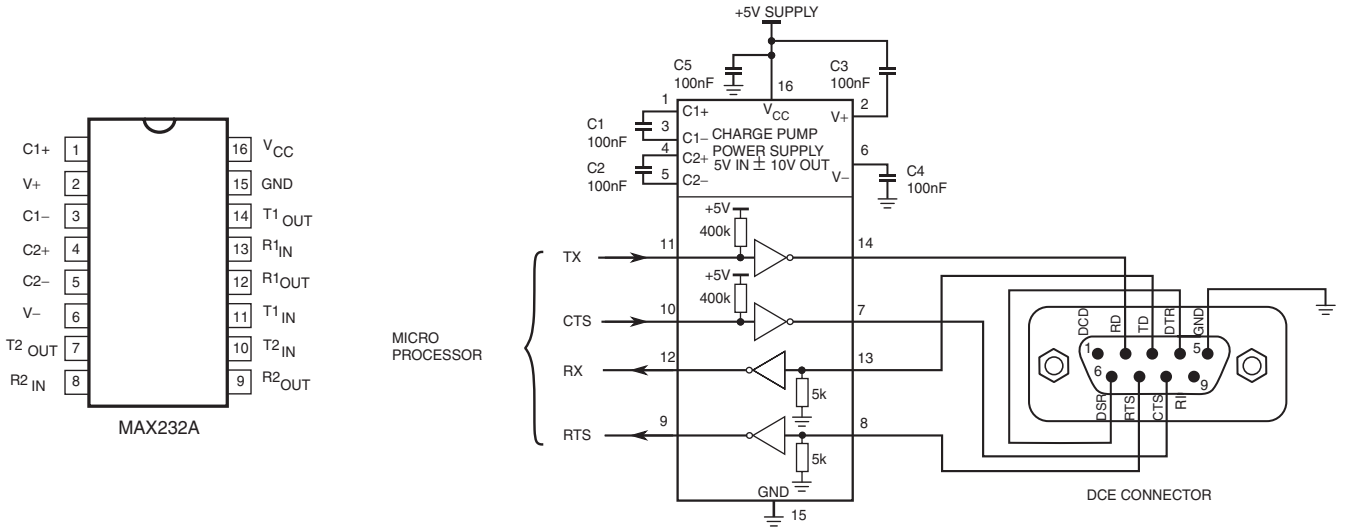


Figure 14.7

EIA/TIA-232E dual line driver/receiver circuit: **(a)** MAX232A chip and **(b)** schematic for 9-pin DCE application.

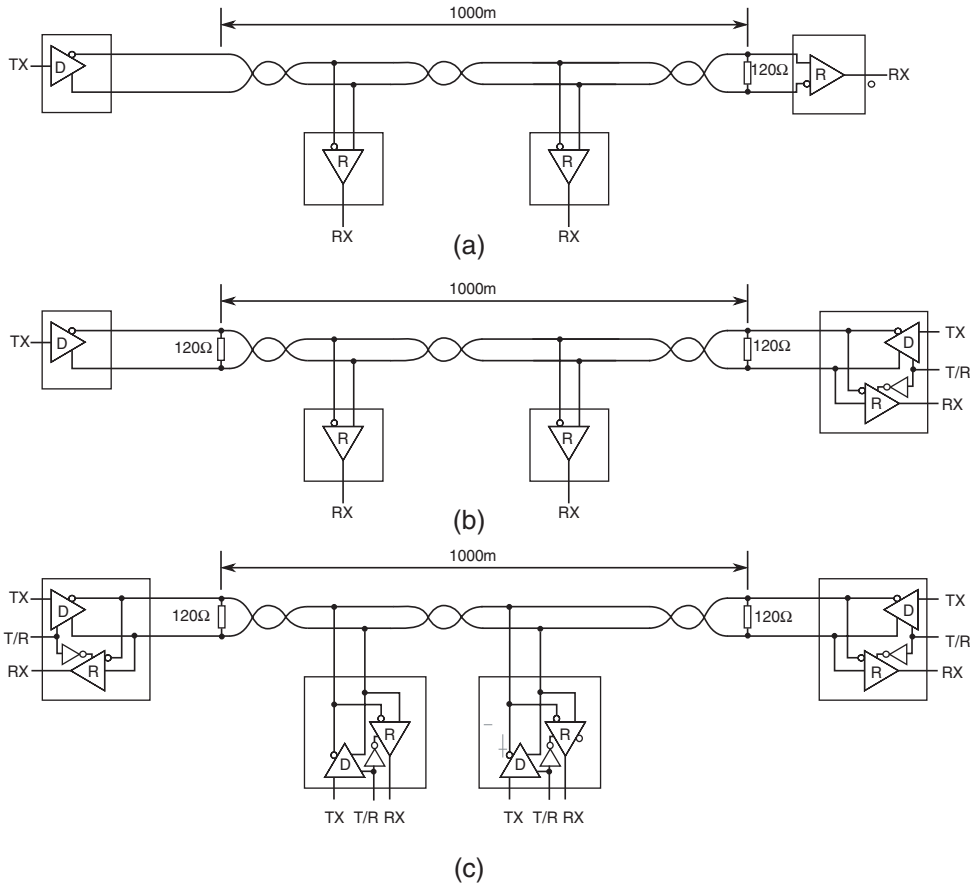


Figure 14.8

Differential signalling serial links: **(a)** RS-422 multi-drop, and **(b)** and **(c)** RS-485 half-duplex links. Line drivers are labelled 'D' and line receivers 'R'.

pair cables terminated at each end with 120 Ω resistors. RS-422/RS-485 links can be configured to provide unidirectional RS-422 (Figure 14.8a), half-duplex RS-485 (Figure 14.8b and c) and full duplex signalling. Using 24AWG twisted pair cable, link distances of 1 km are possible; over shorter distances, speeds around 10 Mbps can be achieved – a general rule is that the maximum data rate multiplied by the cable length in metres should be less than 10^8. RS-422/RS-485 also provides multi-drop capability, i.e. one transmitter to many receivers.

Differential signalling with screened twisted pair cable can provide good performance in electrically noisy environments like factories and vehicles. Standard transceiver ICs for RS-422/RS-485 are manufactured by several companies. These chips can be used to transmit signals like clocks and reference frequencies between pieces of equipment as well as being used for data transmission.

Asynchronous data transmission works well over cable connections which are usually a very benign environment. It does not take much noise to affect the edge detection in line or radio receivers and for this reason radio links that are used to carry data tend not to use simple start bit/stop bit framed transmission.

Wireless links

It is often inconvenient and sometimes impossible to run cables between pieces of equipment; TV remote controls and similar devices require wireless links.

Infra-red

Infra Red Data Association (IRDA) physical layer communications are based on the asynchronous data frame from a standard UART; in fact there are many combined RS-232/IRDA solutions used in mobile devices. The IRDA standard calls for a pulse of no more than 3/16 of a bit period to be transmitted from the half-bit position of the low bits of the data frame. The pulses range from 78.55 μs at 2400 bps to 1.63 μs at 115.2 kbps. The interface works by using a clock 16 times the bit rate to produce the pulses starting 7 cycles from the falling edge of the data waveform. The IRDA receiver detects the transmitted pulses and stretches them to provide correctly timed waveform for the UART.

The infra-red pulses of the IRDA data frame follow the form shown in Figure 14.9. The data rate may be between 2400 bps and 115.2 kbps using this type of pulse sequence. The IRDA standard provides for error correction and packet control protocol abstraction layers between the raw data of the link and the application that is transmitting or receiving data.

These protocols allow the seamless handling of noisy links by error correction and requests to re-send data. The optical environment that IRDA is designed to work in is not particularly noisy but the transmission does have to cope with the link being interrupted by objects passing between the transmitter and receiver. This calls for a packet structure and the ability to re-send data until it gets through.

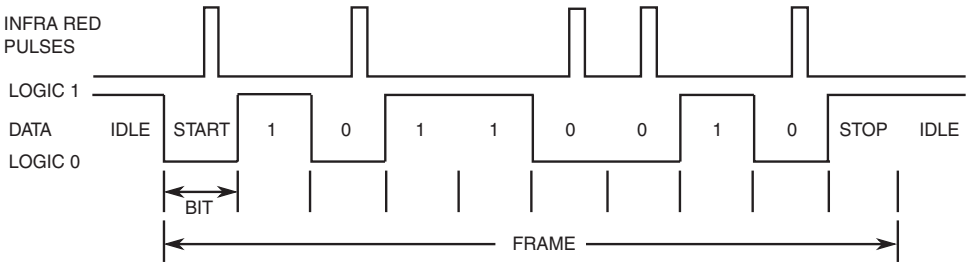


Figure 14.9

Typical IRDA data frame: 3/16 bit wide pulses transmitted to represent logic 0 bits.

Audio frequency signalling

For slow data links an audio frequency shift modulation can be used. A modem (modulator demodulator) at each end of the link converts the serial data to a sequence of tones and then back to voltage levels. Modems used for communication over radio or telephone lines were commonly connected via RS-232 ports of the connected devices; it is now more common to find either internal modems or USB connected ones. Baud rates up to 2400 bps can be supported by AFSK modulation of the type shown in Figure 14.10; higher bit rates tend to use phase shift keying PSK rather than FSK or AFSK. AFSK systems can be used to store data on conventional audio cassette tape in the way that home computers like the Sinclair Spectrum and BBC computers did in the 1980s. Some industrial systems still use this type of data storage since it is cheap and very robust.

Base-band signalling

One of the problems for systems using noisy channels like radio links or acoustically coupled audio transmission over the public telephone network

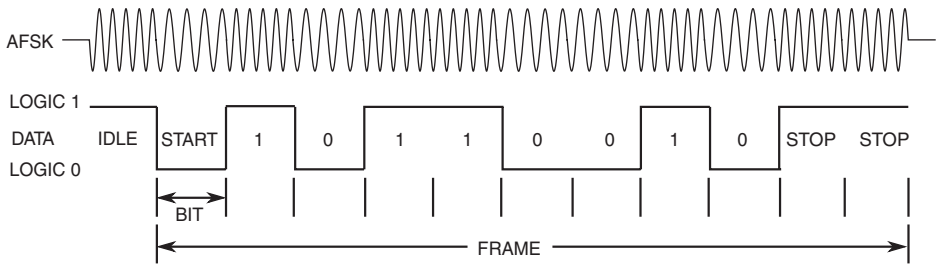


Figure 14.10

AFSK data frame: 8N2, 300 bps, low-tone 1200 Hz, high-tone 1500 Hz, 300 Hz shift.

is the difficulty of detecting the start of the data. Usually it is necessary to transmit a pilot tone or training data pattern for several frame periods before the start of the real data to ensure that the demodulator and serial receiver are able to synchronize to the data. Another requirement is to make the bit boundaries more clearly defined. The raw serial data shown in Figure 14.11 is described as **NRZ** (non return to zero)signalling because if two 1-bits are together in the data stream there is no change between them.

Most asynchronous systems that communicate over noisy channels or shared cables use return to zero (RZ) types of signalling.

Non return to zero signalling has problems, which makes it unsuitable for use over wireless links, like infra-red or short-range radio, used for

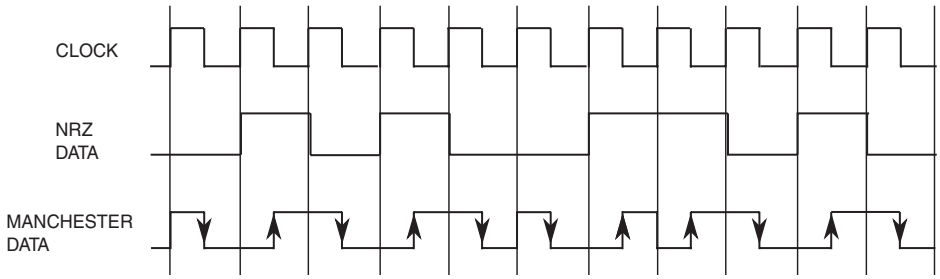


Figure 14.11

Manchester encoding – clock information is carried with the data.

remote control. This is due to the fact that long runs of ones or zeros can make recovery of the timing information from the signal difficult and in the case of simple AM radio receivers the decision threshold of the data detector depends on the average DC level of the data being 50% of the peak in order to demodulate the pulses with the correct widths.

The solution to this problem is to ensure that there is one transition in every bit period of the data; this makes the signal self-clocking and greatly aids decoding.

Manchester encoding, which is widely used in applications from wired Ethernet to wireless remote controls, is typically generated by XORing the clock and data signals. This has the effect of ensuring that there is at least one transition in every bit period. There are several advantages to using Manchester encoding, including easier recovery of the clock signal and maintenance of an average DC level near 50%.

Infra-red remote controls often use Manchester encoding to transmit data. This reduces the effects of ambient light at the receiver; and the data is modulated onto a 38 kHz carrier signal, so bursts of infra-red pulses 26 μs wide are transmitted to represent the high periods of the Manchester encoded data. The infra-red receiver output is bandpass filtered to reject continuous sources of ambient light as well as 50 Hz and its harmonics from mains lighting. The output from the bandpass filter is rectified to recover the Manchester data.

Short-range license-exempt radio systems using radio transmitter and receiver modules are often a convenient way of providing telemetry data links, and remote controls like key fob remote keyless entry (RKE) systems are common for car central-locking systems.

Radio links are very susceptible to noise, whether caused by other radio systems, intentional emitters, emissions from electrical and electronic equipment like motors and computers that are unintentional, or natural phenomena like lightning and other sources of atmospheric noise. Noise affects different types of modulation in different ways; low-cost systems often use amplitude modulation (AM) or, more correctly, on–off keying (OOK) of the carrier frequency. Frequency shift keying (FSK), often called frequency modulation (FM), in this context is more reliable in the presence of noise but the transmitters and receivers are more complex and

expensive and the current consumption is often higher, which makes AM attractive for small battery-operated transmitters like those used in key-fob RKE systems.

Remote keyless entry-type transmitters typically transmit between 16 and 64 bits of data at about 1 kbps data using 433.92 MHz either AM or FM, giving a range of 10–20 m with a suitable receiver.

On–off keyed transmitters based on surface acoustic wave (SAW) resonators are the simplest and most commonly used, consisting of just a single-stage transistor oscillator, which is turned on and off by the data stream. PLL-based transmitters are more expensive and generally use FM rather than AM.

Transmitter modules are available from a number of vendors for licence-exempt frequencies used in the UK. Although the EU has harmonized regulations for short-wave radio devices under the RTTE directive there are still national variations in frequency allocation and output power that must be observed. The most common transmitters intended for use in the UK are 433.92 MHz AM and FM units; Figure 14.12 shows typical connections for one such transmitter. The type and gain of the antenna that may be used is regulated by law, but generally a small resonant loop or a $\frac{1}{4}$-wave whip are recommended. Some transmitter modules have built-in antennas, and usually these may not be modified without breaking the type approval for the module.

Cheap receiver modules based on super-regenerative detectors are available but should be avoided because they are generally unreliable in the presence of out-of-band interference. The more expensive receivers of this type use an SAW filter before the detector and their performance can be good, but super-heterodyne receivers are to be preferred for stability and reliability reasons.

Receiver and transmitter modules are designed to be used as black boxes; feed them with power and data and connect an aerial and they should just work – so long as the manufacturer's recommendations for grounding, power supply decupling and aerial design are followed this is usually the case. Unfortunately, just because the radio link works does not imply that data can be reliably transmitted over it; the choice of data encoding method and data decoding algorithm are critical to successful deployment of a link.

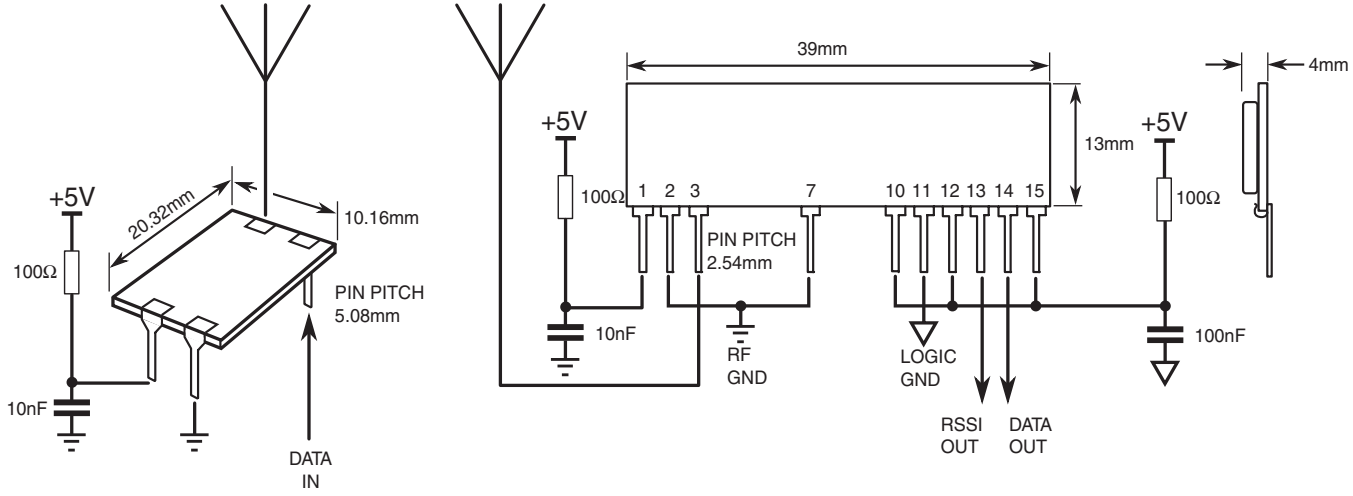


Figure 14.12

Radio modules: typical low-cost 433.92 MHz transmitter and receiver.

Some vendors have taken the module concept a step further and include microcontroller, transmitter and receiver in a single unit; the module typically has an RS-232 port and the whole link is transparent to the user – it can be considered to be a wire.

In order to establish the bit clock and start of data it is usual to transmit a preamble pattern to ensure that the data slicer in the radio receiver is operating at the average level of the transmitted signal. Typically the data slicer consists of two low-pass filters which integrate the signal from the AM or FM detector with different time constant. One is fast enough to respond to the shortest feature of the data being received and the other is typically ten to twenty times slower so that it settles on the average value of the transmitted signal. The outputs of these two filters are compared by a comparator, the output being high when the faster filter's output is the greater of the two. In Figure 14.13 the effect can be seen: when no data are being received the data slicer output 'chatters' randomly as it tries to slice the background noise; the preamble pulses charge the slow integrator and the output pulses become clean and well defined. The synchronizing bit can be clearly seen.

Error detection and correction

Most data communication systems are two-way links and, like IRDA or other packet switching systems, requests to re-send data can be used to improve the resistance to interference. Using error-detection coding like parity bits and cyclic redundancy checks that there is a compromise between the length of the data that form a unit for error checking and the number of data that are added by the error-checking system, and the speed of data transmission that can be achieved in the presence of noise. The problem for one-way links like remote controls is different because the receiver can not ask for retransmission and it is usual to either add redundancy to the transmission, for example transmit every message three times, or to use error-correcting codes. An example of an error-correcting code is the use of an increased decoding distance between valid symbols, making it harder for the receiver to confuse one symbol for another in the presence of noise; for example, in Figure 14.11, if a code is used that ensures that the only valid symbols are at least 5 bits apart (referred to as a Hamming distance of 5) the errors of up to 2 bits in any symbol can be corrected. The effect of this is to make messages longer, since 16 symbols take 256 raw codes and

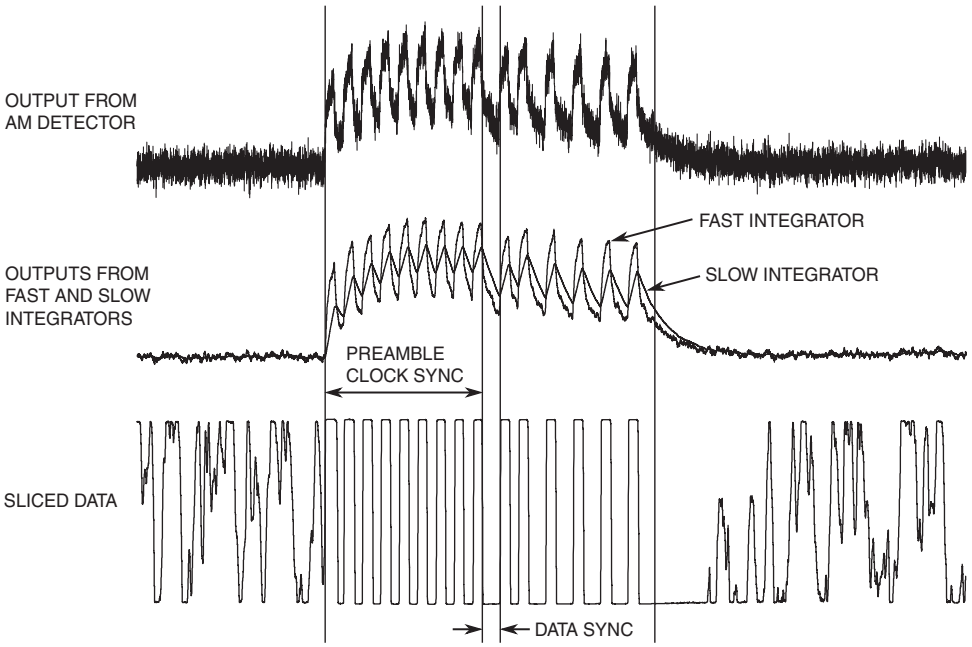


Figure 14.13

Recovering data from a demodulated AM signal; the data slicer output can be seen to chatter randomly when no valid signal is being received.

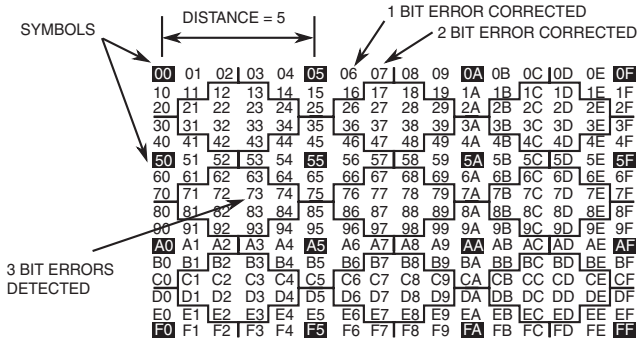


Figure 14.14

Error-correcting codes, 16 symbols out of a possible 256 allowing correction of up to 2 bits per symbol.

so two symbols (16 bits) are required to represent each of the 128 ASCII codes shown in Table 14.5 – this makes messages twice as long.

Error-correction schemes always make a message longer, but the type of error correction used depends on the noise environment; sometimes redundancy is more effective than error correction, particularly in the presence of burst noise which can destroy more than a few bits at a time.

Useful websites

RS232 and RS485 transceivers

Maxim Integrated Products
www.maxim-ic.com
Linear Technology
www.linear.com
Texas Instruments
www.ti.com

Short-range radio modules

Low Power Radio Association
www.lpra.org
RF Solutions
www.rfsolutions.co.uk
Telecontrolli
www.telecontrolli.com

CHAPTER 15

MICROCONTROLLER APPLICATIONS

Introduction

Microcontrollers provide the building blocks that support almost every device and piece of equipment that defines modern life. Mobile phones, digital cameras and media players are obvious examples. Larger pieces of domestic equipment like dish washers, washing machines and central heating controllers are today expected to be controlled by microcontrollers too. Other less well known applications include rechargeable batteries for some mobile phones, which use a microcontroller to ensure safe charging and authenticate that the battery is genuine, ink cartridges for printers that report use by date and ink level to the printer, and disposable temperature loggers the size of a £1 coin which can be attached to pallets of food or packages containing perishable medicines in refrigerated transport to show that they have not been exposed to out-of-specification conditions.

Microcontroller applications are categorized into three basic types:

1. stand-alone devices that interact with people, but do not need to interact with other systems, for example thermometers, pedometers, calculators and hand-held translators;

2. devices that interact with their environment or users and other systems, for example infra-red remote controls, media players, mobile phones, computer mice and burglar alarm sensors;

3. devices that perform automated functions when they are triggered and may optionally be connected to other equipment, like car washes, automatic wheel-balancing machines and CD players.

Microcontrollers are available in a range of sizes, complexities and with different on-chip peripherals. Most 8- and 16-bit micros are packaged in plastic dual-in-line or surface-mount packages with pin counts between 14 and 68 pins. At the extreme ends of the range are devices like ATMEL AT91RM9200 ARM9 based parts in 208-pin fine pitch quad flat pack and the PIC 10F200 in a 6-pin SOT23-6 package (Figure 15.1).

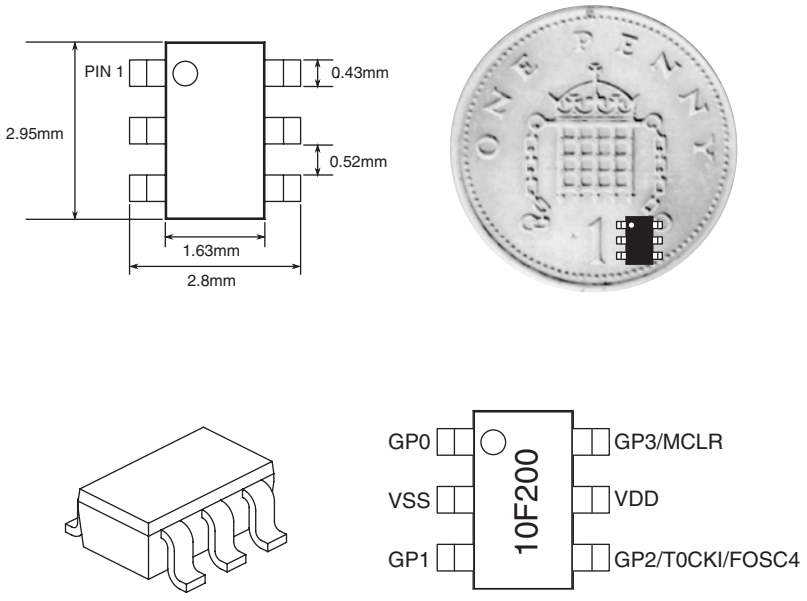


Figure 15.1

Microcontroller chips can be very small, like the Microchip PIC 10F200 in a 6-pin SOT-23 package shown here compared with a 1p piece.

Very small packaged microcontrollers, in standard surface-mount format, open up a range of product applications that would not have been possible without resorting to chip-on-board or ceramic substrate multi-chip hybrids; this makes small-volume production for custom application much more attractive.

Configuration

Microcontrollers are programmable; they execute a stored program. They also need configuration, and this is one of the differences between microcontrollers and microprocessor systems where multiple chips are assembled on a PCB and configuration can be achieved by choice of chips and interconnection.

The configuration choices depend on the peripherals available and how they are to be used. Typically the choices include internal or external oscillator for the processor clock, selecting pin functions between general purpose digital I/O and peripherals like UARTS and ADCs. There is also usually the choice of enabling an on-chip watchdog timer with its own independent oscillator. It is usual to have the configuration data in the assembler source file or in one of the project files for a C project. The example given below is typical of a PIC16 configuration line, in this case setting code protection off, which is a security feature; turning it off allows the code to be read back after programming. The watchdog timer is disabled, the clock oscillator is set up for a crystal and the power-up timer is enabled.

```
__CONFIG _CP_OFF & _WDT_OFF & _PWRTE_ON & _XT_OSC
```

The power-up timer holds the processor in reset to ensure that the crystal oscillator is running before the processor starts to execute; crystal oscillators often take several milliseconds to start up owing to the very high Q of the crystal resonator. The RC oscillator can start in the equivalent of a couple of clock cycles.

Clock

Some microcontrollers can select between internal RC and external RC oscillator and one or more speeds of crystal oscillator, in their configuration. Obviously it is important to ensure that the correct choice is programmed; however, the factory default is usually the internal RC oscillator. The internal RC oscillator has the advantage of making extra I/O pins available.

Some microcontrollers allow switching between oscillator types under software control; generally this would be done to improve power consumption while performing non-critical tasks.

The options for external clocks are shown in Figure 15.2. The standard Pierce crystal oscillator circuit uses two I/O pins and when clock accuracy is required this is an easy way of achieving it. Employing an external clock source like a packaged oscillator, as shown in Figure 15.2c, would normally have the internal oscillator in crystal oscillator mode, using it as a buffer circuit, and for this reason the OSC2 pin should be left unconnected since loading it might affect the clock signal.

The external RC oscillator simply consists of a resistor and a capacitor with an open-drain FET typically used to discharge the timing capacitor and the resistor to charge it. The signal at the OSC1 pin will be a sawtooth waveform oscillating between the high and low thresholds of the Schmitt buffer. In this case, and in the case of the internal RC oscillator, the OSC2 pin can be configured to output the divide-by-4 internal processor clock or be used a general purpose I/O. The external RC clock allows stability of about 0.5% to be achieved with arbitrary R and C values. This can be convenient in some circumstances.

Internal RC oscillator

The internal RC oscillators of microcontrollers are usually factory trimmed by measuring the clock frequency and determining the value that needs to be written to a DAC that sets the current or capacitor value to achieve a frequency near nominal. Microchip offer 37 kHz, 4 MHz and 8 MHz internal oscillators in a variety of their parts; most common, however, is the single option of a 4 MHz oscillator. Published data for the PIC12C50X suggests that the calibrated oscillator will be between 3.65 MHz and 4.28 MHz at 5 V and 25°C, a spread of –8.75% to +7% around 4 MHz, and a graph in the data sheet shows a negative temperature coefficient of about 1 kHz per degree. The graph shown in Figure 15.3 is the result of what we measured and although based on a sample of one it is interesting – the oscillator was 6.4% high at 25°C so just inside the limit, and the frequency deviation has a quadratic shape which is probably indicative of a band gap voltage regulator, without curvature correction, providing the supply to the RC oscillator circuit in the processor. The main thing to note, however, is just

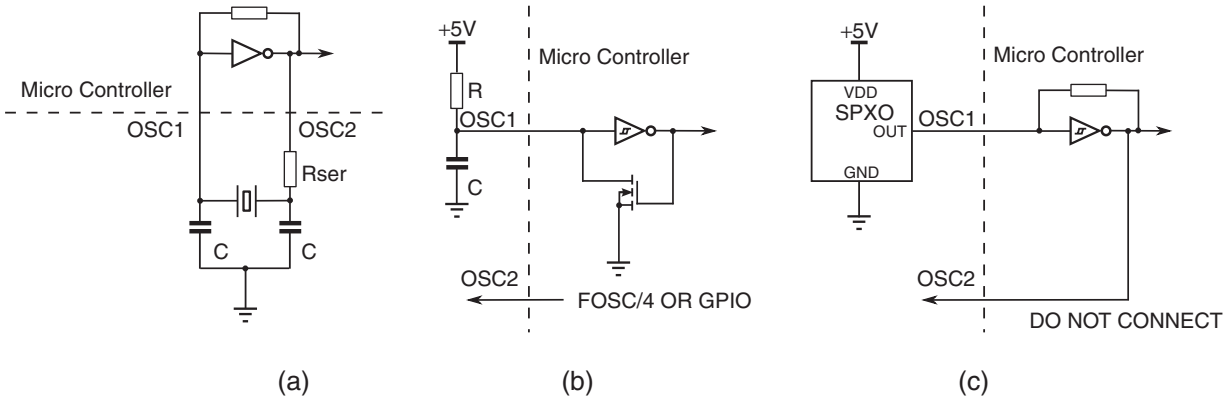


Figure 15.2

Microcontroller clocks: **(a)** crystal or ceramic resonator, **(b)** external RC oscillator and **(c)** external clock, e.g. crystal oscillator module.

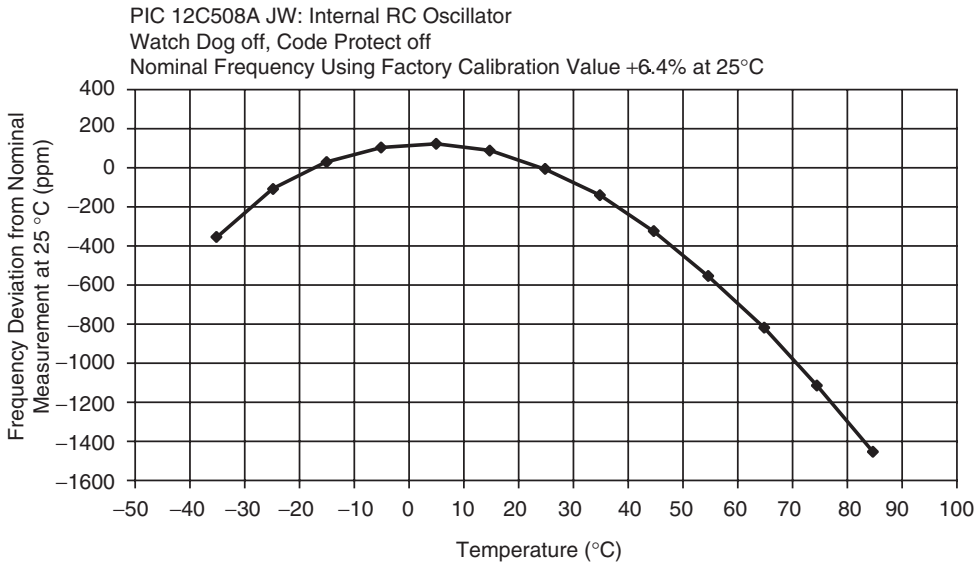


Figure 15.3

Performance of the internal oscillator of a microcontroller over the operating temperature range.

how good the temperature performance of the oscillator is in the region between –20°C and +30°C where the deviation from the 25°C is less than 200 ppm or 0.02%.

Watchdog and sleep

Watchdog timers are designed to reset the processor, to put it into a known state, if it appears to have crashed. The watchdog may be an external chip or one of the internal peripherals of the processor. In operation the watchdog consists of a timer that takes typically a couple of seconds to reach its terminal value, so the software running on the processor must reset the timer at a shorter interval than the period of the timer to avoid a reset being forced. A re-triggerable monostable based on the NE555 timer could be used as a watchdog, but there are quite a number of dedicated watchdog chips on the market often combining the watchdog with other functions like real-time clocks or voltage monitors.

Internal watchdog counters have to run from their own oscillator since the processor in sleep mode might have switched off the clock; however, they can often have prescaler values set in software which allow the time out to be varied from a few tens of milliseconds to several seconds. It is wise to ensure that the watchdog timer is reset several times in its period to avoid oscillator accuracy problems.

Most microcontrollers have a *sleep* instruction that causes the processor to be put into a low-power mode and stops the processor clock. The processor can be woken from this state by a variety of means typically including changes on input pins and the watchdog timer timing out. When the processor wakes up, it starts to execute the program at the instruction following the sleep instruction; if it is woken by something that caused an interrupt then the interrupt service routine will execute followed by the instruction after the sleep instruction. The sleep function can be used in devices like remote controls for TVs and DVD players to save battery life by putting the processor into a low-power state until a button is pressed.

The watchdog timer can be used to wake a processor from its sleep state, and, to enable consistent and maximum sleep time, the sleep instruction typically resets the watchdog timer. When the watchdog times out a reset is generated so the processor does not execute the instruction following the sleep instruction.

The sleep instruction can also be used to shut down the processor until an external reset takes place; for example, if a microcontroller is used to carry out the power-up set-up of a system and is not required after this has been completed, then by disabling the interrupts and watchdog the sleep instruction can be used to turn off the processor until the next time the equipment is turned on. This kind of behaviour can be used to program FPGAs and frequency synthesizer chips at power-up.

Power-up reset

At power-up it is important to ensure that the processor starts in a known state, and in order to achieve this a reset circuit in the processor clears registers like the program counter, port data and direction registers, interrupt flags, etc. It is usual for this reset circuit to be driven by an input pin,

although many microcontrollers can operate with entirely internal reset functions making the pin available for general purpose I/O.

In situations when the power supply starts slowly or when the power supply voltage may dip or slowly decline, as in the case of discharging batteries, it is often necessary to provide external reset circuitry to ensure that the processor is held in the reset condition until the power supply is stable enough for operation.

A reset circuit consisting of a resistor and capacitor (Figure 15.4a), can provide a delay of a few milliseconds at start-up. The diode is required to prevent the capacitor discharging through the protection diode of the reset input in the event that the power supply line is pulled low.

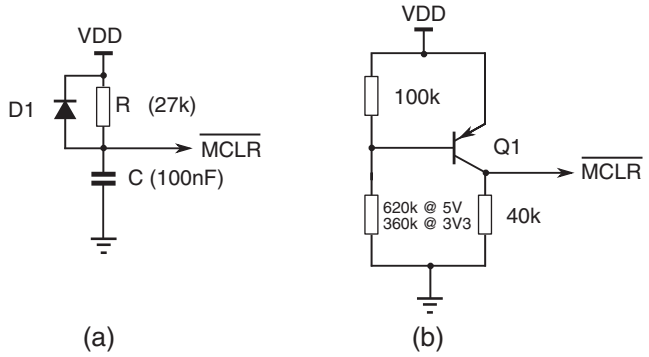


Figure 15.4

Power-up reset circuit **(a)** and brown-out circuit **(b)** recommended by Microchip for the PIC16 family.

The brown-out circuit shown in Figure 15.4b is designed to let the 40k resistor pull the reset pin low if the power supply drops below a given value. This can be important when running a microcontroller with a crystal oscillator at the high-frequency end of the operating range, because at low voltages the CMOS circuits are slower and operation may become uncertain as the frequency of the clock reaches the speed limit of the gates that make up the processor. In operation the transistor will be turned on so long as about 0.6 V is available across the 100k resistor. So if the supply voltage

drops below the voltage necessary to keep the transistor turned on the processor will be reset, and held in reset until the voltage rises again.

The simple circuits shown in the Figure 15.4 are suitable for most common applications but where specific requirements make more precise reset necessary application-specific reset chips are available from several vendors. These are sometimes combined with watchdog chips, and are often described as *microcontroller supervisors*.

Setting up I/O ports

The ports of a microcontroller can be set up by software; some functions, however, may be determined by the configuration settings of the processor. A common example of this is the function of the second oscillator pin (OSC2) which may be used as general purpose I/O when an RC oscillator is used.

Because microcontroller inputs are CMOS they are high-impedance inputs. Unused pins should, where possible, be configured as outputs; if they must be left as inputs or cannot be configured, they should be tied to ground or supply. This is because CMOS inputs have a tendency to drift into the undefined region between the one and zero thresholds where both the top and bottom FET are conducting and this causes a significant increase in the power supply current of the chip, shortening the battery life of battery-operated equipment and possibly causing voltage drops in the internal chip power distribution, which may affect performance.

Typically, microcontrollers offer 8-bit-wide I/O ports, even when they use a 16-bit or wider internal data bus. Figure 15.5 shows the main features of an I/O port pin. The pin can be configured for input or output under software control by writing to the data direction register.

When configured as an output, writing a 1 to the data latch will cause the driver to pull the pin; high writing a 0 pulls the pin low. Reading the port pin value when it is configured as an output depends not just on the data that was written to the pin but on the load. If the output pin is shorted to VDD or GND, the pin will be read as the shorted value rather than

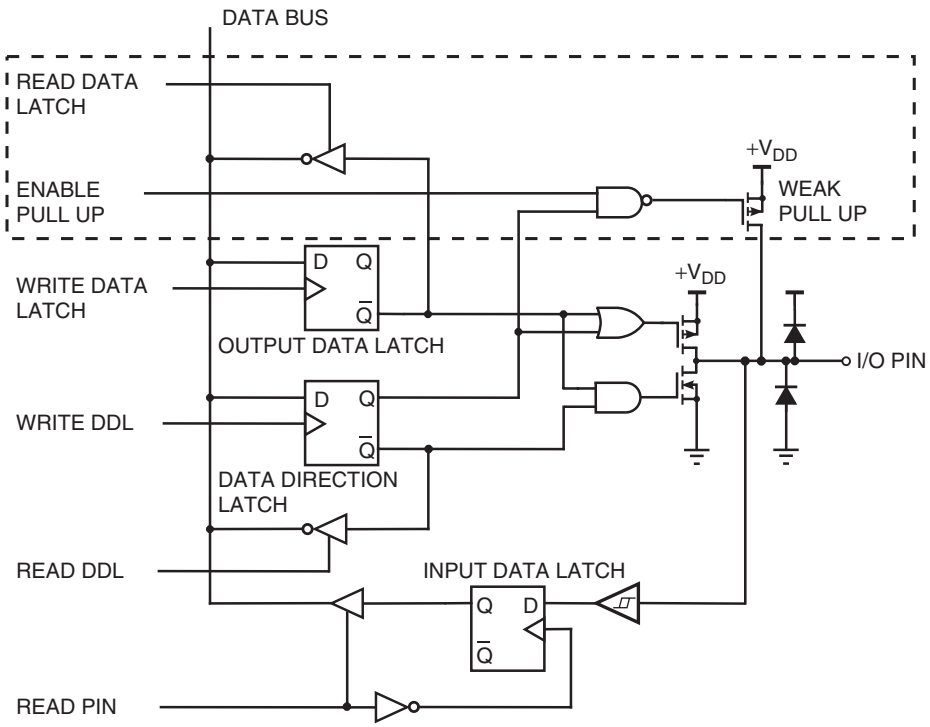


Figure 15.5

Typical microcontroller I/O port; note that weak pull-up and ability to read the data latch output are not features of all microcontrollers.

the data that were written to it. If the pin has a large capacitive load, it may take a significant length of time for the pin to change state when it is driven. Some microcontrollers provide a separate register to allow the data latch to be read directly, as shown in Figure 15.5; this can make control of the port easier because the set value can be read back.

It is often useful to have open drain outputs since this allows wired-AND connections as used in some data buses for handshake signals, and it is also useful for applications like the row drivers for a keypad. Some microcontroller ports have open drain outputs but often this is limited to a single pin, for example the RA4 pin on the Microchip PIC16F series

of microcontrollers. Open drain ports can be emulated by setting the data register to zero and using the data direction register to set the port to be an input for logic 1 and an output for logic 0 output. This means that when the data is 0 the output driver is enabled and since the data register is set to 0 the output pin is pulled low, when the data is a 1 the port becomes an input and the pin floats just as an open drain port would behave. There is an issue with some microcontrollers in that, like the PIC12C and 16F parts, the data direction register is in an alternate register bank requiring extra instruction cycles to switch to it before the port value can be changed. This limits the speed at which the pin can be toggled to about a third the speed for the port used conventionally; this is however not usually a problem.

When the port is configured as an input the output driver transistors are turned off, and a buffer, often of the Schmitt variety to improve the response to slowly changing and noisy inputs, drives the data input of a latch. In order to ensure that changes on the pin do not occur while the microcontroller is reading the internal data bus the latch is triggered before the data bus read cycle occurs. Microcontroller inputs are almost always synchronous although not all data sheets show this explicitly. Sometimes an on-chip weak pull-up for the port pin is provided; this is typically implemented with a high on-resistance p-channel MOSFET. The weak pull-up provides roughly between 20 kΩ and 1 MΩ pull-up resistors. This value may vary considerably from batch to batch, so the exact value of weak pull-up resistors should not be relied on. The weak pull-up feature is very useful, however, because it allows the implementation of inputs like the keypad (see Figure 12.12b) without any additional components apart from the switches.

Setting the port up as an output automatically disables weak pull-ups where these are provided, which is another advantage of using this arrangement. External resistors would either have to be taken account of in the design or disconnected with extra external circuitry.

The implementation of pin change detection for generating interrupts is shown in Figure 15.6. The state of the input is compared once each clock cycle with the last value that was read, and if there is a difference the output of the XOR gate will be high and an interrupt can be caused. Additional logic, not shown in Figure 15.6, disables the circuit when the pin is configured as an output.

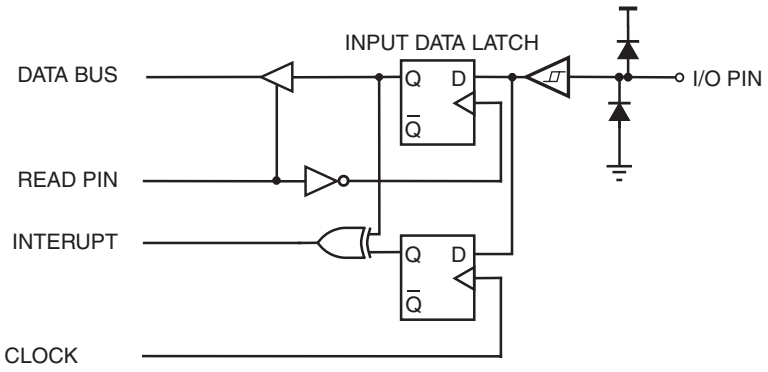


Figure 15.6

One implementation of interrupt on pin change input circuit.

Integrated peripherals

It is possible to use a microcontroller which only has conventional digital I/O, but there are many advantages in having more complex peripherals like counters, serial interfaces or analogue-to-digital converters. This is partly because the size of the peripheral on chip hardly changes the size of the chip but interfacing to the same peripheral off chip will require several times the board area, interconnect on the PCB and at least one extra IC package as well as decoupling capacitors, etc. If on-chip peripherals or software solutions using conventional digital I/O can meet the desired specification it is always better to use them because of the size and cost advantages.

Counter timer

One of the simplest peripherals to implement is the counter timer. There are many possible options but most follow a standard pattern. A counter that has a parallel load facility forms the heart of most counter timer circuits. The counter is typically an up-counter but the parallel load facility allows it to be used in applications that would usually require a down-counter by writing a start value that is the desired number of counts less than the maximum value. For example, 246 plus 10 counts would cause the timer to

roll over to zero. The counter will usually be able to generate an interrupt when it reaches zero.

The counter can have inputs from the processor clock or an external pin or even an external oscillator using a crystal. Figure 15.7 shows a generic counter arrangement in which the input is from an I/O pin and an XOR gate allows the selection of low-to-high or high-to-low transitions being counted. A prescaler block allows for a set number of edges to be required before the counter is incremented, so, for example, setting the prescaler to 64 would mean that the counter required 16 384 edges to count from 0 to roll over to 0 again causing an interrupt.

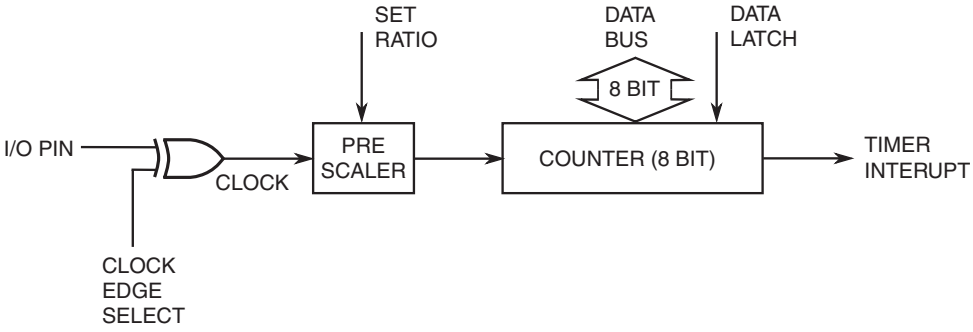


Figure 15.7

General purpose counter timer.

Pulse width modulator

A more complex counter peripheral is the pulse width modulator. This consists of a counter with additional comparator and a data latch and output flip-flop. In operation the counter counts continuously and every time it rolls over to zero the output flip-flop is set and the output pin is high. The comparator compares the count value with the data value in the duty cycle latch and when they match the output flip-flop is reset and the output pin pulled low. The general arrangement is shown in Figure 15.8.

If the duty cycle register contains 64, for example, the output will be high for the first quarter of each cycle of the counter, 64/256. PWM counters

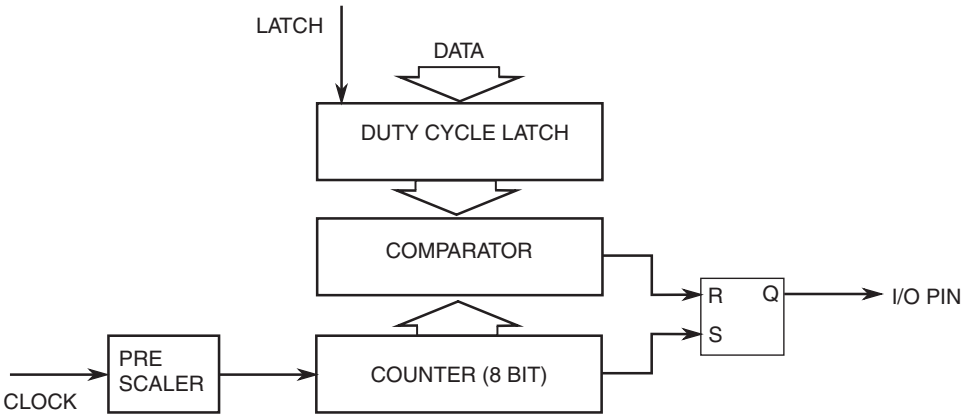


Figure 15.8

PWM module using an 8-bit counter.

are usually 8 or 16 bit; some are implemented with two comparators, the second one being for the period rather than just counting up until the count rolls over to zero. The nature of this type of PWM generator means that the resolution is inversely proportional to the repetition rate since if the period is set to 8 counts the smallest step is 1/8 but the repetition is 32 times higher than one with 256 steps.

Serial interfaces

- **UART/USART**

One of the commonest peripherals is a serial interface, often a combined synchronous asynchronous serial port that can be configured as required. The serial interface consists of shift registers clocked by a baud-rate generator, parity-bit generation and checking and transmit, receive and status registers. It is usual for the receiver to cause an interrupt once it has received a byte and the transmitter to interrupt once it has transmitted a byte. Serial port implementations often differ considerably between different vendors' microcontrollers, partly as a result of the combination of features made available. Generally, in operation, the baud rate generator, number of data, parity and stop bits and transmit enable must be set up by writing to

a control register and the transmitter and receiver must be selected by setting bits in a peripheral control register. This will usually override the status of the data direction registers of the I/O pins used but this is not always the case. A status register will have flags for *transmit shift register empty*, *receive register full*, *receive checksum error*, *receive framing error* and, if the receiver has a first-in first-out buffer, there will be a *receive buffer overrun* flag. The transmit shift register empty and receive register full will usually have interrupts associated with them.

Once the baud rate and other configuration settings have been set, transmission of data is effected by writing the data to the transmit data register, and waiting for an interrupt or the transmit register empty flag to be set in the status register before writing the next byte. Received data can be read from the receive register once the receive register flag is set in the status register; the program can wait for an interrupt or poll the status register at regular intervals to determine if data is waiting.

Baud rate generation is performed by a programmable divider and not all baud rates can be achieved for any given microcontroller clock frequency. It is usual to use a crystal oscillator for the clock when asynchronous serial communications are required because the clock usually needs to have better than ±2.5% accuracy. Vendor data sheets usually give recommended division ratios for common crystal frequencies to best match the standard baud rates. 9600 bits per second is a very popular rate because it can be generated from a 4 MHz crystal divided by 4 and then 104, and a 10 MHz crystal divided by 4 and then 255.

- **SPI/I^2C Bus**

Serial data buses are often used inside equipment to connect microcontrollers to peripherals and non-volatile memory devices. There are three main types of serial bus in common use, all of which have several variants. These are three-wire bus, SPI bus and I^2C bus.

The three-wire serial bus is a synchronous system using a clock signal to synchronize the data bits and a latch signal to synchronize the bytes or blocks of data.

The operation is very simple (see Figure 15.9). The data line is set and the clock line is pulsed high for a short period; then the next data line is set to

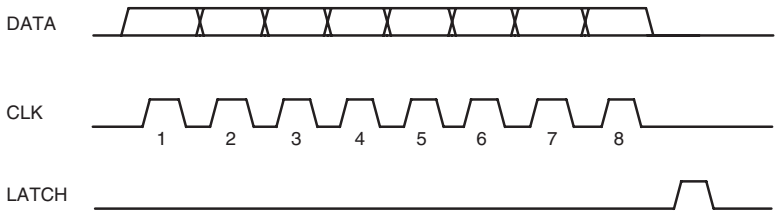


Figure 15.9

Three-wire serial, data, clock and latch.

the next bit and the clock pulsed again. Once the desired number of bits has been transferred, a latch or chip-select signal is pulsed to load the data in parallel from the receiving shift register in the device. This system tends to be used for one-way communications such as writing configuration data to frequency synthesizers and display segment data to serial LED displays. The three-wire link can be easily implemented using standard 74HC164 serial-in parallel-out shift register and 74HC574 octal latch, and this is a convenient way to add I/O to a microcontroller.

A fully synchronous bus, the SPI bus is similar to the three-wire bus but does not use a separate latch signal, counting the clock cycles instead to determine when all the data has been clocked into the shift register. The SPI bus is commonly used with E^2 memory and ADC chips. There is one master, the microcontroller and multiple slaves; the master selects which slave it is addressing with extra lines called slave select lines. SPI is a three-wire bus in the sense that as well as the clock (SCLK) there is a data line for data from the master to the slave (MOSI) and a separate data line for data from the slave to the master (MISO). The SPI bus has a limitation in that it requires a separate slave select line for every slave; this means that while it is useful for connecting one or two devices to a microcontroller it becomes inefficient in complex systems with many peripherals.

The I^2C bus, developed by Philips and protected by their patent, allows multiple devices to be connected to a two-wire bus. It has developed into a standard with many vendors buying licenses to use the I^2C interface for their microcontroller, E^2 memory and other devices. Originally it was developed for internal communications in consumer electronic and

telecommunications equipment, the typical application being a television receiver, with remote control, digital PLL tuner, E^2 memory to store channel, video and audio settings and a multi-page teletext decoder. Each device connected to the bus has a unique address which is composed of two parts: a device identity which is unique to the type of IC and is assigned by the I^2C bus standards committee at Philips, and an instance address which is usually set by the address pins of the device, thus allowing several identical devices to be used on the same bus. More than 50 vendors are listed as having subscribed to the I^2C standard but there are still some devices marketed as being I^2C compatible; these may have the same ID code as other devices since they have not been officially assigned by the standards committee – in itself this is not usually a problem because the address bits can be set to discriminate between devices. However, you should be aware that if the I^2C logo is not displayed on the data sheet it is worth taking extra care in checking the specification.

The I^2C bus uses two wires: SCL, the clock line; and SDA the data line. These are connected to open drain drivers and CMOS inputs of the devices connected to the bus. The lines are pulled up to the common supply rail by resistors or current sources depending on the implementation. Figure 15.10 shows the general arrangement of the bus connections and the driver receiver circuits in each device.

The open drain connection allows wire ANDing of signals, which means that each wire of the bus can be used to signal in both directions from master to slave and from slave to master.

Initially both lines are high, that is none of the devices has an active output; this condition must have existed for at least 4.7 μs (based on 1.8 V to 2.5 V supply voltages) to ensure that the state machines of all devices are reset and a transaction can be correctly initiated. The master, usually the microcontroller, always initiates a transaction. This is done by pulling the SDA line low while the SCL line remains high and then pulling the SCL line low after 4 μs. Figure 15.11 shows a 2-byte transaction.

Once a start condition has arisen the master can send an 8-bit block of data. SDA is allowed to change only while the SCL line is low. Data is clocked on the rising edge of the SCL line and after 8 bits have been clocked out the master releases the SDA line and reads it after the 9^{th} rising edge of

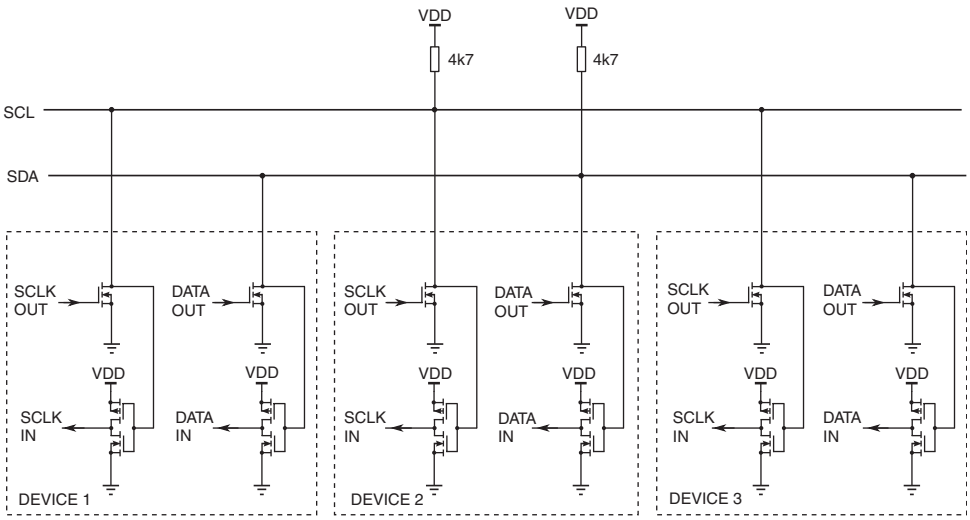


Figure 15.10

Connection of devices to the I²C serial bus.

the clock. The slaves must hold SDA low during the high period of the 9th clock pulse to indicate that they have acknowledged the data transfer – this acknowledge is required before the master will try to transmit further data.

If an acknowledge is received by the master it will continue to transmit data bytes. If no more data needs to be transmitted, or the acknowledge is not detected, the master generates a stop condition by pulling SDA low, releasing the clock and then releasing SDA after a delay.

The master always generates the clock; if a command causes a device to send data back to the master then the master generates the acknowledge conditions.

The format of data transmissions using the I²C bus is defined by the standard. The first byte that the master transmits after a start condition determines the mode that is being used. There are several options but the most frequently used is the 7-bit address format. This consists of the 7-bit address of the device being addressed, made up from its unique code and the settings of its address pins, followed by a read/write bit which determines

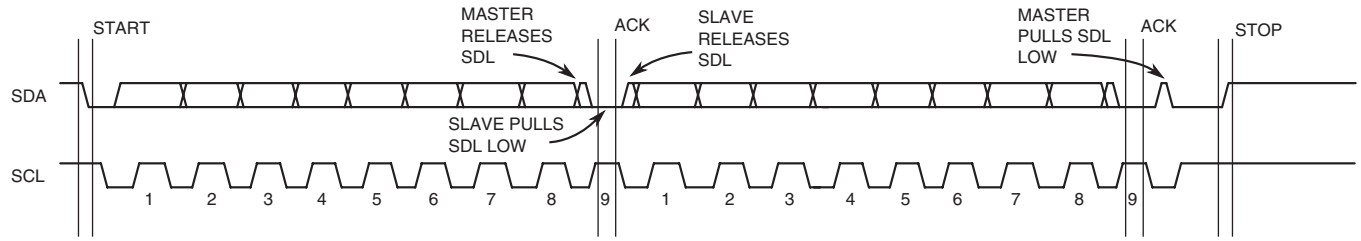


Figure 15.11

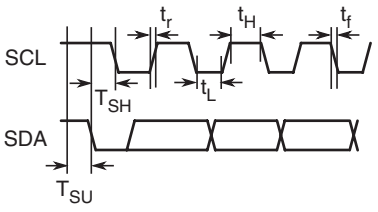
I^2C clock and data waveforms.

whether the master is writing to or reading from the slave. An example of writing the wiper position of a digital potentiometer would be a 3-byte transmission. The first byte addresses the potentiometer chip to be written, the second byte is a command byte setting which channel of the chip is to be written, and the third byte would be the wiper setting for the selected channel.

When the master reads data from a device it first addresses it and sends it control information, the address to read, for example. After the slave acknowledges the control instruction the master generates a start condition followed by the slave address with the read/write bit set to read. Once the slave has acknowledged this byte subsequent clock cycles will clock data out of the device being read; the master acknowledges each byte until the last byte is received, at which time the master does not acknowledge and sets the bus stop condition on the following clock cycle, terminating the transaction.

Figure 15.12

I^2C bus timing.



	VDD 1.8V to 2.5V	VDD 2.5V to 5.5V
T_{SU}	> 4.7µs	> 0.6µs
T_{SH}	> 4µs	> 0.6µs
t_L	> 4.7µs	> 0.6µs
t_H	> 4µs	> 0.6µs
t_r	<300ns	<300ns
t_f	<300ns	<300ns

*minimum bus free time is equal to T_{SU}

The main timing criteria for the I^2C bus are shown in Figure 15.12. These are based on normal and fast mode of the Version 2.0 I^2C specification issued in 1998; the majority of microcontroller peripherals are compatible with this version of the specification.

The I^2C bus implementation is electrically very similar to the PS2 keyboard and mouse interface of the PC. With the same basic software techniques required for I^2C devices a PC keyboard can be interfaced to a microcontroller – this can be a very convenient way of adding a user interface to equipment since PC keyboards are cheap and readily available.

Interrupts

Not all microcontrollers support interrupts; devices like microchips 16C54 and 10F200 for instance do not have interrupts and in many applications interrupts are not required. Interrupts do make many programming tasks more efficient and allow a microcontroller to do a task more slowly than would be required if interrupts were not available.

Interrupts have to be set up in software. Usually a fixed interrupt vector is provided in program memory and the instructions to jump to the address of interrupt service routine can be programmed there. At power-up the processor will start executing at the reset vector, I/O status, and interrupts will be disabled, so one of the first tasks for any program is to configure the processor to use the desired peripherals and interrupts.

It is usual to have a global interrupt enable, and individual enables for the peripherals that can generate interrupts. When an interrupt occurs a flag will be set in the interrupt flag register (Figure 15.13); this allows the interrupt service routine to determine what peripheral caused the interrupt. If interrupts are disabled the flag will still be set so polling the interrupt flag register can be used to determine the status of the peripherals when interrupts are disabled.

In microcontrollers with only one level of interrupts the global interrupts are disabled once an interrupt has occurred, and they are only re-enabled when the return from interrupt instruction is executed, returning execution to the main program. Some processors have multiple interrupt levels, and in this case a higher priority interrupt may interrupt a lower priority service routine in exactly the same way that the lower level interrupt interrupted the main program code.

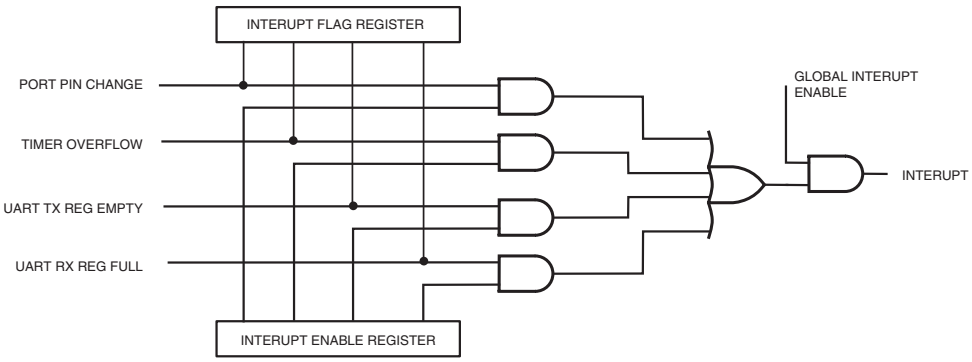


Figure 15.13

Interrupt sources, enable and flag registers.

The first thing that the service routine has to do is to determine what the source of the interrupt was by reading the interrupt flags. It is usually necessary explicitly to clear the flag in software, and not doing so can cause a variety of problems ranging from ambiguous interrupts with more than one flag set to recursive interrupts causing stack overflows and processor crashes. Interrupt should always be treated with care and a particular aspect is handling watchdog timeouts. If an interrupt service routine does not check the status of the watchdog, or often more simply just reset the timer, a timeout can occur during an interrupt and reset the processor. This kind of problem can be hard to diagnose because it can appear random since, by nature, an infrequent interrupt service routine occurring just at the point that the watchdog timeout period occurs will not happen very often. Fortunately, interrupt service routines are usually quite short, in the order of 100 μs to 5 ms, whereas watchdog timeout periods may be several seconds in length, giving the main program plenty of chances to reset the watchdog.

Implementing serial output in software

Asynchronous serial outputs are often useful for providing diagnostic data or for capturing measurement information from a system. It is often the case, however, that the processor either does not have a U(S)ART or

the U(S)ART is being used by other functions of the system like E^2 memory. Implementing serial output in software is easy provided the required clock stability can be achieved using software to shift the data bits successively to the output pin and a delay loop or interrupt to time the bit widths. Listing 15.1 shows an example of an implementation using a PIC16 processor with a 4 MHz clock and one port pin. The byte to be sent is passed to the routine in the variable **stx**; the code generates the start bit followed by 8 data bits and a stop bit. A TTL to RS-232 converter chip like the MAX232A from Maxim would allow connection to a PC serial port. The delay loop generates the bit times and if a different clock frequency is used this is the only part of the program that would require modification.

Listing 15.1 Software RS-232 9600 bits/s transmit routine

```
;--------------------------------------------------------------------------
;RS232 tx sbit is bit counter, stx is data to tx, output is PORTA bit 2
;Transmit data at 9600 bits per second.
;Call routine with byte to transmit in stx, returns with stx undefined
sbit         EQU      0x1A              ;RS232 serial bit tx counter
stx          EQU      0x1B              ;RS232 tx byte
delay_count  EQU      0x1C              ;counter for delay loop

serialtx

                movlw 0x08              ;8 data bits, no parity
                movwf sbit
                bcf   PORTA, 2          ;start bit 100us low
                call  delay
                call  delay

serialbit
                rrf   stx, 1            ;rotate bit to transmit into carry
                btfsc STATUS, C         ;carry flag - test flag
                bsf   PORTA, 2          ;if it is a 1 output high
                btfss STATUS, C
                bcf   PORTA, 2          ;if 0 then output low
                call  delay             ;100us bit period
                call  delay

                decfsz sbit, 1
                goto  serialbit
```

```
                bsf   PORTA, 2        ;stop bit 100us high
                call  delay
                call  delay

                return

;delay of 50us including call/return
;14 times round the loop + the calls = 50us
;using 4MHz XTAL

delay

                movlw 0x0E
                movwf delay_count
delay_loop
                decfsz delay_count, 1
                goto delay_loop
                return
;-------------------------------------------------------------------------
```

Converting binary data to ASCII hex

Transmitting data from a microcontroller to a PC or other host is very useful, but the data is often binary and not easily used without writing some sort of program to deal with it at the PC end of the link. It is often useful to be able to view data with a simple terminal emulator which is a text-based application, therefore converting the data to ASCII hex in the microcontroller before transmitting it can be very useful. An example of conversion of unsigned 8-bit data to ASCII hex in a PIC assembler is shown in Listing 15.2.

Listing 15.2 Converting unsigned 8-bit binary number to ASCII hex

```
bin     EQU     0x1C     ;binary value for convert
tmp     EQU     0x1D     ;temporary register
hh      EQU     0x1E     ;high nibble of HEX number
hl      EQU     0x1F     ;low nibble of HEX number
;Convert one byte unsigned to 2 bytes ascii hex, and transmit
;Call with value in bin, returns with values in hh and hl

bin2hex
```

```
            movlw 0x30          ;start with ascii for zero
            movwf tmp

hnibble                         ;convert high nibble first
            swapf bin, 0        ;swap nibbles and put result in w
            andlw 0x0F          ;mask off bits
            addwf tmp, 1        ;add to ascii 0x30
            movlw 0x3A          ;if number greater than 09
            subwf tmp, 0        ;result will be more than 0x39
            btfss STATUS, C
            goto  hhex
            movlw 0x07          ;if number greater than 09 then
            addwf tmp, 1        ;add offset of 7 to ascii code
hhex                            ;to skip chars between 0 and A
            movfw tmp
            movwf hh

lnibble                         ;now do the low nibble in the same way

            movlw  0x30
            movwf  tmp
            movfw bin
            andlw 0x0F          ;mask off bits
            addwf tmp, 1        ;add to ascii 0x30
            movlw 0x3A          ;if the number is greater than 09
            subwf tmp, 0
            btfss STATUS, C
            goto  lhex
            movlw 0x07
            addwf tmp, 1

lhex
            movfw tmp
            movwf hl
            return
```

The serial transmit routine of Listing 15.1 could be called once the hex character is encoded rather than storing it as hh or hl; this would be done for example by storing it in stx and calling the serial transmit routine:

```
movwf stx
call serialtx
```

Useful websites

Philips I2C-bus specification
www.semiconductors.philips.com/i2c

Microcontrollers

Microchip
www.microchip.com
Atmel
www.atmel.com
Zilog
www.zilog.com

CHAPTER 16

DIGITAL SIGNAL PROCESSING

Introduction

The development of programmable digital logic and microprocessors and the availability of fast, reliable memory has made digital signal processing an attractive option in many applications that have hitherto depended on analogue techniques as well as opening up new possibilities that have not previously been viable.

Functions such as filtering, peak detection and frequency analysis can be performed numerically, often at lower financial cost – as well as in terms of board area, component count and set-up time – than for their analogue equivalents. Digital devices are essentially drift-free and calibration and alignment functions can be implemented in software very easily.

Digital signal processing solutions also offer the potential for in-the-field re-configuration or even self-adapting which can enhance maintenance and reduce the cost of ownership of products.

This chapter does not cover the detail of mathematics required to analyse and design digital filter and signal processing implementations or algorithms but is intended to convey the key points and implementation compromises that apply to digital signal processing using microcontrollers and DSP engines.

Low-pass and high-pass filters

Analogue filters are usually linear time invariant systems and are characterized by the fact that they obey the principle of superposition, that is the output resulting from applying the sum of a set of independent signals to the input of a filter is the same as summing the outputs obtained for all the signals applied independently. This being the case, we can characterize a filter by its step response – the output response to an infinitely fast rise time step at the input.

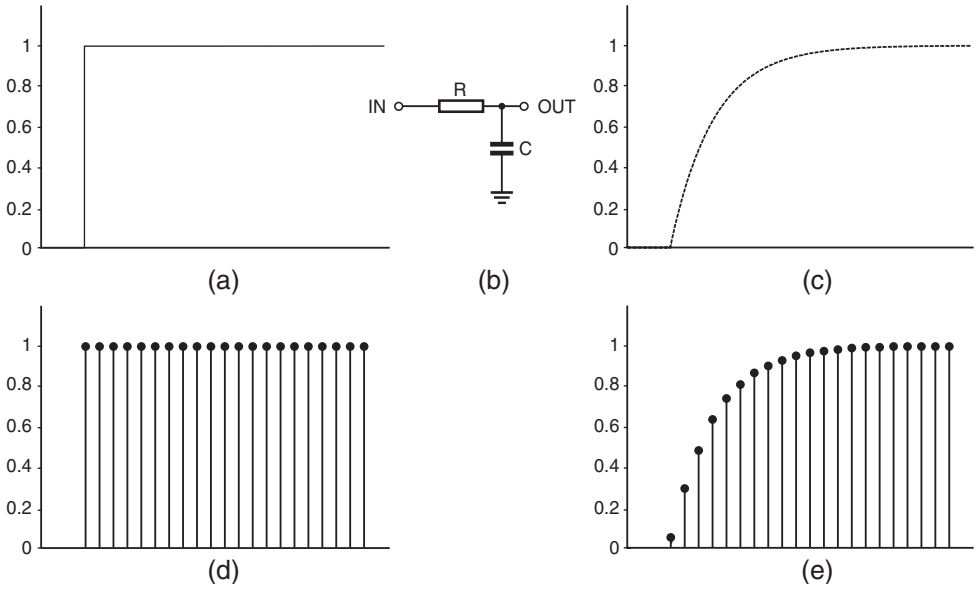


Figure 16.1

Step input signal **(a)**, RC low-pass filter **(b)**, output response **(c)**, sampled input **(d)** and output **(e)** signals.

Figure 16.1 shows the response of a simple RC network, a low-pass filter, to a unit step at its input. We saw earlier in Chapter 2 that the output can be considered to have settled to the final value after approximately 5 time constants of the RC network.

If we sample the input and output signals with an ADC the sampled data might look as shown in Figure 16.1. As one would expect of a low-pass filter, the output signal does not follow the input instantly but grows towards the final value over time. A low-pass filter is in fact averaging the input with respect to time. If we were to average the converted data with a moving average function over 5 samples, for example:

$$y_n = \frac{1}{5}\sum_{n-4}^{n} x_n$$

the output y at time n is equal to the sum for the last 5 input samples $x_n \ldots x_{n-4}$ divided by 5. The input and output data are represented in Table 16.1. The moving average grows linearly in response to a step input and although useful in some circumstances does not match the RC low-pass filter very well.

Table 16.1 Simple low-pass filtering by averaging

Input step	Moving average	Binary weighted series
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
1	0.2	0.5
1	0.4	0.75
1	0.6	0.875
1	0.8	0.9375
1	1	0.96875
1	1	0.984375
1	1	0.984375
1	1	0.984375
1	1	0.984375
1	1	0.984375

A better approximation to the RC low-pass filter characteristic is a binary weighted sequence. As shown in the table this grows more slowly as it

gets nearer the final value, (Figure 16.2) just like a real RC low-pass filter output. The binary weighted series used here is:

$$y_n = \frac{1}{2}x_n + \frac{1}{4}x_{n-1} + \frac{1}{8}x_{n-2} + \frac{1}{16}x_{n-3} + \frac{1}{32}x_{n-4}$$

which is easy to implement in a simple microcontroller because the only operations required are shifts and adds; in fact this can be quite efficiently implemented because each coefficient is the previous one shifted once to the right.

The time constant of the filter depends on the number of terms in the filter function as well as the sampling rate.

High-pass filters can be implemented in the same way. A high-pass filter is effectively a change or edge detector; the output is zero unless the input is changing at a frequency above the cut-off frequency (Figure 16.3).

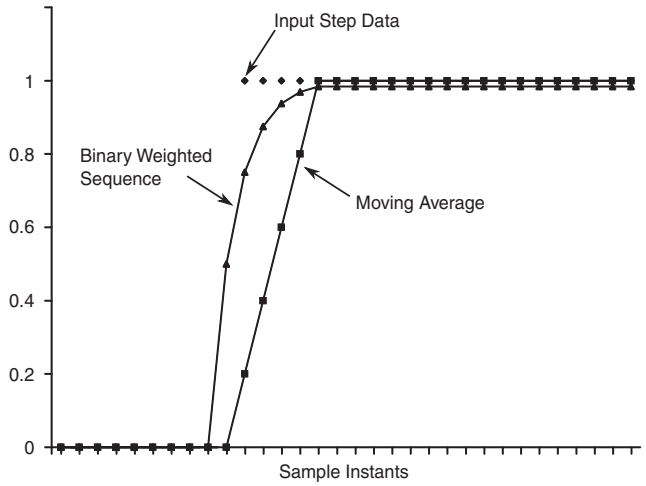


Figure 16.2

Moving average and binary weighted step response.

Considering a series of samples at regular time intervals an edge detection function needs to determine whether the current sample is different from

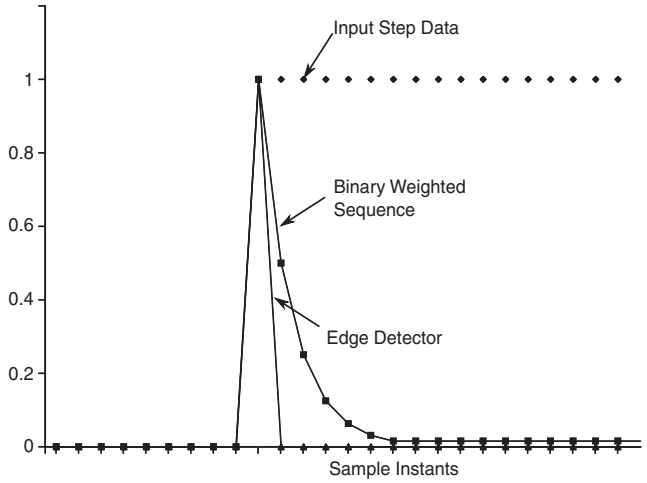


Figure 16.3

Edge detector and high-pass filter response.

the one that preceded it, so simply subtracting the last sample from the current one would detect a change.

$$y_n = x_n - x_{n-1}$$

This basic edge detector can be made more like a real RC high-pass filter by using a binary weighted sequence to control the rate of decay of the step response, in the same way that it controlled the rate at which the low-pass filter converged on a steady value. In this case the filter needs to respond to rapid changes immediately and recover exponentially from the effect.

$$y_n = x_n - \frac{1}{2}x_{n-1} - \frac{1}{4}x_{n-2} - \frac{1}{8}x_{n-3} - \frac{1}{16}x_{n-4} - \frac{1}{32}x_{n-5} - \frac{1}{64}x_{n-6}$$

Again this binary weighting of the sequence is very easy to implement in a microcontroller. Sampling rate and the values of the coefficients determine the cut-off frequency of the filter. Figure 16.3 shows the step response and Table 16.2 the input and output values.

Because this high-pass filter implementation is asymmetric, truncation of the series results in an offset at the output equal to the last term of

Table 16.2 Edge detecting high-pass filter

Input step	Edge detector	Binary weighted series	Modified binary weighted series
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
1	1	1	0
1	0	0.5	0
1	0	0.25	1
1	0	0.125	0.5
1	0	0.0625	0.25
1	0	0.03125	0.125
1	0	0.015625	0.0625
1	0	0.015625	0.03125
1	0	0.015625	0
1	0	0.015625	0
1	0	0.015625	0

the series. A way round this can be to double the value of the last coefficient as shown.

$$y_n = x_n - \frac{1}{2}x_{n-1} - \frac{1}{4}x_{n-2} - \frac{1}{8}x_{n-3} - \frac{1}{16}x_{n-4} - \frac{1}{32}x_{n-5} - \frac{1}{32}x_{n-6}$$

Truncation always introduces an error. The size of this error and the best way of dealing with it depends on the length of the series, the application and cost in processor clock cycles and memory of using a longer series or a different filter implementation. Frequency domain effects of truncation will be covered later in this chapter.

Simple low-pass and high-pass filter functions like the ones described can be very useful for rejecting mains frequency noise in microcontroller applications.

Finite impulse response (FIR) filters

Implementing the filters of the previous section produces a type of structure called a finite impulse response or causal filter. The output of this type of filter is a direct result of the input stimulus and its output will always settle to a steady value if the input does not change; it is unconditionally stable, meaning that it will never oscillate. FIR filters are very straightforward to implement in conventional microcontroller programs.

Figure 16.4 shows a block diagram of the arrangement of a generic FIR filter. The coefficients h_0 to h_n multiply the n input samples; these results are then summed to provide the output.

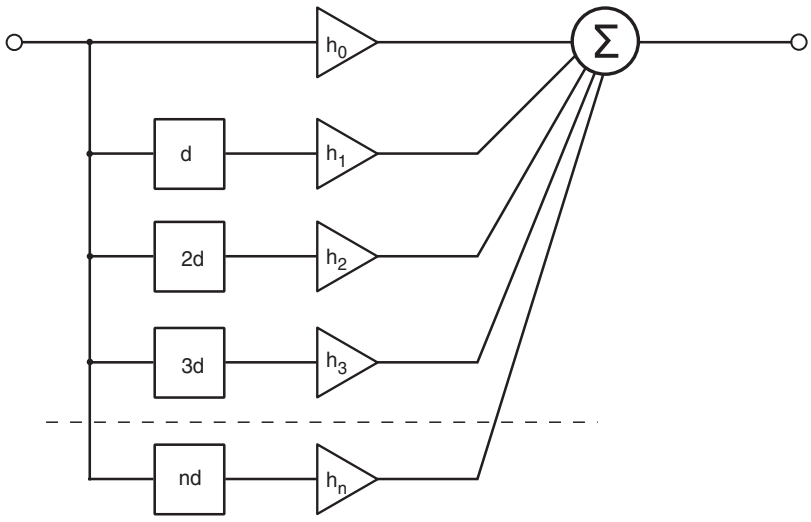


Figure 16.4

Finite impulse response (FIR) filter block diagram.

The mathematics is simple to implement but care over the full scale range and scaling is essential. A typical implementation might use an 8-bit ADC and 8-bit DAC (Figure 16.5) but to ensure that there is no loss of resolution during the process of adding and multiplying, data must be handled internally in wider words such as 24 or 32 bits.

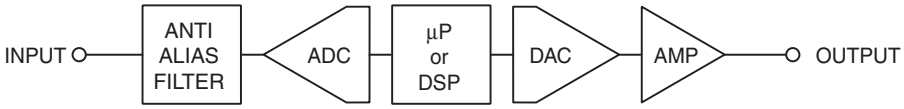


Figure 16.5

Typical DSP system block diagram.

Quantization

We have looked at filters implemented with simple binary weighted series. These are useful for implementation in simple microprocessors without hardware multipliers, or floating point numbers, where shifts and adds are often all that is practical. Accurate filters with steep cut-off and large stop band attenuation often need to be implemented with more resolution than is available in 8-bit integer microcontrollers. Using non-binary power or even fractional coefficients and wide data words, DSP engines are the answer.

Saturated arithmetic

Chips designed specifically for digital signal processing like the Analog Devices Black Fin ADSP-BF or the Microchip dsPIC30F series parts have hardware designed for efficient implementation of signal processing algorithms. Generally this consists of hardware multiplier accumulators, ring buffer memory usually referred to as a barrel shifter and instructions to perform saturated arithmetic. Typically this hardware uses wider words than the general-purpose arithmetic logic unit or processor data bus, 16-bit multiply and 40-bit wide accumulator register for example.

In conventional microprocessor arithmetic, adding two numbers the sum of which is larger than the register results in the carry flag being set and the output data wrapping round. For example, 8-bit unsigned addition of 250 and 7 would result in an answer of 1 and the carry flag being set. This would not be helpful in filter algorithms so saturated arithmetic is used, This means that when a result is larger than the register, the register is set to its maximum value; it is saturated, because any further numbers added to it will not increase its value. This may have the effect of clipping a signal but this is much more likely to be acceptable than sudden full-scale swings in the signal resulting from wrapping.

Conventional microcontrollers that do not support hardware for saturated arithmetic require additional programming effort, and hence clock cycles, to implement it in software.

Truncation

Thus far we have been looking at filter functions in the time domain but usually filters are specified in the frequency domain; while the filter can be completely specified in either the time or frequency domain, it is not always easy to see how the two relate.

Figure 16.6 shows the frequency responses of three different implementations of a low-pass filter. The dashed plot (Figure 16.6a) is of a 4-term filter very similar to the ones considered earlier. The other two traces show filters with the same design of pass band and stop band edges but with more terms (16 and 64) in the series. It can be clearly seen that the more terms in the series there are, the less ripple there is in the pass band and stop band of the filter, the 64-term filter providing an essentially flat response.

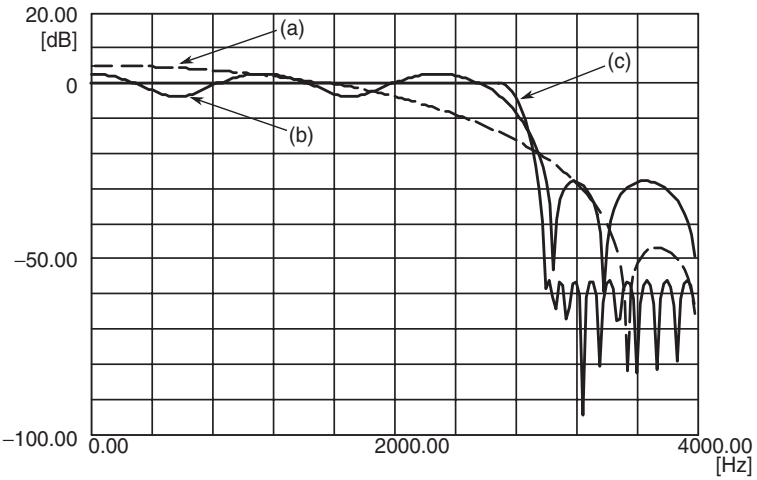


Figure 16.6

Finite impulse response (FIR) low-pass filter frequency response: **(a)** 4 terms, **(b)** 16 terms and **(c)** 64 terms.

The origin of this ripple in the pass band is truncation of the series, usually referred to as the *Gibbs effect*; this is the effect that is observed when a set of odd, harmonically related sine waves are added to create a square wave (Figure 16.7). When fewer harmonics are summed (Figure 16.7a) the square wave has more ripple.

This leads from the fact that the time series that makes the filter is related to the frequency response of the filter by the Fourier transform. The Fourier transform is also a very useful tool implemented in DSP and it can be used to determine the frequency content of a signal.

The series used to implement digital filters is almost always truncated, so minimization of ripple in the pass band cannot be simply based on using longer and longer series. Window functions that modify the coefficient of the series in such a way as to minimize the ripple can be developed. The filters that we have looked at so far have used a rectangular window; that is the coefficients employed are calculated as if the series was not truncated. If we take the truncation into account, a window function based on the truncation can be applied to determine how to modify the coefficients to minimize the ripple in the resulting filters' frequency response.

Bandpass and notch filters

Bandpass and notch, or band stop filters, can also be implemented digitally with similar structures and algorithms to those of high- and low-pass filters. Care to make the frequency response as symmetrical as possible will usually make the implementation as a FIR filter easier. Typically, bandpass filters require longer sequences than low- or high-pass filters; Figure 16.8 shows the frequency response of a 40 coefficient filter.

Infinite impulse response (IIR) filters

Infinite impulse response filters, also known as recursive or acausal filters, have several advantages over FIR filters but they are not unconditionally stable. This type of filter has feedback from previous output states as well

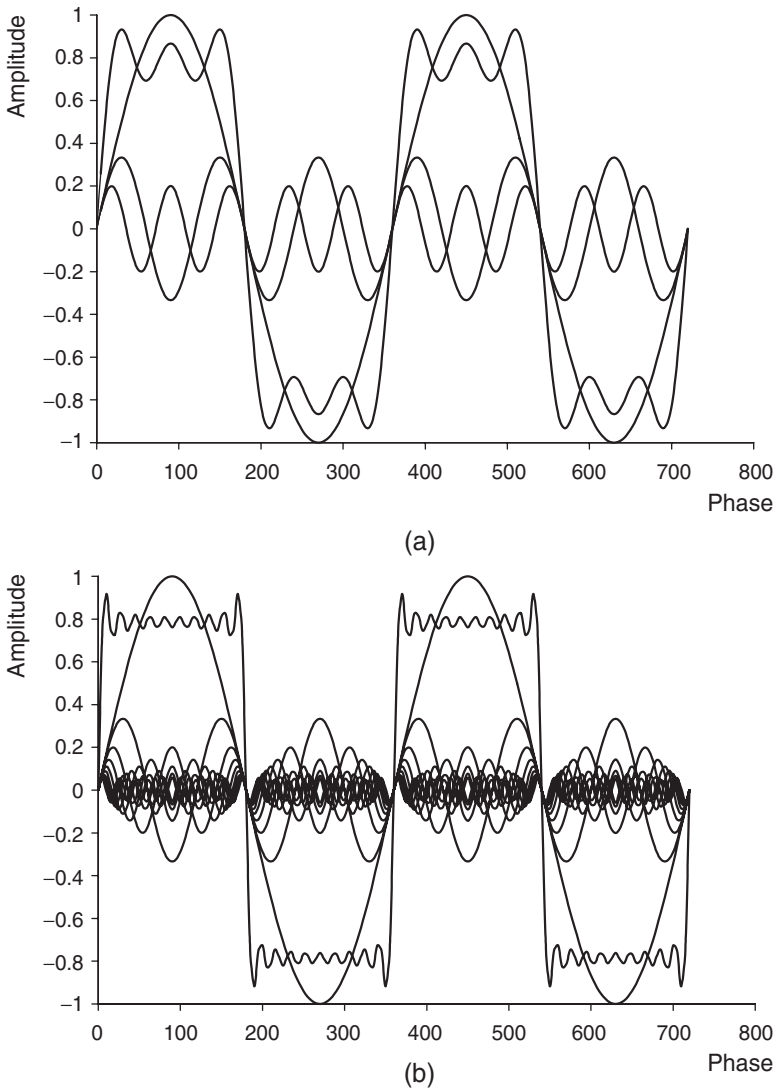


Figure 16.7

The effect of truncating a Fourier series: **(a)** 3 terms and **(b)** 10 terms of a square wave.

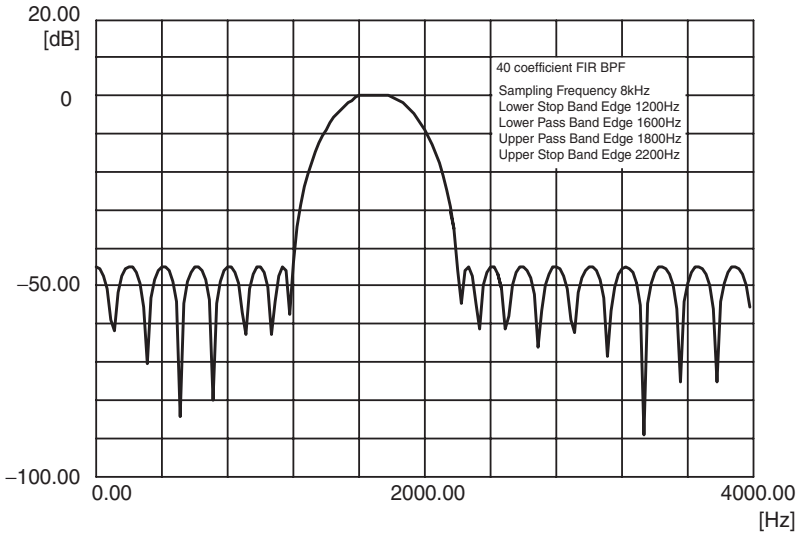


Figure 16.8

Frequency response of a 40 coefficient FIR bandpass filter.

as passed input ones: they are recursive (Figure 16.9). The name infinite impulse response refers to the potential for the recursive structure to oscillate; this is the same as being acausal, meaning that the output may change when the input does not.

Infinite impulse response filters do have advantages in implementation cost, often requiring less memory and shorter sequences for the same performance. In addition, IIR filters can achieve steeper transitions between passband and stop band than can FIR filters. They are, however, more difficult to design since they are not inherently stable and for this reason should be treated with caution.

Other applications

Digital signal processing is not limited to implementation of filters. It has many advantages in other processing options such as using the discrete or fast Fourier transforms to determine the frequency content of a signal, or entire subsystems like quadrature modulators and demodulators.

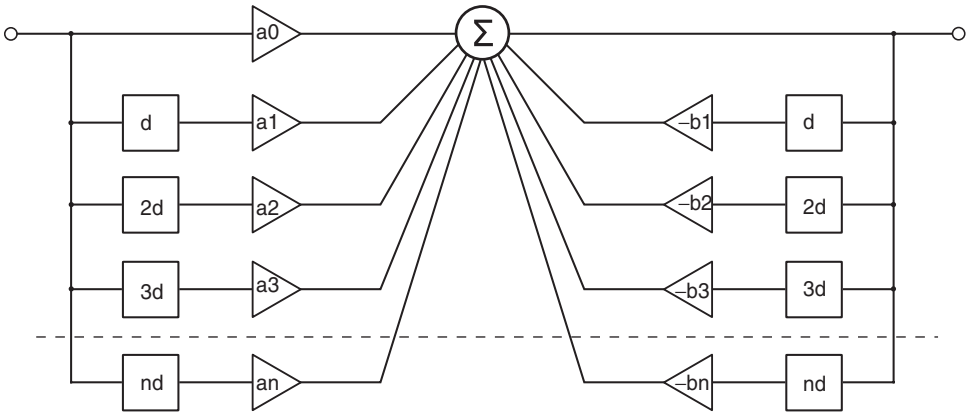


Figure 16.9

Infinite impulse response (IIR) filter block diagram.

Entire systems like modems, digital TV receivers and MP3 players can be almost entirely implemented as digital signal processing systems.

Such design practice is very good for manufacturers, allowing them to put system-specific software in standard hardware, for instance allowing the same physical hardware to serve as a set-top digital TV receiver in countries with different TV standards requiring different DSP implementations of the decoder.

Digital signal processor chips like the Analog Devices Black Fin are becoming common in consumer devices like cameras that require both digital signal processing and conventional microcontroller functions. There is also a trend towards implementing custom hardware DSP functions in FPGAs – this offers the designer several advantages, the main one being execution speed. Hardware implementations without software are often more reliable, or at least more testable.

Design tools

Vendors offering DSP processors, like Microchip, Analog Devices and Texas Instruments, provide tool suites to develop DSP applications on

their hardware. Restricted versions of these tools are made available with demo or application boards from the processor family in question. There are also several third-party tool vendors that include DSP tools in their tool chains.

Further reading

Hamming, R.W. (1998) *Digital Filters*, 3rd edn, Dover Publications.
Microchip, AN852. Implementing FIR and IIR Digital Filters Using PIC18 Microcontrollers.
Microchip, AN616. Digital Signal Processing with the PIC16C74.

CHAPTER 17

COMPUTER AIDS TO CIRCUIT DESIGN

Introduction

Most designs start with a sketch – often on the back of an envelope. In order for them to evolve and reach production some communication is necessary. In some cases giving the sketch to a craftsman could be sufficient; however, this is not generally the case!

The development and evolution of electronic systems can be speeded up by the use of the right tools. It is important to recognize which tools are suitable and to understand their limitations. There is an oft-quoted phrase 'garbage in – garbage out' (GIGO) meaning if you put bad or erroneous data into a computer it will give you back bad answers.

The design process usually consists of four recursive stages: the initial problem; the idea for the solution; implementation, testing and refinement of the idea; and deployment of the solution. If you work entirely on your own and have a good memory, possibly you would never need to document any of the stages; however, as soon as other people or systems become involved some kind of communication is necessary to convey your ideas to others.

Until quite recently a sketch on the back of an envelope would have been converted to a pen-and-ink drawing on paper of some standard size and filed so that when necessary it could be copied and distributed to communicate the idea. The users of the drawing would probably have marked changes and corrections on it and sent it back to the drawing office to be completely redrawn. The main problems were the time taken, redrawing not being very efficient, and keeping everyone updated with the new drawings.

Document control becomes critically important in a production environment where out-of-date drawings or procedures can result in expensive mistakes. It is generally a bad idea for uncontrolled photocopying of production documentation to be allowed, because once an uncontrolled document exists it is difficult for it to be found and updated. Usually a serial number or other marking is stamped on controlled drawings and a register of users is maintained so that the drawings can be identified, replaced with new versions and the old copies recovered for destruction – the marking is usually designed so that photocopying does not reproduce it fully, colour or embossed stamping or dot matrix perforation being used, making it obvious to the user if an uncontrolled document is being handled.

Using computers as drawing and documenting tools improves the editing speed but makes the uncontrolled use of out-of-date information worse. The drawing, editing and updating process is greatly improved because drawings do not need to be completely re-drawn, and documents do not need to be re-typed whenever errors are corrected or modification or additions are required. However, because the documents can be easily communicated and stored there is a greater possibility of someone using out-of-date documents. The potential for sensitive data to be copied and removed from a company is much more significant than it was; a 1 GB USB flash memory device, smaller than your thumb, fits on a key ring but can hold as much data as an estate car full of printed documents – 250 000 sheets of A4!

Schematic capture

Schematic capture packages allow the design to be edited and updated and quickly communicated by e-mail, but they provide many more functions that are not provided by the paper drawing system that they replace. The advantage is conveyed by the name, i.e. schematic *capture* not schematic *drawing*. The difference is that the details of the design are available to the computer as well as being communicated by the lines and symbols of the drawing to the human reader. Schematic capture communicates the topology and component information of the circuit, and the software can use this information for bill-of-material generation for budget and purchasing, for net-list generation for circuit simulation, and for PCB design and checking.

There are many good schematic capture programs on the market including free and open source programs, so rather than concentrate on any particular program we will cover the underlying concepts and how to use the features that all provide.

Schematic capture packages are graphical front ends for component and material databases. When the user selects and places a part on the schematic the data attached to the symbol provides several different types of information: electrical parameters for simulation; physical shape for PCB design and mechanical assembly; manufacturer or stores part number for procurement and costing; and links to electronic copies of data sheets or manufacturers' websites.

In Figure 17.1a a schematic of an oscillator is shown, drawn as an illustration using a computer drawing package as a replacement for pen-and-ink drawing. The version of the schematic shown in Figure 17.1b was exported from a schematic capture package; the extra details shown next to each component are stored in a database. Reports like parts lists can be generated from this database and Table 17.1 shows the parts list for the oscillator.

Libraries

Schematic capture packages rely on libraries of component symbols to enable efficient maintenance of the schematic and ensure consistency of the data stored in the design database. When the user selects a component from the library and places it on the schematic sheet the data relating to the component becomes available to the system. The schematic software can copy the data from the library into the database for the schematic or it can just use a reference to the library entry and reload the data when it needs it. Typically, systems that use references to the library provide the option of embedding the data when the file is saved. This is useful when schematics have to be sent to other users who may not have access to the same libraries.

Library management is critical to the maintenance and accuracy of schematics. Errors in libraries can be hidden from the user, not becoming obvious until PCB design or manufacture, which can cause delays and be very costly. Figure 17.2 shows the typical relationship between schematic

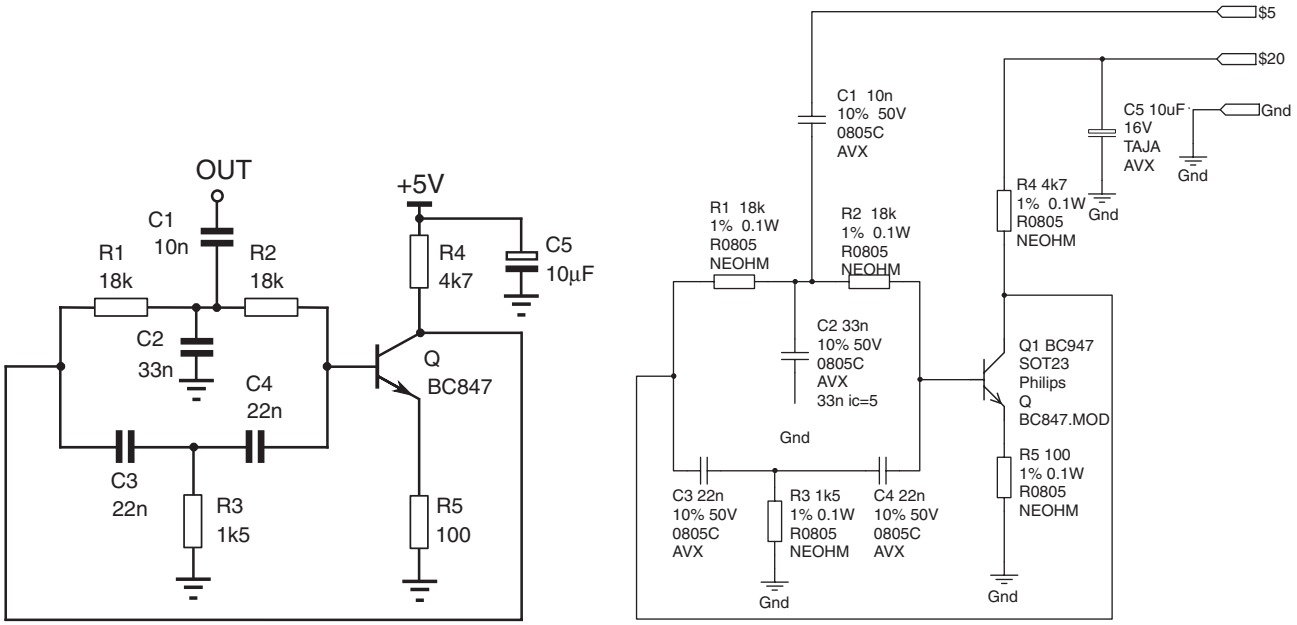


Figure 17.1

(a) Schematic of a twin-T oscillator as drawn for this book; **(b)** Pulsonix 4.0 schematic capture.

Table 17.1 Parts list printout for schematic of Figure 17.1b

Ref.	Value	Qty.	Part	Manufacturer	Description	Footprint
C1	10n	1	C MLCC 50V	AVX	Capacitor, Surface Mount, Multi-Layer Ceramic	SM0805
C2	47n	1	C MLCC 50V	AVX	Capacitor, Surface Mount, Multi-Layer Ceramic	SM0805
C5	10u	1	C Tantalum 16V	AVX	Capacitor, Surface Mount, Solid Tantalum	TAJA
C3	22n	2	C MLCC 50V	AVX	Capacitor, Surface Mount, Multi-Layer Ceramic	SM0805
C4						SM0805
Q1	BC847	1	NPN	Philips	Bipolar Transistor, Surface Mount	SOT23
R1	18k	2	R 0.1W	SMTF NEOHM	Thick Film Surface Mount Resistor	SM0805
R2						SM0805
R3	1k5	1	R 0.1W	SMTF NEOHM	Thick Film Surface Mount Resistor	SM0805
R4	4k7	1	R 0.1W	SMTF NEOHM	Thick Film Surface Mount Resistor	SM0805
R5	100	1	R 0.1W	SMTF NEOHM	Thick Film Surface Mount Resistor	SM0805

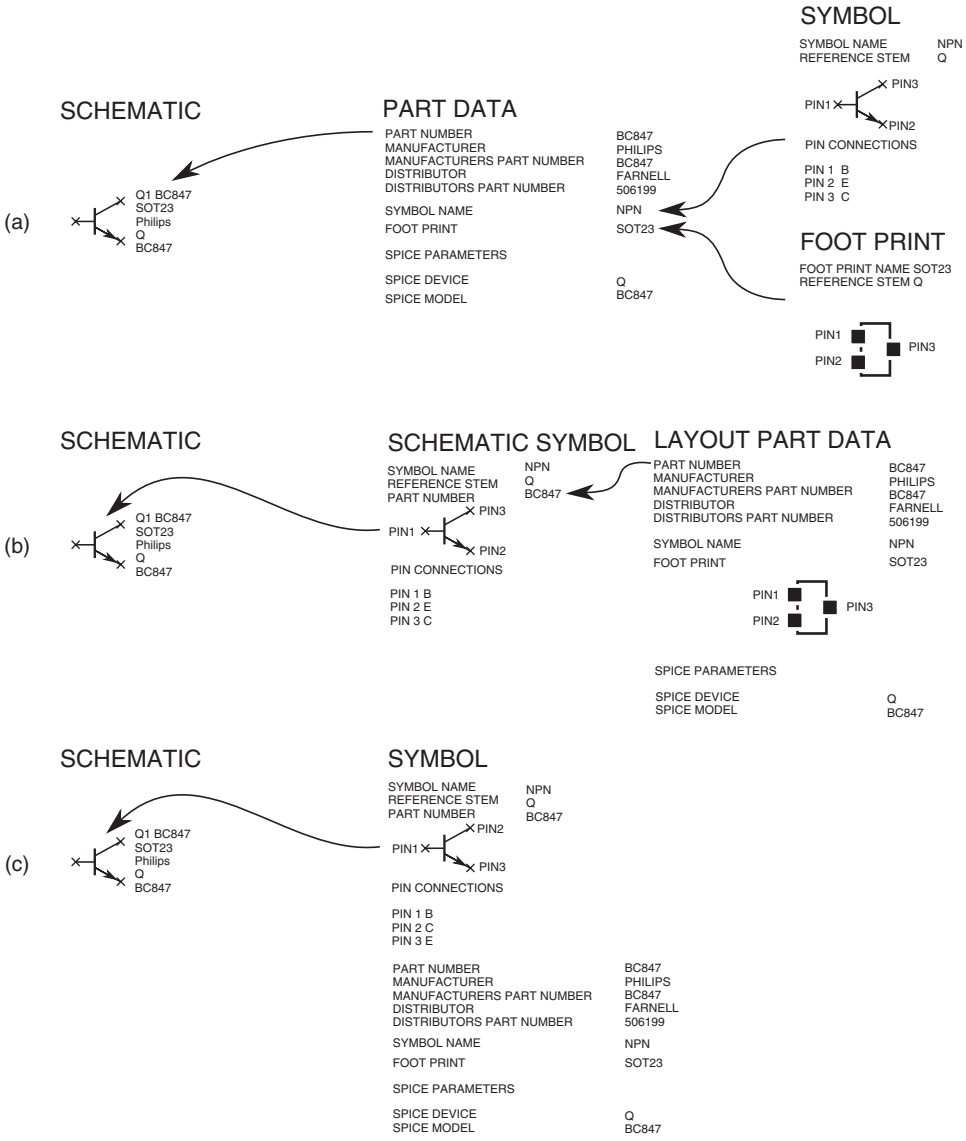


Figure 17.2

Three ways in which symbol and component data may be placed in a schematic: **(a)** part-library driven, **(b)** symbol-library driven and **(c)** symbol library only.

drawing and library data. There are three common arrangements which cover most schematic capture packages. The first, shown in Figure 17.2a, is part-library driven; that is, when a part is placed in the schematic the part database contains the name of the schematic symbol to use. This is common in integrated schematic capture and layout packages. The second, shown in Figure 17.2b, is symbol-library driven; that is, the part database record is referenced by data in the symbol which is placed in the schematic. The part database is separate from the symbol database and may be maintained by an external application like manufacturing resource planning (MRP) software that controls purchasing and kitting for production.

The third method, shown in Figure 17.2c, relies on the symbol library to provide all the component data; this is common in stand-alone schematic capture applications, particularly those used as the front end for simulation programs like SPICE.

In all three of these methods the critical data is the symbol pin assignment. Typically the graphical symbol pins are numbered, in this example 1, 2 and 3, and the database provides mapping from these numbers to the pin function information that is needed by a simulator and the footprint pin information that is required by a layout program.

Most schematic capture packages are shipped with libraries containing many component symbols; these libraries are generally not reliable enough for use in a production environment but they give a good set of templates for building your own libraries. Some CAD vendors provide at additional cost validated libraries with maintenance contracts; these can be useful but careful examination of what exactly is on offer is very important.

Library maintenance is critical to accuracy and efficiency of drawing. When designing library symbols and defining parameters, it is important to be consistent in style and to double-check everything. Ideally, two people should check each other's symbols because it can be very easy to miss your own errors, particularly if there are a large number of new components to add to the library. A swapped power pin on an IC symbol, for example, can cost many hours of PCB fault-finding and thousands of pounds of scrap if you are unlucky.

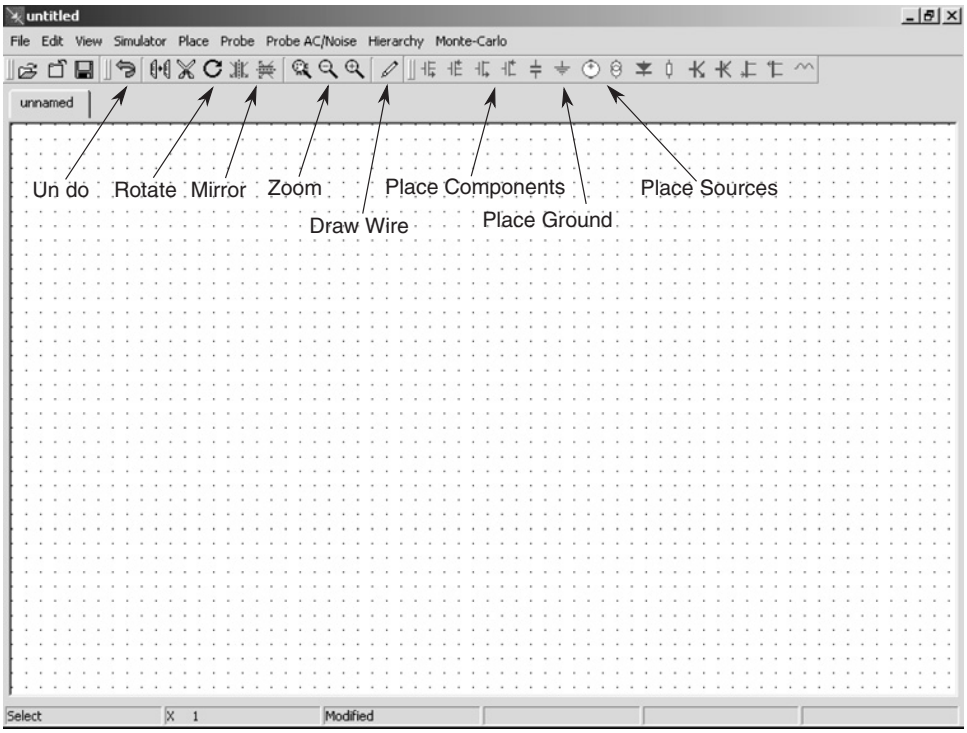
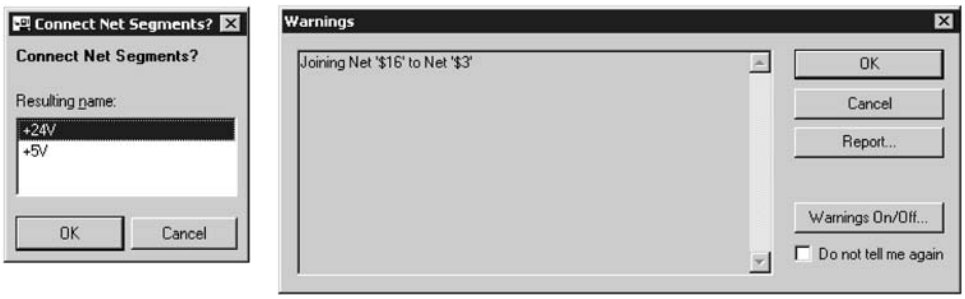


Figure 17.3

Schematic capture programs usually have tools for wiring and component placement on a tool bar like SIMetrix 5.2 shown here.

Connections

Once symbols have been placed on the schematic page using a graphical editor (Figure 17.3), the connections between them that define the circuit need to be added. Electrical connections are drawn between the pins of the components, and all the pins connected to each other are said to *belong to the same net*. When a connection is added between two previously unconnected pins, a net name or number is assigned to the connection, and when a connection is made between a pin and an existing net or between two existing nets most schematic capture packages produce a warning message (Figure 17.4). This is very helpful because it reduces the number of



(a) (b)

Figure 17.4

Connect nets warning message: **(a)** Eagle Layout 4.1.4 Light and **(b)** Pulsonix 4.0.

errors that the operator can make without realizing. One of the most useful applications of this kind of warning is found when moving blocks of circuitry; for example if you want to drag part of a circuit to make room for an extra component, the software will warn if any new connections could be accidentally made by the dragging action. Packages that don't provide this sort of warning leave the user open to all sorts of unintended connections appearing in the drawing.

Where nets are connected, the connection is marked with a dot, as shown in Figure 17.5. One of the biggest causes of confusion when reading schematics is the 4-wire connection because it can look like a crossover particularly if the drawing has been photocopied or, worse still, faxed. For this reason it is recommended always to use two offset 3-wire junctions, which makes the situation clear. The size of the junction dot can vary from CAD package to CAD package; the dot size can be quite small, as can be seen in Figure 17.5b.

Net names

When the pins of components are joined by drawn wires they are connected, and a net name is assigned to the net that connects them. Usually the automatically assigned net name is something like 'N0030' or '$20' depending on the naming scheme used by a particular package. The name can be made

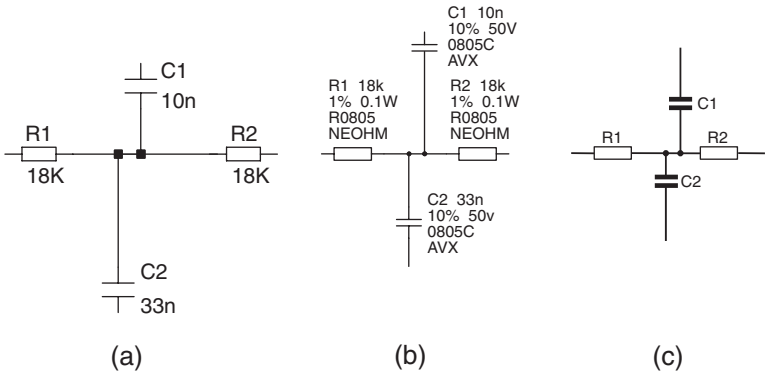


Figure 17.5

To avoid confusion never draw a 4-wire junction; use two 3-wire junctions offset from each other. **(a)** Simetrix 5.2, **(b)** Pulsonix 4.0 and **(c)** as drawn for this book.

meaningful as long as it is unique to the net, so a name like 'logic_supply' or '+5V' may be used. Spaces are usually **not** allowed in net names so the under-score character is used. This is because when the net list is exported the space character is often used as the delimiter to indicate the end of a field like net name or component name. As a general rule it is good practice to avoid the use of spaces in names whether for files, symbols or nets.

Schematic capture packages have facilities for showing the name of a net next to it (Figure 17.6), or label symbols which can be used to show the net name, or force a net name. The ground symbol usually forces the name of the net to which it is attached to be '0'.

Virtual wiring

Using net names in a schematic to link component pins or sections of nets can be very useful when schematics become complex. It is common to use hidden pins and virtual wiring to connect the power and ground pins of logic devices in schematics; this reduces the 'clutter' and allows the user to see the important connections. Connections can be made via symbols, as shown in Figure 17.6a, or by setting the names of the nets directly, as shown in Figure 17.6c. It is very important to make sure that net

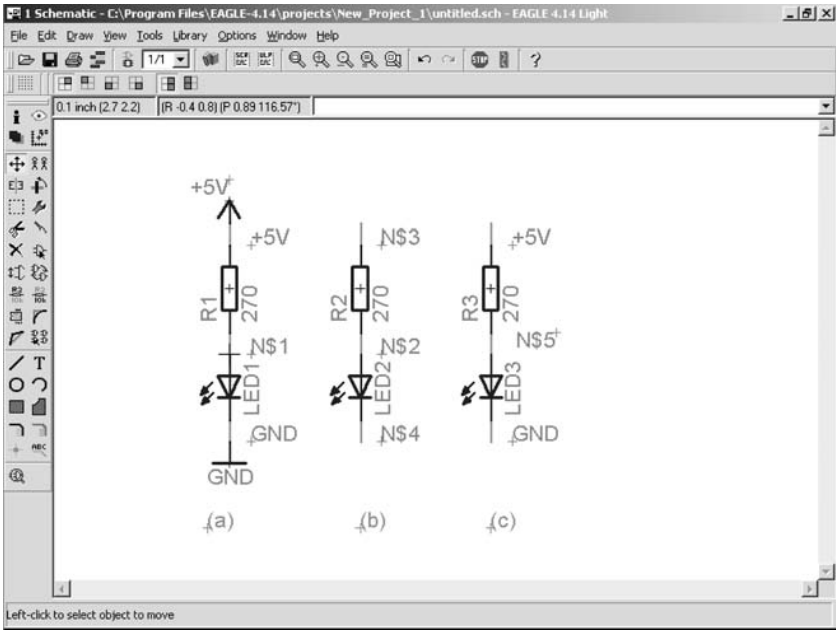


Figure 17.6

Net connections, in Eagle 4.1.4 Light – net names displayed, **(a)** LED and resistor connected between +5 V power and ground symbols, **(b)** LED and resistor connected between nets N$3 and N$4, **(c)** LED and resistor connected between +5 V power and ground nets.

names are displayed if they are to be used to implement virtual connections. Invisible connectivity on a schematic can cause as many problems as errors in library symbols. Most schematic capture packages provide electrical rule-checking tools to check that power and ground pins are not shorted and, if specified in symbols, outputs are not shorted together, etc.

Multi-page schematics can be connected by using the same net names or they can be hierarchical with explicitly defined connectivity between them. Schematic capture packages that allow hierarchical schematics usually use port pins to define the connections between levels in the hierarchy; nets not connected to port pins are not passed through the hierarchy.

Microprocessor and logic systems often have groups of signals that need to travel together between several components. These groups of connections,

known as **buses**, can be conveniently made as virtual connections or drawn as buses – consisting of a single, usually thicker wire – to which multiple labelled links are made (Figure 17.7). One of the advantages of using buses is that all the wires of a bus will share the same attributes, so rules for the PCB layout may be defined and attached to the bus. For instance it is possible to specify that all the wires of a bus are the same electrical length; the PCB layout software can then adjust the routing accordingly. If you look at the tracks on a PC motherboard you will often see meanders in

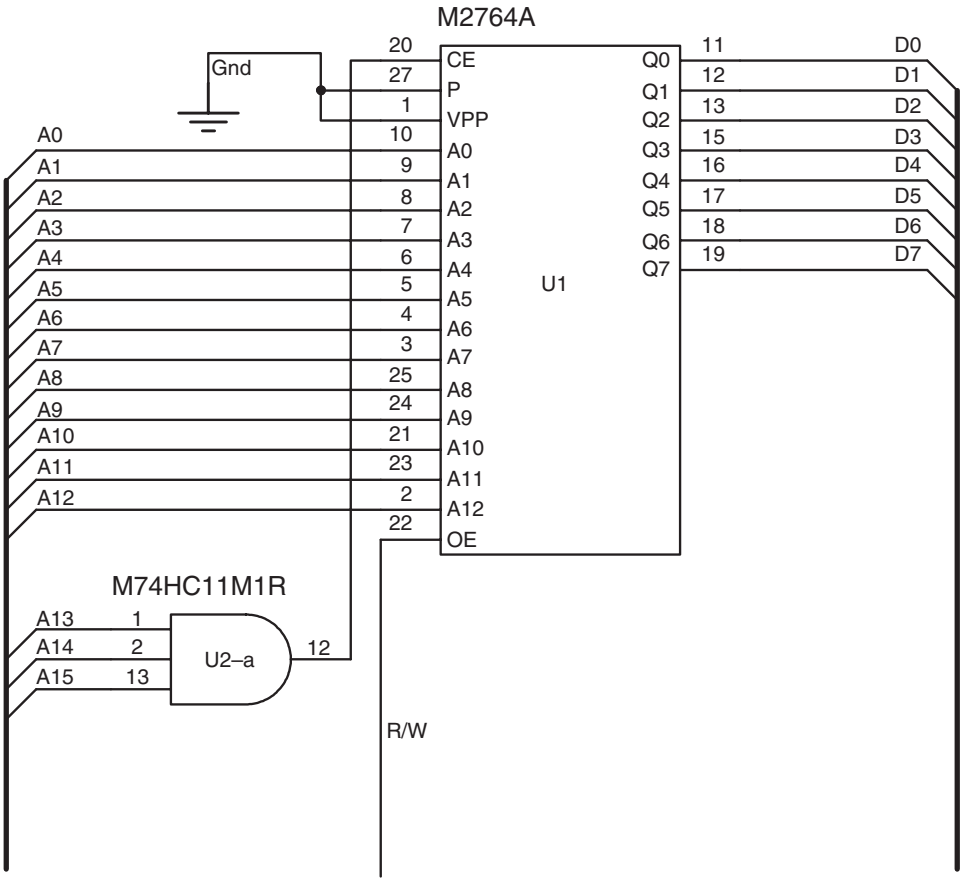


Figure 17.7

Using buses to carry multiple nets in a schematic.

tracks on the inside radius of a bend in a bus – these add electrical length to the track with the purpose of achieving delay equal to that in the long tracks on the outside radius of the bend.

Net lists

The net list is a representation of the electrical connectivity of the design. There are many formats for net lists but the two main ones are **PADS** for PCB design and **SPICE 3** for simulation.

PADS is a professional PCB design tool (http://www.mentor.com/products/PCB/pads/) and its net list format has become a de-facto industry standard supported by many other vendors.

SPICE, Simulation **P**rogram with **I**ntegrated **C**ircuit **E**mphasis, is a circuit simulator and was developed at the University of California, Berkeley (http://bwrc.eecs.berkeley.edu/Classes/IcBook/SPICE/). Its net list files are constructed to convey both connectivity and component data to the simulator; it also is a de-facto standard and many commercial programs are based on it or use its file format.

The format of the net list is clear from the example given in Table 17.2 (see later). It starts with ***PADS-PCB*,** which tells PCB software what format to expect; the first section of the file after the ***PART*** label lists the part footprints, by reference. The ***NET*** section contains the connections between the parts, each net being listed followed by a list of the component pins connected to the net; these are in dot separated form, reference dot pin. ***SIGNAL*** is used to indicate the net name. Some packages define the widths of different classes of track such as power and signal; in this case the label for the net may be ***POWER*** or ***TRACK50OHMS*** rather than ***SIGNAL***. This allows the schematic to define how the track layout is designed, by setting track widths and clearances between them. The last entry in the file is ***END***.

The net name $20 can be seen to connect to pins of R4 and C5 and this can be verified by looking at the schematic in Figure 17.1b, where $20 is shown as the power connection.

Listing 17.1 PADS format net list representing the schematic of Figure 17.1b.

```
*PADS-PCB*
*PART*

C1 0805C
C2 0805C
C3 0805C
C4 0805C
C5 TAJA
Q1 SOT23
R1 R0805
R2 R0805
R3 R0805
R4 R0805
R5 R0805

*NET*
*SIGNAL* $0
C3.2 C4.1 R3.2
*SIGNAL* $3
Q1.1 R2.2 C4.2
*SIGNAL* $5
C1.1
*SIGNAL* $18
C3.1 R1.1 Q1.3 R4.1
*SIGNAL* $20
R4.2 C5.2
*SIGNAL* $31
Q1.2 R5.2
*SIGNAL* $32
C2.2 R1.2 C1.2 R2.1
*SIGNAL* Gnd
C2.1 R3.1 R5.1 C5.1
*END*
```

The SPICE net list format is different from that used for PCB connectivity. Each component is listed with its connections in a way that communicates the part reference, pin connections and part value in a single line. The first line of a SPICE net list is the title of the circuit; this is a comment

and in SPICE net list files comments begin with '*' and are ignored by the simulator. The line **C3 $18 $0 22n** defines the connections of C3 as being to net $18 and net $0, and its value is 22 nF. SPICE uses a special net name '0' for the ground connection, SPICE net lists must always have a net '0' otherwise the simulator will be unable to run, since it refers all voltages to the ground net. Placing a ground symbol creates the '0' net.

Note that there are still some SPICE-2 based simulation tools in use. SPICE-2 net names *must* be numeric; they cannot include characters other than 0–9. The SPICE net list device syntax is generally:

<Device> <net> <net> … <net> <value> [<parameter>]

and the control syntax is similar:

<.control> <options> [<parameters>]

Required values are in angled brackets <> and optional ones between square brackets [ ].

Listing 17.2 SPICE net list generated from the schematic shown in Figure 17.1b.

```
* Twin T oscillator

C3 $18 $0 22n
C2 0 $32 33n ic = 5
R1 $18 $32 18k
Q1 $18 $3 $31 BC847
C1 $5 $32 10n
R2 $32 $3 18k
C4 $0 $3 22n
R3 0 $0 1k5
R5 0 $31 100
R4 $18 $20 4k7
C5 0 $20 10u
```

Note that the circuit in Listing 17.2 would not be able to be usefully simulated as there are no voltage or current sources defined; some source of power is required to bias the transistor. The text "ic = 5" after the value

of C2 forces the initial conditions for transient simulation, the capacitor will start the simulation charged to 5 V; using initial conditions like this can be useful for simulation of oscillators which might otherwise not start.

Printing

In addition to the net list, parts list and other reports, schematic capture software generates printed schematic drawings. In order to track these drawings and maintain them in a useful way we need to add information to be printed on the sheet, and this is usually done as part of a page frame. Figure 17.8 shows a typical page frame; this has fields that are automatically

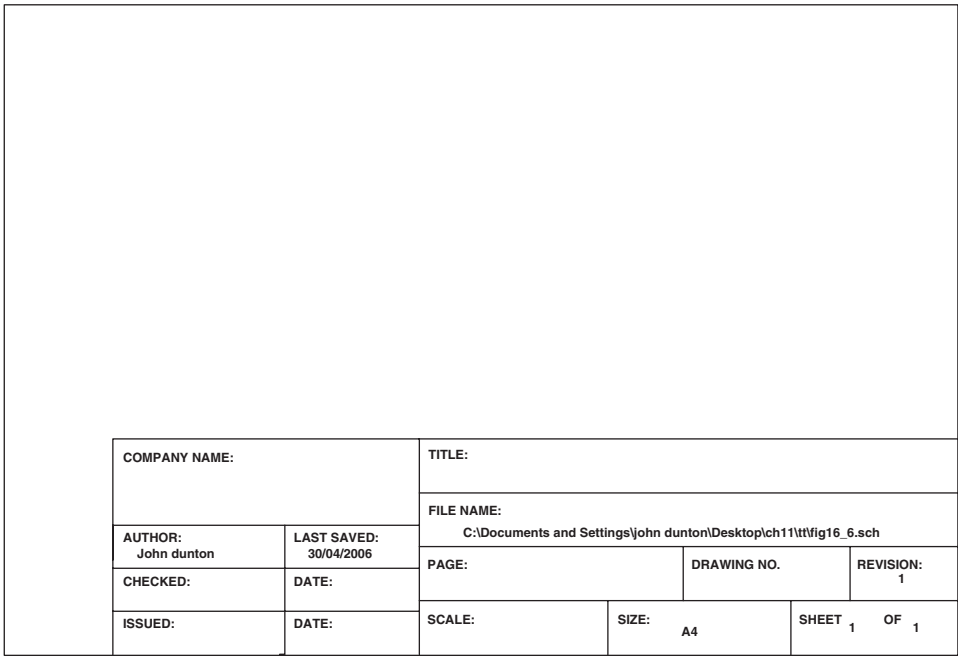


Figure 17.8

Typical schematic page frame.

filled in by the software, such as the file name and last saved date, and spaces for text such as the drawing title and company name.

Using the drawing frame can save time in filing and identifying the sheets of a drawing pack. Complex equipment can often have many hundreds of pages of schematic and other data such as set up and test instructions, and by careful application of the drawing frame and layout of the schematic the use of this data can be made much easier.

Simulation

It is possible to design and build circuits that work first time, but there are drawbacks to working this way unless you only want one of something. Determining the performance of a circuit based on a single sample is not possible once any complexity is reached. Circuit simulation tools allow the designer to assess the effects of temperature, supply voltage variation and component tolerance on the design, identifying critical sensitivities and component tolerances in advance of building the product. While circuit simulation provides the designer with valuable checking and validation tools, they should not be viewed as a replacement for understanding the operation of the components or circuit. The development of a circuit on paper should provide a solid starting point, which can be refined by simulation, answering performance and tolerance questions that would be inefficient and prone to numerical errors if carried out using full exploration on paper.

Most circuit simulation of analogue circuits is done with SPICE or one of the many derivative programs. The original version of SPICE was written in FORTRAN and handled only analogue simulation. X-SPICE was written to add digital simulation, and other extensions have been added to improve the simulation of MOSFETs. Some commercial versions include analogue behavioural modelling such as Laplace functions and polynomial equations.

SPICE is a set of simulators for solving network equations, the type of simulation required being specified in the net list file. The last release of the University of California, Berkeley – which developed SPICE – was version 3F4, which is the basis of many commercial products and is

available also as C source code for UNIX computers under the BSD license. SPICE was developed when punch cards were still in use for programming and data entry of computers so a SPICE list file is still sometimes referred to as a SPICE deck (as in deck of cards). The SPICE 2G6 source code, FORTRAN, is also available; it can be compiled and run on Linux and BSD machines without modification.

Analysis

The three main analysis types are DC, AC and Transient; other options are Noise analysis, related to AC, and Sensitivity and Monte Carlo options which use the other analyses to generate their data sets.

DC analysis solves networks of resistors and semiconductors for DC bias conditions or operating point; capacitors are ignored and inductors are treated as short circuits, if specified temperature effects on resistance and semiconductors are included. The DC analysis is used in two ways, either to specify the operating point of active devices for other analysis types or, swept, to calculate the effects of changing temperature or supply voltage on a circuit's operating point.

AC small signal analysis treats the circuit as linear and independent of signal voltage. It uses the values of capacitors and inductors as well as resistors and the small signal parameters of semiconductor devices as calculated at their operating point established by a DC analysis. AC analysis is useful for establishing the frequency characteristics of filters and amplifiers, but since it is a linear analysis it does not take account of the non-linearity or amplitude limiting of semiconductor devices. The default AC input amplitude is 1 V, and output voltages of 1000 V when the supply voltage of a circuit is 5 V simply means that for a small enough input signal there is a gain of 1000 between input and output.

Transient analysis calculates the response of the circuit to changes in voltage and current in time; it takes account of the non-linearity of the devices and provides accurate amplitude and frequency information. The transient analysis is preceded by a DC analysis to determine the initial operating points of semiconductor devices in the circuit.

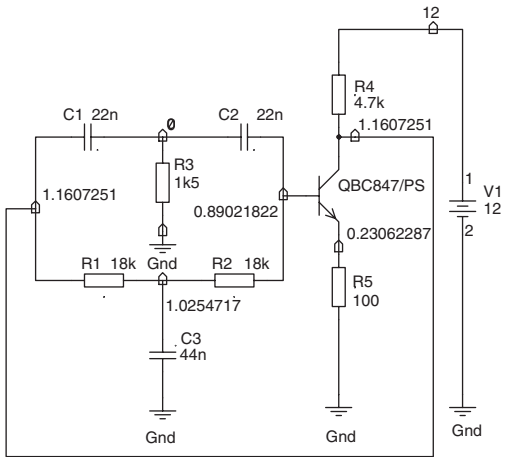
DC Analysis

Using once again the example of the single transistor twin-T oscillator, Figure 17.9 shows the schematic with bias annotations and the SPICE deck that was generated from the schematic. The simulation calculates the DC operating point of the circuit, allowing quiescent power dissipation and small signal gain of the transistor to be estimated; this is an essential first step for other analyses. Operating point analysis can be used to solve quiescent conditions in resistor networks and amplifier bias circuits. In order to investigate the operating point of a circuit at several temperature or supply voltages one could run an operating point analysis at each point; it is more convenient to use the DC sweep analysis which in effect does multiple operating points one after another. DC voltage or current sweeps can be combined with temperature, value or tolerance sweeps.

The power supply, or voltage source is device V1 in the net list; its value is 12, meaning 12 V. Units are optional since the device can only have voltage specified for it. SPICE recognizes multipliers like **k** for 1000 so *R1 $18 $5 18k* is an 18k resistor connected between nets $18 and $5. Note that SPICE is not case sensitive, that is k and K are the same; in SPICE either M or m mean milli (1/1000), and 1 000 000 is represented by meg or MEG (or Meg, mEG, etc). This can be a cause of much confusion to new users accustomed to M and m being different (note that, in a SPICE deck, 1 MΩ resistors are good conductors).

Table 17.2 SPICE decade multipliers (scale factor abbreviations)

T	t	Tera	10^{12}
G	g	Giga	10^{9}
Meg	meg	Mega	10^{6}
K	k	Kilo	10^{3}
m	M	milli	10^{-3}
u	U	micro	10^{-6}
n	N	nano	10^{-9}
p	P	pico	10^{-12}
f	F	femto	10^{-15}




```
* Twin T Oscillator

C1 $18 $0 22n
C2 $0 $3 22n
C3 0 $5 44n
R1 $18 $5 18k
R2 $5 $3 18k
R3 0 $0 1k5
Q1 $18 $3 $21 QBC847/PS
R4 $18 $20 4.7k
V1  $20 0 12
R5 0 $21 100
.op

.temp 27

.model QBC847/PS npn (IS=7.59E-15  VAF=73.4  BF=480  IKF=0.0962  NE=1.2665
+         ISE=3.278E–15  IKR=0.03   ISC=2.00E-13  NC=1.2  NR=1  BR=5  RC=0.25  CJC=6.33E–12
+         FC=0.5  MJC=0.33  VJC=0.65  CJE=1.25E–11  MJE=0.55  VJE=0.65 TF=4.26E-10
+         ITF=0.6 VTF=3 XTF=20 RB=100 IRB=0.0001 RBM=10 RE=0.5 TR=1.50E-07)
* From Philips SC04 "Small signal transistors 1991"
```

Figure 17.9

Circuit with bias annotations and the SPICE deck used for the operating point simulation.

Temperature sweep

Most components have a certain level of temperature dependence. Semiconductor devices are generally more sensitive to temperature than are metals or insulators and it is often necessary to compensate for or reduce the effect of the temperature coefficient of a component. The base-emitter voltage of a transistor or forward drop of a diode is often used as a temperature sensor. Figure 17.10 shows a SPICE temperature sweep of a diode-connected transistor with a constant 10 μA current source. The collector voltage decreases at about 2 mV per °C, and is reasonably independent of transistor type or current, thereby making a convenient temperature sensor.

SPICE provides voltage and current sources that work with operating point and DC analysis; these are referred to as independent sources because they are controlled directly by user-supplied parameters. SPICE also has controlled or dependent sources which have four terminals, two inputs and two outputs, and a scaling parameter; the output is set to be the input multiplied by the scaling parameter. The controlled sources are listed in Table 17.3.

Controlled sources can be used to represent 'perfect' devices; for instance a perfect op-amp might be modelled by a voltage-controlled voltage source with a multiplier of 10 MEG. In fact in the interests of speed, many semiconductor companies produce models that are described as *macro models,* usually based on the Boyle op-amp model (Boyle, IEEE JSSP 1974). The advantage of these models is simulation speed and the fact that they don't give away any secrets in the design of the op-amp IC in question. However, for accurate noise and transient performance, transistor-level models are required and macro models are not generally sufficient for purpose. Some versions of SPICE (HSPICE for example) provide the facility for vendors to encrypt their models so that users can operate them but are unable to investigate their internal circuitry.

SPICE simulators work by solving matrix representations of differential equations; these matrices sometimes cannot be solved, or at least do not

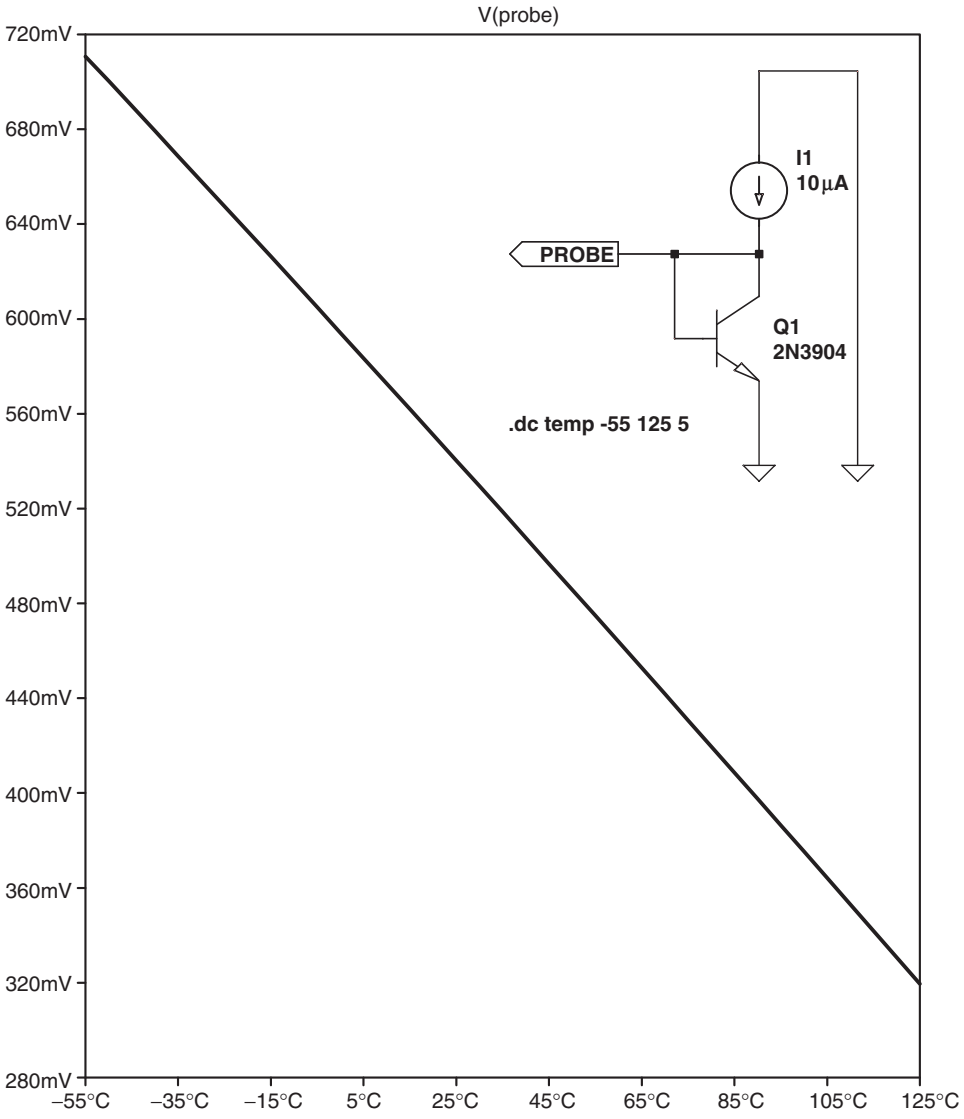


Figure 17.10

Temperature sweep of a diode-connected transistor, with a temperature-independent 10 µA current source, using LTSPICE. The *.DC* control line defines a temperature sweep from —55° to 125° in 5°C steps.

Table 17.3 SPICE-dependent sources

Device letter	Netlist syntax	Function
E	E <in+> <in−> <out+> <out−> <multiplier>	Voltage-controlled voltage source
G	G <in+> <in−> <out+> <out−> <multiplier>	Voltage-controlled current source
H	H <in+> <in−> <out+> <out−> <multiplier>	Current-controlled voltage source
F	F <in+> <in−> <out+> <out−> <multiplier>	Current-controlled current source

have a unique solution and in this case the simulator will output an error message, for example:

> Singular matrix. This may be due to a floating node or a loop of voltage sources and/or inductors. In particular, check node/pin <name>. Check also that there is a ground node. Try setting ".options noopiter"

The commonest reasons for this type of error are: shorting a voltage source with an inductor (remember that inductors are treated as short circuits in DC analysis), putting two voltage sources in parallel, or leaving out the ground symbol on the schematic that sets the '0' net name.

The *noopiter* option tells the simulator not to try to calculate the DC operating point. Generally, however, the error message means that the circuit has errors, so it is worth checking carefully before trying *noopiter*.

AC Analysis

In AC analysis the simulator sweeps the frequency of a constant amplitude source. The active devices are modelled as their small signal transfer functions at the operating point determined by preceding operating point

analysis. The bias voltage changes due to AC signals are ignored and it is quite possible for a transistor with a 5 V supply to have a theoretical 3000 V peak-to-peak waveform at its collector terminal, meaning simply that the gain of the system up to that point would give that ratio of output-to-input signal for inputs that were small enough not to cause non-linear effects in any active devices. AC analysis completely ignores any non-linear effects.

AC analysis is useful for establishing the gain and phase transfer functions of passive circuits and small signal effects in active devices. AC analysis is suitable for modelling filters, impedance matching and transmission line circuits as well as small signal analysis or linear amplifiers.

AC analysis of the twin-T filter used in the oscillator circuit is shown in Figure 17.11. The schematic and SPICE deck show that the network is driven by AC source V1, which has a 1 V output. The *.AC* control line sets a decade sweep, 25 points per decade from 10 Hz to 100 kHz. The graph shows the amplitude and phase at the output node, net $3, relative to the AC source. The probes, dB and Phase, shown on the schematic generate the *.graph* lines in the SPICE deck.

Noise analysis is closely related to AC analysis and models the noise sources in the active devices and resistors (capacitors and inductors are usually assumed to be noise free). An output node must be defined and the noise from the devices in the network is calculated for that node over the range of frequencies.

Transient analysis

Transient analysis of active circuits most closely resembles their physical operation. Using good device models and correctly selected analysis parameters yields accurate small and large signal, non-linear, results. The simulator calculates the network equations at each time step, using the results of preceding steps and solving the differential network equations. Transient analysis can take a lot of processor time. Small time-steps required to give great accuracy mean that many iterations are required; setting a maximum 20 ps step for 10 µs means that at least 500 000 steps will be calculated. The print step sets when the data from the simulation are stored, typically many fewer than the calculation steps.

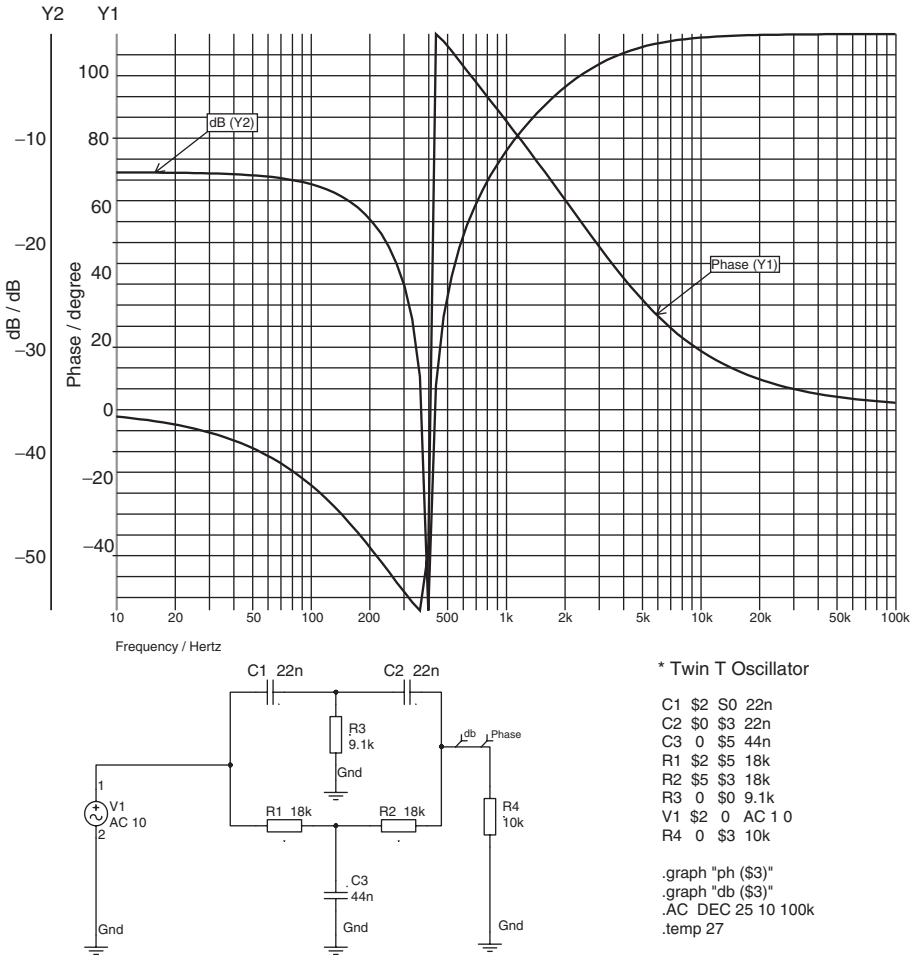
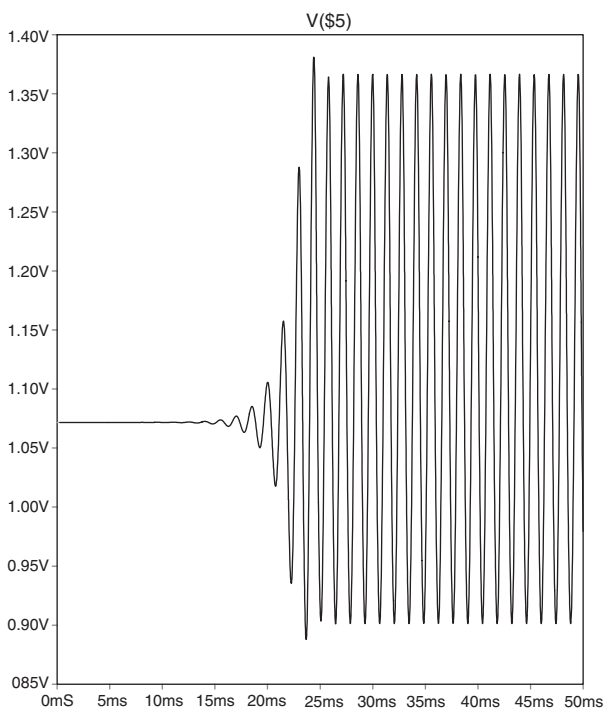
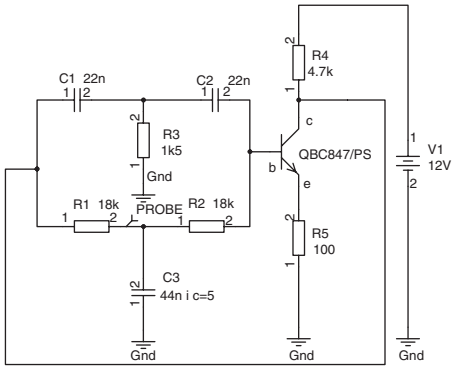


Figure 17.11

AC analysis, gain and phase plot, schematic with Probes and SPICE net list.

The results of the transient analysis of the twin-T oscillator are shown in Figure 17.12; the length of time the oscillator takes to start up is about 20 ms in this case. Oscillator start-up is due to noise being selectively amplified in the circuit until eventually the signal is large enough to drive the amplifier out of the linear region. Once oscillating, the loop gain is precisely 1 but at start-up it may be 3 or 4; so long as the initial loop gain is greater than 1 the oscillator will eventually start.






```
* Twin T Oscillator

C1 $18 $0 22n
C2 $0   $3  22n
C3  0   $5  44n  ic=5
R1  $18  $5  18k
R2  $5 $3 18k
R3  0  $0  1k5
Q1  $18  $3  $21 QBC847/PS
R4  $18 $20 4.7k
v1  $20 0 12
R5  0  $21  100

.op
.TRAN 50u 50m 0 100n

.model QBC847/PS npn (IS=7.59E–15  VAF=73.4  BF=480  IKF=0.0962  NE=1.2665
+       ISE=3.278–15   IKR=0.03  ISC=2.00E–13  NC=1.2  NR=1  BR=5  RC=0.25  CJC=6.33E–12
+       FC=0.5  MJC=0.33  VJC=0.65  CJE=1.25E–11  MJE=0.55  VJE=0.65  TF=4.26E–10
+       ITF=0.6  VTF=3  XTF=20  RB=100  IRB=0.001  RBM=10  RE=0.5  TR=1.50E-07)
*From Philips SC04 "Small signal transistors 1991"
```

Figure 17.12

Transient analysis of the twin-T oscillator circuit, plotted at the junction of C3 and R1 (net $5). The SPICE net list shows the .TRAN analysis line and initial conditions given for C3.

Transient analysis results are time domain samples; if one needs to look at the frequency domain performance, post-processing of the data using Fast Fourier Transform (FFT) provides a frequency spectrum similar to using a spectrum analyser or the FFT software of a digital oscilloscope.

In order for FFT post-processing to give useful results a large number of samples – several thousand – from the transient analysis is required (Figure 17.13); the amplitudes and the widths in frequency of spectral artefacts will depend on the amount of data available for the FFT.

Monte Carlo analysis is used to examine how tolerances stack and so affect the operation of a circuit. Take for example two 10% 1k resistors, and assume that their values are not correlated, and the distribution of values within the tolerance range are the results of a manufacturing process with a normal (Gaussian) distribution table (Table 17.4). The graph in Figure 17.14 shows that out of 20 samples, most of the samples, 17, will give between 2.42 and 2.56 V output at 5 V input but there are 3 outlying values between 2.33 V and 2.62 V.

Monte Carlo analysis should be used with care; unless you know that your component supplies conform to a Gaussian distribution of values within the tolerance range the results can be misleading. To this end SPICE simulators allow tolerance range, matching and component lot to be specified.

The distribution that a particular device exhibits will depend on the manufacturing process, but also on the way that the batch has been processed and, possibly, selected. For instance, for some devices, if you buy 5% parts there will be no parts with better than 1% tolerance in the distribution because the manufacturer has already selected those parts out to sell at a premium as 1% parts – if you buy these 1% parts, their distribution is likely to be uniform rather than Gaussian.

Many recent versions of SPICE have facilities for digital and mixed signal analysis, using transient analysis. Similarly to simulating analogue circuits, the time step needs to be carefully chosen, to ensure accurate results. Figure 17.15 shows a ripple counter simulated over 50 cycles of the clock. The glitches resulting at the decode output can be seen following the valid decode pulses (see race hazards in Chapter 9). For complex digital simulation, event driven simulation is preferred. Conventional transient analysis can be very slow; some versions of SPICE support

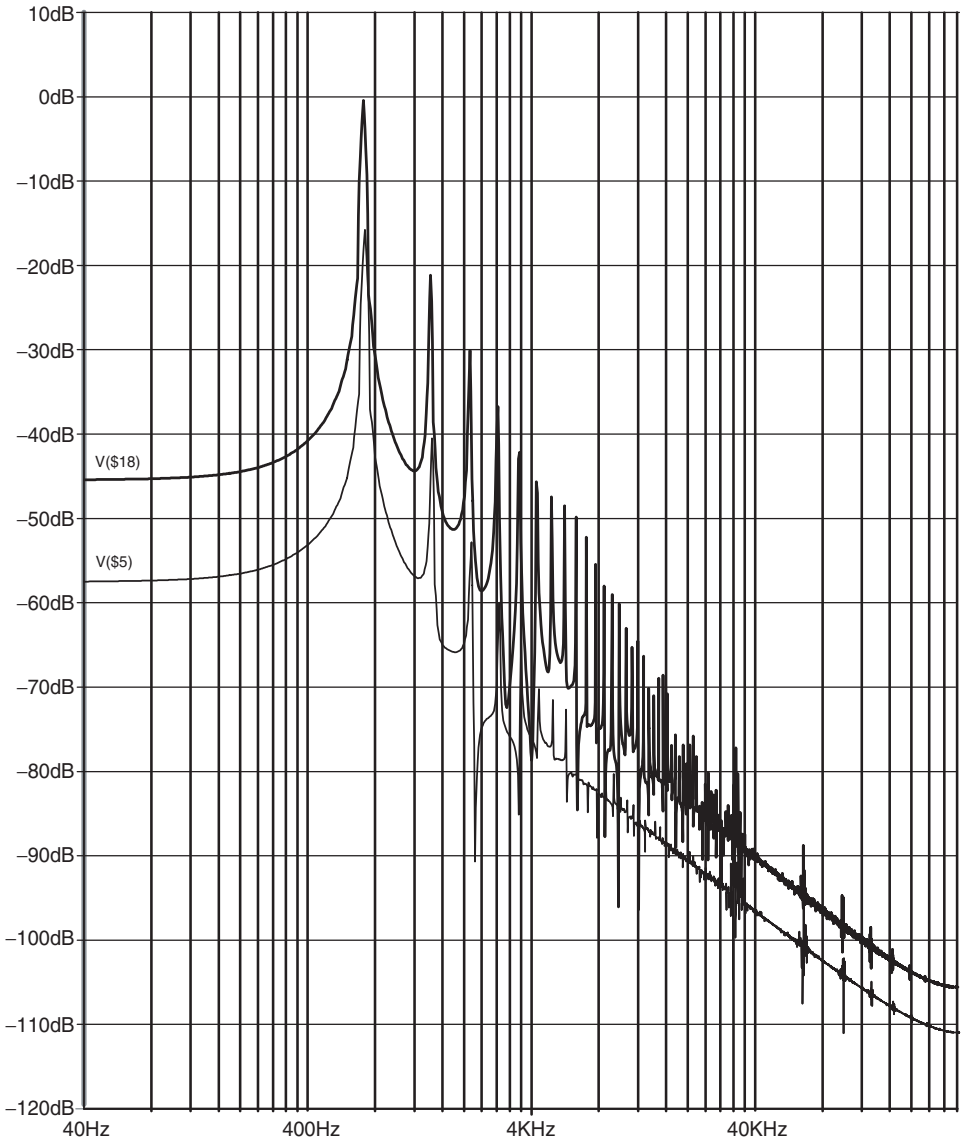


Figure 17.13

Post-processing options include FFT plot of the transient data. Plotting FFT for data between 25 ms and 50 ms at the junction of C3 and R1 and at the collector of Q1 shows better spectral purity; approximately 6 dB is available from the junction of C3 and R1, at the cost of reduced amplitude.

Table 17.4 Types of distribution that Simetrix SPICE supports; names with an L suffix allow for devices from the same Lot with various level of matching

SPICE name	Distribution type
GAUSS	Gaussian (3-sigma)
GAUSSL	Gaussian (3-sigma)
GAUSSE	Gaussian logarithmic (3-sigma)
GAUSSEL Gaussian logarithmic (3-sigma)	
UNIF	Uniform
UNIFL	Uniform
UNIFE	Uniform logarithmic
UNIFEL	Uniform logarithmic
WC	Worst case
WCL	Worst case
WCE	Worst case logarithmic
WCEL	Worst case logarithmic

both types of simulation, the choice being determined by the circuit to be simulated.

Digital, not mixed signal, simulation is often better performed by digital simulations tools of the type that are intended to support the design of VHDL/Verilog code for implementation in FPGAs. Tools provided by FPGA vendors and EDA companies include digital schematic capture and event driven simulators.

PCB Layout

Once a schematic has been entered and checked, possibly simulated, the task of laying out a PCB can begin. Library management is just as important for layout tools as it is for schematic capture. It is all too easy to miss errors in component footprints and end up with expensive scrap PCBs.

Integrated schematic and layout packages offer features like cross-probing, so selecting a net or component in the schematic will highlight the

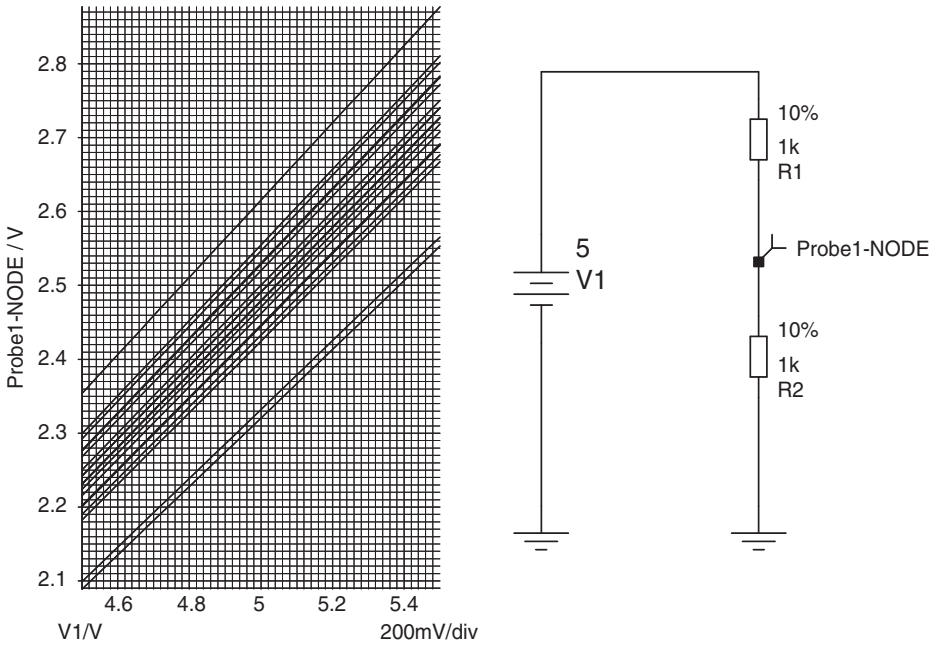


Figure 17.14

Monte Carlo analysis of two 10% resistors, showing how, based on a normal distribution of values within tolerance range, the output voltage might be affected in production.

associated copper in the layout, and vice-versa. Integrated tools provide many advantages but stand-alone tools can be used, for instance using the schematic capture tool from a simulation package to generate a net list for a different layout tool.

When importing a net list, the most important thing to do is to ensure that the schematic library and PCB footprint library match so that the connections of the schematic symbol are correctly mapped to the pins of the component footprint.

The Pulsonix dialogue shown in Figure 17.16 is typical of the tools provided by integrated schematic capture and PCB layout packages for managing the mapping between symbols and components. These types of tool allow effective management of alternative footprints for symbols, for example allowing both surface mount and pin-through hole versions of the same

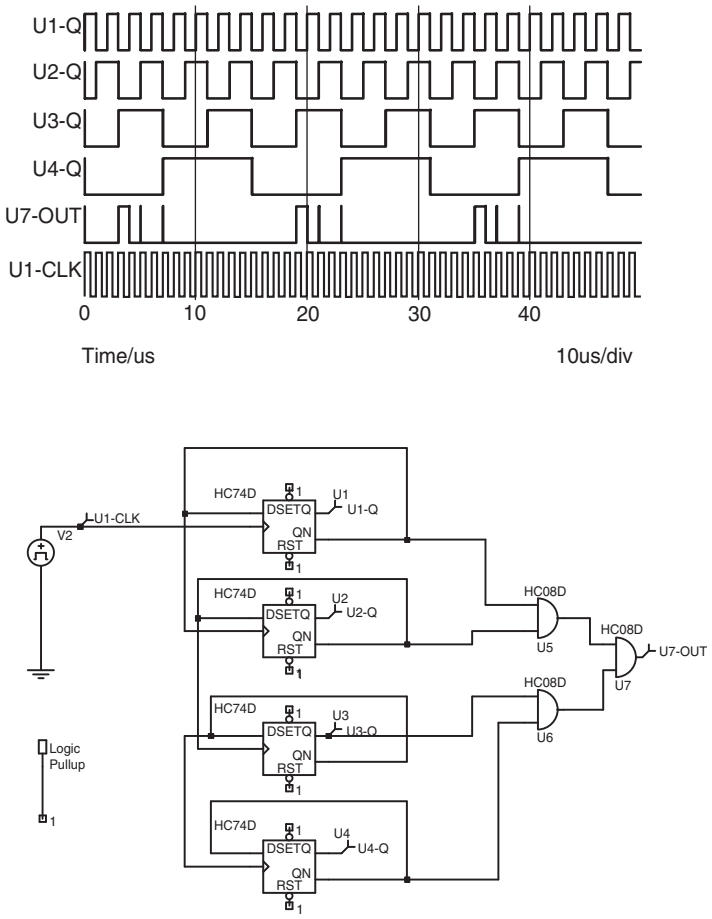


Figure 17.15

Simulation waveforms for a digital counter circuit **(top)** and the circuit **(bottom)**.

component to be selected without changing the symbol. Many surface mount components are available in differently sized versions of the package, for example SOT23 and SOT323 or SO8N and SO8W.

The imported net list data must be supported by libraries with correctly related pin assignments. This is a good reason to use an integrated schematic capture and layout package with good library management tools. Figure 17.17 shows the components and connections between them

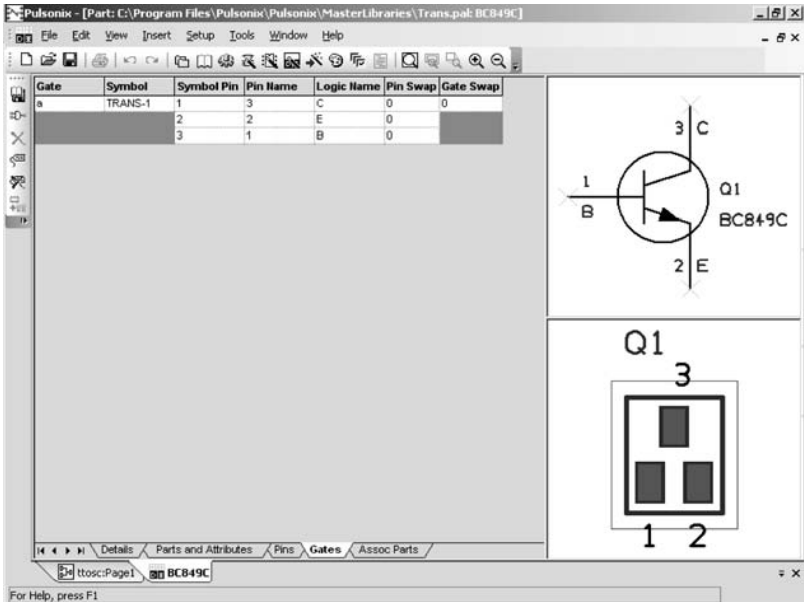


Figure 17.16

Pulsonix library part dialogue allows mapping of schematic symbol connections and component footprint pins.

resulting from importing the net list of Listing 17.1. The lines between the component that indicate the connections are called *rat lines* and the display of components and connections is called a *rat's nest*, implying that the completed layout will bring order to the chaos. Note that the transistor connections of Q1 in Figure 17.17 are wrong, the mapping in the layout package and the schematic capture package not being matched – there are no error checking tools for this sort of error so library management is critical. The pin definitions for SOT23 transistors are unfortunately not standard between different manufacturers; many use the numbers shown in Figure 17.16 but others do not; one common version swaps 1,2,3 for 2,3,1, which is what happened in Figure 17.17.

Laying out the PCB design (Figure 17.18) usually proceeds in three stages: define the board area, which is usually determined by the mechanical characteristics of the housing for the product; place the components within the board area, and attempt to make the connections with tracks. Often this is an iterative process with some movement of components within the board

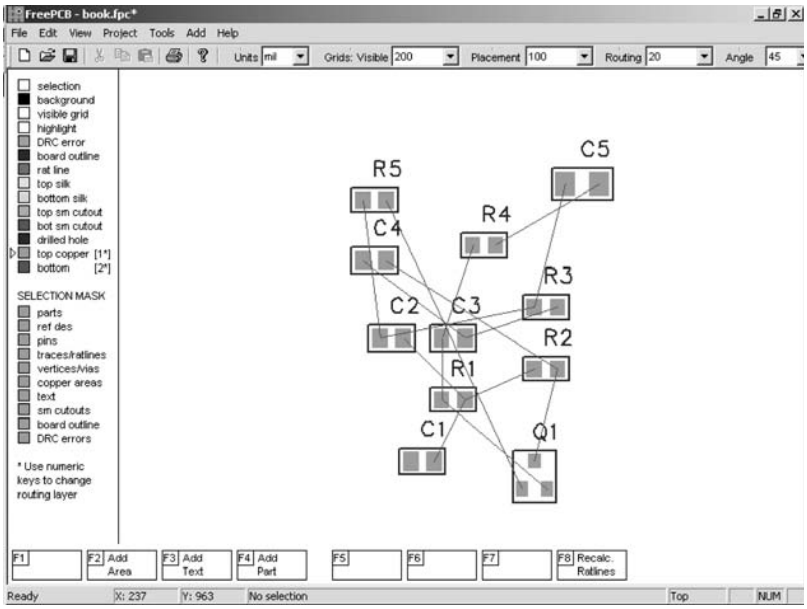


Figure 17.17

Importing the net list into a layout package like FreePCB shown here results in the component footprints connected by a 'rat's nest' of lines, which show the electrical connectivity of the net list.

area being required to achieve a workable layout. The layout of a PCB is usually much easier on a double-sided board, that is one with copper on both sides; single-sided boards are useful but do not offer the high-frequency performance that can be achieved with a *ground plane*. The effect of the ground plane is to reduce the resistance and inductance of connections made to it. It also acts as a shield and helps to control the impedance of the tracks that run across it. Connections between the layers of a PCB are made with plated drill holes called *vias*. Vias are usually added automatically by the design software if the user connects a track from one layer to another.

Circuits that operate with highest frequencies below about 100 kHz can be laid out on single- or double-sided PCBs without significant differences in performance although good grounding and minimizing of current loops are always essential. At frequencies above about 10 MHz ground planes become essential to reduce the radiation of radio frequency signals from tracks.

PC motherboards and graphics cards which use high-frequency clocks and have many interconnecting tracks between processor, memory and I/O, are usually built on PCBs of 12 or more layers with most of the tracks buried; the ground and power plane layers provide screening as the outer layers of the board.

Multi-layer boards present test and design challenges. Test difficulties derive from the need to probe buried interconnects and it is always advisable to get the PCB manufacturer to test the continuity of tracks and the isolation between nets as part of the final inspection testing before shipping the PCBs for assembly.

Design rules

PCB layout packages provide interconnect and separation checking facilities. Like a word processor's spell checker these can be either interactive, highlighting errors as soon as they occur, or run on demand by the user as required. The major rules deal with spacing between copper primitives like pads and tracks, the drill holes and the board edge. Figure 17.18 shows the definitions for some common rules as well as some conventions (Table 17.5).

Most PCB design packages are grid-based, that is components, tracks and holes can be placed only on a set spacing, or grid. This makes interconnection easier if the grids are chosen to make pads line up. It is usual for component placement to be on a relatively coarse grid and track and via placement to be much finer.

Silk-screen printed component placement legends provide useful information for assembly, test and maintenance, showing the orientation of components and their reference designators. Figures 17.19 and 17.20 show some common legend outlines for pin through hole and surface mount components.

When designing a silk-screen symbol or placing component designator information on a PCB it is important to avoid placing the markings over solder areas of pads. This is because the paint that is screen printed to

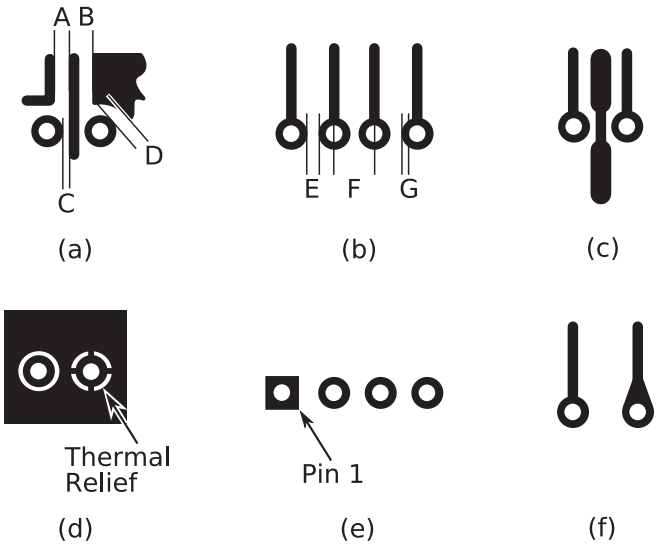


Figure 17.18

Layout rules and conventions: **(a)** track to track, track to pad and plane to pad, **(b)** pad to pad, drill to pad and hole to hole, **(c)** neck track to pass between pads, **(d)** thermal relief pad to plane, **(e)** pin 1 indicated by square pad, and **(f)** tear drop pad to track connection.

Table 17.5 Design rules (see also Figure 17.18)

Rule	Default value		Typical minimum		Fig. 17.18 reference
	thou	mm	thou	mm	
Minimum track width	10	0.254	8	0.203	
Track to track	10	0.254	6	0.152	A
Track to copper area	10	0.254	8	0.203	B
Track to pad	10	0.254	8	0.203	C
Pad to copper area	20	0.508	8	0.203	D
Pad to pad	20	0.508	8	0.203	E
Minimum drill hole	25	0.635	20	0.508	
Hole to hole	50	1.27	50	1.27	F
Minimum pad for hole	8	0.203	5	0.127	G
Board edge to copper	20	0.508	10	0.254	
Board edge to drill hole	50	1.27	25	0.635	

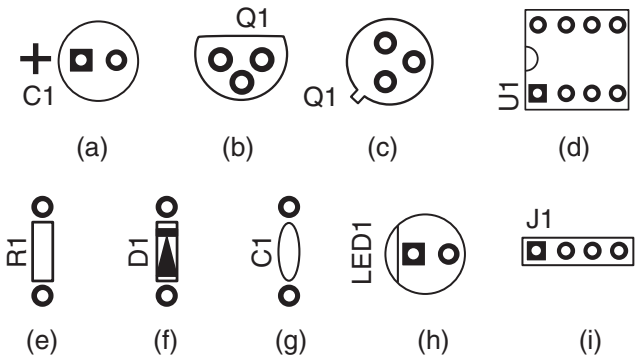


Figure 17.19

Typical silk-screen legends for common through hole components: **(a)** electrolytic capacitor, **(b)** TO92 transistor, **(c)** TO18 transistor, **(d)** DIP8 IC, **(e)** axial resistor, **(f)** diode, **(g)** disc capacitor, **(h)** 5 mm LED, and **(i)** 4-pin header.

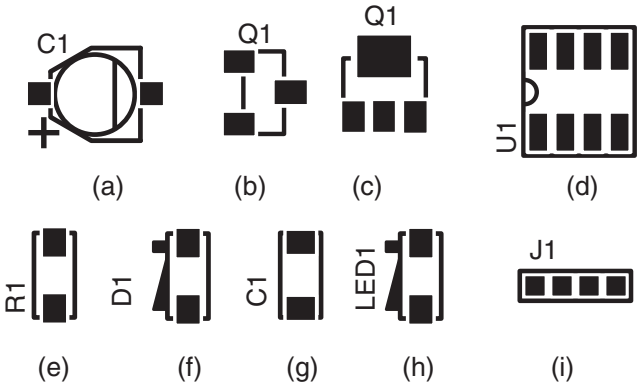


Figure 17.20

Typical silk-screen legends for common surface mount components: **(a)** electrolytic capacitor, **(b)** SOT23 transistor, **(c)** SOT89 transistor, **(d)** SO8 IC, **(e)** 0805/0603/0402 resistor, **(f)** SOD323 diode, **(g)** 0805/0603/0402 capacitor, **(h)** 0805 LED, and **(i)** 4-pad header.

make the legend can prevent the solder from flowing on the pad and cause poor joints or even prevent the joint being made altogether. Some PCB design software will generate warnings if legend and pad or drill data coincide.

Once a PCB layout has been completed (Figure 17.21) it is good practice to run design rule checks to ensure that nothing has been omitted or forgotten. It is very easy when editing a large layout to turn off layers for a better view of some part of the layout or turn off the interactive warnings while trying to route some difficult tracks – the warnings can clutter the screen.

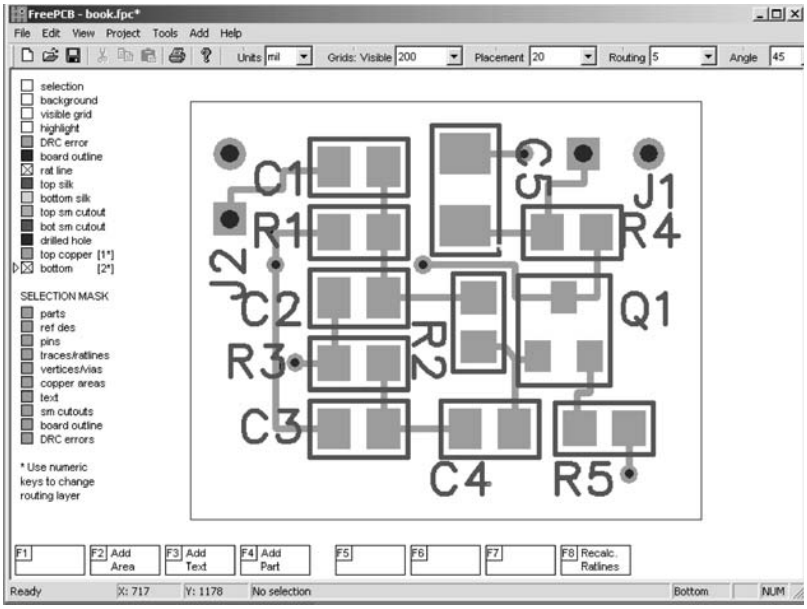


Figure 17.21

PCB layout completed in FreePCB – PCB boarder, power and signal connectors added.

Design rule checking usually includes electrical connectivity, spacing and minimum sizes of pads as well as manufacturing and production orientated checks like areas of isolated copper, test points under components and test points not on a grid. Many flying probe test machines have set minimum

steps sizes. A typical Design Rule Check dialogue is shown in Figure 17.22. As well as the checks it offers the option of saving a report file and it is good practice to save such reports and include them in the project information along with the manufacturing files.

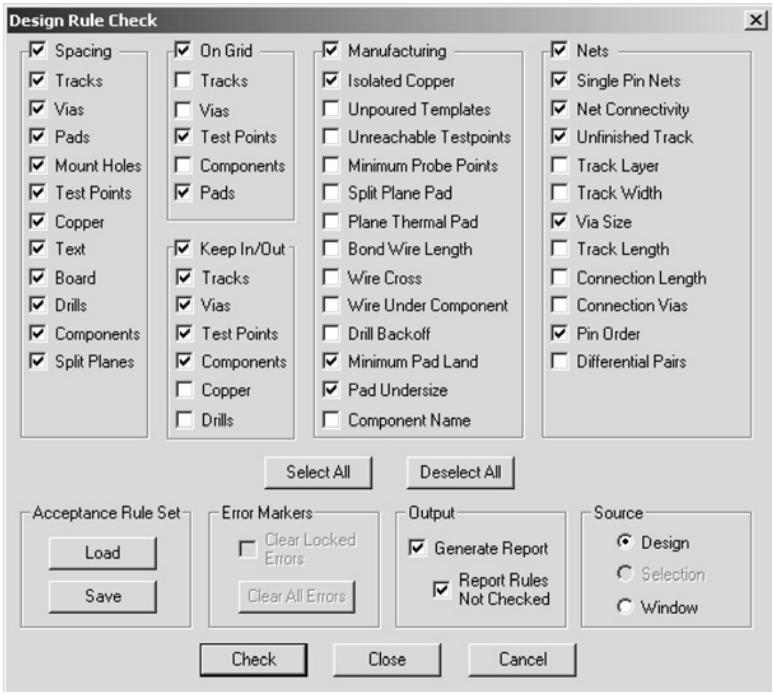


Figure 17.22

Design Rule Check Options menu (Pulsonix 4) allows selection of the rules to be checked; values to check can be entered in other menus or loaded from a file.

After passing all the design rule checks the layout is ready for manufacturing. In order to do this the layout data must be exported from the PCB design package in a format that can be understood by PCB manufacturer's equipment. The industry standard formats defining the copper areas of PCB layout are EIA standard RS274D and the extended RS274X; both are also referred to as Gerber data files, because they are descended from the data format used by Gerber Scientific Industries photo plotters. Drill data is usually exported in Excellon Numerical Control format, which is generally

similar to the RS274 data files. All these formats are ASCII text files and can be edited by hand if necessary (see Appendix D).

It is always advisable to check the manufacturing files before sending them out for production. Simply printing the PCB layout on a laser printer is not sufficient because the computer-aided manufacturing (CAM) file output process generates data that may not be identical to that available from a printer – for example, thermal relief of connections to ground planes or generation of tear drops where tracks connect to round pads (Figure 17.18d and f). Thermal relief of pads connected to planes or very wide tracks is required to stop the large area of copper acting as a heatsink and making soldering to the pad difficult.

Gerber and NC drill file checking

There are a number of Gerber and drill file viewers, both commercial and under the GPL, that may be downloaded. Many of the commercial packages allow free operation with a restricted database size and, possibly, editing features disabled. The advantage of these viewers is that you can load the CAM files generated by the PCB design software layer by layer, including the automatically generated data like solder mask files.

Figure 17.23 shows the CAM data generated for the twin-T oscillator PCB including solder mask and drill data, plotted using the ViewMate 9 program from Pentalogix. The solder mask data is plotted as negative data; that is, the holes in the mask are plotted rather than the masked areas; solder mask is the green protective layer on commercial PCBs that prevents solder splashing from one track to another. Solder masks are essential for machine assembled re-flowed or wave soldered PCBs.

Desktop routing machines

Rapid prototyping of PCBs can be achieved at low cost using a laser printer, a pack of overhead transparency films, an ultraviolet light box, and developer and etch tanks containing sodium hydroxide and ferric chloride solutions respectively. This is a messy process and often health and safety procedures mean that there is nowhere suitable for the developing and etching to be done since it requires fume extraction and running water.

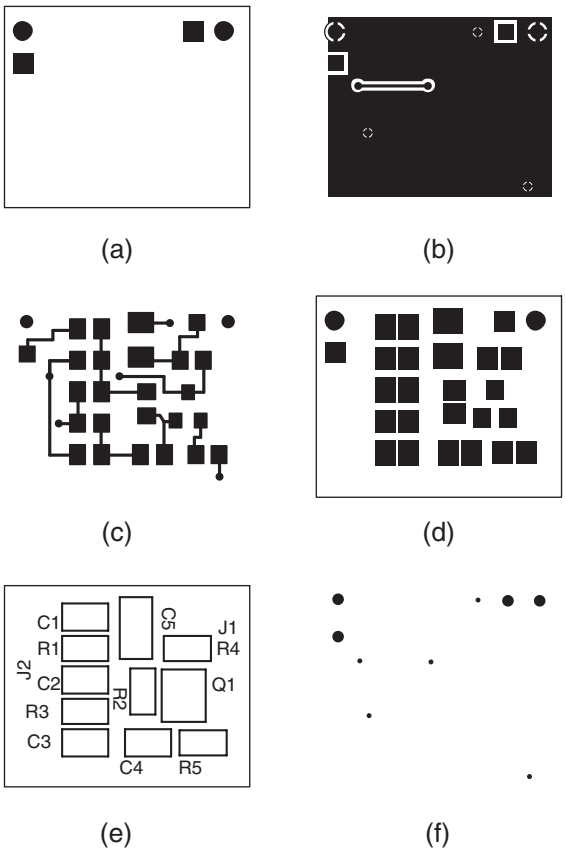


Figure 17.23

The PCB manufacturing files should be checked with a Gerber/drill file viewer before being sent to the PCB manufacturer, or plotted on film, for production. Check plots can be printed and reviewed. **(a)** Bottom solder mask layer (mirrored), **(b)** bottom copper layer (mirrored), **(c)** top copper, **(d)** top solder mask, **(e)** top silk-screen legend, and **(f)** drill holes.

Desktop routing machines that can be programmed with the Gerber and NC drill data files exported from PCB software are available from a number of vendors. These machines can be very effective for one-off single-sided or double-sided PCB prototypes, saving development time and tooling charges. They use very high speed milling cutters, up to 20000 rpm, to remove unwanted copper from the PCB. The milling cutters, if properly set up, last for about 20 linear metres of cutting, which is typically a single euro card circuit size (160 mm × 100 mm) double-sided PCB.

At the time of writing a desktop router with an A4 size milling bed costs about £5000. The cost of a milling cutter is around £8 to £20 depending on width; compare this with the cost of a one-off double-sided PCB from a PCB manufacturer on a 5 day turn-round of about £60 or 24 hour turn-round of £300.

Useful websites

Schematic capture

Pulsonix	**wwwb.pulsonix.com**
TinyCAD	**ww.tinycad.com**
LTSPICE	**www.linear.com**
Eagle	**www.cadsoft.de**

SPICE

Simetrix	**www.catenna.uk.com**
WinSPICE	**www.winspice.com**
PSPICE	**www.orcad.com**
LTSPICE	**www.linear.com**

Layout

Pulsonix	**www.pulsonix.com**
Easy PC	**www.numberonesystems.co.uk**
Eagle	**www.cadsoft.de**
Free PCB	**www.freePCB.com**

Gerber viewers

Gerbv	**www.geda.org**
ViewMATE	**www.pentalogix.com**
GCPreview	**www.graphicraft.com**

Rapid prototyping routing machines

LPKF	**www.LPKF.de**

This page intentionally left blank

CHAPTER 18

CONNECTORS, PROTOTYPING AND MECHANICAL CONSTRUCTION

Hardware

The hardware of electronics, meaning the chassis, covers, cabinets, switches and other external features is, for the manufacturer of consumer electronics, almost as important as the circuitry within, because the external appearance is all that the casual user can go by to judge the system. For military contract work, adhering to specifications is what counts, but on any score hardware cannot be neglected, though it very often is. The task is made much easier when there is a company policy of using some particular form of hardware such as standardized board and cabinet sizes. The most difficult decision is on how to package some piece of equipment that is, for the moment, a one-off product, particularly if there is any chance that other items will follow. A lash-up that works in the laboratory may not work in the same way when packaged as a consumer product, so the final packaging should be considered once a prototype is working.

The prospective user of a piece of electronics equipment first makes contact with the design when he/she tries to connect it to other equipment and to whatever power supply is used. Mains-operated equipment for domestic or office use will usually have a connected and well-tethered mains cable of the correct rating, preferably with a correctly fused three-pin plug, if it is intended for the UK market. The option is to use a BS/IEC-approved three-pin plug on the chassis, with a lead that has a matching socket at one end and a suitable mains plug at the other. Mains cables can be obtained as a standard item with the BS/IEC fitting at one end and various UK or other plugs at the other. If additional equipment has to be driven, BS/IEC sockets can be used to allow power to be taken from the main unit rather than

from a set of additional mains cables. Only low-power and double-insulated equipment should ever use the two-pin form of 'cassette-recorder' leads, and only if this is enforced by considerations of space or expense. Such connections should never be used for professional equipment.

Industrial equipment should use one of the approved industrial connectors, almost certainly to the BS4343 specification for equipment to be used in the UK. For domestic electronics, mains connections are about as standardized as anything in the use of terminals ever attains, but there is a bewildering variety of styles available for low-voltage connectors and for signal connectors. The primary aim in choosing connections should be to achieve some coordination with the equipment that will be used along with yours, and this means that you need to have, from the start of the design stage, a good idea of what amounts to *de facto* standards. If no such standards exist, try to resist making new ones because there are far too many already.

Take, for example, the low-voltage supply connector which is used for portable stereo players and for some calculators. Even in this very restricted range of applications there are two sizes of plug/socket, the 2.1 mm and the 2.5 mm, with some manufacturers using the centre pin as earth and others using the centre pin as the supply voltage pin (usually 6 V), so there can be four variations on this design alone. Power supplies for such equipment usually cope by using a lead that is terminated in a four-way jack plug and which at the supply end is fitted with a reversible plug, often with no polarity markings. Getting the correct polarity is very often more a matter of luck than good instructions, so if this type of connector is used there should be a clear indication of polarity and also some protection against use of the wrong polarity (a diode in series) if the voltage drop can be tolerated.

The main confusion, however, exists among signal connectors, and the only possible advice here is to try to stay with industry standards for comparable equipment. For educational equipment, for example, the 4 mm plug and socket is almost universal, and for connecting UHF signals to a domestic TV receiver the standard coaxial plug and socket type should be used. Connections to TV receivers that are being used as monitors often use the SCART (standard connector for audio, radio and television) form of Euroconnector, but for other connections, particularly for computer monitors, standards can vary widely. Fortunately, the almost universal adoption of the IBM PC standards in computing, except in education where they are most needed, ensures reasonable uniformity.

The largest range of connectors is found in the RF and video ranges, with audio coming a close second. RF connectors are used for radio transmitters, including CB and car telephone uses, and a variety of VHP and UHF work. Although these are virtually all coaxial in design they offer a wide range of fittings, whether bayonet- or screw-retained. The wide range of connectors reflects the wide range of VHF and UHF cables, so you cannot necessarily use any type of connector with any type of cable. Generally, connectors are made to work with a limited range of cables, though in some cases the range can be extended by using adapters.

The range of RF connectors is intended to match the range of RF cables, of which Table 18.1 shows a summary of the better-known cable types. RF cables are identified by various designations, but the RG set, of US origin, is the best known world-wide. The main measurable features of an RF cable are the characteristic impedance to which the cable must be matched in order to minimize reflections, the capacitance per metre length, and the attenuation in decibels per 100 feet length at various RF frequencies. The attenuation per 10 m length is sometimes quoted. All of the cable groups illustrated are coaxial and the characteristic impedance is virtually always either 50 Ω or 75 Ω, but the attenuation characteristics differ considerably from one cable type to another. Most of the RF cables are rated to withstand high voltages between inner and outer conductors, often exceeding 20 kV.

Cables that have similar characteristics can be grouped, and Table 18.2 shows groups that are in use at the time of writing. These groupings of equivalent cables should not be taken as indicating perfect equivalence, and if you are selecting RF cable you are always advised to check with the manufacturer's data.

The various connectors are likely to be used for other than RF cable connections, of course, and there is a wide range of other coaxial cables which will fit one connector type or another and to which the connectors will match well. These applications include audio, video and digital network signal applications.

The BNC range of connectors covers both 50 Ω and 70 Ω types which are manufactured for an assortment of cable sizes. All feature a bayonet locking system and a maximum diameter of about 15 mm, and both solder and crimp fittings are available. The standard range of BNC connectors

Table 18.1 Attenuation of RF cables

Cable designation	1.0	10	50	100	200	400	900	1000	3000	5000
RG6A, 212	0.26	0.83	1.9	2.7	4.1	5.9	6.5	9.8	23.0	32.0
RG8 MINI, 8X		1.1	2.5	3.8	5.4	7.9	8.8	13.0	26.0	
LMR-240	0.24	0.76	1.7	2.4	3.4	4.9	7.5	7.9	14.2	18.7
RG8, 8A, 10A, 213	0.15	0.55	1.3	1.9	2.7	4.1	7.5	8.0	16.0	27.0
9913, 9086, 9096			0.9	1.4	1.8	2.6	4.2	4.5		13.0
4XL8IIA, FLEXI 4XL			0.9	1.4	1.8	2.6	4.2	4.5		13.0
LMR-400			0.9	1.2		2.5	4.1	4.3		
LMR-500			0.7	1.0		2.0	3.2	3.4		
LMR-600			0.6	0.8		1.4	2.5	2.7		
8214		0.6	1.2	1.7	2.7	4.2		7.8	14.2	22.0
9095			1.0	1.8	2.6	3.8	6.0	7.5		
RG9, 9A, 9B, 214	0.21	0.66	1.5	2.3	3.3	5.0	7.8	8.8	18.0	27.0
RG11, 11A, 12, 12A, 13, 13A, 216	0.19	0.66	1.6	2.3	3.3	4.8		7.8	16.5	26.5
RG14, 14A, 217	0.12	0.41	1.0	1.4	2.0	3.1		5.5	12.4	19.0
RG17, 17A, 18, 18A, 218, 219	0.06	0.24	0.62	0.95	1.5	2.4		4.4	9.5	15.3
RG55B, 223	0.30	1.2	3.2	4.8	7.0	10.0	14.3	16.5	30.5	46.0
RG58	0.33	1.2	3.1	4.6	6.9	10.5	14.5	17.5	37.5	60.0
RG58A, 58C	0.44	1.4	3.3	4.9	7.4	12.0	20.0	24.0	54.0	83.0
RG59, 59B	0.33	1.1	2.4	3.4	4.9	7.0	11.0	12.0	26.5	42.0
RG62, 62A, 71A, 71B	0.25	0.85	1.9	2.7	3.8	5.3	8.3	8.7	18.5	30.0
RG62B	0.31	0.90	2.0	2.9	4.2	6.2		11.0	24.0	38.0
RG141, 141A, 400, 142, 142A	0.30	0.90	2.1	3.3	4.7	6.9		13.0	26.0	40.0
RG174	2.3	3.9	6.6	8.9	12.0	17.5	28.2	30.0	64.0	99.0
RG178B, 196A	2.6	5.6	10.5	14.0	19.0	28.0		46.0	85.0	100
RG188A, 316	3.1	6.0	9.6	11.4	14.2	16.7		31.0	60.0	82.0
RG179B	3.0	5.3	8.5	10.0	12.5	16.0		24.0	44.0	64.0
RG393, 235		0.6	1.4	2.1	3.1	4.5		7.5	14.0	21.0
RG402		1.2	2.7	3.9	5.5	8.0		13.0	26.0	26.0
RG405								22.0		
LDF4-50A	0.06	0.21	0.47	0.68	0.98	1.4	2.2	2.3	4.3	5.9
LDF5-50A	0.03	0.11	0.25	0.36	0.53	0.78	1.2	1.4	2.5	3.5

Note: Attenuation values given are in dB per 100 feet

Table 18.2 Cable groups and equivalents

Group	Typical cable designations
C	RG-58, 58A, 58C, 141, 141A; Belden 8219, 8240, 8259, 8262, 9201, 9203, 9310, 9311; CommScope BWC-195, BWC-195R, 0268; Times LMR-195
C1	RG-55, 55A, 55B, 142, 142A, 142B, 400/U; Alpha 9055, 9055B, 9223; Belden 8219, 9907, 83242, 84142
C2	Alpha 9848; CommScope BWC-200, BWC-200R; Times MR-200, MSI-22
D	RG-59, 59A, 59B, 62, 62A, 62C/U; Alpha 9059, 9062, 9830, 9840, 9845; Belden 8221, 8241, 9169, 9204, 9228; Canare LV-61S; CommScope 5555, 5560, 5563
E	RG-8, 8A, 213/U; Alpha 9008, 9213; Belden 8267, 9251, 9880, 89880; Intercomp 4082, 22132; Times AA-4478
F	RG-9, 9A, 9B, 214/U; Alpha 9214; Belden 8242, 8268
I	RG-8/U type; Belden 9913; CommScope BWC-400, BWC-400R, Cushcraft Ultralink TL93605; Times AA-5886, AA-6146, LMR-400, LMR-400 Ultraflex; T-COM-400
L1	CommScope BWC-500, BWC-500R; Times LMR-500, LMR-500 UltraFlex; T-COM-500
L2	CommScope BWC-600, BWC-600R; Times LMR-600, LMR-600 UltraDlex; T-COM-600
P	Antenna Specialists K214, Pro-Flex 800; Times AA-3096
PL	Belden 89913
X	RG-8X; Belden 9258; CommScope BWC-240, BWC-240R; Micro 8/U; Remee 1600; Saxton 8315; Times LMR-240, LMR-240UltraFlex

can be used in the frequency range up to 4 GHz (absolute maximum 10 GHz) and with signal voltage levels up to 500 V peak. Terminations and attenuators in the same series are also obtainable, and there is also a miniature BNC type, 10 mm diameter, with the same RF ratings, and a screw-retained version, the TNC couplers. The connectors of this family offer a substantially constant impedance when used with the recommended cables.

The SMA series of connectors are to BS9210 N0006 and MILC-39012 specifications, and are screw retained. This provides more rigidity and improves performance under conditions such as vibration or impact. The voltage rating is up to 450 V peak, and frequencies up to 12.4 GHz on flexible cable and up to 18 GHz on semi-rigid cable are usable. The VSWR

(voltage standing wave ratio, ideally equal to 1.00) which measures reflection in the coupling is low at the lower frequencies but increases linearly with frequency. A typical quoted formula for semi-rigid cable coupling is $1.05 + 0.006f$ with f in gigahertz, so for a 10 GHz signal the VSWR would be $1.05 + (0.006 \times 10) = 1.11$. The body material is stainless steel, gold plated, with brass or beryllium–copper contacts, and a PTFE insulator. Operating temperature range is –55°C to +155°C.

The SMB (sub-miniature bayonet) range is to BS9210 and MIL-C-9301 2B specifications and has a 6 mm typical diameter, rated for 500 V signal peak in the frequency range up to 4 GHz. The VSWR is typically around 1.4 for a straight connector and 1.7 for an elbowed type, and both solder and crimp fittings are available. SMC (sub-miniature screw) connectors are also available to the same BS and MIL specifications. The older UHF plug series are also to MIL specifications, but with the limited frequency range up to about 500 MHz. These are larger connectors, typically 19 mm diameter for a plug, with screw clamping, and they are particularly well suited to the larger cable sizes. They are often also used for video signal coupling. There are adapters available for every possible combination of RF connector, so total incompatibility of leads should never arise. This is not a perfect solution, however, because the use of an adapter invariably increases the VSWR figure, and it is always better to try to ensure that the correct matching plug/socket is used in the first place.

Video connectors

Video connections can make use of the VHP type of RF connectors, or more specialized types. These fall into two classes, the professional video connectors intended for use with TV studio equipment, and the domestic type of connector as used on video recorders to enable dubbing from one recorder to another or from a recorder to a monitor so that the replayed picture can be of better quality than is obtainable using the usual RF modulator connection to the aerial socket of a TV receiver. Many video recorders nowadays use nothing more elaborate than an audio-style coaxial phono connector for their video as well as for their audio output.

For studio use, video connectors for a camera may have to carry a complete set of signals, including separate synchronizing signals, audio telephone

signals for a camera operator, and power cables as well as the usual video-out and audio-out signals. Multiway rectangular connectors can be used for such purposes, with 8-way connections for small installations and 20-way connections for editing consoles and similar equipment. These connectors feature very low contact resistance, typically 5 mΩ. For smaller equipment, circular cross-section connectors of about 17 mm overall diameter are used, carrying ten connectors with a typical contact resistance of 14 mΩ and rated at 350 V AC.

Audio connectors

Audio connectors start with the remarkably long-lived jacks which were originally inherited (in the old 0.25 inch size) from telephone equipment. Jacks of this size are still manufactured, both in mono (two-pole) and stereo (three-pole) forms, and either chassis mounted or with line sockets. Their use is now confined to professional audio equipment, mainly in the older range, because there are more modern forms of connectors available which have a larger contact area in comparison with their overall size. Smaller versions of the jack connector are still used to a considerable extent however, particularly in the stereo form. The 3.5 mm size was the original miniature jack, and is still used on some domestic equipment, but the 2.5 mm size has become more common for mono use in particular. Figure 18.1 shows mono and stereo forms of the smaller jacks.

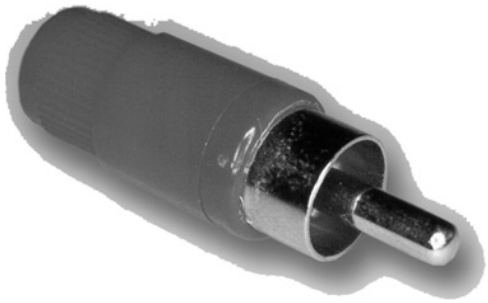
Figure 18.1

Mono **(a)** and stereo **(b)** miniature jack plugs. (Photo courtesy of Alan Winstanley.)

One of the most common forms of connector for domestic audio is still the phono connector, also called the RCA connector (Figure 18.2), whose

Figure 18.2

The RCA or phono type of plug. (Photo courtesy of Alan Winstanley.)



name indicates its US origins. Phono connectors are single channel only, but are well screened and offer low-resistance connections along with sturdy construction. The drawback is the number of fittings needed for a two-way stereo connection such as would be used on a stereo recorder, and for such purposes DIN plugs are more often used, particularly in European equipment. Many users prefer the phono type of plug on the grounds of lower contact resistance and more secure connections.

The European DIN (Deutsches Industrie Normallschaft, the German standardizing body) connectors use a common shell size for a large range of connections from the loudspeaker, two-pole type to the eight-way variety. Though the shell is common to all, the layouts (Figure 18.3) are not.

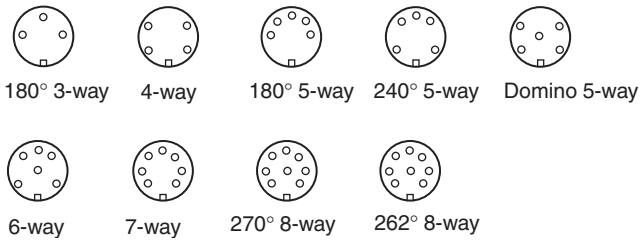


Figure 18.3

The format of the DIN connector family.

The original types are the three-way and the 180° five-way connectors, which had the merit of allowing a three-way plug to be inserted into a five-way socket. Later types, however, have used 240° pin configurations

with five, six and seven connectors, four-way and earth types with the pins in square format, and a five-way domino type with a central pin, along with the eight-way type which is configured like the seven-way 240° type with a central pin added. This has detracted from the original simplicity of the scheme, which was intended to make the connections to and from stereo domestic audio equipment easier.

The more crowded layouts of plugs and sockets are notoriously difficult to solder unless they have been mechanically well designed, using splayed connectors on the chassis-mounted sockets and to some extent also on the line-mounted plugs. Use of the five-way 240° type is recommended for audio equipment other than professional-grade equipment, but only where signal strengths are adequate and risk of hum pickup is minimal. Latched connectors can be obtained to avoid the possibility of pulling the connectors apart accidentally. For low-level use, phono plugs are preferable. Figure 18.4 shows the appearance of typical DIN plugs and a socket.

Figure 18.4

DIN two-pin, three-pin and five-pin plugs and five-pin socket. (Photo courtesy of Alan Winstanley.)

For professional (or high-quality domestic) audio equipment, the XLR series of connectors provides multiple connections with much superior mechanical quality. These are available as three-, four-, or five-pole types and they feature anchored pins and no loose springs or set screws. The contacts are rated to 15 A for the three-pole design (lower for the others) and can be used for a maximum working voltage of 120 V. Contact resistance is low, and the connectors are latched to avoid accidental disconnection. There is a corresponding range of loudspeaker connectors to the same high specifications, A variety of other connectors also exists, such as

the EPX series of heavy-duty connectors and the MUSA coaxial connectors. These are more specialized, and would be used only on equipment that is intended to match other items using these connectors.

Computer and other digital signal connections have, at least, reached some measure of standardization on the PC type of machines (IBM clones and compatibles) after a period of chaos. The use of edge connectors should be confined to internal connections because edge connectors are much too fragile for external use though they are still present on such items as digital camera memory strips which can be removed for reading by the computer.

At the time of writing, connections to computers are now mainly by serial connectors, of which the most common are the USB (universal serial bus), Firewire (IEEE 1394) and the Ethernet network connector. The older Centronics parallel and serial (typically RS-232 or RS-423) port connectors are rapidly becoming redundant and very few peripherals still use the RS type of serial port though many printers still feature the Centronics connector as well as USB.

The Centronics connector (Figure 18.5) is used mainly for connecting a computer to a printer, and it consists of a 36-contact connector which uses flat contact faces. At one time, both computer and printer would have used identical fittings, but it is much more common for the 36-pin Centronics

Figure 18.5

A typical Centronics socket.

socket to be used only on the printer. At the computer a 25-pin sub-miniature D-connection is used, usually with the socket chassis-mounted. In a normal connection from computer to printer, only 18 of the pins are used for signals (including ground). The shape of the body shell makes the connector irreversible.

The same 25-pin D-connector can be used for serial connections, but more modern machines are 9-pin sub-miniature D-sockets for this purpose, since no more than 9-pin connections are ever needed. For other connections, such as to keyboards, mice and monitors, DIN-style connectors are often used, though the sub-miniature D-type connectors are also common. The D-type connectors are widely available in a range of sizes and with a large range of accessories in the form of casings, adapters and tools, so their use for all forms of digital signals is strongly recommended.

There are now standard DIN fittings for edge connectors, including the more satisfactory indirect edge connectors that have superseded the older direct style. The indirect connectors are mounted on the board and soldered to the PCB leads, avoiding making rubbing contact with the board itself.

Lastly, the multiway connector for computing use often utilizes the IDC type of connector – the letters stand for Insulation Displacement Connector. This is supplied in two halves, with one half containing the metal connector pins and the small v-grooves for each wire of the multiway cable (Figure 18.6). The cable is laid over these grooves, taking care to ensure that the marked end of the cable is located on the connector for pin 1.

The cable should be pressed into place by hand to check that each wire is in its correct position, the other half of the shell then being fitted on. The two halves are then clamped tightly together until they lock. During this action, the edges of the grooves penetrate the insulation of each wire in the cable, making the connection. Use of a specialized clamp is best for making up connectors, but a mole wrench can be used if you need only a few connectors made up.

The main problems with IDC connectors are caused by failure to locate the cable precisely in the holder, and if a large number of cables are to be made up, a combined jig and clamp tool is necessary. Sometimes imperfect cable insulation will cause shorting or an open circuit, and each cable should

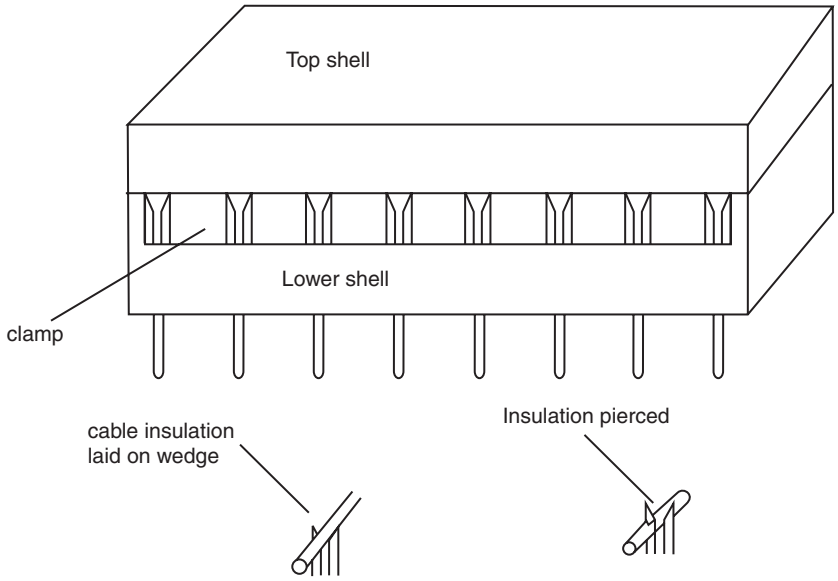


Figure 18.6

The IDC type of connector which is used extensively with ribbon cables.

be tested for continuity and for shorts once a connector has been put on each end.

Control knobs and switches

There is as great a variety in control knobs and switches as in terminals and connectors. Control knobs for rotary potentiometers are available in a bewildering range of sizes and styles, mostly using grub-screw fastening, though a few feature push-on fitting. For all but the least costly equipment, a secure fastening is desirable, but the traditional grub-screw is not entirely satisfactory because it can work loose and cause considerable delay when knobs have to be removed for servicing work. A more modern development is collet-fitting, using a split collet over the potentiometer shaft which is tightened down by a nut. The recess for the collet nut is then covered by a cap which can be colour coded or moulded with an arrow pointer. This is

a much more satisfactory form of fitting. A more specialized form of knob allows multi-turn use, so 10 to 15 turns of the dial will be needed to rotate the potentiometer shaft from one end stop to the other. These multi-turn dials can use digital or analogue readouts and are normally located by a grub-screw with an Allen (hexagon) head.

Switches

Switches are required to make a low-resistance connection in the ON setting, and a very high-resistance insulation in the OFF setting. The resistance of the switch circuit when the switch is on (made) is determined by the switch contacts, the moving metal parts in each part of the circuit which will touch when the switch is on. The amount of the contact resistance depends on the area of contact, the contact material, the amount of force that presses the contacts together and, also, in the way that this force has been applied.

If the contacts are scraped against each other in a wiping action as they are forced together, the contact resistance can often be much lower than can be achieved when the same force is used simply to push the contacts straight together. In general, large contact areas are used only for high-current operation and the contact areas for low-current switches as used for electronics circuits will be small. The actual area of electrical connection will not be the same as the physical area of the contacts, because it is generally not possible to construct contacts that are precisely flat or with surfaces that are perfectly parallel when the contacts come together.

A switch contact can be made entirely from one material, or it can use electroplating to deposit a more suitable contact material. By using electroplating, the bulk of the contact can be made from any material that is mechanically suitable, and the plated coating will provide the material whose resistivity and chemical action is more suitable. In addition, plating makes it possible to use materials such as gold and platinum which would make the switch impossibly expensive if used as the bulk material for the contacts. It is normal, then, to find that contacts for switches are constructed from steel or from nickel alloys, with a coating of material that will supply the necessary electrical and chemical properties for the contact area.

Switch ratings are always quoted separately for AC and for DC, with the AC rating often allowing higher current and voltage limits, particularly for inductive circuits. When DC through an inductor is decreased, a reverse voltage is induced across the inductor, and the size of this voltage is equal to inductance multiplied by rate of change of current. The effect of breaking the inductive circuit is a pulse of voltage, and the peak of the pulse can be very large, so arcing is almost certain when an inductive circuit is broken unless some form of arc suppression is used.

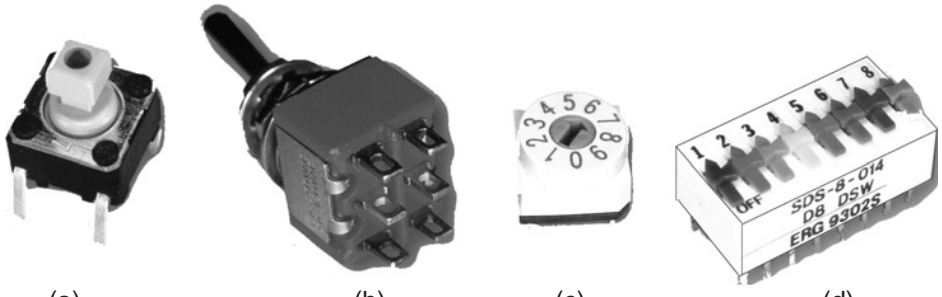
Arcing is one of the most serious of the effects that reduce the life of a switch. During the time of an arc very high temperatures can be reached both in the air and on the metal of the contacts, causing this metal to vaporize and be carried from one contact to the other. This effect is very much more serious when the contacts carry DC, because the metal vapour will also be ionized, and the charged particles will always be carried in one direction. Arcing is almost imperceptible if the circuits that are being switched run at low voltage and contain no inductors, because a comparatively high voltage is needed to start an arc. For this reason, then, arcing is not a significant problem for switches that control low voltage, such as the 5 V or 9 V DC that is used as a supply for solid-state circuitry, with no appreciable inductance in the circuit. Even low-voltage circuits, however, will present arcing problems if they contain inductive components, and these include relays and electric motors as well as chokes. Circuits in which voltages above about 50 V are switched, and particularly if inductive components are present, are the most susceptible to arcing problems, and some consideration should be given to selecting suitably rated switches, and to arc suppression, if appropriate.

The normal temperature range for switches is typically −20°C to +80°C, with some rated at −50°C to +100°C. This range is greater than is allowed for most other electronic components, and reflects the fact that switches usually have to withstand considerably harsher environmental conditions than do other components. The effect of very low temperatures is due to the effect on the materials of the switch. If the mechanical action of a switch requires any form of lubricant, that lubricant is likely to freeze at very low temperatures. Since lubrication is not usually an essential part of switch maintenance, the effect of low temperature is more likely to alter the physical form of materials such as low-friction plastics and even contact metals.

Flameproof switches must be specified wherever flammable gas can exist in the environment, such as in mines, in chemical stores, and in processing plants that make use of flammable solvents. Such switches are sealed in such a way that sparking at the contacts can have no effect on the atmosphere outside the switch. This makes the preferred type of mechanism the push-on, push-off type, since the push button can have a small movement and can be completely encased along with the rest of the switch.

Switch connections can be made by soldering, welding, crimping or by various connectors or other plug-in fittings. The use of soldering is now comparatively rare, because unless the switch is mounted on a PCB which can be dip-soldered, manual assembly will be required at this point. Welded connections are used where robot welders are employed for other connection work, or where military assembly standards insist on the greater reliability of welding. By far the most common connection method for panel switches, as distinct from PCB-mounted switches, is crimping, because this is very much better adapted for production use. Where printed circuit boards are prepared with leads for fitting into various housings, the leads will often be fitted with bullet or blade crimped-on connectors so that switch connections can be made.

Figure 18.7 shows a selection of switches that are commonly found on equipment.



(a) (b) (c) (d)

Figure 18.7

Switches: **(a)** PCB push button, **(b)** toggle, **(c)** 10-way PCB rotary, **(d)** 8-pole PCB on-off. (Photos courtesy of Alan Winstanley.)

Cabinets and cases

The variety of cabinets and cases is as wide as that of the other hardware components. Small battery-operated equipment can be housed in plastic cases, particularly one-off or developmental circuits, but for the production of equipment for professional use, some form of standard casing will have to be used. As so often happens, industry standards have to be obeyed. The old 19-inch rack standard for industrial equipment has now become the IEC 297 standard, with cabinet heights designated as U numbers – the corresponding millimetre measurements are shown in Table 18.3. These cabinets can be supplied with panels, doors, mains interlocks, top and bottom panels, fan plates, and supports for chassis, providing ample space for internal wiring and cooling. Internal chassis in the form of racks and modules can be fitted, usually in 3U and 6U sizes.

Table 18.3 The standard set of cabinet dimensions (U-set)

U-number	Height	Width	Depth
40U	1920	647	807
34U	1620	600	600
27U	1290	600	600
20U	998	600	639
12U	619	600	639
6U	230	88*	160
3U	95	88*	160

Note: Typical dimensions in millimetres. Smaller cases can be specified as 10E width (38 mm) or 20E width (88 mm)

Smaller units are accommodated in instrument cases, of which the range is much larger. There is a range of cases which will fit the 19-inch units from the standard rack systems, so identical chassis layouts can be used either in racks or in the smaller cases. The more general range of casings covers all sizes from a single card upwards, and also down to pocket calculator sizes. Cases can be obtained with carrying handles for enclosing portable instruments, or for bench or desk use. Many casings can be obtained in tough ABS plastics or in die-cast metal form. Some further degree of standardization is emerging, so far as European equipment is concerned, as a

Table 18.4 EIA, IEC and MIL standards for equipment racks

EIA-310-D	Cabinets, racks, panels, and associated equipment
IEC 60297-1	Mechanics for racks – 19-inch common standard (DIN41494
IEC 297-x	Dimensions of panels and racks
MIL-STD-901D	Shock tests. h.i. (high-impact) shipboard machinery, equipment, and systems, requirements for shock
MIL-STD-167-1	Mechanical vibrations of shipboard equipment (type 1 – environmental; type 2 – internally excited vibration)
MIL-STD-5400	Electronic equipment, aerospace, general specifications
MIL-STD-810E	Environmental engineering considerations and laboratory tests (humidity)
MIL-STD-46IE	Requirements for the control of electromagnetic interference (emi/emc) characteristics of subsystems and equipment
MIL-STD-704F	Aircraft electric power characteristics
MIL-STD-1275B	Characteristics of 28-volt DC electrical systems in military vehicles

DIN standard 43700 for small cases and boxes along with plug-in modules that fit inside. Table 18.4. shows the various applicable standards for racks of electronic equipment.

Handling

There is an important distinction between transistors and ICs in the sense that transistors are always soldered into circuits, but ICs can be either soldered to a printed circuit board or plugged into holders that are soldered to the board. The plug-in construction for ICs has been used extensively in the past for small computers, particularly when the design made it possible to expand the memory of the machine by plugging in more memory ICs. The use of IC holders is not necessarily helpful either from the reliability point of view or that of servicing, however. From the production point of view, the use of IC holders means that another step is needed – insertion of ICs – before the board is ready for use. Currently, only the processor unit is plugged into a PC motherboard with all other chips soldered, either to the motherboard or to daughterboards that are inserted into holders.

PCBs for circuit use can make use of a large number of optional construction and attachment methods. The board material may be fibreglass

(higher cost and quality) or phenolic (lower cost), or exotic materials for military applications. The conventional PCB uses an acid-etched copper-coated board, usually 1.6 mm thick, with all the components mounted on the plain (non-copper) side and wire connections taken through holes to be soldered to the copper, and clipped flush. This type of structure is easier for servicing attention and for repair actions. The double-sided version has copper tracks on both sides with plated-through holes (PTHs or *vias*) to make connection between tracks on opposite sides. A few boards use gold-plating for contacts. Small PCBs known as daughterboards are intended to be plugged into a large motherboard in order to allow for flexible use (such as adding new facilities to a computer).

A PCB using surface mount devices (SMD) has very small components attached by small tags on the copper side of the board. Holes may be used, mainly vias to connect between copper on one side and copper on the other. The size of a circuit is smaller than that on a conventional board, and automated assembly is easier. Servicing, however, is much more difficult and some SMD components are so small that they can easily be mistaken for blobs of solder. Most PCBs in commercial use contain a mixture of conventional components and SMD components, making assembly more difficult. Where the components are predominantly SMD, a double-sided board will have its SMD components mounted on both sides of the board. The ultimate in difficulty is the multi-layer board, a sandwich of boards using more than two copper layers. This is the standard form of construction for computer motherboards.

Less usual types of circuit mounting include the flexible PCB, used for computer keyboards of the membrane type (as distinct from the array of switches type), and for circuit boards that must be fitted into awkward spaces (such as in camcorders, clocks and watches). Boards for some high-volume applications such as musical novelties (greetings cards, calendars) can have one or more IC dies attached directly (chip on board, COB), with the bond wires welded directly to the PCB tracks.

The use of insertion tools in production makes the operation relatively easy, but for small-scale runs, insertion of ICs without inserting tools can easily cause pin damage. The most common form of damage is for one pin which has not been correctly lined up to hit the side of the holder and be bent under the body of the IC as it is inserted. Visual inspection often fails to spot this, and it is only when pin voltages are read using a logic probe that

the disconnection will be noticed. This mishandling is even easier to do when an IC is being replaced, and several manufacturers take the view that it is more satisfactory to solder-in all ICs, since the reliability of ICs is such that replacement is unlikely to be needed.

There is a lot to be said for this, because holders other than the expensive ZIF (zero insertion force) type are by no means completely satisfactory for high-volume applications other than computers, and as much time can be spent in removing and inserting ICs from and into holders as would be spent in snipping the pins, removing the remains and soldering in a new IC. A further point is that inserting an MOS IC into a holder is more likely to cause electrostatic damage than soldering the IC. If all of the pins are not placed in contact with the holder at the same time, there is a risk that one pin which is isolated could be touched. Though the built-in diodes of MOS ICs will generally protect against damage, the risk is higher than that of soldering. It is possible to solder-in an IC with a wire wrapped around all the pins high up on the shanks. This shorts all of the pins together, preventing any risk of electrostatic damage while the IC pins are being soldered. This is particularly useful when an IC is being replaced during servicing, because the pins will generally be soldered one at a time in such a case. The shorting wire can be removed after all the soldered joints have been checked.

The tendency now on computers is to use surface mounting wherever possible, and the minimum of socketing. The memory of the machine will be housed in strips designated as DIMM (dual inline memory module) which use edge connections. The main board (or motherboard) of the computer will contain holders for DIMM units, and the memory ICs themselves use surface-mounted memory chips that are soldered in place. This allows the memory to be inserted and changed by plugging in single units rather than a large number of chips. At one time, adding 1 Mb to the memory of a PC computer could involve plugging in 36 chips (256 × 1-bit chips, using 8 bits for each memory byte and 1 parity bit for checking the integrity of the other 8 bits). Modern practice is to insert one DIMM strip with a memory slice in the range 256 MB to 1 GB. On all machines, processors are mounted in zero insertion force (ZIF) sockets to make replacement (with an upgraded unit) easier.

Desoldering of an IC is needed only if there is some doubt about a fault. If the chip is known to be faulty, desoldering is a waste of time – it is always

easier and less harmful to other ICs to snip the pins of the defective IC and then remove them one by one from the board with a hot soldering iron and a pair of pliers. On the few occasions when a chip has to be desoldered and kept in a working state, the use of some sort of desoldering tool is helpful. Using the type of extended soldering iron bit that covers all the pins of an IC is by far the most satisfactory method, but a separate head is needed for each different size of IC. This technique is particularly useful if the chip is to be tested in a separate tester or in another circuit. It is seldom satisfactory to unsolder an IC and expect to test it in a circuit that makes use of holders, because the presence of even only a film of solder on the pins of an IC makes it very difficult to insert into a holder.

If the need to remove an IC by desoldering is a task that is seldom required, when it does arise it can be tackled by using desoldering braid. This is copper braid which is laid against a soldered joint. When a hot soldering bit is held against the braid, the solder of the joint will melt and will be absorbed by the braid. The braid can be removed, leaving the joint free of solder. The piece of braid, which is now full of solder, can then be cut off, and another piece of braid used. This is a slow business, since it has to be repeated for each pin of the IC, but the printed circuit board is left clean and in good condition. The snag is that a hot iron is being applied to the board many times, and this can cause overheating of other components.

Heat dissipation

Heat dissipation from ICs and transistors is a critical feature of many circuits, and failure to dissipate heat correctly can be the cause of many failures. It is not generally appreciated that the memory boards of computers can run hot, and that by placing one board over another it is possible to reduce the cooling to such an extent as to cause failure. The remedies that are commercially available include copper heatsinks that fit closely over the individual memory chips in a DIMM strip, and the use of cooling fans. Cooling of computers in general is good, and some recent designs use heat-pipe techniques for the processor (which can be dissipating 100 W or more); others depend on massive heatsinks with a fan to remove hot air. Computer cases now provide for fans other than that provided in the switch-mode supply unit, with fans blowing cool air into the case in addition to fans used for extraction. Care over component placement is needed to ensure that all this cooling reaches the hot spots where it is needed.

The main heat dissipation problems on any type of equipment arise when power transistors of ICs are in use, and are attached to a heatsink. Once again, well-designed original equipment gives very little trouble unless there has been careless assembly of transistors or ICs onto heatsinks. This can also be a cause of a newly repaired circuit failing again in a very short time. The problem arises because the flow of heat from the collector of a transistor to the metal of a heatsink is very similar to the flow of current through a circuit. At any point where there is a high resistance to the flow of heat, there will be a 'thermal potential difference' in the form of a large temperature difference. This can mean that the heatsink body feels pleasantly warm, but the collector junction of the transistor is approaching the danger level. The problem arises because of the connections from one piece of metal to another.

Any roughness where two pieces of metal are bolted together will drastically increase the thermal resistance and cause overheating. If either the transistor/IC or the heatsink has any trace of roughness on the mounting surface, or if the surface is buckled in such a way as to reduce the area of contact, some metalwork with a fine file and emery paper will be needed. In addition, silicone heatsink grease should always be used on both surfaces before they are bolted together. If the grease is not pressed out from the joint when the joints are tightened, this is an indication that contact is not good. In addition, though, the heatsink grease is quite a good heat conductor, and will greatly reduce the thermal resistance where two metal surfaces are bolted together. Failure to use heatsink grease, either at original assembly or later when a transistor or IC is replaced, will almost certainly cause trouble. One of the drawbacks of poor heatsinking is that the problem it causes may be seasonal, with the upshot that the equipment behaves flawlessly throughout the 364-day British winter, only to expire mysteriously on the day of summer.

Constructing circuits

The methods that are used industrially to construct circuit boards in large numbers are quite different to the methods that have to be used for experimental or one-off circuits. One factor that is common to all constructional methods, however, is that the circuit diagram is a way of showing connections which does not give any indication of how the components can

be physically arranged on a board. The main difference between circuit diagrams and layout diagrams is that on a layout diagram any crossing of connecting leads has to be avoided. A component such as a resistor or capacitor can cross a circuit-connecting track because on a conventional single-sided printed circuit board (PCB) the component will be on the opposite side of the board from the track. The simplest circuits to lay out are discrete transistor amplifiers; the most difficult are digital circuits in which each chip contains a large number of separate devices and which are generally laid out on double-sided boards with SMD components so as to make the geometry of the boards possible. Double-sided boards require dummy pads (vias), which are used only to connect a lead on one side with a lead on the other, and if a layout is not a good one, there may even need to be connections made by wire leads.

Small-scale circuits can be laid out manually, using cardboard cutouts of the components, with connections and internal circuits marked, on a large sheet of tracing paper or on transparent plastic which has been marked with a pattern of dots at 0.1 inch centres. The designing is done looking at the component side of the board, and will start by roughing out a practicable layout which does not require any track crossovers (but see Chapter 17 regarding computer layout programs). At this stage, it is important to show any interconnecting points, using edge connections or fixed sockets, because it must be possible to take leads to these connectors without crossovers. This can be quite difficult when the connection pattern is fixed in advance, as for example when a standard form of connection like a stereo DIN socket or a Centronics printer socket is to be used.

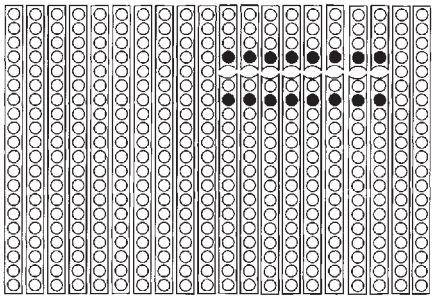
This layout can then be improved, with particular attention paid to durability and servicing. The positions of presets and other adjustments will have to be arranged, along with points where test voltages can be measured, so that servicing and adjustment will be comparatively easy. Signal lines may have to be rerouted to avoid having some lines running parallel for more than a few millimetres (because of stray capacitances), or to keep high-impedance connections away from power supply leads. It is normal, however, for bus lines on microprocessor circuitry to run parallel. Some tracks that carry RF may need to be screened, so earthed tracks must be provided, possibly on each side, to which metal screens can be soldered. Components which will run hot, such as high-wattage resistors, will have to be mounted clear of other components so that they do not cause breakdown because of overheating in semiconductors or capacitors. One way of

achieving this is to use long leads for these components so that they stand well clear of the board, but this is not always a feasible solution if several boards have to be mounted near to each other. More extensive circuits can be planned by computer (see Chapter 16) so that the PCB pattern is printed out after details of each component and each join in the circuit have been typed into the computer. Even computer-produced layouts, however, may have to be manually adjusted to avoid having unwanted stray capacitances appearing between components. Either method will have to take account of the physical differences that can exist between similar components, such as length of tubular capacitors and the difference between axial and radial lead positions.

The simplest form of construction for a one-off circuit is the **matrix stripboard**. This can be obtained in a wide range of sizes, up to 119 × 455 mm (4.7 × 18 inches approx.). The traditional types of stripboard in both 0.15-inch and 0.1-inch pitch are still available. These are always single-sided, and are suited mainly to small-scale analogue circuitry for lower frequencies (a few MHz). For digital circuits there is a range of Euro-card prototyping boards, either single- or double-sided. For connections between strips on opposite sides, copper pins are available which can be soldered to each track, avoiding the difficulties of making soldered-through connections on such boards. Boards can also be obtained in patterns such as the IBM PC expansion card or the Apple expansion card forms. Tracks can be cut by a 'spot-cutter' tool, which can be used in a hand or electric drill, thus allowing components like DIL ICs to be mounted without shorting the pin connections (Figure 18.8). Once these cuts have been made, the components can be soldered onto the board and the circuit tested.

Figure 18.8

A matrix stripboard with tracks separated by drilling.



Arrangement of components for really small-scale, low-frequency circuits can be tested in advance by using a solderless breadboard, which allows components to be inserted and held by spring clips, using a layout which is essentially the same as for matrix stripboards. The circuit needs to be marked out in nodes and a separate strip of connector assigned to each node. For complex circuits, this can be more easily done if the node numbers are also marked on the strips, using typists' erasing liquid such as Tippex to make a white surface on which pencil marks can be made. The components can then have their leads bent and cut to fit, and can be soldered into place. Components such as DIL ICs can be mounted without any need to trim leads, but remember to cut the strips that would otherwise short out pins.

Soldering onto stripboard is usually easy, but some components may need some heat-shunting. Avoid old pieces of stripboard whose copper surfaces have become tarnished, because good soldering onto such surfaces is not easy, and is likely to need a hot iron with risk to semiconductor components. Always use clean board, tinned if available. The layout that is used on these boards can be used also as a pattern for manufacturing PCBs for mass production.

A method of constructing one-off boards using either conventional or SMD components is illustrated in Figure 18.9.

The basic method employed is to place the components on the piece of copper clad board, and, using an OHP pen or similar, to mark the pin positions. You then need to work out how to connect the pins that you want connected by creating islands in the copper, and then mark the outlines of these islands with the pen. If the components are leaded devices such as the TO92, resistors and capacitors, you can plan on paper and copy the layout approximately on to the copper, but multi-pin devices and certainly ICs need to be placed on the copper to achieve the accuracy required.

It is preferable to use a straight edge to the islands because it makes cutting easier. Use a Stanley knife with two blades fitted side by side; standard and heavy duty blades give about 0.6 mm and 1.1 mm parallel lines respectively. Score along the outlines that you have drawn using a stainless steel rule to guide the blade. The score must be heavy enough to cut the copper down to the base material, easy enough on 0.5 oz and 1 oz material but you may need to score lines two or three times on 2 oz and 4 oz material.

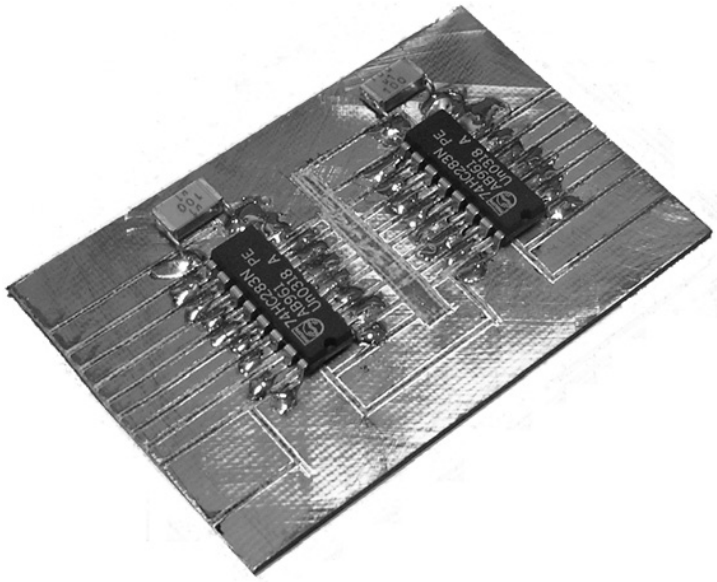


Figure 18.9

A PCB cut into tracks using a Stanley knife. (Photo courtesy of John Dunton.)

Now use a soldering iron to heat the copper between the conductor lines and peel away the unwanted copper strip with a pair of fine point tweezers. The copper is held down with an adhesive, usually an epoxy which melts/loses its adhesive properties at 200°C or so. Small areas of copper can be removed by heating and then applying a lateral pressure with the soldering iron tip, sliding the copper foil and breaking the adhesive bond.

For larger-scale work, PCBs must be produced which will later be put into mass production. A board material must be used which will be strong and heat resistant, with good electrical insulation. The choice will usually be between plastic-impregnated glass fibre board, or SRBP (synthetic resin bonded paper), though some special-purpose circuits may have to be laid out on ceramic (like porcelain) or vitreous (like glass) materials in order to cope with high-temperature use and flameproofing requirements.

The pattern of the circuit tracks has to be drawn on a piece of copper-laminated board, using a felt-tipped pen which contains etch-resistant ink.

For experimental uses, circuit tracks can be applied directly by using etch-resistant transfers. Standard patterns include lines of various thicknesses, IC and transistor pads, and pads for mounting connectors, together with a variety of curves, dots, triangles and other patterns. Another option is to work from a transparency of the pattern, using light-sensitive, etch-resistive material which is then 'developed' in a sodium hydroxide solution. When the ink or other etch-resist is dry, the board is etched in a ferric chloride bath (acid hazard – wear goggles, gloves and an apron) until all of the unwanted copper has been removed. The layout of a circuit onto copper laminate board should be started by making a drawing on tracing paper or transparent film. Components, or cardboard cut-outs, can then be placed on the drawing to show sizes, and to mark-in the mounting pads to which the leads will be soldered. This is done with the drawing representing the component side of the board, but the tracks can now be drawn in as they will exist on the copper side – this means that the actual appearance of the component side will be the mirror-image of your layout.

The drawing will then have to be transferred to the copper. This can be done manually, using an etch-resistant ink as described earlier, or by photographic methods. The board can then be etched, thoroughly cleaned and drilled, the components then being mounted and soldered into place. For the hand-made etched board, all traces of the resist material and the etching solution have to be removed by washing and scrubbing with wire wool.

The board can then be drilled, using a 1 mm drill, and the components inserted. The final action is soldering, using an iron with a small tip. The boards are heat resistant, not heatproof, so soldering should be done fairly quickly, never keeping the iron in contact with the copper for too long. Excessive heating will loosen the copper from the plastic board, or burn the board, and if the copper has been cleaned correctly and all component leads are equally clean, soldering should be very rapid. Chemical tinning solutions can be used to treat the copper of the board so that soldering can be even more rapid. Remember that excessive heat will damage not only the board but the more susceptible components like semiconductors and capacitors. The time needed to obtain a good soldered joint should not exceed a few seconds.

Boards for larger-scale production are undrilled and completely covered with copper, which will then be etched away into the pattern of connections that is needed. In a mass-production process, the pattern of etch-resist is

placed onto the copper by a silkscreen printing process. The copper is then etched in baths which are maintained at a constant high temperature, and the boards are washed thoroughly in water, followed by demineralized (soft) water, and then finally in alcohol so as to make drying more rapid. The holes are then drilled for the component leads. For mass production, all of these processes are completely automated, and the assembly of the components onto the board and subsequent soldering will also be totally automatic.

Circuit boards for commercial use have been of almost a stereotyped pattern until the comparatively recent rise in the use of surface-mounted components. The standard board backing materials are either SRBP (synthetic resin-bonded paper) or glass and epoxy resin composites, with copper coating. Many suppliers offer such boards with the copper already coated with photo-resist, saving considerable time and effort for small batches. These precoated boards must be stored carefully, preferably at low temperatures between 2°C and 13°C, and have a shelf-life which is typically 1 year at 20°C. The maximum allowable temperature is 29°C.

Board sizes now follow the Eurocard standards of 100 mm × 160 mm, 100 mm × 220 mm, 233.4 mm × 160 mm, and 233.4 mm × 220 mm; and there are also the older sizes of 203 mm × 95 mm (8 × 3.75 inches) and 304.8 mm × 457.2 mm (12 × 18 inches). Boards can be obtained with edge-connecting tongues already in place – these must, of course, be masked when the main board is etched. When boards are bought uncoated, photo-resist can be sprayed as an aerosol for small-scale production or R&D applications.

Many commercial PCBs, particularly for computer or other digital applications, are double-sided, with tracks on the component side as well as on the conventional track side. Where connections are needed between sides, plated-through holes are used. These are holes which have copper on each side and which have been electroplated with copper so that the holes have become partly filled, making a copper contact between the sides. These connections are strengthened when the board is soldered. The use of double-sided board is particularly important for digital circuits where a single-sided board presents difficulties because of the need to cross leads. The use of a well-designed, double-sided board can solve these problems, but care needs to be taken over capacitances between tracks that are on opposite sides of the board.

Soldering and unsoldering

Connections to/from electronic circuits can be made by wire-wrapping, crimping, welding or by soldering, but soldering is still the most common connection method for all but specialized equipment. The basis of soldering is to use a metal alloy, formerly of tin and lead, that has a low melting point, to bond to metals that have a much higher melting point. In a good soldered joint, the solder alloy penetrates the surface of each metal that is being bonded, forming a join that is both mechanically strong and of low electrical resistance. The action of soldering is straightforward, so much so that it is normally automated in the form of wave-soldering baths, but manual soldering still causes problems. Most of these problems can be traced to inadequate preparation of the surfaces that are to be soldered, particularly when very-small-gauge wires are to be soldered.

The other main causes of soldering problems concern the under- or over-use of the soldering iron or other source of heat. To start with, the surfaces that are to be soldered must be clean. The main enemy in this case is oxidation of copper surfaces after storage, and any attempts to solder to a copper surface that is dark brown in colour, instead of the bright gold–red of freshly cleaned copper, will be doomed. For a one-off soldering action, metal surfaces can be cleaned mechanically, using a fine emery paper, but for mass production some form of acid cleaner or mechanical scrubbing equipment is needed. The tracks of printed circuit boards are particularly vulnerable to oxidation because they have a large surface area. Stranded wire is particularly difficult to clean well, and if a new surface can be exposed by cutting a few inches from the end of a stranded cable this will make connection much easier.

In making a connection using stranded wire, the conventional wisdom at one time was that the end of a stranded cable should always be ‘tinned’, meaning coated with solder, because it was mechanically connected (such as to a mains plug), so as to avoid the possibility of strands causing a short-circuit. This practice is now frowned on because where some flexing of the wire can happen, as on most connectors, the wire will eventually snap where the tinned section joins the untinned wire. In addition, the clamping screw often works loose as the tinned wire shrinks. A good compromise is to tin only the last millimetre or so of a stranded wire, and insert the wire into a holder so that any fixing screw bears on the untinned wire.

Dirt and oxidation are the main causes of soldering problems but not the only ones. Good soldering requires the solder to be melted onto the whole area of the join, and this is possible only if all of this area is at a high enough temperature. Insufficient heat flow will cause the solder to solidify before it penetrates the other metal, making a join that looks soldered but which is really only a blob of solder placed on the surface. Such a joint is not mechanically strong and is electrically unreliable.

The other extreme occurs when too much heat is used, and the temperature becomes high enough to oxidize the solder and the other metals. Such a joint will often look secure, but it is 'dry'; the solder has penetrated but has been boiled out again, leaving a film of oxide with a high resistance. Joints of this sort are a very potent source of trouble because they are difficult to detect and remedial action is problematic.

Choice of soldering equipment is important, and it starts with the soldering iron. Despite the name, the tip of the iron is made of copper (sometimes iron plated), and most types of iron come with a selection of different tips for different grades of work. Note that if you use the iron-plated type of bit, it should never be filed to clean it because this will remove the iron and allow the remainder of the tip to disintegrate quickly. A lot of work with modern components can be carried out with a miniature electric iron of 12 W to 25 W rating, but there are still a few actions, such as soldering thick bus-bars or metal strips (particularly silver-plated strip lines), that need more heating power than a small iron can provide. Thermostatically controlled 50 W irons are very useful, but a few soldering jobs are easier with the larger irons that were commonly used in the days of valve circuits. These large irons are rare now, and a useful substitute is the gas-flame torch, using miniature butane cylinders. These need to be used with care, because the gas flame is at a very much higher temperature than the tip of an electric iron. The miniature gas burners are very useful when no source of mains power is available because they are lighter and more compact than battery-powered electric irons. Gas torches can also be used for brazing, which is very much the same as soldering but using a silver–copper–zinc alloy, which has a higher melting point. Brazed joints are used where mechanical strength is important, such as chassis joints; the main application on the small scale is to modelling with metal.

The choice of solder was easier in the past. The standard type of solder for electronic purposes was formerly 60% tin and 40% lead, in the form of a

small-diameter tube filled with the flux compound. Flux is a jelly or paste which combines several actions, lowering surface tension to allow solder to spread, protecting surfaces from oxidation, and (to some extent) helping to keep surfaces clean. Typically, the melting point of the solder will be around 190°C, requiring a temperature at the tip of the soldering iron of around 250°C. Virtually all soldering requirements were formerly fulfilled by this type of solder but from July 2006 solder containing lead will be banned throughout the EU. The ban is already operating in Japan, but it is uncertain that other competing countries in the Far East will follow suit unless the EU places an embargo on boards containing lead in their soldered joints.

PCBs for military and other high-reliability requirements will continue to use lead-based solders because these show up significantly in tests involving vibration and thermal cycling.

The new soldering alloys use metals such as silver or copper in place of lead, but these require higher soldering temperatures: 20°C to 40°C higher. This makes the reflow window (the temperature gap between melting point of solder and the temperature at which components will be damaged) for a solder bath much more critical. Soldered joints look very different when the new alloys are used, with a dull appearance and a steep contact angle between metal and solder. Suitable fluxes are also needed. Many of the new alloys are subject to patent restrictions, and there is a confusing variety of lead-free solders available, with little guidance for the user. Most of the new alloys use a higher percentage of tin. The use of lead-free solder is quite pointless unless the PCB metal and the component leadout wires are also lead-free. When a board has to be repaired it will be important to know which type of solder was used in its original construction, though the 99C alloy (99.7% tin and 0.3% copper) can be used for all lead-free hand soldering. In general, however, it is preferable to use the same solder type for repairs as was originally used because the reliability of mixed solders is problematic.

Hobby electronics is not excluded from the lead-free requirement, and Maplin Electronics can supply 99C flux-cored solder and also irons with a tip temperature of 270°C. Look at the website www.maplin.co.uk for more information.

For more specialized work, aluminium solder and silver solder can be used. Aluminium solder is an alloy that contains about 2% silver in a solder that is much richer in lead than the standard formulation, and this raises the melting temperature to around 270°C, requiring a higher iron temperature or the use of a gas-torch. Another option is a solder with 10% silver content which will provide joints of significantly lower resistance. This has a melting point of around 370°C, which is considerably more than most electric irons can provide, and more than most PCBs or components can endure for more than a few seconds. For some small-scale assembly work, solder pots can be used. These are containers that are electrically heated, usually with some provision for scraping the top surface of the molten solder to keep it clean. These pots cannot be used with flux-cored solder, and the usual procedure is to keep the pot topped up with 60/40 (or lead-free equivalent) solder pellets. The use of a solder pot follows the same pattern as the use of large solder baths in production machinery.

Typically, a conventional soldered join is made as follows. The materials to be joined are thoroughly cleaned, but avoiding the use of corrosive acids that could cause problems later. The joint is secured mechanically if this is possible (two wires can be hooked together, for example) so the presence of the solder is not essential for mechanical strength. The iron is allowed to come up to working temperature, and the tip is cleaned, either by wiping it with a slightly damp cloth or by rubbing it on a proprietary cleaning block such as the Multicore TTC1. The iron is then placed on the joint and after a short pause the (flux-cored) solder is also placed against the joint. The solder should melt and spread over the joint – it can sometimes be an advantage to move the tip of the iron to help the solder to spread. When the solder forms a thin bright film, remove both the solder and the iron and allow the joint to cool without disturbing it.

For an excellent guide to soldering techniques, see the excellent and informative website of Alan Winstanley, which the author considers to be essential viewing:

www.epemag.wimborne.co.uk/solderfaq.htm

Photos from Alan Winstanley's guide are quoted in the following section on desoldering.

Common faults start with trying to work with the iron alone, coating it with solder and then applying it to the joint. This virtually ensures that most joints are made with little or no flux, so dry and fragile joints result. Another common fault is to dab the iron and the solder on a joint, withdrawing the iron before the solder has spread. This will cause dry and unreliable joints. The opposite, leaving the iron in place until the solder turns dull, will cause severe oxidation and an equally unreliable joint. Continuing to feed solder so that molten blobs of solder drip from the joint is another fault, usually indicating that the metals have not been cleaned thoroughly, thus causing the solder not to spread.

Desoldering

Very few components on a printed circuit board are contained in sockets, and therefore components have to be removed by desoldering. A few components can be removed easily, the simplest being resistors or capacitors whose bodies lie horizontally, parallel to the board. It is easy to apply the hot iron to one joint, pull the wire out of the hole in the PCB, and then use the iron on the other lead to pull out the component entirely. It is not so easy to unsolder vertically mounted components or components with a large number of short leads, such as ICs. For such components you must remove the solder from each joint and then pull out the component (not while you are using the iron). The danger in these desoldering actions is that you will overheat the PCB or the surrounding components, making repairs difficult or impossible.

The two main methods of small-scale desoldering are the use of solder braid or the use of a desoldering pump. Solder braid is a form of stranded copper, around 2 mm diameter and usually supplied on a reel. The braid is wound around the joint, and the tip of the hot iron is cleaned, ensuring that it has only a thin film of solder. If a hotter iron temperature can be used than is used for soldering, the process is made easier. The hot iron tip is applied to the braid (not to the solder) and the iron is held in place until the solder of the joint is seen to run into the braid (a form of capillary action) (Figure 18.10). The iron is then removed, followed as quickly as possible by taking away the braid. This should leave the joint almost totally free of solder, with only a thin bright film showing. The solder-filled braid is snipped off, and the process is repeated for each joint (with a rest between each to allow the PCB to cool) until all the joints

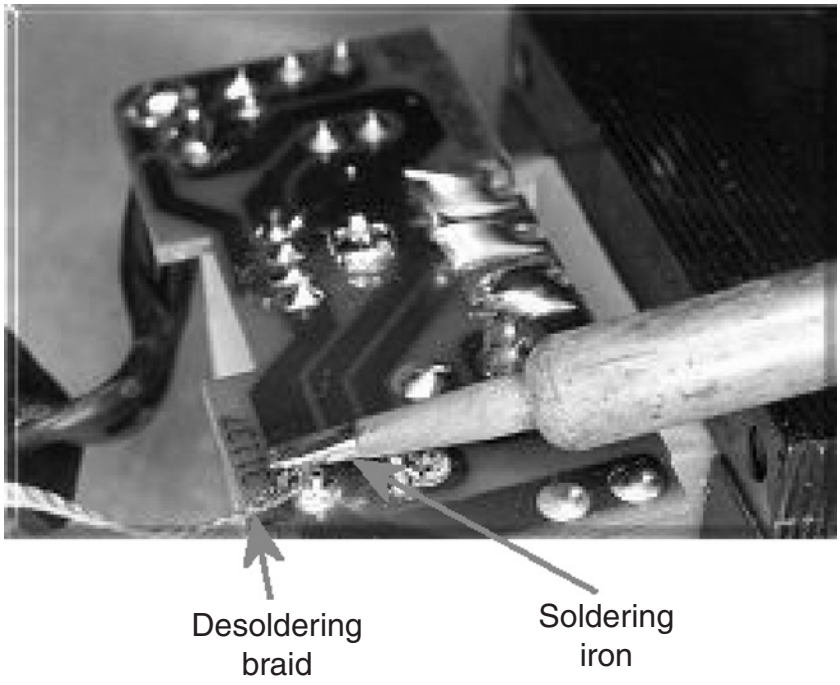


Figure 18.10

Using desoldering braid. (Photo courtesy of Alan Winstanley.)

have been treated. The component can then be pulled out. Any component that has been desoldered should be treated as faulty and not reused.

The desoldering pump is a small cylinder and piston arrangement with a spring-loaded plunger. The plunger is forced in, and the nozzle of the pump held to the joint. The iron is used to melt the solder, and pressing a catch on the pump allows the plunger to return, sucking the solder into the pump. Solder fragments can be removed after several desoldering actions. The pump is easier to use, particularly for close-packed components, and will soon justify its higher cost if a large amount of desoldering needs to be done. The nozzle of the pump can be either of aluminium or of high-temperature plastic such as Teflon, and nozzles are interchangeable. Figure 18.11 shows a desoldering pump in use.

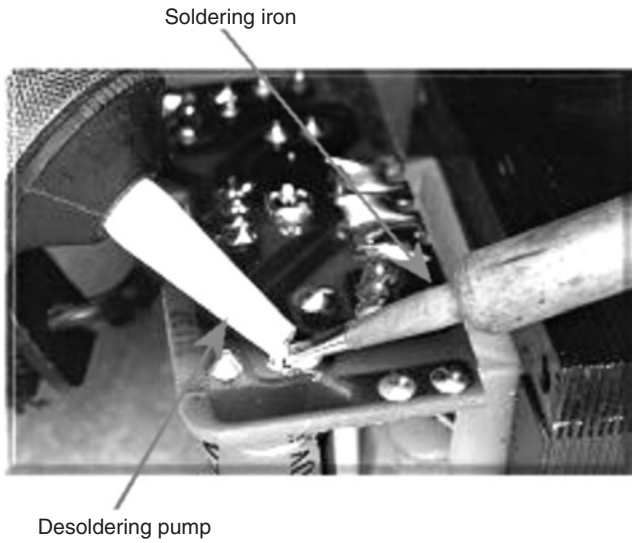


Figure 18.11

Using a solder sucker (desoldering pump). (Photo courtesy of Alan Winstanley.)

Other soldering tools

Heat-shunt tweezers are useful when soldering semiconductors that are susceptible to damage by overheating. They are seldom used in production work, because the flow-soldering methods that are used allow very little heat to flow, and an air blast will cool the components rapidly. For hand soldering and desoldering, thermal tweezers can be clipped onto the wire leads between the body of the semiconductor and the soldered joint. The effect is to allow heat to take the path of least resistance, into the thicker tweezers rather than along the thinner wire, so preventing the temperature inside the semiconductor material from rising to a value which would destroy a junction. Heat shunts are particularly valuable when desoldering, because the iron often has to be applied for longer, and when components with very short leads are used (as in RF circuits). Components other than semiconductors with long leads do not normally need heat shunting.

Soldering toolkits are useful if you do not already have an assortment of tools. Typical sets contain a reamer for cleaning inside plated-through holes

and a hook for reaching awkward places. A brush is useful for removing dust and solder fragments, and a fork can be used to hold wires in place. A scraper and a knife are both useful in cleaning wires and PCB tracks. A set of Swiss files is also useful for cleaning work. Avoid the use of acid solutions and old-fashioned flux solutions such as Baker's Fluid.

For SMD soldering you should use a temperature-controlled iron, with a solder that has a silver content to improve flow. A flux pencil is a convenient way of ensuring flux coverage, and a pair of No. 7 curved tweezers is very useful for holding components (but be careful not to solder the tweezers to the SM device).

Wiring up connectors can be tedious, and for all but the simplest and least-used connectors it is worthwhile making up some form of jig that will hold wires in place while they are being soldered. The DIN plugs and sockets are particularly fiddly items, especially when stranded cables are being used, and it is advisable to tin the ends of the stranded wires to keep the strands together. Preferably, each connection should be clamped into place using small-nosed pliers and then soldered, but some DIN connectors offer no easy way of doing this. Anything you can do to avoid having to juggle simultaneously with a (hot) connector, a cable, solder and a soldering iron, must be welcome. A jig is also useful for soldering connections to D-type plugs of the 9-pin or 25-pin type as used for computer serial cables and also for printer leads for PC machines, though computer applications generally use IDC connectors.

Care needs to be taken when soldering cables into coaxial connectors, whether of the TV or the BNC type. It is only too easy to melt the insulation on the cable and to soften the insulation in the connector if the iron is applied too long or in the wrong place. If the connector insulation softens you may find that the inner pin position has changed, and that the connector is unusable. One useful tip is to connect a plug into a socket if you are about to solder either (or both) so that one acts as a heatsink for the other.

This page intentionally left blank

CHAPTER 19

TESTING AND TROUBLESHOOTING

Introduction

The troubleshooting of circuits and systems is a skill that has to be acquired by practice; this chapter is aimed at providing pointers to some common techniques and pitfalls. In order to test a circuit you first have to know what it should do if it is working correctly, and the circumstances under which it is expected to perform to specification. It is also necessary to understand the performance and limitations of the test equipment.

One further point is that wherever mains circuits are involved extra caution and care is essential. It is very easy to kill yourself and others with poor electrical safety. If in doubt get help from someone who is qualified and experienced. Do not assume that experience with electronic circuits provides the experience necessary for dealing with mains voltages. See the section headed Mains work in this chapter.

Test equipment

Test leads

Good quality test leads make an enormous difference to the speed and efficiency of troubleshooting and testing. A range of standard power connector to 4 mm leads that will plug into a laboratory power supply will ensure safe and easy connection of devices with DC power sockets. Ideally these leads should be clearly colour coded. Where sockets may have either centre pin polarity make sure that the connectors are clearly marked. Signal leads are just as important and using the correct connectors and cable type can often

be essential for correct operation of a piece of equipment. It is important too that test leads are well maintained and labelled. Never put broken or damaged leads back into circulation, repair them or label them as broken. It is very frustrating when trying to test a piece of equipment to have to debug the test leads as well.

Interconnecting cables like printer cables, null modems or hard disk cables should be treated in the same way as test leads and kept in good order as it can take hours to find faults that may not even relate to the work in hand.

Power supplies and battery packs

Much repair and maintenance work can be completed using the power supplies built into, or provided with, equipment. However, it is very much easier and usually safer to use a variable voltage power supply with an adjustable current limiter, often referred to as a *laboratory power supply*; this can protect equipment with a fault from being destroyed by high fault currents. Variable voltage power supplies, often 2–30 V output, usually have meters or LED displays to indicate the set voltage and the current being drawn. It is good practice to leave these power supplies with the output voltage turned down to its minimum setting after use, particularly when several people are using the equipment; there are few things more irritating than connecting 27 V to your 5 V circuit because the last time you used the power supply you were charging a truck battery with it.

Many laboratory power supplies use four terminals, that is separate output and sense terminals for each connection (Figure 19.1). This can be very useful if the leads from the supply are long and the supply currents are high. However, these terminals are usually connected by shorting links that are held in place by the screw-down terminals that are also used to connect wires without plugs to the supply. Leaving the screw-down caps loose can cause the output voltage to be very badly regulated and often to have glitches to maximum supply voltage if the supply is shaken or bumped. This can cause malfunction of, or damage to, the circuits being powered, so always check that the sense links are tightly clamped by the screw caps. Another precaution when using power supplies is to ensure that the output is switched off or circuits under test are disconnected when the mains switch is turned on; under some circumstances the power supply

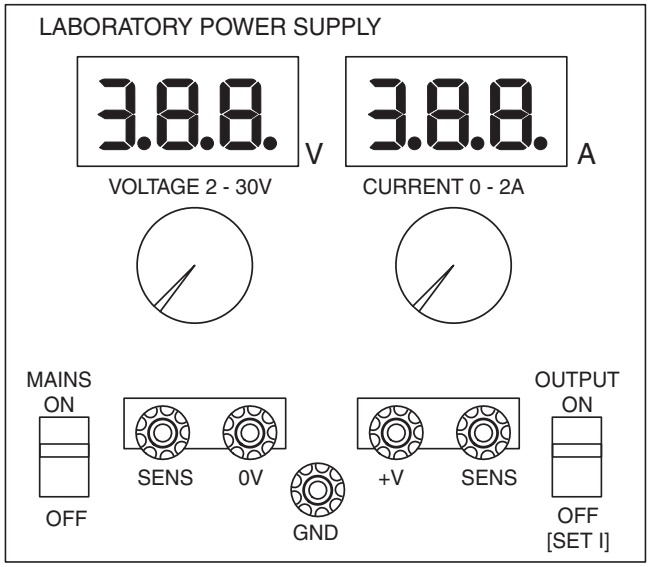


Figure 19.1

Laboratory power supply; note the links between output and sense terminals.

output can overshoot before coming under control, producing an output transient up to the maximum output voltage.

The voltage and current meters of variable power supplies are good indicators of output conditions but should not be relied on for measurement purposes.

While variable voltage power supplies with current limiters are very useful, they can sometimes be electrically noisy, and a battery pack with 4 mm terminals providing ±4.5 V can be very useful for breaking earth loops and providing a low-noise supply for testing sensitive circuits like microphone preamplifiers and short-range radio transmitters.

Digital multimeters

Digital multimeters are available with price tags of between £5 and £5000 – so what is the difference between them? Very cheap meters – hand-held

multi-tester type meters – are great for checking fuses and measuring the terminal voltage of a battery, or the resistance of a 5% resistor to verify that you read the colour code correctly; they can also be useful for monitoring supply current, but at 3.5 digits resolution they are strictly indication-only devices. At the other end of the price range are calibration grade meters with 6.5 digits; typically, though, a 4.5 digit meter will be sufficient for most troubleshooting testing.

When measuring voltage with a multimeter, be aware that the input impedance changes with range, because a resistive divider is used to scale the input to an ADC, so measurements and errors may well change across the range boundaries.

The current measurement range of a multimeter is usually implemented as a low-resistance shunt, 0.1 Ω for the 10 A range, and higher values for the lower current ranges. The current through the shunt produces a voltage drop between the current range socket and the common socket (Figure 19.2). In cheap meters the top end of the shunt is often connected to the voltage input by a high-value resistor, so if you leave the leads plugged in the wrong sockets and make a measurement unusual results may be observed. When using a meter to measure supply current and another to measure supply voltage remember to take the drop across the current-sensing resistor into account and measure the supply voltage directly at the terminals of the device being measured rather than at the input to the ammeter.

Digital multimeters typically measure other quantities such as resistance capacitance, frequency, the forward voltage of diodes and temperature as well as volts and amps. Functions like transistor testers that measure h_{FE} are also common and useful for 'go'–'no go' testing and, with care, for gain grouping of low-frequency transistors. Capacitance measurement is usually at a low frequency such as 1 kHz and should be viewed as indicative only. Beware of electrolytic capacitors; ensure that they are fully discharged before measuring them since they can store enough energy to destroy the input circuits of a multimeter. Electrolytic capacitors should always be handled with care, particularly the high-voltage types which can store a lethal shock. Temperature measurement is usually performed using a k-type thermocouple probe; multimeters that provide this option are usually good to 1°C accuracy, which should be sufficient for most troubleshooting jobs.

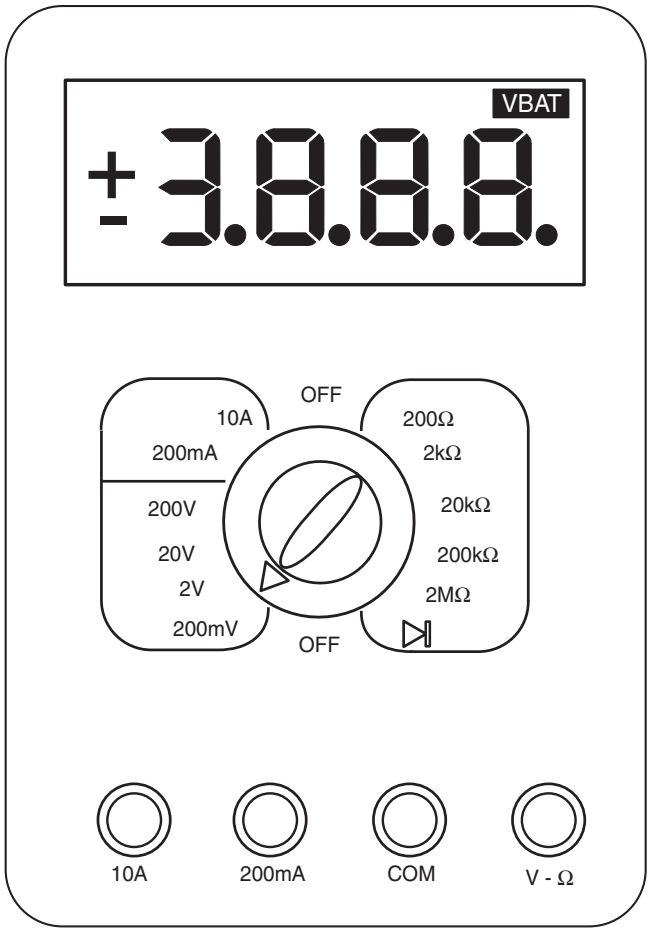


Figure 19.2

Typical low-cost hand-held DC multimeter, with volt, amp, ohm and diode test ranges.

Analogue meters are becoming quite rare now since digital meters offer so many advantages both in terms of the functions that can be built in and the accuracy of readout. With an analogue meter it can be easier to see trends, or drifts such as the change in supply current as an amplifier heats up. In the last few years relatively cheap hand-held meters with optically isolated RS232 interfaces have become available; when connected to a PC these can be used to capture readings every half second or so, so the viewing of trends is very easy as a graph can be drawn in real time.

LCR meter

Inductance, capacitance, and resistance meters, like multimeters, can be cheap or very expensive. Component measurement to tight tolerances is expensive but for most applications a meter costing the same as a good quality hand-held multimeter will be sufficient. If possible choose one that measures inductance and capacitance at more than one frequency, typically 1 kHz and 100 kHz or 1 MHz.

Oscilloscope

Oscilloscopes are the mainstay of circuit testing because circuits are usually designed to operate on signals that change with time; very few, apart from power supplies, are designed not to change with time.

Like multimeters, analogue oscilloscopes are being replaced by digital ones; also, since oscilloscopes tend to be more expensive and less portable devices, there are many 'virtual' oscilloscope interfaces for Windows- and Linux-based computers. These are often USB interfaced units containing fast ADC and sample memories that, under control from the PC, can perform like an oscilloscope.

An advantage of digital oscilloscopes is their ability to measure the input signal characteristics directly, for instance peak-to-peak voltage or frequency. Figure 19.3 shows a screen capture from a digital oscilloscope (TDS3200) in which the measurements and settings of the channels are clearly displayed.

Oscilloscopes and their probes must be understood to get the best out of them; both of these devices are critical to making good measurements, and misusing a probe can result in very misleading data. Oscilloscope inputs are typically either high-impedance 1 MΩ in parallel with 20 pF or low-impedance 50 Ω. Inputs of 50 Ω are usually found only on high-specification units. Inputs of 1 MΩ can be used for 50 Ω measurements by attaching a 'T' piece with a 50 Ω terminator to the front input panel connector to provide a 50 Ω termination for the signal being measured. Remember, however, that there is a capacitor, typically 20 pF, in parallel with the input so in this case at 159 MHz the 20 pF will have a reactance of 50 Ω.

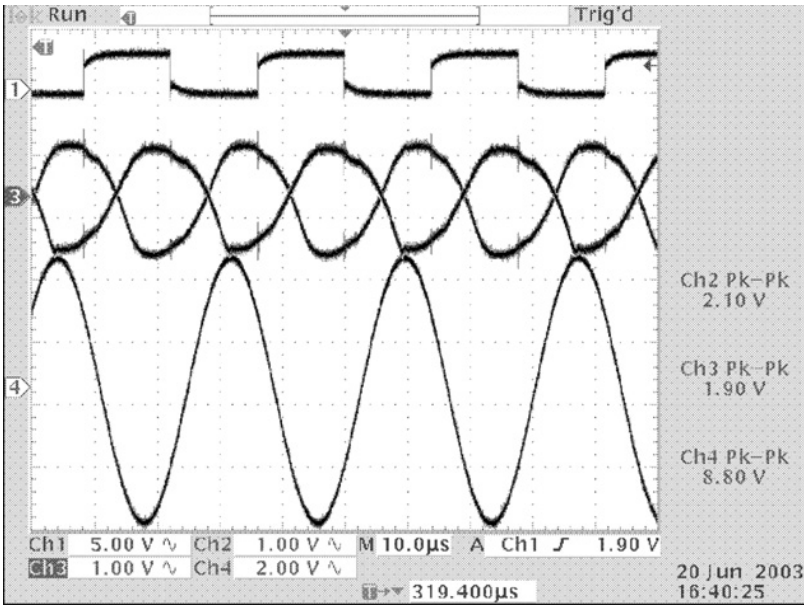


Figure 19.3

Typical oscilloscope display, showing channel amplitude settings and measurements.

Oscilloscope probes must be matched with the oscilloscope input that they are used with and, to this end, frequency compensation presets are provided, at one or other end of the probe lead, marked (b) in Figure 19.4.

The compensation matches the resistive divider to the oscilloscope input, which otherwise can look like a low-pass filter. A square wave test signal with fast rise-and-fall times is provided on the front panel of the oscilloscope to assist in the adjustment of the compensation preset. Probes should be checked (Figure 19.5) using this signal from time to time, and particularly if they have been used with different oscilloscopes to ensure that they still match.

Oscilloscope probes are usually of the switchable type, with $\times 1$ and $\times 10$ ranges and sometimes a reference position selectable on the probe. The reference position grounds the probe circuit locally so that the DC level can be adjusted on the oscilloscope display. Most probes isolate

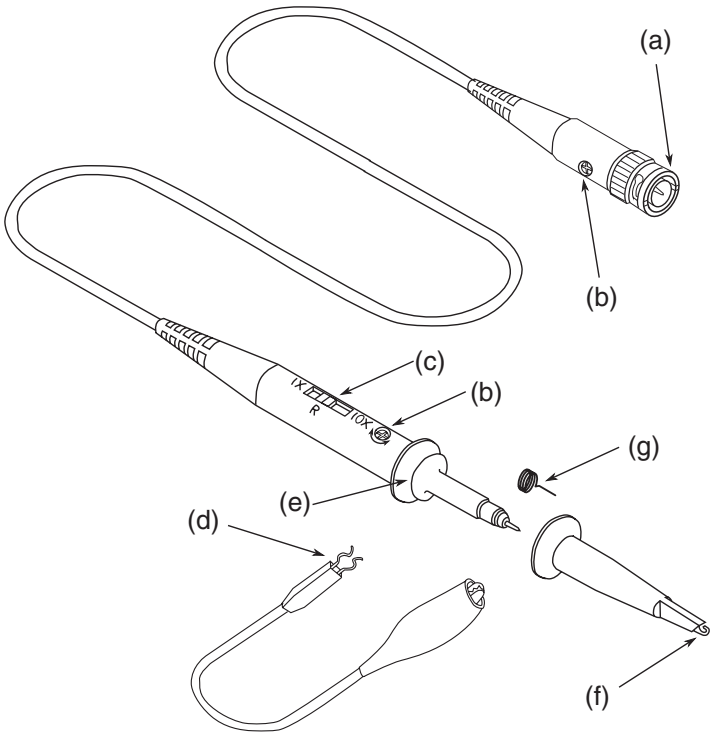


Figure 19.4

Oscilloscope probes: **(a)** BNC connector plugs into probe socket on oscilloscope front panel, **(b)** frequency compensation preset – see text, **(c)** range switch, **(d)** ground lead, **(e)** ground lead connection below finger guard, **(f)** probe hook and **(g)** spring ground connection.

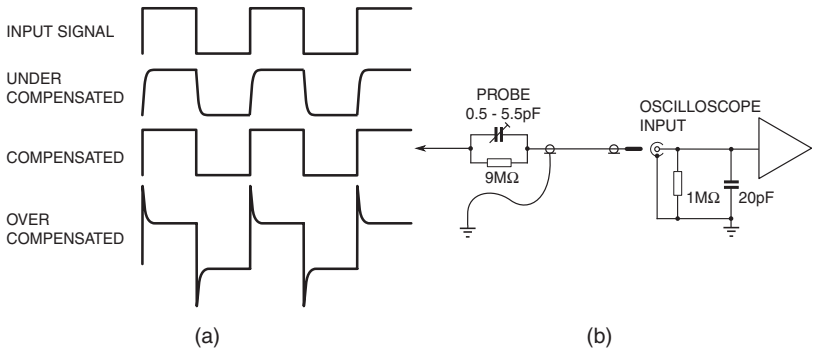


Figure 19.5

Oscilloscope probe compensation: **(a)** waveforms and **(b)** schematic of a × 10 passive probe.

the tip when the reference position is selected to avoid dangerous short circuits.

Analogue oscilloscopes usually require the user to work out the effect of the probe on signal amplitude; that is, if the oscilloscope is on a 1 V per division range with a ×10 probe, it is read as 10 V per division. Digital- and PC-based instruments usually allow the probe factor to be entered in a menu and then correct all readings to that value. Be careful with switchable probes; if the readings seem unreasonable, for example 30 V in a 5 V circuit, the oscilloscope is probably set for a ×10 probe and the probe is set for ×1.

Many modern digital oscilloscopes have coded sockets and probe plugs so that if you plug a suitably coded ×10 probe or switchable probe into the oscilloscope the display and cursor readouts will be automatically scaled to match.

When measuring signals with fast edges, that is most digital and switching signals, and any analogue circuit with frequencies above about 10 MHz, the ground lead that clips onto the probe is too long to be useful. The inductance that it represents causes undesirable ringing and amplitude errors, so some probes come with a spring ground that fits to the collar of the probe with the hook removed (Figure 19.4g). If one is not available it can be made with tinned or silver plated 20swg copper wire. This will make it possible to connect to grounds close to the probe tip, within 5 mm to 10 mm, reducing the inductance of the ground connection dramatically. Some designers wisely leave a hole in the solder-resist of ground plane close to test points for this purpose. For signal frequencies above about 80 MHz it is worth considering using test sockets rather than test pads on a PCB, or proprietary parts designed to fit a particular oscilloscope probe. Consult the probe manufacturer, or use general-purpose constant impedance miniature RF connectors like the SMA or MCX series. The latter can be used with an appropriate probe adapter.

Connecting an oscilloscope probe to a circuit changes the circuit by the addition of the probe capacitance and resistance between the measurement point and ground. Using ×10 probes helps to some extent in that the probe capacitance is relatively low, approximately 10 pF, and the resistance is high – but the probe still changes the circuit conditions. An example of a probe affecting the circuit is frequency shift or even the stopping of a crystal

oscillator when it is probed because the circuit capacitance may be doubled by attaching the probe. Very low input capacitance probes with FET buffer amplifiers are available at costs often equal to the cost of an oscilloscope. A useful circuit for a DIY FET probe is given in *Troubleshooting Analogue Electronic Circuits*, Butterworth-Heinemann.

In addition to the voltage probes described there are also current probes available in the form of toroidal transformers or ferrite ring clamps with Hall effect sensors embedded in them. These are similar in action but will be calibrated in mV/mA or V/A depending on the current range. Some oscilloscopes allow the calibration factor and units of such probes to be entered so that cursors read mA rather than mV but generally you will need to convert the measurement data manually.

Signal generator

Some equipment cannot be tested without an input signal. Very often a signal generator will make measurements easier than will the sensor or system that is used in actual operation. Applying a single frequency at controlled amplitude to an audio amplifier, for example, allows measurement of gain, and changing the frequency can explore the bandwidth. Good stable signal sources are costly but much more can be learned about a circuit's performance if you can trust the input signals. Pulse generators and function generators which provide sine, square and triangle waveform outputs are very useful. A common failing of the cheaper ones is noisy amplitude or level control potentiometers but hopefully this will become less common as digitally controlled pots and microprocessor-controlled interfaces reach the cheaper instruments. A very useful companion to a signal generator is a log-step attenuator, either 600 Ω for audio work or 50 Ω or 75 Ω for RF and video signals.

Low-cost signal generators often have a 20 dB attenuator switch on the front panel. This can lead to trouble if pressed without thought as to the consequences. Many audio signal generators can deliver up to 30 V peak-to-peak, so switching the front panel attenuator from −20 dB to 0 dB could change your input signal from 3 V peak-to-peak to 30 V peak-to-peak and potentially destroy the circuit under test. Where such risks are present it can be useful to use a fixed inline attenuator on the output of the signal generator to guard against accidents.

Temperature testing

Intermittent faults that show up after a device has been switched on for a while or appear to be random in nature can often be tracked down by measurement and forcing of temperature. Many faults in electrical equipment are in actuality mechanical failures of solder joints or cracks in tracks or wires. Most of the latter are sensitive to expansion and contraction as temperature changes. A hairdryer and a can of CFC-free freezer spray are very useful tools for tracking these faults down.

Gentle application of heat from a hairdryer about 15–20 cm away from the circuit board can bring it up to 70°C quite quickly and this is often enough to enable detection of cracked resistors and dry solder joints. Freezer sprays can be directed at individual components and can quickly bring the temperature down to −20°C or below; overheating voltage regulators can be brought out of thermal shut down temporarily at least, and sources of heat can be quickly identified by dint of watching the frost dissipate from an area of circuit board as it heats up in operation.

A thermometer or multimeter with thermocouple probe is called for when measuring the temperature of heatsinks and semiconductor packages or the ambient temperature inside equipment. Data-logging multimeters are very useful for observing the temperature rise of operating equipment, and often the temperature will correlate with the failure mode and speed diagnosis.

Mains work

Testing and debugging mains equipment is exactly the same as for any other electronic circuit except that mistakes can kill you.

Always use earth leakage circuit breakers to protect you from mains-powered equipment under test. Remember that mains equipment *should* be designed to fail safe but *never* assume that it **is** safe.

Earth-leakage circuit breakers (ELCBs) of 10 mA should be used with any mains-operated equipment that is being tested. Figure 19.6 shows a typical consumer market plug-in unit. Always use the test button on an earth leakage trip to ensure that it will break the circuit in the event of an earth

Figure 19.6

A typical plug-in earth leakage circuit breaker and plug-in mains power meter.

fault – although failure is rare you don't want to find out by getting a possibly lethal shock.

When carrying out any work that involves operating mains-powered equipment an isolating transformer can help to make the equipment safer. Because the equipment is isolated from the mains there is no ground path back to the local mains neutral, so lethal ground currents are unlikely. Many engineers say that when working on mains voltage equipment protected by an isolating transformer you should keep one hand in your pocket, out of the way of potentially completing a high-voltage circuit across your chest from one part of the equipment to another. This is good advice, and even better advice if the use of an isolating transformer is not possible.

It is not just the risk of shock that makes mains equipment potentially dangerous; the available energy is much higher than in most low-voltage devices. Short out a 5 V circuit with a screwdriver and you might blow the

fuse but you are unlikely to do any damage; short out a 400 V circuit and you can weld the screwdriver to the contacts and produce a very loud bang.

Plug-in mains power meters have become available at low cost (Figure 19.6), and they are very useful for testing equipment to see if the power supply is drawing current from the mains – without opening things up and connecting meters to mains wiring.

Testing

Troubleshooting is best performed by understanding the circuit, and following a divide-and-conquer technique. Break the problem down logically and use every symptom and characteristic of the ailing circuit to help point to the fault. It is an oft-said truth that the commonest cause of apparent failure to operate is flat or missing batteries, unconnected or switched-off mains supply, or a blown fuse, so the first thing to do when a piece of equipment fails to show any sign of life is to check the obvious. If the equipment has a separate power supply, it is worth checking that the supply is providing voltage. Power supply circuits, particularly the 'wall warts' and inline power supply blocks, are becoming common to consumer equipment and are less reliable than most other parts of these systems, This is mostly because they are very compact, poorly ventilated and run hot deliberately to make most efficient use of the (ferrite) magnetic core of the transformer as well as cost and weight constraints that limit the size of the heatsink for switching power supply transistors. This high temperature often results in drying out of electrolytic capacitors, and thermal cycling due to repeated switching on and off can cause the failure of solder joints owing to thermal coefficient of expansion differences between solder, component leads and PCB material.

Electronic component failure is the least likely cause of a piece of properly designed equipment failing; almost all consumer electronic devices have mechanical components of constructions that are much more prone to failure. The classics are power supply and headphone sockets either becoming detached from the solder joints that connect them to the PCB, or the contacts being permanently bent by repeated use. Also in this set is the noisy volume control potentiometer that gives no sound over most of its range and then works well at very high volumes due to the track being

either worn out or dirty. Motors are another difficult component; most CD players and video recorders that fail in service are victims of mechanical damage to a switch or of a motor failing; usually it is the rotor contacts, or a broken wire.

The best way to improve your troubleshooting skills is practice and understanding how equipment should work and how it is failing to work, and then how to fix it, so building your experience.

Further reading

Horowitz, P and Hill, W. (1989) *The Art of Electronics*, Cambridge University Press.

Pease, R. (1993) *Troubleshooting Analogue Circuits*, Newnes: Butterworth-Heinemann.

APPENDIX A

STANDARD METRIC WIRE TABLE

Standard metric wire table

Diameter (mm)	Resistance (Ω/m)	Current rating (mA)	Weight (g/m)
0.025	35.1	2.3	0.000044
0.032	21.4	3.7	0.000072
0.040	13.7	5.8	0.000112
0.050	8.8	9.1	0.000175
0.063	5.5	14.5	0.000278
0.080	3.4	23.4	0.000449
0.100	2.2	36.5	0.000701
0.125	1.4	57.1	0.001096
0.140	1.1	71.6	0.001375
0.160	0.86	93.5	0.001795
0.180	0.68	118.3	0.002272
0.200	0.55	146.1	0.002805
0.250	0.35	228.3	0.004384
0.280	0.28	286.3	0.005499
0.315	0.22	362.4	0.006959
0.400	0.14	584.3	0.011222
0.450	0.11	739.5	0.014203
0.500	0.088	913.0	0.017534
0.56	0.070	1140	0.021995
0.63	0.055	1450	0.027837
0.71	0.043	1840	0.035356
0.75	0.039	2050	0.039452
0.80	0.034	2340	0.044887
0.85	0.030	2640	0.050673
0.90	0.027	2960	0.056810
0.95	0.024	3300	0.063298
1.00	0.022	3650	0.070136

The values of resistance per metre, current rating, and weight per metre have all been rounded off. Only the smaller gauges are tabulated, representing the range of wire gauges which might be used in constructing RF and AF transformers. Weight per metre is calculated for copper wire with no insulation.

This page intentionally left blank

APPENDIX B

ARITHMETIC AND LOGIC INSTRUCTIONS TABLE

Arithmetic and Logic instructions

Mnemonic	AVR	PIC16	Z8	8051	68HC11
Add, no carry	ADD R_d,R_f	ADDWF f,d	ADD d,s	ADD a,D	ADDA, ADDB
Add with carry	ADC R_d,R_r	——	ADC d,s	ADDC a,D	ADCA, ADCB
Sub, no carry	SUB R_d,R_r	SUBWF f,d	SUB d,s	SUBB a,D	SBA, B,A
Sub with carry	SBC R_d,R_f	——	SBC d,s	——	SBCA, SBCB
AND	AND R_d,R_f	ANDWF f,d	AND d,s	ANL a,r	ANDA, ANDB
OR	OR R_d,R_f	IORWF f,d	OR d,s	ORL a,r	ORAA, ORAB
EXOR	EOR R_d,R_f	XORWF f,d	XOR d,s,	XRL a,r	EORA, EORB
1s complement	COM R_d	COMF	f,d	COM d	COMA, COMB
2s complement	NEG R_d	——	——	——	NEGA, NEGB
Set bits	R_d,K	BSF f,b	——	SETB bit	BSET
Clear bits	CBR R_d,K	BCF f,b	——	CLR bit	CLRA, CLRB
Increment	INC R_d	INCF f,d	INC, INCW	INC a	INCA, INCB
Decrement	DEC R_d	DECF f,d	DEC, DECW	DEC a	DECA, DECB
Test zero or negative	TST R_d	——	——	TSTA, TSTB	
Clear register	CLR R_d	CLRF, CLRW	CLR d	CLR a	CLRA, CLRB
Set register	SER R_d	——	——	SETB, c	BSET bits

Test and Branch instructions

Jump unconditional	JMP k	GOTO k	——	AJMP a	JMP
Jump conditional	BR(X)	BTFS(X)	JP c,d	J(X)	B(X)
Call subroutine	CALL k	CALL k	CALL d	(X)CALL	BSR/JSR
Return from sub.	RET	RETURN	RET	RET	RTS
Disable int.	CLI	——	DI		

Load and Save instructions					
Copy register	MOV R_d, R_r	MOVF	LD(X)	MOV a, r	LD(X)
Load data	LD(X)	MOVLW k	LD(X)	MOV a, r	LDA, LDB
Store data	ST(X)	MOVWF f	LD(X)	MOV r, a	STAA, STAB
Read port	IN R_d, P	——	——	IN data	——
Out to port	OUT P, R_r	——	——	OUT data	——
Push register	PUSH R_r	——	PUSH s	PUSH addr	PSH(X)
Pop register	POP R_d	——	POP d	POP addr	PUL(X)
Return from int.	RETI	RETFIE	IRET	RETI	RTI
Bit and Bit-test instructions					
Logical left shift	LSL R_d	——	——	——	LSL(X)
Logical right shift	LSR R_d	——	——	——	LSR(X)
Rotate left + carry	ROL R_d	RLF f, d	RL(X)	RL(X)	ROL(X)
Rotate right + carry	ROR R_d	RRF f,d	RR(X)	RR(X)	ROR(X)
Arith. Shift right	ASR R_d	——	SRA d	——	ASR(X)
Set carry flag	SEC	BSF f	SCF	SETB bit	SEC
Clear carry flag	CLC	BCF f	CCF*	CLR bit	CLC
Other instructions					
No operation	NOP	NOP	NOP	——	NOP

*CCF is *complement* carry flag (clears only if flag set).

This page intentionally left blank

APPENDIX C

Hex record formats

Many microcontrollers can be programmed with development programmers that interface directly with an integrated development environment like Microchip's MPLAB, via USB, serial or parallel ports. In production, devices are usually programmed with generic EPROM memory programmers and production programming equipment that includes testing of the programmed part at the extremes of the specified operating temperature and supply voltage range.

Programming memory devices often requires the data to be made available in a format that can be read by the programming equipment or software running on a host PC. Contract electronic manufacturers and semiconductor memory programming service companies usually require the data to be programmed in one of two standard formats, either Intel Hexadecimal Object File Format or Motorola S-record Format. The Microchip MPLAB Assembler has the option of exporting Intel HEX in 8-bit or 32-bit format.

Intel HEX files conform to a set format; the file is made up of records or groups of data, each record usually being on a separate line. Records always begin with the ':' which is called the record mark. Records are made up of fields; that is the data within the record is grouped according to its meaning.

The Record Length is 1 byte; therefore the maximum value is FF and the maximum number of bytes of data in the record is 255 bytes or 510 HEX digits. It is more usual to use 8-, 16- or 32-byte record lengths to make viewing and printing easier.

Table C.1 Intel hex object record format

Record mark	Record length	Offset	Record type	Data or information	Checksum
:	LL	OOOO	TT	NN NN NN . . . NN NN NN	CC
: 1 byte	00–FF 1 byte	0000–FFFF 2 bytes	00–05 1 byte	00–FF bytes N bytes	00–FF 1 byte

The **offset** is a 2-byte value, maximum value is 65535; for data records the offset indicates the address in the memory where the first byte of the data record should be programmed. The offset of the following record is thus the record length of the previous record added to its offset.

The **record type** can be one of the following values:

- data record;
- end of file record;
- extended segment address record;
- start segment address record (Intel x86 processors only);
- extended linear address record;
- start linear address record.

The length of the info or data field is determined by the record length byte. The record ends with a checksum byte which is the twos complement of the 8-bit sum of all the bytes excluding the record mark and the checksum byte. To check the record, the 8-bit sum of all the bytes including the checksum must equal zero, the sum of a number and its twos complement.

The **end of file record** always has a record length of 00 since it does not contain data and the offset is 0000 for the same reason; the record type is 01 and so the checksum must be FF, since 01 + FF = 00.

The extended segment address record contains a value that is used to set the segment that the offset address in data records appears in; this allows access to 1 Mbyte in 16 pages of 64k each. The format of the record is given in Table C.2. The default extended segment address is 00. The start segment address record, 03, and start linear address record, 05, are used only for Intel x86 architecture processors.

Table C.2 Extended segment address record

Record mark	Record length	Offset	Record type	Upper segment base address	Checksum
: 1 byte	02 1 byte	0000 2 bytes	02 1 byte	0000–FFF0 1 byte	00–FF 1 byte

The start linear address record sets the high bytes of the base address; this allows a 32-bit address range by addressing 64k pages each of 64k bytes. The default base address is 0000. The format of the linear address record is given in Table C.3.

Table C.3 Extended linear address record

Record mark	Record length	Offset	Record type	Upper linear base address	Checksum
: 1 byte	02 1 byte	0000 2 bytes	04 1 byte	0000–FFFF 2 bytes	00–FF 1 byte

An example of an intel HEX file produced by MPLAB is shown below as listing C.1.

Listing C.1 Hex listing of assembly for microchip 16F84A

```
:020000040000FA
:020000000C28CA
:080008008C0003088D000D08B7
:0800100083008C0E0C0E0900A8
:02400E00F73F7A
:00000001FF
```

The first line sets the extended linear address as zero – which is the fault value. The next 4 lines are data records, the last line is the end of file record.

Motorola S record file format

The Motorola S record format is similar to the Intel HEX record format.

Table C.4 Motorola S record format

Record Mark	Record length	Address	Data	Checksum
SN	LL	OOOO	NN NN NN . . . NN NN NN	CC
S0–S9 1 byte	0–FF 1 byte	0000–FFFF [FF] [FF] 2, 3 or 4 bytes	00–FF bytes N bytes	00-FF 1 byte

The S record type can take one of 10 values:

S0 Block header record

S1 Data record with 2-byte address

S2 Data record with 3-byte address

S3 Data record with 4-byte address

S5 Number of S1, S2 or S3 records in block – no data field

S7 Termination record for block of S3 records

S8 Termination record for block of S2 records

S9 Termination record for block of S1 records.

The checksum is the least significant byte of the ones compliment of the sum of the values making up the record length, address and data fields.

The header record, S0, usually has a zero address and the data field can contain arbitrary data; the checksum is calculated from the record length, address and data fields.

The 2-byte address record, S1, contains data; the address field indicates the absolute load address of the first byte of data. The maximum record length is 255 bytes with the address length and the record length byte included. The maximum number of data bytes is therefore 250 to 252 bytes. The S2 and S3 records are similar except that the address fields are longer.

The termination records, S7, 8 and 9 do not contain data but the address field can contain the entry point address for execution of object code if the records are being downloaded to an emulator, usually the address record is zero for EPROM programming.

Listing C.2 Example S record file

```
S0030000FC
S10B0000FF00FF00FF00FF0003
S9030000FC
```

The first line is the header (it contains no data); the second line contains 8 bytes of data, 2-byte address and 1-byte record length, thus the record length is 11 bytes. The terminator record contains a zero address.

This page intentionally left blank

APPENDIX D

Gerber data format

Gerber data (RS274D) and extended Gerber (RS274X) provide an effective and generic way of providing manufacturing data for use by PCB manufacturers; the files are human readable ASCII text which means that they can be modified and generated by hand when necessary.

Gerber data files were designed to convey control information to GSI photo plotters and as such the data structures that they use are based on the way that the hardware of the Gerber Scientific Photo Plotters worked.

Essentially the photo plotter was a flat bed onto which a piece of photographic film was attached; the plotter then moved a light source over the film, turning it on and off as required to expose areas of the film. In order to make different width, a wheel with holes of different diameter in it was placed between the light source and the film; this wheel, called the aperture wheel, could be rotated by a stepper motor to select the required hole. If a pad was required rather than a track the source could be flashed, producing a mark on the film that was of the same shape as the hole in the wheel. The number of apertures that the machine could select from was limited by the size of the wheel. Since the wheel was fixed to the machine a table of apertures was usually provided to allow the PCB layout software to export data using the available apertures.

At its simplest, Gerber data consists of program control instructions, machine set-up instructions and tool and movement (aperture and lamp) instructions; all instructions or codes terminate with the character '*' ASCII $2A_H$.

Control instructions are M-codes; these were used to stop the machine to allow the operator to change the aperture wheel, etc. and to indicate the end of the program. The program 'stop' and 'end of program' commands are often used together to indicate the end of file.

M00 Program Stop

M01 Optional Stop

M02 End of Program

The **tool instructions** are D-codes (sometimes called draft codes); these define the state of the light source and selection of the aperture.

D01 Lamp on; moves following a D01 will draw lines

D02 Lamp off; moves following a D02 will not draw lines

D03 Flash lamp; this draws pads

D10 – D999 select a pre-defined aperture; older machines may not accept codes above D72

The **G-codes** specify how to use the coordinate data provided, that is the interpolation, units, etc.

G00 Move

G01 Linear interpolation

G02 Clockwise circular interpolation (not supported by all plotters)

G03 Anti-clockwise circular interpolation (not supported by all plotters)

G04 Ignore data block – used for comment text

G10 Linear interpolation 10 × scale

G11 Linear interpolation 0.1 × scale

G12 Linear interpolation 0.01 × scale

G36 Start polygon fill (not supported by all plotters)

G37 Complete polygon fill (not supported by all plotters)

G54 Change tool

G70 Use inches

G71 Use mm

G90 Absolute format

G91 Incremental format.

Coordinate data defines the position of the aperture over the film. Whenever coordinate data are fed into the machine the aperture is moved to the new coordinates. Coordinates are provided in the form X0000Y0000*, which would send the aperture to the sheet origin, in absolute mode or cause no movement in incremental mode. The move command (G00) is usually assumed when coordinate data are given unless another command is in force.

Listing D.1 This draws a box 50000 × 50000 using aperture D10 with top right corner at 100000,100000, before returning the aperture head to the origin

```
G70*
G90*
G01*

G00X0Y0*
G54D10*
X100000Y100000D02*
X100000Y50000D01*
X50000Y50000D01*
X50000Y100000D01*
X100000Y100000D01*
X0Y0D02*
M00*
M02*
```

The polygon fill codes G36 and G37 provide an easy way of generating fills, for ground planes, the alternative is filling the area using overlapping drawn lines, but this can make the Gerber file very large if complex polygons are to be filled. Polygon filling became better supported when laser raster plotters replaced the original aperture XY plotters.

Listing D.2 This draws a filled box 50000×50000 using G36/37 polygon fill, with top right corner at 200000,100000, before returning the aperture head to the origin. Note that no aperture needs to be selected because this is determined by the plotter itself.

```
G70*
G90*
G01*

G00X0Y0*
G36*
X200000Y100000D02*
X200000Y50000D01*
X150000Y50000D01*
X150000Y100000D01*
X200000Y100000D01*
G37*
X0Y0D02*
M00*
M02*
```

Circular interpolation is not well supported by hardware plotters so is not recommended for use, and will not be covered here.

RS274X

RS274X or extended Gerber adds the ability to embed the aperture table into the file, which became necessary when raster plotters started to replace wheel plotters. It also adds several other features; the main ones are covered below.

RS274X parameters are usually placed at the beginning of a file to define how the data in the file is to be handled by the plotter; they are delimited

by a % sign at the beginning and end of each command block. The block also has the usual Gerber terminator * before the last % sign.

MO mode

MO defines the units of dimensions, IN for inches and MM for millimetres.

```
%MOIN*%

%MOMM*%
```

LN layer name

The layer name information can be placed in the file using the LN command. Up to 77 ASCII characters may follow the LN, excluding '*', which is the terminator. For example:

```
%LNTop_Copper*%

%LNBottom_Copper*%
```

LP layer polarity

The layer polarity statement determines whether a layer is plotted as black or a cut-out. It takes either a **C** for clear or a **D** for dark as parameters. This is useful when using multiple layers to define tracks crossing ground planes in cut-outs, so you use a dark for the ground plane, clear for the cut-out and a dark for the track.

```
%LPC*%

%LPD*%
```

FS format statement

The FS statement takes the parameters L or T for leading or trailing zeros omitted in data; A for absolute or incremental coordinate data; Xn and Yn for the positon of the X and Y data decimal points.

```
%FSLAX23Y23*%
```

The X23 and Y23 parameters indicate that the X and Y coordinate data have 2 digits before and 3 digits after the decimal point. If inches were in use, the largest dimension that could be plotted would be 99.999 inches.

AD aperture description

The AD statement is used to add aperture data to the file. The parameters are the D code to be defined, e.g. D10 followed by the aperture type C for circular, R for rectangle, O for oval. The apertures may be specified as hollow; that is they may have a hole inside, and this hole must be specified to be smaller than the aperture. The following apertures defined are D10 circular solid 0.005 diameter, D11 circular hollow, 0.005 with 0.002 hole, D12 circular 0.005 with square hole 0.002 × 0.002. The rectangular apertures are similar, D13 solid square 0.005 side, D14 solid rectangle 0.005 × 0.002 and D15 hollow square 0.005 side with 0.002 round hole.

```
%ADD10C,0.005000*%

%ADD11C,0.005000X0.00200*%

%ADD12C,0.00500X0.00200X0.00200*%

%ADD13R,0.005000X0.005000*%

%ADD14R,0.005000X0.002000*%

%ADD15R,0.005000X0.005000X0.002000*%
```

Examples

The example of Gerber data below produces the plot at Figure D.1 – not all functions are supported by the viewer so the square hole specified in D12 is plotted as a circular one; square holes in pads are an unusual requirement, and this is something to be aware of if using features of RS274X. Check

with your photo plotting bureau or PCB manufacturer. CAD packages usually allow circular interpolation (hardware arcs) and polygon fill to be turned off.

Listing D.3 Example of Gerber file

```
%FSLAX26Y26*%
%MOIN*%
%ADD10C,0.005000*%
%ADD11C,0.005000X0.00200*%
%ADD12C,0.00500X0.00200X0.00200*%
%ADD13R,0.005000X0.005000*%
%ADD14O,0.005000X0.002000*%
%ADD15R,0.005000X0.005000X0.002000*%

G70*
G90*
G01*
G00X0Y0*
G54D10*
X100000Y100000D02*
X100000Y50000D01*
X50000Y50000D01*
X50000Y100000D01*
X100000Y100000D01*
X0Y0D02*

G36*
X200000Y100000D02*
X200000Y50000D01*
X150000Y50000D01*
X150000Y100000D01*
X200000Y100000D01*
G37*

X0Y0D02*
G54D11*
X5000Y110000D03*
G54D12*
X15000Y110000D03*
G54D13*
X25000Y110000D03*
```

```
G54D14*
X35000Y110000D03*
G54D15*
X45000Y110000D03*

M00*
M02*
```

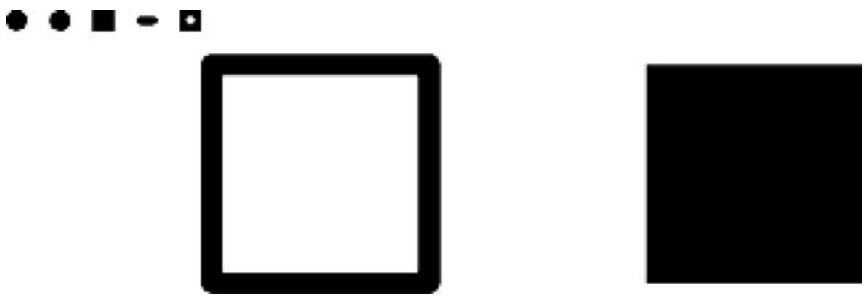


Figure D.1

The result of plotting the Gerber file of listing xyz.3.

The layer polarity settings, LPC and LPD allow cut-outs to be made in planes, as shown in the example below.

Listing D.4 Example of Gerber File, showing the use of dark and clear layers

```
%FSLAX26Y26*%
%MOIN*%

%ADD10C,0.005000*%
%ADD16C,0.001000*%

G70*
G90*
G01*
%LPD*%

G36*
X300000Y100000D02*
```

```
X300000Y50000D01*
X250000Y50000D01*
X250000Y100000D01*
X300000Y100000D01*
G37*

%LPC*%
G54D10*
X210000Y160000D02*
X280000Y75000D01*

%LPD*%
G54D16*
X210000Y160000D02*
X280000Y75000D01*

M00*
M02*
```

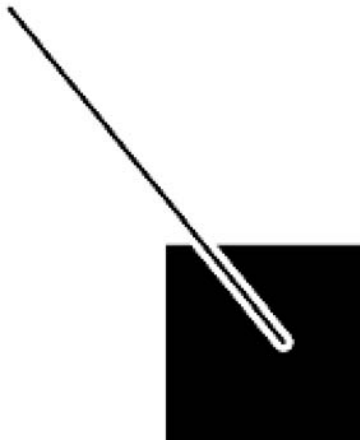


Figure D.2

The result of plotting the Gerber file of listing D.4.

This page intentionally left blank

APPENDIX E

Pinout information links

http://www.kingswood-consulting.co.uk/giicm/
Pinout finder (with a very odd name)

http://www.etechhelponline.com/
Semiconductor data sheets arranged by manufacturer

http://www.xs4all.nl/~ganswijk/chipdir/pinusr/index.htm
Part of a chip directory with ICs arranged by number

http://www.chipdocs.com/index.html
Full documentation for semiconductor devices (need to register)

http://www.datasheetlocator.com/
Free service for finding datasheets from specified manufacturer (more than 860 manufacturers listed)

http://www.ee.washington.edu/circuit_archive/parts/cross.html
Transistor cross-reference database

http://www.technick.net/public/code/pinouts.php
Pinouts for connectors, etc.

http://pinouts.ru/
Pinout links for all types of hardware

This page intentionally left blank

APPENDIX F

SMT packages and guides

http://www.radio-electronics.com/info/data/smt/smt_packages.php
List of standard package types

http://www.nickc.com/packageinfo.htm
Packages with dimensions and outline drawings

http://www.engineeringlab.com/landpatterns.html
Land patterns for SMT, subscription required

http://www.dprg.org/tutorials/1999-07a/
Beginner's guide to SMT

http://www.intel.com/design/packtech/ch_09.pdf
Solder reflow guidance

http://www.smta.org/index.cfm?EXPAND_ALL=TRUE
Surface Mount Technology Association

http://smt.pennnet.com/Articles/Article_Display.cfm?&Section =Articles&SubSection =Display&ARTICLE_ID=84836
Moisture sensitivity of SMT components

http://download.micron.com/pdf/technotes/tn0011.pdf
Recommendations for BGA assembly

http://www.smartgroup.org/linklist.asp
European SMT trade association

http://ww1.microchip.com/downloads/en/AppNotes/DS-00598a.pdf
Effect of heat on plastic packaging during SMT soldering

http://www.standardics.philips.com/packaging/handbook/pdf/pkgchapter5.pdf
SMD handling, soldering and desoldering

http://www.circuittechctr.com/guides/4-7-3.shtml
BGA pad repair

INDEX