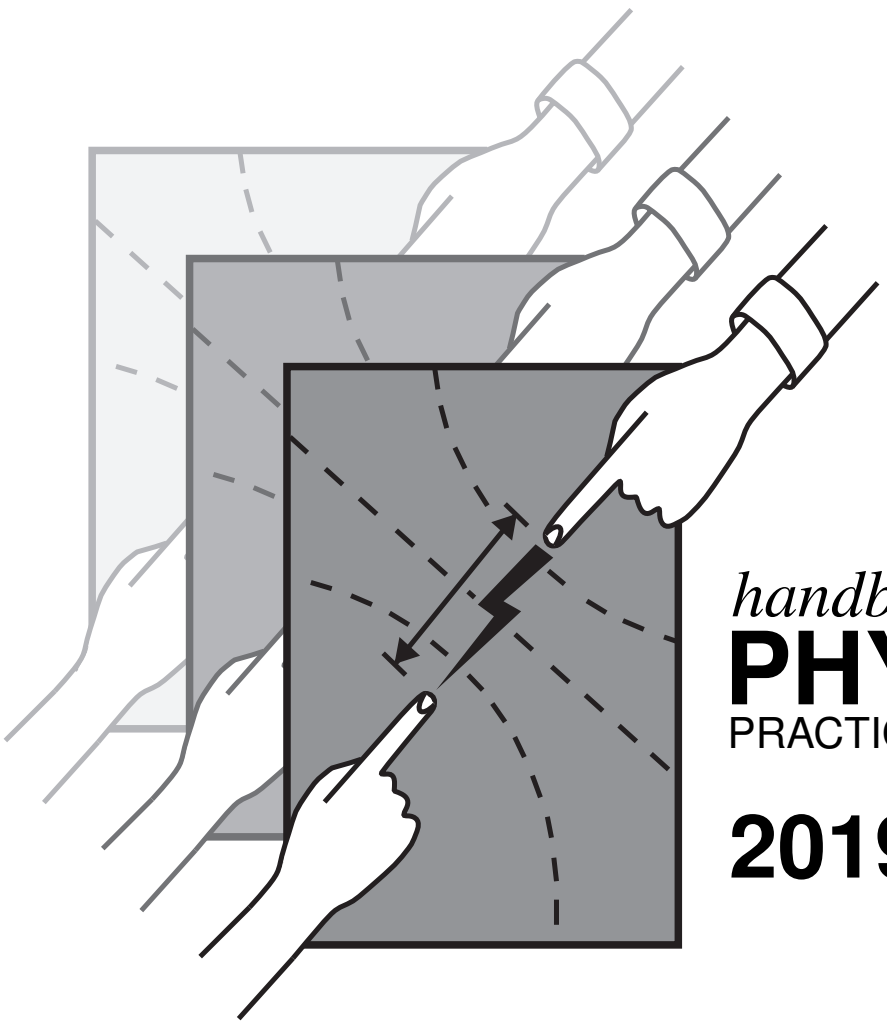
Oxford Physics



handbook of the

PHYSICS

PRACTICAL COURSE

2019–2020

Electronics Manual

UNIVERSITY OF OXFORD

PREFACE *to the November 2019 edition*

The purpose of this manual is to provide concise theoretical and technical background material to support the electronics laboratory work in the Oxford Physics course. You should read the basic material in chapter 1 before you come to the laboratory for the first time. The theory in chapters 2, 3, and 4 will also be needed for the first year practical sessions. The instruments you will use and techniques of measurement are described in chapters 6, 7, and 8. Chapters 1, 2, and 3 cover all the theory needed to answer prelims questions on circuits. Compared with an electronic engineer's view, a physicist's view of electronics will involve a greater awareness of the fundamental theory and physical processes behind the models (equivalent circuits) used to describe real circuits and devices, but a lesser awareness of technological detail. The material presented has been chosen with this in mind. The derivation of circuit concepts from Maxwell's equations given in chapters 21 and 22 cannot be properly appreciated until most of the electromagnetism theory in part A has been covered (which might not be until the end of second year). The derivation is important for a proper understanding of circuit theory, not least to bring out the approximations involved, and you should look at it when you have the background. It is missing from most textbooks.

We thank Guy Peskett for writing this manual and modifying it over a number of years. Please report any errors or omissions to labhelp <labhelp@physics.ox.ac.uk>, so that it can be improved for future years.

2012–2019 Electronics Course Manual *version 1.3*

1 Linear Circuit Theory And How To Apply It

1.1 What is a Circuit?

A circuit is an electromagnetic system that can be discussed to the desired accuracy without considering loss of energy by radiation. Usually this means systems whose largest dimension is much smaller than the wavelength of radiation at the frequency at which the system is being excited. A radio transmitter is likely to be a circuit but the aerial it is feeding cannot be.

Given the sizes of typical hardware and the speed of light the range of frequencies over which the circuit approximation applies turns out to be from 0 Hz up to a few 10^9 Hz. The size criterion means that in a circuit the effects of any changes in the distributions of charge or current are felt instantaneously throughout.

1.2 Kirchhoff's laws

Extending the work of Ohm to enable the currents in networks of resistances to be found systematically Kirchhoff published his two laws in 1845, 28 years before the first edition of Maxwell's treatise on electromagnetism. Maxwell's theory shows that the laws have a wider application than Kirchhoff envisaged and apply to circuits with components other than resistances. The laws will be discussed below in the light of Maxwell's theory.

1.2.1 Kirchhoff's first law (K1)

Consider the fragment of circuit shown in Figure 1.1 below containing two parallel conducting plates (a capacitor C) and a junction of three wires. S is the surface of the region of space containing the fragment of circuit being discussed.

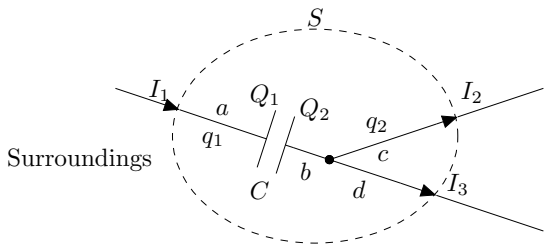


Figure 1.1: Charges on a circuit fragment.

When the circuit is excited there will be (generally time dependent) charges on the conductor surfaces. These are shown in the diagram as Q_1 and Q_2 on the facing surfaces of the plates of the capacitor, a charge q_1 on wire a and the outside surface of the left hand plate of the capacitor and a charge q_2 on wires b, c, d, and the outside surface of the right hand plate of the capacitor. The charges q_1 and q_2 can be thought of as being on the plates of small capacitors the other plates of which are in the surroundings. Typically these *stray* capacitors will have capacitances of a few pF. The capacitors found in circuits operating at frequencies up to a few MHz will usually have much larger capacitances than this.

Kirchhoff's first law (K1) follows from a statement of the conservation of charge which says that the sum of the currents entering a region of space containing a fragment of a circuit is equal to the rate of accumulation of charge on the fragment, in our case:

$$I_1 - I_2 - I_3 = \frac{\mathrm{d}(Q_1 + Q_2 + q_1 + q_2)}{\mathrm{d}t} \tag{1.1}$$

Now Gauss tells us that the charges on the facing surfaces of the plates of a capacitor are always equal and opposite i.e. $Q_1 + Q_2 = 0$, so we are left with

$$I_1 - I_2 - I_3 = \frac{\mathrm{d}(q_1 + q_2)}{\mathrm{d}t} \tag{1.2}$$

The net current flowing into the fragment is only that needed to change the charges q_1 and q_2 on the stray capacitors. It is hard to be definite about the frequency at which this net current needs to be taken into account but it can safely be ignored at audio frequencies leading to the usual statement of Kirchhoff's first law

$$I_1 - I_2 - I_3 = 0 \tag{1.3}$$

It is worth noting that it also follows from Gauss' law that the current supplying charge to the inside surface of the left hand plate of C and the current taking charge away from the inside surface of the right hand plate are always equal. It is as if the current flows through the capacitor.

1.2.2 Kirchhoff's second law (K2)

Kirchhoff's second law is concerned with the forces acting on electrons in circuits. Before getting to the law we must say something about these forces.

Batteries exert forces as a result of chemical action. They tend to pull free electrons[1] into terminals we call positive and push them out of terminals we call negative. Consider a battery without anything connected to it. There will be an excess of free electrons on its negative terminal and a deficit (positive charge) on its positive terminal. The charges will have built up until the Coulomb forces $\mathbf{F}_{\text{charges}}$ they exert on each free electron is equal and opposite to the force generated by the chemical action.

Next consider connecting to the battery a length of wire bent into a tight hair-pin shape so that the area of the loop formed is negligible. The charges on the battery terminals apply forces to the free electrons at the ends of the wire. The forces are then transmitted rapidly by mutual repulsion to all the free electrons in the wire causing a current to flow. The chemical action pushes electrons through the battery replenishing the charges on the its terminals.

The flowing electrons lose momentum through collisions with the fixed positively charged ions which make up the bulk of the wire. The average rate of loss of momentum by an electron is the average drag force $\mathbf{F}_{\text{drag}}$ on it (Newton) - there is *resistance* to the flow.

For the materials of which wires are made the average rate of loss of momentum by an electron is found to be very close to proportional to its average momentum, i.e. the drag force is very close to proportional to the current. The constant of proportionality depends on the material and temperature of the wire (Ohm).

Finally consider the situation when the wire connected to the battery is a perfect conductor (no drag force) but the loop formed has significant area. The current will increase producing an increasing magnetic flux through the loop causing an electric field (Faraday) and a force on a flowing electron $\mathbf{F}_{\text{changing flux}}$ which opposes (Lenz) the force from the battery.

If we consider only systems made of linear materials[2] Maxwell's theory tells us that these three types of force, due to charges built up, collisions, and changing linked magnetic flux are the only ones that a circuit can produce to oppose an applied force.

Batteries generate steady forces. In a circuit excited by an oscillating applied force, oscillating opposing forces will arise and these will be out of balance with the applied force by an amount depending on the inertia of the electrons. The inertia affects the time between the collisions the free electrons make with the wire atoms and therefore the time for the average flow and the average drag force to be established. Now the time between collisions is very short, for copper it is about 10^{-14} seconds, so for frequencies up to a few 10^9 Hz the drag force, and (given circuits are small in the sense described in the introduction) the other two forces as well, can be assumed to build up to oppose the applied force instantaneously - we can ignore the electrons inertia.

Summarising, we say that *in a circuit the total opposing force provoked (due to charges built up, drag, and changing currents) is in balance with the applied force at all points and all instants*, i.e.

$$\mathbf{F}_{\text{applied}} + \mathbf{F}_{\text{charges}} + \mathbf{F}_{\text{drag}} + \mathbf{F}_{\text{changing flux}} = 0 \tag{1.4}$$

This force balance equation becomes Kirchhoff's second law (K2) when it is converted[3] into a statement about what happens in a *path* in a circuit:-

$$\mathcal{E} = \Delta V_{\text{cap}} + \Delta V_{\text{res}} + \Delta V_{\text{ind}} \qquad \text{(K2)} \tag{1.5}$$

The terms are no longer forces and are given new names.

$\mathcal{E}$ is called the *electromotive force* or *emf* applied to the path. The quantity on the right-hand side is called the total *voltage drop* in the path.

[1] In a typical metal of the order of one electron in each atom becomes free leaving the atom positively ionised

[2] For example, a linear conductor is one in which the current density is exactly proportional to the electric field. Such a conductor is said to obey Ohms law

[3] See chapter 21 when you have completed your second year electromagnetism course

(i) emf

Calling the voltages at the ends of a path in a circuit to which an emf is applied V_1 and V_2, so $V_1 - V_2$ is the total voltage drop, K2 can be written

$$\mathcal{E} = V_1 - V_2 \tag{1.6}$$

a symbolic version of which is shown in Figure 1.2. The emf is represented by a symbol consisting of a circle with

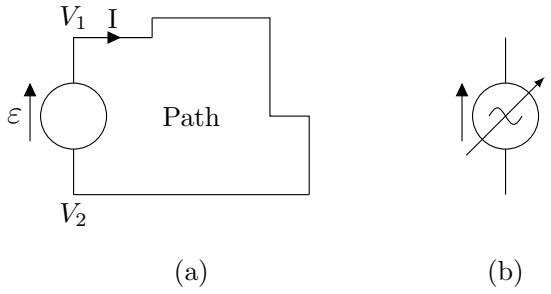


Figure 1.2: Applied emf and total voltage drop

an arrow alongside it showing the direction in which a positive $\mathcal{E}$ tends to drive a current. Note that V_1 is at the head of the emf arrow. (There is always a perfectly conducting path through an emf even when it is turned down to zero.) A diagonal arrow over the circle, see symbol for sine wave emf (b), is used to indicate that the magnitude and/or the frequency of an emf is manually adjustable.

Emf has units of work/charge which are volts not Newtons so use of the word force is slightly unfortunate. The value of an emf in volts is numerically equal to the work done in Joules when 1 coulomb of charge flows in the direction of the emf. It follows that the power delivered by an emf is $\mathcal{E}I$ where I is the current through the emf in the direction shown. In terms of the total voltage drop in the path the power delivered is $(V_1 - V_2)I$.

(ii) voltage drop

The three possible types of contribution to the total voltage drop, deriving from drag, charges built up, and changing magnetic fields, are called the *resistive*, *capacitive*, and *inductive* voltage drops respectively. Unfortunately you won't be able to derive their functional forms[1] until you have completed your Part A electromagnetism theory. For your work on the first year circuits course you will simply have to learn them and take them on trust. They are set out in the next section. Questions involving them are set in the prelims exam.

1.3 Catalogue of voltage drops

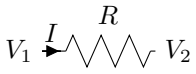
The resistive, capacitive, and inductive voltage drops turn out to be proportional to the current, the time integral of the current, and the time derivative of the current respectively, (all linear relations). The definition of each voltage drop consists of a symbol, a labelling convention, and an equation. The positive direction of the current I is indicated with an arrow. Note how the direction of the current arrow relates to the order in which the differences $(V_1 - V_2)I$ are taken. The symbols are used in the construction of models of circuits, of which more later.

1.3.1 Resistance

A resistive voltage drop is represented by the symbol shown in Figure 1.1. The voltage drop is proportional to the current and with the labelling shown is given by equation 1.7 where R = constant, units ohms (Ω).

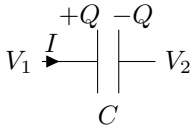
1.3.2 Capacitance

A capacitative voltage drop is represented by the symbol shown in Figure 1.2. The charges Q and $-Q$ are on the facing surfaces of the plates. The voltage drop is proportional to the time integral of the current and with the labelling shown is given by equation 1.8 where C = constant with units farads (F), Q_0 is the value of Q at $t = 0$, and as labelled $I = +\frac{\mathrm{d}Q}{\mathrm{d}t}$.



$$V_1 - V_2 = IR \qquad \text{(Ohm's law)} \tag{1.7}$$

Table 1.1: Definition of resistance



$$V_1 - V_2 = \frac{1}{C}\int_{-\infty}^{t} I\,\mathrm{d}t = \frac{Q_0}{C} + \frac{1}{C}\int_{0}^{t} I\,\mathrm{d}t = \frac{Q}{C} \tag{1.8}$$

Table 1.2: Definition of capacitance

1.3.3 Self inductance

An inductive voltage drop is represented by the symbol shown in Figure 1.3. The voltage drop is proportional to the time derivative of the current and with the labelling shown is given by equation 1.9 where L = constant, units henries (H).

Additional definitions and remarks

1.3.4 Ideal interconnection

To use the symbols given above to build models of real circuits we need a symbol to represent a conducting path with no voltage drop between its ends regardless of the current flowing to link the symbols together. We call this an *ideal interconnection*. It is represented simply by a plain line as shown in Figure 1.4.

1.3.5 Coupled inductances (Transformers)

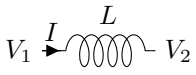
When two inductances are positioned so that the magnetic field due to a current in one links with the other an emf is induced in one when the current changes in the other. A representative symbol for coupled inductances is shown in Figure 1.3. Each path has two symbols, one representing the self inductive voltage drop, the other the induced emf caused by the changing current in the other path. In equations 1.11, L_p and L_s are constants, units henries (H). The same constant M, called the *mutual inductance*, units Henries, appears in both branches.

1.3.6 Ideal opamp

An ideal opamp is an emf which has the same time dependence as the voltage drop between two input terminals but is very much larger. The symbol consists of an emf ($\mathcal{E}_A$) enclosed in a triangle as shown in Figure 1.4. The input terminals draw no current.

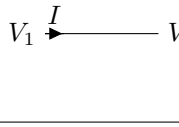
A rise in the voltage V_1 applied to the *non-inverting input*, labelled +, causes V_{out} to rise, a rise in V_2 causes V_{out} to fall. The defining equation and K2 are given in equations 1.12. The constant of proportionality A is positive and very large, essentially infinite, and V_{ref} is usually the zero of voltage. (See the remarks on assigning voltages in section 1.6).

[1] Look at chapter 22 next year.



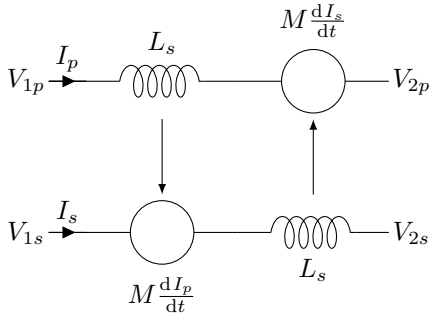
$$V_1 - V_2 = L\frac{\mathrm{d}I}{\mathrm{d}t} \tag{1.9}$$

Table 1.3: Definition of self-inductance



$$V_1 - V_2 = 0 \quad \text{for all } I. \tag{1.10}$$

Table 1.4: Definition of ideal interconnection



$$V_{1p} - V_{2p} = L_p \frac{dI_p}{dt} \pm M\frac{dI_s}{dt}$$
$$V_{1s} - V_{2s} = L_s \frac{dI_s}{dt} \pm M\frac{dI_p}{dt} \tag{1.11}$$

Figure 1.3: Definition of mutual inductance

1.4 Representation of signal sources

When a source of emf is connected to a circuit and a current flows the voltage drop at the terminals of the source is found to be less than the emf and the difference is usually proportional to the current flowing. To allow for this sources are represented as emfs in series with resistances called *internal resistances*, see Figure 1.5(a)[1]. This depiction of a source is called a *Thevenin* representation.

There is a formal alternative to this representation of sources based on the concept of an element which produces the same current regardless of the voltage across it (the *dual* of an emf). The symbol is –◯–, and it is usually called a *current generator*. It appears in parallel with the internal resistance r_{int} as shown as R_{int} in Figure 1.5(b). This depiction of a source is called a *Norton* representation. It is left as an exercise to show that the internal resistance is the same in both representations and that $\varepsilon = I r_{\text{int}}$.

The currents and voltage drops we calculate in a circuit are the same whether we use the emf representation (b) or the current generator representation (c) for the source.

Which representation we use is just a matter of convenience. If the components in the circuit connected to the source are in series, version (a) usually leads to simpler equations, if the components are in parallel a simpler analysis will result from using version (b).

1.5 Equivalent circuits

Physicists trying to understand a part of the real world begin by constructing a *model* of it. In electronics the models used are called *equivalent circuits*.

An equivalent circuit is a drawing showing all the paths along which significant current is thought to flow. The places where currents divide or come together are called *nodes* and are emphasised with black dots. The paths between neighbouring nodes are called *branches*.

The symbols introduced above are used to indicate the kinds of voltage drop present in a branch. In a branch where the force opposing the applied emf is a superposition of several different forces that branch will show the relevant symbols in series.

In principle the paths shown in an equivalent circuit will include not only the obvious ones like the wire or conducting film in a resistor, the wire in an inductor, and paths containing a capacitor but also the capacitative and resistive leakage paths within the components, between the components, and between the components and the surroundings. In addition there may be significant mutual inductances between different parts of the circuit.

[1] After chapter 3 internal resistances may be generalised to internal impedances.

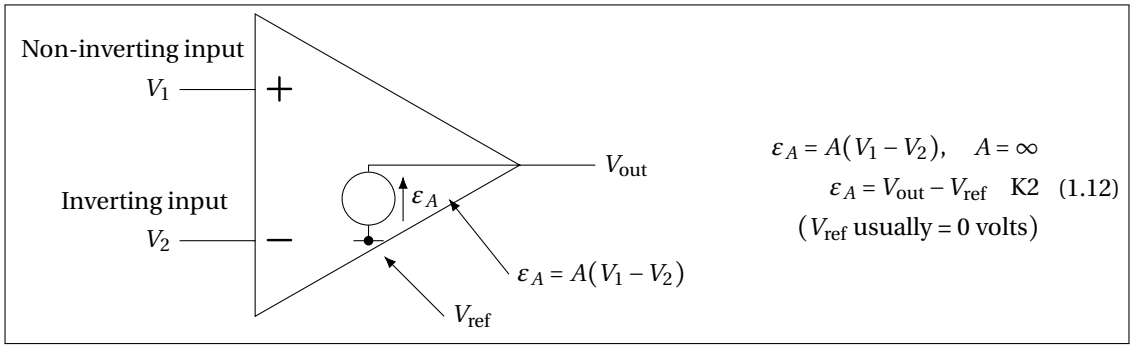


Figure 1.4: An ideal opamp

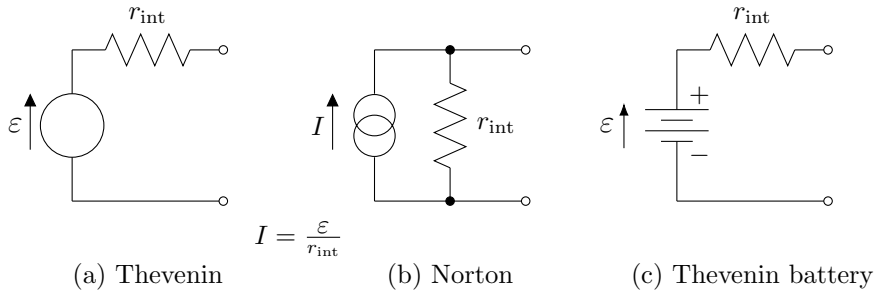


Figure 1.5: Representations of signal generators.

The number of these *stray* effects we need to include depends on the accuracy we desire from our model and the highest frequency we are considering. More current flows along small capacitance paths and induced emfs are larger at high frequencies so the number of strays that need to be included increases with frequency.

For circuit components used at audio frequencies the equivalent circuits shown in column 2 of Table 1.5 will usually be adequate and any capacitances and mutual inductances between components can usually be ignored. At high frequencies the equivalent circuits of inductors can be quite complicated. See e.g. Figure 1.6.

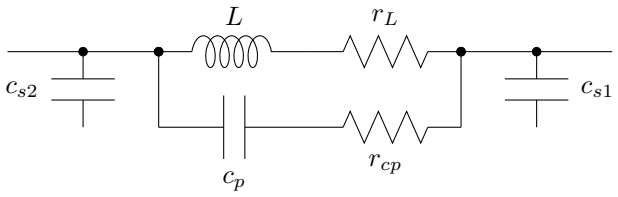


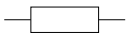
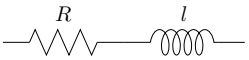
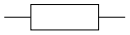
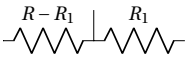
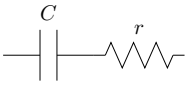
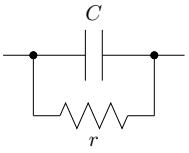
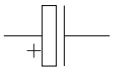
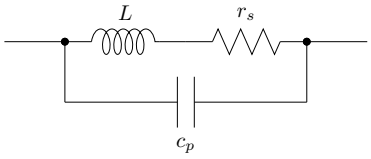
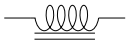
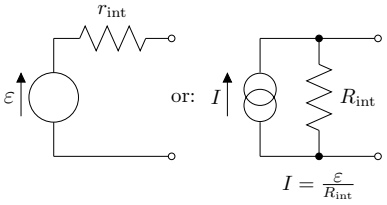
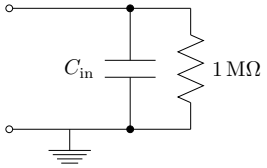
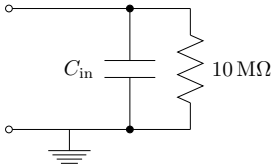
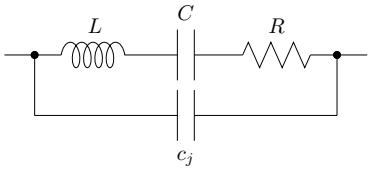
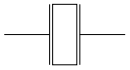
Figure 1.6: High frequency equivalent circuit of an inductor

It helps that manufacturers try to make components which are "pure" over their intended operating frequency range. For example, at audio frequencies inductors are often wound on magnetic cores. This reduces the number of turns required (and therefore the resistance), confines the magnetic field of the coil which prevents it inducing voltages in other parts of the circuit, and shields the coil from external magnetic fields. Sometimes the layers of the winding are spaced by insulating sheets to reduce the capacitances between them.

1.6 Using an equivalent circuit

Having constructed an equivalent circuit, which is the hardest and most interesting part of the process of trying to understand a real circuit, the next steps are:-

1. Assign a voltage to each node.

Component	Typical audio frequency equivalent circuit *	Circuit diagram symbol
Connecting wire		
Metal film resistor		
Wire-wound resistor		
Potentiometer		
Metallised film capacitor		
Electrolytic capacitor		
Laminated core inductor		
Signal source		
Scope input		
DVM input		
Quartz crystal		

∗ Lower case labels used for strays.

Table 1.5: Audio frequency equivalent circuits and circuit diagram symbols for various components.

Label one of the nodes, the node connected to earth if there is one, zero volts, and label every other node with a voltage with respect to this. For an ideal opamp assign the same voltage to both input terminals. (The output voltage is finite, A is infinite, so the difference in voltage between the input terminals must be zero).

2. Assign a current to each branch.

   Label each branch with a current indicating its positive direction with an arrow. It is helpful to label the currents so that K1 is satisfied. All these current and voltage labels are arbitrary choices (except that duplication must be avoided).

3. Write down the equations for the voltage drop in each branch. (This is preferable to the usual recommendation of dealing with complete loops made up of several branches because it avoids redundancy and the chance of missing out a branch).

4. Eliminate unwanted variables from the equations and obtain an equation for the wanted variable.

5. Solve the equation. (The technique used depends on form of applied emf).

The whole procedure is best illustrated with an example.

Imagine we have a circuit on a bench which we can see consists of an inductor (with an iron core), a capacitor, a resistor, an audio frequency signal generator, and an oscilloscope and we want to understand what it does. We begin by drawing a *circuit diagram* showing how the components are connected. Suppose it's as in Figure 1.7 (see Table 1.5 column 3 for the symbols). The next step is to turn the circuit diagram into an equivalent circuit.

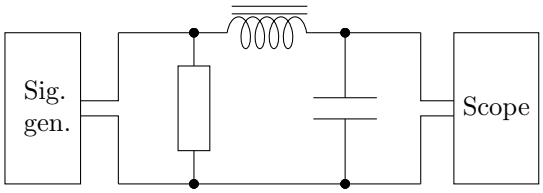


Figure 1.7: Circuit diagram of example.

We do this by replacing each symbol in the circuit diagram with the corresponding mini-network in column 2 of Table 1.5.

Finally we must decide if any capacitances or mutual inductances *between* the components need to be included. We will assume it is not necessary here (it rarely is at audio frequencies). The result is the equivalent circuit shown in Figure 1.8. The signal generator is represented by an adjustable emf in series with a resistance,

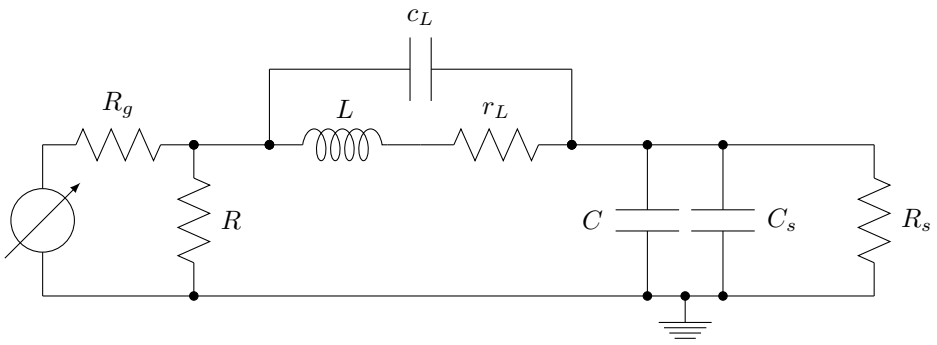


Figure 1.8: Equivalent circuit of example.

the resistance and parallel capacitance of the inductor are included as well as its inductance, and the capacitance and resistance of the oscilloscope input are shown. The connecting wires have been assumed to be ideal interconnections.

Figure 1.9 shows an example of labelling. Note that Kirchhoff's first law has been satisfied at each node by appropriate choice of current variables. Note also that C and C_s can be combined. Next, taking account of the

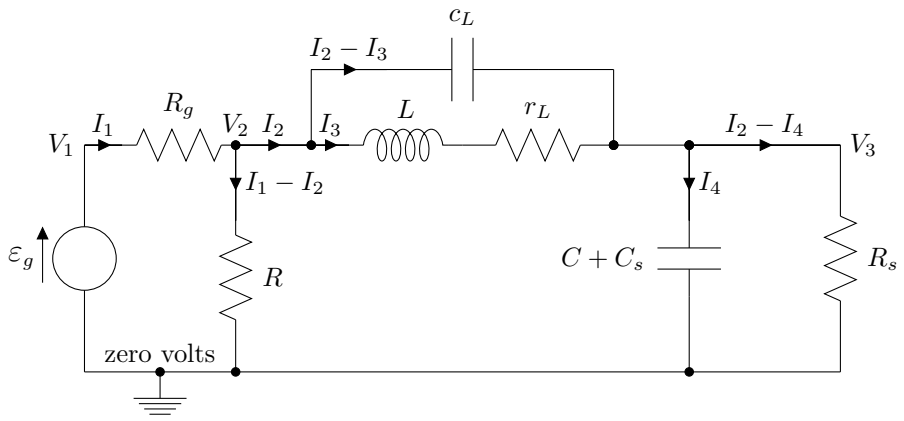


Figure 1.9: Equivalent circuit shown in Figure 1.8 with labelling of currents and voltages added.

labelling, write down K2 for the total voltage drop:

$$\mathcal{E}_g = V_1 - 0 \tag{1.13}$$

and the voltage drop for each branch

$$V_1 - V_2 = I_1 R_g \tag{1.14}$$

$$V_2 - 0 = (I_1 - I_2)R \tag{1.15}$$

$$V_2 - V_3 = L\frac{\mathrm{d}I_3}{\mathrm{d}t} + I_3 r_L \tag{1.16}$$

$$V_2 - V_3 = \frac{1}{c_L}\int_{-\infty}^{t} (I_2 - I_3)\,\mathrm{d}t \tag{1.17}$$

$$V_3 - 0 = \frac{1}{C + C_s}\int_{-\infty}^{t} I_4\,\mathrm{d}t \tag{1.18}$$

$$V_3 - 0 = (I_2 - I_4)R_s \tag{1.19}$$

No node equations are needed because K1 has been satisfied at each node by the labelling.

These equations enable us to eliminate unwanted variables and discover the magnitude and time dependence of any chosen variable. How we proceed depends on the time dependence of the applied emf $\mathcal{E}_g$.

If $\mathcal{E}_g$ is a dc emf which was switched on at $t = -\infty$ (i.e. a steady state has been reached) the currents will be constant (zero for those to capacitors). Inductances can be considered to be short circuits and capacitances open circuits so the network will reduce to one of resistances only and the result can be found by simple algebra.

If $\mathcal{E}_g$ has a time dependence which can be constructed from abrupt steps of voltage turn to chapter 2. The result is found by deriving and solving a differential equation.

If $\mathcal{E}_g$ has a harmonic (sine or cosine) dependence on time turn to chapter 3. The result is found with the aid of complex algebra.

If $\mathcal{E}_g$ is a random voltage, for example white noise, turn to chapter 20. The result is found in terms of mean square values.

1.7 Circuit diagrams

So far our discussion has been largely about equivalent circuits because these are the models we use when we want to understand what a circuit does.

If we wish to do maintenance or fault finding, equivalent circuits are inappropriate. *Circuit diagrams* in which each component is represented by a single symbol are more useful. Some such symbols are shown in column 3 of Table 1.5. (By 'components' we mean resistors, capacitors, inductors, transformers, plugs, sockets, opamps, transistors, switches, etc. so a great variety of symbols may be encountered).

Circuit diagrams often carry ancillary information such as device type numbers, pin numbering for connectors and device packages, voltages at various points, waveform cartoons, wire colour codes etc.

In this manual the presence of a resistance (zigzag) symbol indicates that a figure is an equivalent circuit. The presence of a resistor (box) symbol indicates that the figure is a circuit diagram.

2 Linear Circuits Excited By Steps And Pulses

2.1 Introduction

The method we use to find out what happens in a circuit when an emf is applied depends on the time dependence of the emf. Here we consider emfs which consist of sequences of abrupt steps. Circuits in which switches are operated are also covered briefly. The method involves the solution of differential equations and is known as "analysis in the time domain".

2.2 Step emfs

We are considering emfs with waveforms of the sort shown at the top of Figure 2.1. This family includes single

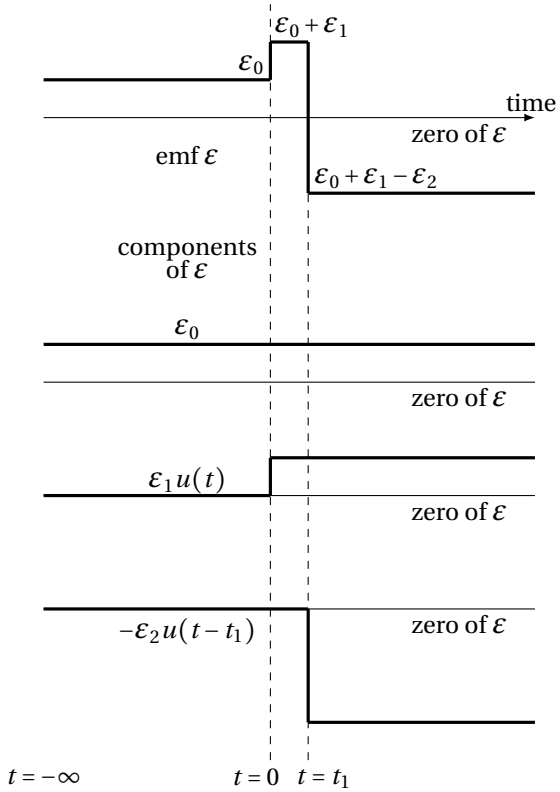


Figure 2.1: A step emf and its decomposition

steps, rectangular pulses, and staircases among others[1]. The figure also makes it clear that such waveforms can be constructed from a steady level and steps from zero. We describe them with the aid of the *unit step function*, $u(t)$. This is a dimensionless function of time which is zero for all time before $t = 0$, makes an abrupt step up to unity at $t = 0$ and remains at unity thereafter. It should be clear from the Figure that the waveform illustrated can be expressed in terms of $u(t)$ as

$$\mathcal{E} = \mathcal{E}_0 + \mathcal{E}_1 u(t) - \mathcal{E}_2 u(t - t_1) \tag{2.1}$$

[1] It also includes regularly repeating waveforms such as squarewaves and it is reasonable to ask if we would use the method about to be described for them. If the time between steps was long enough for the effect of each to have died away before the next one arrived we probably would. If it wasn't we would probably use the method described in chapter 3

2.3 Linear circuits

A linear circuit is one whose output for a given applied emf is simply the sum of the outputs it would give if each component of the emf were applied separately. Such a circuit is said to have the property of *superposition*. For such a circuit doubling the magnitude of an applied emf doubles all the voltage drops and currents in the circuit. All the circuits we will be discussing fall into this category as they are governed by Kirchhoff's laws and the voltage drop relations introduced in chapter 1 which are linear.

2.4 Working out responses

The idea of superposition tells us we can work out the response of a linear circuit to an input by decomposing the input into components, working out the response to each, and then summing them.

As an example consider finding the voltage drop at the output of a linear circuit when the emf $\mathcal{E}$ at the top of Figure 2.1 is applied.

The first steps are to obtain the equivalent circuit, label it, and write down the branch equations as described in chapter 1. From the branch equations a differential equation relating the output voltage to the applied emf is then derived.

Next the outputs for the three components of $\mathcal{E}$ are worked out. First the steady component $\mathcal{E}_0$. All the voltages and currents in the circuit due to $\mathcal{E}_0$ are steady so all the derivatives in the differential equation can be set to zero. The output voltage due to $\mathcal{E}_0$ is then found by simple algebra (or often just by inspection).

Next the output due to the first step $\mathcal{E}_1 u(t)$ is found. Here the general solution of the differential equation for $t > 0$ must be worked out. To find the constant(s) of integration, initial ($t = 0$) and/or final ($t = \infty$) values of the desired voltage drop are specified. These values are usually obvious from inspection of the circuit.

Next the output due to the second step $-\mathcal{E}_2 u(t - t_1)$ is found. The good news is that we don't have to do the analysis again, we simply change the sign, amplitude, and origin of time in the result for the first step.

Finally the total output voltage is obtained by adding the output voltages produced by the three components.

It should be clear from the above that the key task in finding a current or voltage in a linear circuit for inputs like the one in Figure 2.1 is to work out what it is for one step. Some partly worked examples follow[1]. It is suggested you fill in the missing steps for practice.

2.5 Examples for practice

2.5.1 Resistor and capacitor (RC) circuit with a step emf applied

The resistor and capacitor are connected as shown in the circuit diagram below. Our aim is to find the voltage

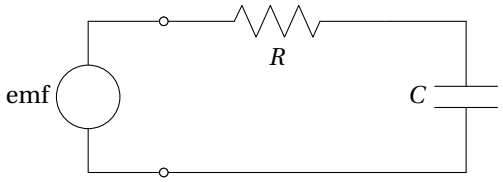


Figure 2.2: RC circuit diagram

drop across the capacitor for an applied step emf. Assuming that at the frequencies involved the resistor just has resistance and the capacitor just has capacitance and that anything connected across C (an oscilloscope for example) has a negligible effect the labelled equivalent circuit (see section 1.6) is as shown in Figure 2.3. Next we write down the voltage drops in the branches (using the definitions given in section 1.3)

$$V_1 - V_2 = IR \tag{2.2}$$

$$V_2 - 0 = \frac{Q}{C} \tag{2.3}$$

where Q is the charge on the inside surface of the upper plate of the capacitor.

Next we write down K2.

$$\mathcal{E}_1 u(t) = V_1 - 0 \tag{2.4}$$

[1] A more advanced method using Laplace transforms (outside the syllabus) is outlined in chapter 23.

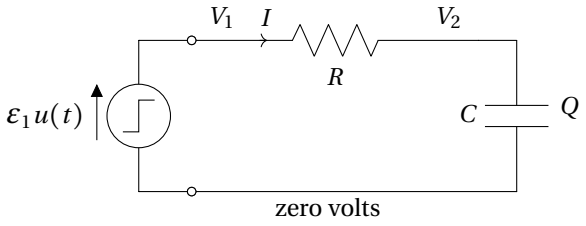


Figure 2.3: RC circuit diagram with added labelling

Next we eliminate V_1 and then, by differentiating equation 2.3 and using $I = \frac{\mathrm{d}Q}{\mathrm{d}T}$, eliminate the other unwanted variable I. The result is :-

$$\mathcal{E}_1 u(t) - V_2 = CR\frac{\mathrm{d}V_2}{\mathrm{d}t} \tag{2.5}$$

We are interested in $t > 0$ for which $\mathcal{E}_1 u(t)$ is constant ($= \mathcal{E}_1$) so we can integrate this equation yielding:-

$$\ln(V_2 - \mathcal{E}_1 u(t)) - \ln A = -\frac{t}{CR} \tag{2.6}$$

where we have written the constant of integration $\ln A$. Rearranging yields

$$V_2 - \mathcal{E}_1 u(t) = Ae^{-\frac{t}{CR}} \tag{2.7}$$

To find A we ask what is V_2 at $t = 0$? (the initial condition). Well, the capacitor is uncharged just before the step, $t = 0-$, and therefore must still be uncharged just after the step, $t = 0+$ because the step is abrupt and there is no time for the charge to change. So $V_2 = 0$ at $t = 0+$ giving us $0 - \mathcal{E}_1 u(t) = A$. Using this we find

$$V_2 = \mathcal{E}_1 u(t)\left(1 - e^{-\frac{t}{CR}}\right) \tag{2.8}$$

The capacitor charges up until its voltage drop balances the applied emf. The quantity CR is called the *time constant.*

2.5.2 CR circuit with a pulse emf applied

It is left as an exercise to show that when C and R are interchanged the output for a step input $\mathcal{E}_1 u(t)$ is given by:-

$$V_2 = \mathcal{E}_1 u(t)e^{-\frac{t}{CR}} \tag{2.9}$$

Consider now that the applied emf consists of two steps forming a pulse of height $\mathcal{E}_1$ and width t_1 i.e. $\mathcal{E}_1 u(t) - \mathcal{E}_1 u(t - t_1)$. Using the result above the output will be

$$V_2 = \mathcal{E}_1 u(t)e^{-\frac{t}{CR}} - \mathcal{E}_1 u(t - t_1)e^{-\frac{t-t_1}{CR}} \tag{2.10}$$

When $t_1 << CR$ this becomes

$$V_2 = \mathcal{E}_1(u(t) - u(t - t_1))e^{-\frac{t}{CR}} \tag{2.11}$$

During the pulse $0 < t < t_1 << CR$ the exponential is close to unity and the output is given by

$$V_2 \approx \mathcal{E}_1(u(t) - u(t - t_1)) \tag{2.12}$$

which is the same as the applied emf, the pulse is transmitted without distortion.

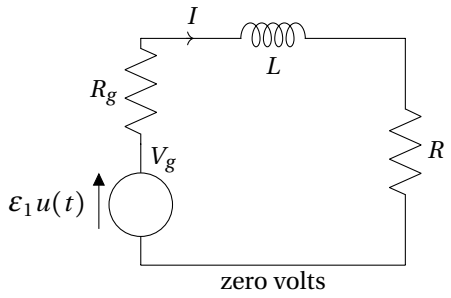


Figure 2.4: LR circuit

2.5.3 An LR circuit with a step emf applied

Figure 2.4 shows an inductance and a resistance in series being excited by a generator of emf $\mathcal{E} = \mathcal{E}_1 u(t)$ and internal resistance R_g. The task is to find the current I There is only one branch equation

$$V_g - 0 = IR_g + L\frac{\mathrm{d}I}{\mathrm{d}t} + IR \tag{2.13}$$

and K2 is

$$\mathcal{E}_1 u(t) = V_g - 0 \tag{2.14}$$

Show that I is given by

$$I = \frac{\mathcal{E}_1 u(t)}{R + R_g}\left(1 - \exp\left(\frac{-(R+R_g)t}{L}\right)\right) \tag{2.15}$$

Hint, to find the constant of integration note that just before the step ($t = 0-$) the current is zero so it will also be zero just after the step ($t = 0+$) because the current in an inductor can only change at a finite rate and the step takes zero time. Here the time constant is $\frac{L}{R+R_g}$.

2.5.4 The x10 Oscilloscope probe

Instead of turning down the sensitivity of an oscilloscope channel when observing signals which do not require the full sensitivity, it can be better to reduce the signal by using an input lead with some components fitted in a probe at the free end. An equivalent circuit representing a signal source, the probe components C_1 and R_1, the cable, and the scope input circuits is shown in Figure 2.5. To reduce the amount of algebra you may assume that

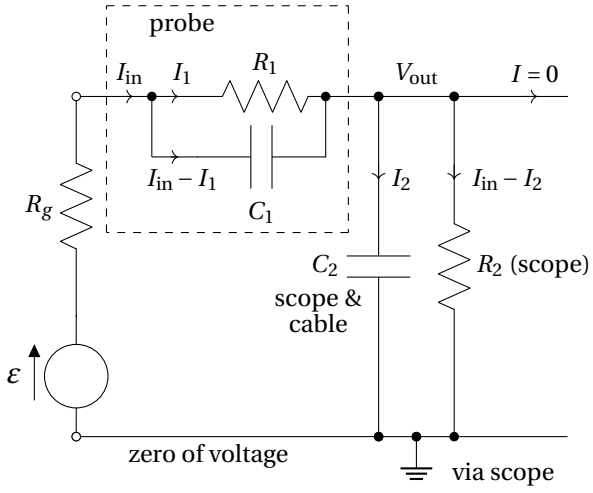


Figure 2.5: Equivalent circuit for analysing scope probe

R_g, the internal resistance of the signal source is negligible. The branch equations are then:

$$V_{\text{in}} - V_{\text{out}} = I_1 R_1 \tag{2.16}$$

$$V_{\text{in}} - V_{\text{out}} = \frac{1}{C_1}\int_{-\infty}^{t} (I_{\text{in}} - I_1)\,\mathrm{d}t \tag{2.17}$$

$$V_{\text{out}} - 0 = \frac{1}{C_2}\int_{-\infty}^{t} (I_2)\,\mathrm{d}t \tag{2.18}$$

$$V_{\text{out}} - 0 = (I_{\text{in}} - I_2)R_2 \tag{2.19}$$

Show that these yield the differential equation

$$\frac{V_{\text{in}}}{R_1} + C_1\frac{\mathrm{d}V_{\text{in}}}{\mathrm{d}t} = (C_1 + C_2)\frac{\mathrm{d}V_{\text{out}}}{\mathrm{d}t} + V_{\text{out}}\left(\frac{1}{R_1} + \frac{1}{R_2}\right) \tag{2.20}$$

Taking the input to be $\mathcal{E}_1 u(t)$ Show that V_{out} is given by

$$\frac{R_2\mathcal{E}_1 u(t)}{R_1 + R_2}\left(1 + \frac{(C_1R_1 - C_2R_2)}{R_2(C_1 + C_2)}\left(e^{-\left(\frac{1}{R_1}+\frac{1}{R_2}\right)\frac{t}{C_1+C_2}}\right)\right) \tag{2.21}$$

Note, ignoring R_g means the two capacitors will charge during the step. Charge (q) will flow from the right hand plate of C_1 to the top plate of C_2 and, since the charges on the plates of a capacitor are always equal and opposite (Gauss), a charge q flows onto the left hand plate of C_1 and from the bottom plate of C_2 to zero volts. No charge flows through the resistors during the step because there isn't any time. It follows that the initial value of V_{out} after the step is

$$V_{\text{out}}(0+) = \frac{C_1}{C_1 + C_2}\mathcal{E}_1 \tag{2.22}$$

The expression for V_{out} shows that it consists of a step and an exponentially decaying part which is either an overshoot spike or a rounding off at the top corner of the step depending on the sign of $C_1R_1 - C_2R_2$. This exponentially decaying part should not be there, the output should just be a step (because the input is just a step). To achieve this C_1 is made adjustable. The correct setting, $C_1R_1 = C_2R_2$, eliminates the spike or rounding and a good step passes to the oscilloscope amplifier.

The advantage of this way of reducing the sensitivity is that it also reduces by a factor of 10 the capacitative and resistive loading effect of connecting the oscilloscope to a circuit. We would not have gained this advantage if the sensitivity had been turned down on the front panel of the oscilloscope.

2.6 Switching

2.6.1 Switching equivalent to applying a step emf

In the arrangement shown in Figure 2.6 the emf of the battery is applied to the circuit when the position of the 2-way switch is changed. With the switch in the original position the emf is replaced by a short circuit. If the position of the switch is changed in much less time than it takes for the circuit to respond, the arrangement behaves exactly like a step emf applied to the circuit.

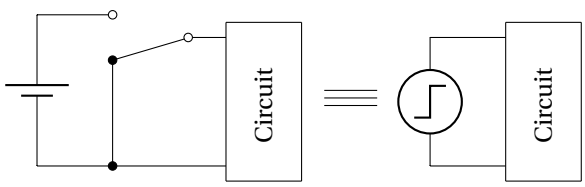


Figure 2.6: Switching equivalent to step emf

2.6.2 Switching not equivalent to applying a step emf

If instead of a 2-way switch a simple on /off switch is used in series with the battery the arrangement is not equivalent to applying a step emf. This is because there is a zero resistance path through the emf when the switch is on and no path when it is off, the circuits are different. Using the 2-way switch there is a zero resistance path in both switch positions.

In the case of the LR circuit analysed earlier the difference would not matter, the initial current in the inductance would be zero before the on/off switch was closed just as it is would be for an emf step preceded by a long initial zero level. However, if a CR circuit was considered instead of an LR circuit it would make a difference. Using the two way switch the charge on the capacitor before the switch was operated would be zero. Using the on/off switch the initial charge on the capacitor could be anything and additional information about what it was just before the switch closure would be required.

2.6.3 Switching within a circuit

Opening and closing switches in a circuit changes the circuit. There will be as many different sets of branch equations and behaviours to work out as there are different combinations of switch positions. The order in which the changes to the switch positions are made obviously has to be known.

The procedure starts by solving for the behaviour of the initial configuration using the method described earlier in this chapter and finding the values of the capacitor charges and inductor currents just before the first switch changes. Assuming this change occurs abruptly these values are then used as initial conditions for the new configuration. This procedure is repeated until all the switch changes have been dealt with.

An example of a circuit with one switch is shown in Figure 2.7. $\mathcal{E}_1 u(t)$ and the switch is closed at $t = \tau$ (after

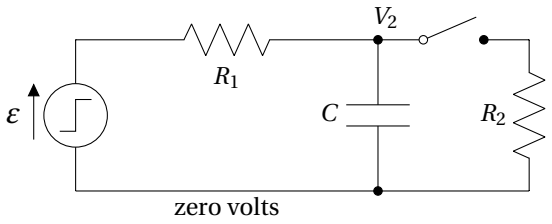


Figure 2.7: An equivalent circuit containing a switch

the step). From section 2.5.1 we know that at time $\tau-$ the voltage across the capacitor is

$$V_c(\tau) = \mathcal{E}_1 u(t)\left(1 - e^{\frac{\tau}{CR_1}}\right) \tag{2.23}$$

Show that for $t > \tau$, i.e. after the switch closure, the voltage across the capacitor is given by

$$V_c(t) = \frac{R_2\mathcal{E}_1}{R_1 + R_2}\left(1 - e^{\frac{t-\tau}{CR_p}}\right) + V_c(\tau)e^{\frac{t-\tau}{CR_p}} \tag{2.24}$$

where R_p is $\frac{R_1R_2}{R_1+R_2}$.

3 Linear Circuits Excited at a Single Frequency

3.1 Analysis in the frequency domain

In Chapter 2 we considered emfs which could be described as linear combinations of unit step functions. Here we consider emfs with precisely repetitive waveforms which Fourier showed can be described as linear combinations of harmonic (sinusoidal or cosinusoidal) functions. The frequencies of the harmonics are integer multiples of $1/T$ (T is the period of the waveform) and their amplitudes and phases depend on the waveform of the emf. Some examples are given in section 3.8.

Knowing the harmonic content of the emf applied to a circuit and knowing that the circuit is linear, the response to the applied emf can be found by summing the responses to its harmonics. The basic skill required is therefore the ability to derive the response of a circuit to a harmonic input. It is called *analysis in the frequency domain* and is the subject of this chapter.

We could proceed as in Chapter 2 and solve a differential equation in which the driving emf is harmonic. Instead, we will describe a simpler method which uses complex algebra.

The method deals only with the steady-state situation after any transient effects generated at switch-on of the signal have died away. Also, it is not the best approach for finding the response to squarewave or similar inputs if they have periods which are so long compared with any time constants or resonances in the circuit that the effect of each edge has died away before the next one comes along. (The time domain analysis in Chapter 2 is the best approach for these cases).

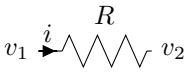
We shall find it convenient to use the *angular frequency*, $\omega = 2\pi f$ which has units of radians per second rather than the ordinary frequency f.

3.2 Voltage drop expressions at angular frequency ω

When a linear circuit is excited at a single frequency all the resulting voltages and currents have the same frequency so it is sufficiently general to substitute for V_1, V_2 and I in the expressions for the voltage drops given in section 1.3 the explicit forms $\hat{v}_1\cos(\omega t+\phi_1)$, $\hat{v}_2\cos(\omega t+\phi_2)$ and $\hat{i}\cos(\omega t+\psi)$ respectively, the hats indicating peak values. It is more useful though to adopt an approach common in other branches of physics and represent such harmonic variables not by the real forms above but by the real parts of the complex exponentials, $\hat{v}_1 e^{j(\omega t+\phi_1)}$, $\hat{v}_2 e^{j(\omega t+\phi_2)}$ and $\hat{i}e^{j(\omega t+\psi)}$ which we will write simply as v_1, v_2 and i. Note the consistent use of lower case for harmonic variables.

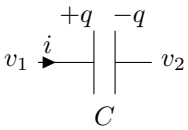
Substituting these forms into the voltage drop expressions in section 1.1.3 yields the single frequency voltage drop expressions:

Resistance:



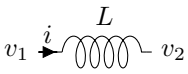
$$v_1 - v_2 = iR \tag{3.1}$$

Capacitance:



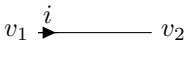
$$v_1 - v_2 = \frac{i}{j\omega C}, \qquad i = j\omega q \tag{3.2}$$

Self inductance:



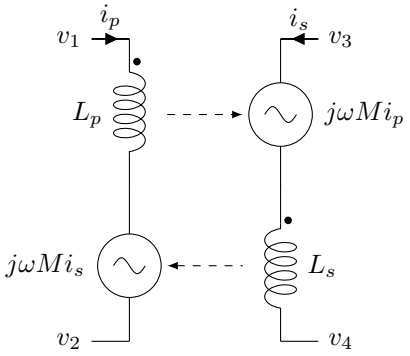
$$v_1 - v_2 = j\omega L i \tag{3.3}$$

Ideal connection:



$$v_1 - v_2 = 0 \tag{3.4}$$

Coupled inductances:



$$v_1 - v_2 = j\omega L_p i_p \pm j\omega M i_s \tag{3.5}$$

$$v_3 - v_4 = j\omega L_s i_s \pm j\omega M i_p \tag{3.6}$$

coefficient of coupling

$$k = \frac{M}{\sqrt{L_p L_s}} \qquad k_{\max} = 1$$

When both coils spiral in the same direction from their starts (indicated by the black dots) the + signs should be taken.

3.3 Circuit analysis using the branch equations

An equivalent circuit is obtained and it is labelled. The branch equations are then written down (using the new voltage drop expressions), and a complex expression relating the two variables (currents or voltages) of interest is found. The ratio of the magnitudes of the two variables and their phase difference can then be extracted. An example is given below.

3.3.1 An RC circuit excited by a signal generator

The equivalent circuit is shown labelled in Figure 3.1 below. We remind you that we use lower case characters to indicate variables represented by complex exponentials. Kirchhoff 2 and the branch equations are:

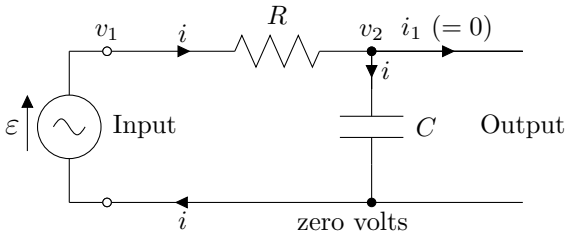


Figure 3.1: RC equivalent circuit

$$\mathcal{E} = v_1 - 0 \quad \text{K2} \tag{3.7}$$
$$v_3 - v_1 = iR_g \tag{3.8}$$
$$v_1 - v_2 = iR \tag{3.9}$$
$$v_2 - 0 = \frac{i}{j\omega C} \tag{3.10}$$

By eliminating unwanted variables from the four equations the following relations (among others) can be easily found. (Check them if you wish).

$$\frac{v_2}{v_1} = \frac{1}{1+j\omega CR} \qquad \frac{v_2}{v_3} = \frac{v_2}{\mathcal{E}} = \frac{1}{1+j\omega CR} \qquad \frac{v_1}{i} = R + \frac{1}{j\omega C} \tag{3.11}$$

Note that the expressions on the right hand sides are functions of the angular frequency ω (which is a given) but are not functions of t. This is because all the currents and voltages in the circuit vary as $e^{j\omega t}$ and this factor cancels out.

A ratio of two voltage differences is called a *voltage transmittance* (T_v) and is dimensionless as is a ratio of two currents, a *current transmittance* (T_i). A ratio of a voltage difference to a current has the units of ohms and is called an *impedance*. Transmittances and impedances play a prominent role in the analysis of linear circuits in the frequency domain.

The ratio of the amplitudes of v_2 and v_1 is the modulus of the transmittance i.e.

$$\frac{\hat{v}_2}{\hat{v}_1} = |T_v| = \frac{1}{\sqrt{1+\omega^2C^2R^2}} \tag{3.12}$$

The transmittance is 1 for $\omega << \frac{1}{CR}$, falls to $\frac{1}{\sqrt{2}}$ at $\omega = \frac{1}{CR}$, and tends to $\frac{1}{\omega CR}$ when $\omega CR >> 1$. At low frequencies the output is the same as the input, at high frequencies the output is less than the input. The circuit is an example of a *low pass filter*.

The phase difference $\phi_2 - \phi_1$ is the argument of the transmittance and is given by $\tan(\phi_2 - \phi_1) = -\omega CR$, so $\phi_2 = \phi_1 - \tan^{-1}(\omega CR)$, i.e. v_2 lags v_1 by $\tan^{-1}(\omega CR)$ radians ($\pi/4$ when $\omega CR = 1$).

Explicitly v_2 is given by:

$$v_2 = \frac{1}{1+j\omega CR}\hat{v}_1 e^{j(\omega t+\phi_1)} \tag{3.13}$$

The real part of v_1 (the actual input voltage) is $\hat{v}_1\cos(\omega t+\phi_1)$, the actual v_2 is the real part of the right hand side i.e.

$$\frac{\hat{v}_1}{\sqrt{1+\omega^2C^2R^2}}\cos(\omega t+\phi_1-\tan^{-1}(\omega CR)) \tag{3.14}$$

Similarly, the real part of i can be found in terms of v_1 from the impedance v_1/i.

3.4 Impedances

The idea of impedance was introduced in the previous section. (There is no corresponding concept in time domain analysis because the variables will in general have different dependences on time.) A general impedance Z is defined by

$$v_1 - v_2 = iZ \tag{3.15}$$

and we will use the symbol shown in Figure 3.2. An impedance has only two connections to other parts of a circuit.

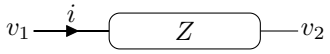


Figure 3.2: Definition of complex impedance.

According to this definition the quantities R, $j\omega L$, and $1/j\omega C$ are the impedances of a resistance R, an inductance L, and a capacitance C at angular frequency ω.

Prelims examiners often ask for a definition of (complex) impedance so here is one:- "The complex impedance of a two terminal network at angular frequency ω is the ratio of the voltage at one terminal minus the voltage at the other to the current entering at the first terminal when the voltages and currents are all expressed as complex exponentials of the form $ae^{j(\omega t+\phi)}$."

3.4.1 Combining impedances

Consider the fragment of equivalent circuit shown in Figure 3.3 comprising a resistance, an inductance and a capacitance connected in series.

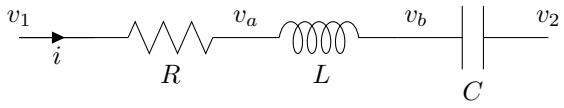


Figure 3.3: RLC in series

Following from the labelling

$$v_1 - v_a = iR \tag{3.16}$$
$$v_a - v_b = j\omega Li \tag{3.17}$$
$$v_b - v_2 = \frac{i}{j\omega C} \tag{3.18}$$

from which it follows that

$$v_1 - v_2 = i\left(R + j\omega L + \frac{1}{j\omega C}\right) \tag{3.19}$$

so the impedance of impedances in series is the sum of their individual impedances.

Consider next a fragment of equivalent circuit comprising a resistance, an inductance and a capacitance in parallel as shown in Figure 3.4. The branch and node equations are

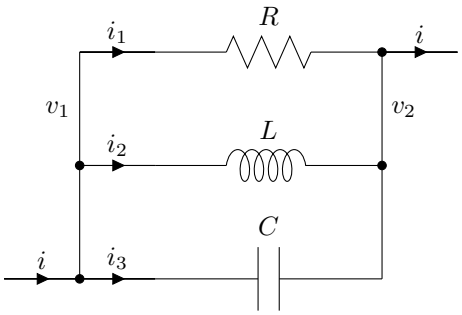


Figure 3.4: RLC in parallel

$$v_1 - v_2 = i_1 R \tag{3.20}$$
$$v_1 - v_2 = i_2 j\omega L \tag{3.21}$$
$$v_1 - v_2 = \frac{i_3}{j\omega C} \tag{3.22}$$
$$i = i_1 + i_2 + i_3 \tag{3.23}$$

from which we can obtain

$$v_1 - v_2 = i\left(\frac{1}{R} + \frac{1}{j\omega L} + j\omega C\right)^{-1} \tag{3.24}$$

so the impedance is the reciprocal of the sum of the reciprocals of the individual impedances.

These formulae enable us to work out the impedance of most two terminal networks. You have probably come across them before for the special case when all the individual impedances are resistances.

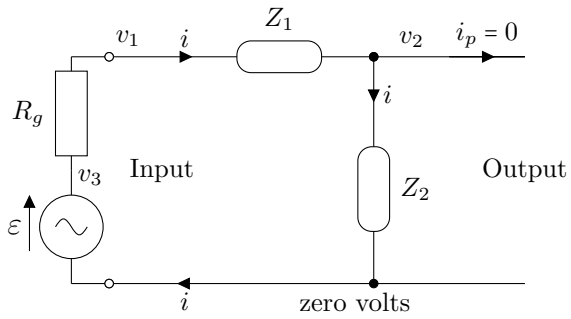


Figure 3.5: Impedances forming a potential divider.

3.4.2 Using impedances

The example in Section 3.3.1 illustrates the straightforward use of the branch equations to find expressions for ratios of emfs, voltages, and currents. Often (but not always) we can avoid most of this labour by simply manipulating impedances. Looking again at the example given in 3.3.1 we see that the circuit is a special case of two impedances arranged as in Figure 3.5 for which the two relevant branch equations are:

$$v_1 - v_2 = i Z_1 \tag{3.25}$$

and

$$v_2 - 0 = i Z_2 \tag{3.26}$$

Eliminating i yields

$$\frac{v_2}{v_1} = \frac{Z_2}{Z_1 + Z_2} \tag{3.27}$$

which you may recognise as the formula for a potential divider.

It is usually acceptable to quote the potential divider formula as a starting point. Plugging the impedances of the R and the C into this immediately yields the transmittance found before (check if you like).

We can also see immediately that the impedance Z of the RC network connected to the generator terminals is

$$Z = R + \frac{1}{j\omega C} \tag{3.28}$$

Many circuits can be analysed quickly if you know the potential divider formula and how to combine impedances. The reason this is quicker than using the branch equations is that you don't have to bother with any of the voltages and currents in the network which would be eliminated in the algebra.

3.4.3 Input impedance

We call the two points at which a network connects to a signal generator the *input terminals* of the network, the impedance v_1/i, (the ratio of the voltage drop across the input terminals to the current flowing into one terminal and out of the other) is called the *input impedance*.

Calculation of input impedance is straightforward, it's just the impedance between the input terminals. In the example above the input impedance of the RC network loading the generator is $R + 1/j\omega C$.

3.4.4 Output impedance

Previously we have used emfs in series with resistances or current generators in parallel with resistances to represent signal sources. More generally we can show impedances rather than resistances, see Figure 3.6. Any linear circuit delivering a signal can be represented like this.

The impedance Z is called the *output impedance* or *internal impedance* of the signal source (both names are used).

There are two ways of calculating the output impedance of a network:-

(i) turn down to zero any emf applied to the input terminals then apply an emf to the output terminals, calculate the current that flows, and form the ratio ϵ/i.

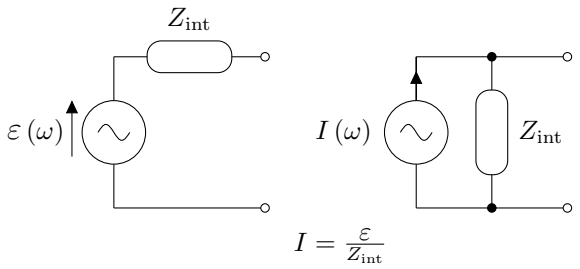


Figure 3.6: Sources with internal impedances

(ii) For a circuit with a load R_L connected (Figure 3.7) and whose voltage transmittance has already been derived the quickest way to find its output impedance is to manipulate the expression for the transmittance into the form

$$\frac{v_{\text{out}}}{\mathcal{E}} = [F]\frac{R_L}{R_L + Z} \tag{3.29}$$

where [F] is a dimensionless function of impedances not including R_L. Z is then the output impedance of the circuit as should be clear from Figure 3.7.

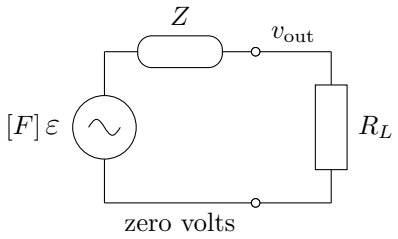


Figure 3.7: Source with an internal impedance with a resistive load added

3.4.5 Power dissipation in impedances

The imaginary parts we add to the real representations of voltages and currents to form complex exponential representations (in order to make calculations easier) do not get mixed in so long as we perform only linear operations. We can simply extract the real part at the end of the calculation. However if we perform a non-linear operation using complex representations, such as forming a product to find a power, care must be taken.

When the current i and voltage drop v in a branch of a circuit are represented by the real variables $\hat{i}\cos\omega t$ and $\hat{v}\cos(\omega t + \phi)$ the power dissipation averaged over a cycle is

$$\frac{1}{T}\int_0^T iv\,\mathrm{d}t = \frac{1}{2}\hat{i}\hat{v}\cos\phi \tag{3.30}$$

where $T = 2\pi/\omega$ and $\cos\phi$ is called the *power factor*.

Note that (i) the average power dissipation is the same for both lag and lead and (ii) if the branch does not contain any resistance, $\phi = \pm\pi/2$ and there is no dissipation.

It is left as an exercise to show that if the current and voltage are represented by (the real parts of) the complex representations $\hat{i}e^{j\omega t}$ and $\hat{v}e^{j(\omega t+\phi)}$ the average power dissipation is given by either $\frac{1}{2}\Re(iv^\star)$ or $\frac{1}{2}\Re(i^\star v)$ where $\Re$ means the real part of and the stars indicate complex conjugates, but *not* by the average value of $\Re(vi)$.

3.5 dc blocking

Sometimes when there are both harmonic and dc voltages applied to a circuit we wish to block the dc component of the current in a branch without affecting the harmonic component. We can achieve this by inserting a capacitor in the branch (which blocks the dc current) and making its capacitance so large that its impedance is negligible.

3.6 Interconnected networks

By a network we mean something very general, it could be a circuit we have put together ourselves or it could be a piece of commercial equipment such as a scope or voltmeter. We represent the boundaries of a network, its terminals, by small open circles.

Usually we are interested in the internal currents and voltages of only one of the networks we are dealing with: for the others (e.g. pieces of test equipment) we just need to know what to expect when we connect them. To achieve this it is not necessary to have their complete equivalent circuits, versions showing only the elements near the terminals are sufficient.

Figure 3.8 below shows a typical system of connected networks represented by such reduced equivalent circuits

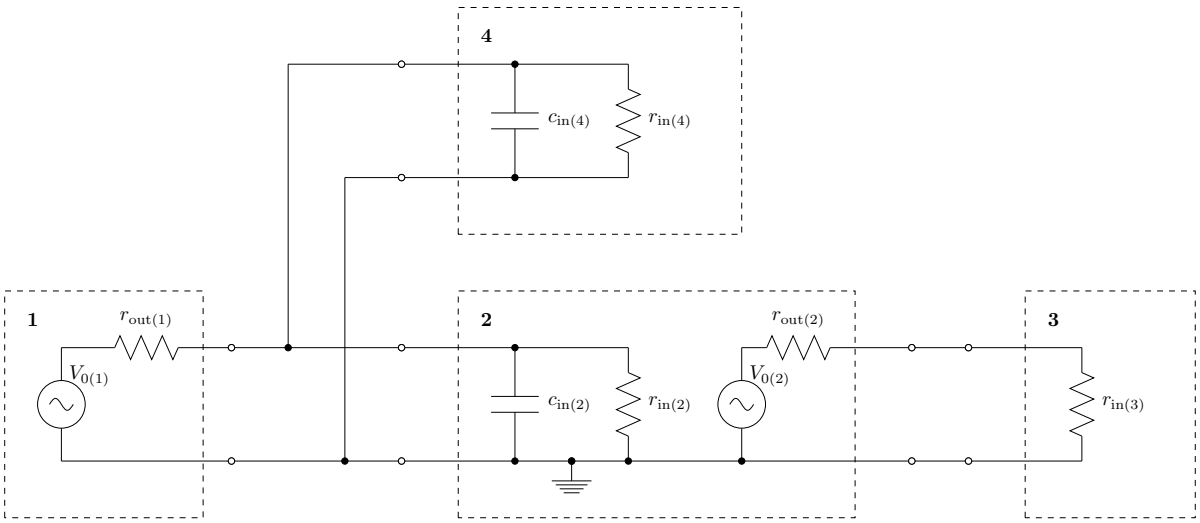


Figure 3.8: Example of connected networks

At the left hand side is a network 1 with output terminals only, (a signal generator or oscillator). In the middle is a network 2 with two input and two output terminals (an amplifier under test). At the right hand side is another network 3 with only two terminals, (the load on the amplifier, a resistance simulating a loudspeaker) shown with its input resistance. Network 4 with only two terminals is a piece of test equipment such as a voltmeter or oscilloscope.

Additional test equipment might be connected to the output of the network under test or perhaps to some internal feature of it thereby defining two more (temporary) terminals.

Voltmeters and oscilloscopes should have high input resistances and low input capacitances. If they draw a significant current from the network to which they are connected then at best the measurement will need to be corrected; at worst the operation of the network will be upset.

3.7 Hints and tips

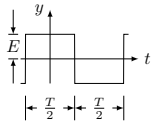
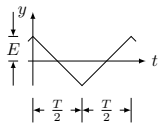
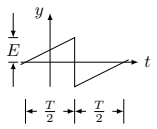
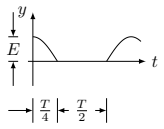
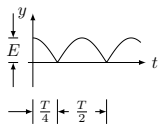
1. Label the currents at nodes in equivalent circuits so that Kirchhoff 1 is satisfied. Only branch equations will then be needed.

2. Always be on your guard against the algebra spreading out. Note that you can often derive all you need directly from a complex expression. If you must multiply top and bottoms of fractions by complex conjugates always simplify as much as possible first.

3. It's a good idea to form dimensionless quantities like $j\omega CR$ in your working to make it easier to spot errors.

4. When no mention is made of any load being connected to a circuit (as often happens in exam questions) state that you are assuming that any load has a negligibly large impedance.

5. An equivalent circuit containing more than five elements will take several pages to analyse.

Waveform		**Harmonic amplitude relative to E**						
		Fundamental	2nd	3rd	4th	5th	6th	7th
Square		$\frac{4}{\pi}$	0	$-\frac{1}{3}\frac{4}{\pi}$	0	$+\frac{1}{5}\frac{4}{\pi}$	0	$+\frac{1}{7}\frac{4}{\pi}$
		(127%)	(0%)	(42.5%)	(0%)	(25.5%)	(0%)	(18.2%)
Triangular		$-\frac{8}{\pi^2}$	0	$-\frac{1}{9}\frac{8}{\pi^2}$	0	$+\frac{1}{25}\frac{8}{\pi^2}$	0	$+\frac{1}{49}\frac{8}{\pi^2}$
		(81%)	(0%)	(9%)	(0%)	(3.2%)	(0%)	(1.6%)
Sawtooth		$\frac{2}{\pi}$	$-\frac{1}{2}\frac{2}{\pi}$	$+\frac{1}{3}\frac{2}{\pi}$	$-\frac{1}{4}\frac{2}{\pi}$	$+\frac{1}{5}\frac{2}{\pi}$	$-\frac{1}{6}\frac{2}{\pi}$	$+\frac{1}{7}\frac{2}{\pi}$
		(63.6%)	(31.8%)	(21.2%)	(15.9%)	(12.7%)	(10.6%)	(9.1%)
Half-wave rectified sine wave		$\frac{1}{2}$	$+\frac{2}{3\pi}$	0	$-\frac{1}{5}\frac{2}{3\pi}$	0	$+\frac{1}{7}\frac{2}{3\pi}$	0
		(50%)	(21.2%)	(0%)	(8.5%)	(0%)	(3.6%)	(0%)
Full-wave rectified sine wave		0	$+\frac{4}{3\pi}$	0	$-\frac{1}{5}\frac{4}{3\pi}$	0	$+\frac{1}{7}\frac{4}{3\pi}$	0
		(0%)	(42.3%)	(0%)	(8.5%)	(0%)	(3.6%)	(0%)
With the origins of time moved $\frac{T}{4}$ to the left	sin or cos?	sin	cos	sin	cos	sin	cos	sin
	change sign?	no	yes	yes	no	no	yes	yes

* + dc component E/π + dc component $2E/\pi$

Table 3.1: Harmonic content of some common waveforms

3.8 Harmonic content of various common waveforms

Any regularly repeating waveform can be built up from a *Fourier series*, that is a series of sine or cosine waves with particular amplitudes and phases and with frequencies which are multiples of the fundamental frequency of the waveform (the reciprocal of its period, T). See Table 3.1 for examples. The sine and cosine waves are known as the *harmonics* of the waveform.

With the origins of time (the y axes) as shown the terms are all cosines, e.g. the square wave is:

$$y = \frac{4}{\pi}E\left(\cos\omega t - \frac{1}{3}\cos 3\omega t + \frac{1}{5}\cos 5\omega t - \ldots\right); \tag{3.31}$$

The changes brought about by moving the origin of time $T/4$ to the left are shown in the lower part of the table. The square wave becomes

$$y = \frac{4}{\pi}E\left(\sin\omega t - \frac{1}{3}\sin 3\omega t + \frac{1}{5}\sin 5\omega t - \ldots\right); \tag{3.32}$$

3.9 Examples for practice

3.9.1 The x10 oscilloscope probe[1]

The response of an oscilloscope with a probe lead to a step emf, showing the value of C_1 needed to give an output without rounding or overshoot, was left as an exercise in section 2.5.1. Here the exercise is to consider an oscilloscope with a probe lead in the frequency domain. An equivalent circuit representing the signal source, the components in the probe, the coaxial cable, and the scope input circuit is shown labelled with harmonic variables in Figure 3.9.

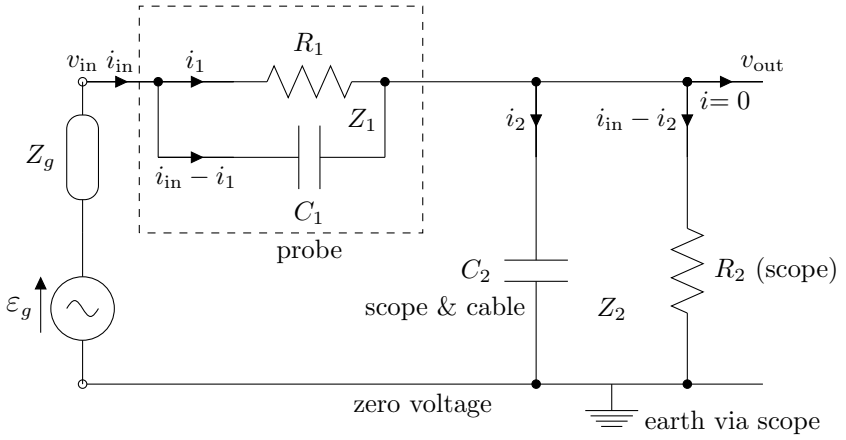


Figure 3.9: Equivalent circuit of oscilloscope probe

Show that the transmittance

$$\frac{v_{\text{out}}}{\mathcal{E}_g} = \frac{Z_2}{Z_1 + Z_2 + Z_g} = \frac{1}{1 + \frac{R_1}{R_2}\frac{1+j\omega C_2 R_2}{1+j\omega C_1 R_1} + \frac{Z_g}{R_2}(1 + j\omega C_2 R_2)} \tag{3.33}$$

Typical values are $R_1 = 9\,\text{M}\Omega$, $R_2 = 1\,\text{M}\Omega$, and C_1 adjusted so that $C_1 R_1 = C_2 R_2, (C_1 = 1/9C_2)$. With these values show that the transmittance is

$$\frac{\frac{1}{10}}{1 + \frac{Z_g}{10R_2}(1 + j\omega C_2 R_2)}. \tag{3.34}$$

Next show that for a plain lead (i.e. without the probe components) the transmittance is

$$\frac{1}{1 + \frac{Z_g}{R_2}(1 + j\omega C_2 R_2)}. \tag{3.35}$$

[1] It multiplies the volts/div scale by 10.

Thus for a given Z_g the variation of the transmittance with frequency and therefore the effect on the harmonic content of $\mathcal{E}_g$ is less using a probe lead. In other words, the waveshape of $\mathcal{E}_g$ is more accurately transferred to the amplifier in the scope.

Finally show that the input impedance of the probe lead is

$$\frac{10R_2}{1+j\omega\frac{C_2}{10}10R_2}. \tag{3.36}$$

i.e. $10R_2$ in parallel with $C_2/10$ which is 10 times better (higher) than it is for a plain lead.

3.9.2 A bandpass filter using a parallel tuned circuit

The voltage transmittance of the equivalent circuit shown in Figure 3.10 has a maximum at one frequency and falls to zero at zero frequency and infinite frequency. It is a type of *bandpass filter*. The imperfections in the inductor and capacitor are represented by the parallel resistance R. If the signal source (of emf $\mathcal{E}_g$) has a significant internal resistance R_g, v_{in} should be replaced with $\mathcal{E}_g$ and R_1 by R_1+R_g.

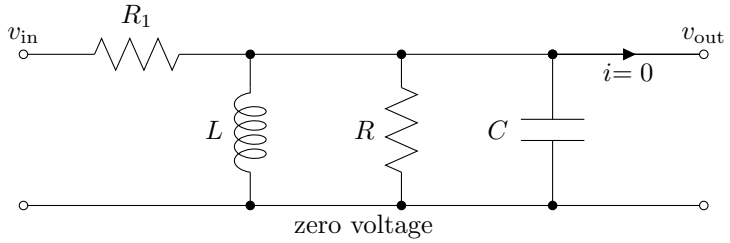


Figure 3.10: Equivalent circuit of a bandpass filter

Show that the voltage transmittance

$$T_v=\frac{v_{\text{out}}}{v_{\text{in}}}=\frac{1}{R_1\left(\frac{1}{j\omega L}+j\omega C+\frac{1}{R}\right)+1}. \tag{3.37}$$

that the magnitude of the transmittance as a function of frequency, its *frequency response*, is given by

$$\frac{\hat{v}_{\text{out}}}{\hat{v}_{\text{in}}}=\frac{1}{\sqrt{\left(\frac{R+R_1}{R}\right)^2+R_1^2\left(\omega C-\frac{1}{\omega C}\right)^2}}. \tag{3.38}$$

and that the phase shift between the output and the input is given by

$$\phi_{\text{out}}-\phi_{\text{in}}=\tan^{-1}\frac{RR_1\left(\omega C-\frac{1}{\omega C}\right)}{R_1+R} \tag{3.39}$$

Examine the expression for T_v further. Show that T_v reaches a maximum value and is real ($\phi_{\text{out}}-\phi_{\text{in}}$ is zero) at $\omega=\omega_0$ (called the *resonant frequency* of the tuned circuit) given by

$$\frac{1}{\omega_0 L}=\omega_0 C \tag{3.40}$$

A measure of the sharpness of the resonance is the *fractional bandwidth* defined as $\frac{\omega_1-\omega_2}{\omega_0}$, where ω_1 and ω_2 are the angular frequencies at which

$$|T_v|=\frac{|T_{v,max}|}{\sqrt{2}} \tag{3.41}$$

This occurs when the real and imaginary parts of T_v are equal in magnitude, i.e. when ω_1 and ω_2 are given by

$$\frac{1}{\omega_1 L}-\omega_1 C=\frac{1}{R}+\frac{1}{R_1} \tag{3.42}$$

and

$$\frac{1}{\omega_2 L} - \omega_2 C = -1\left(\frac{1}{R} + \frac{1}{R_1}\right). \tag{3.43}$$

Hence show that

$$\frac{\omega_1 - \omega_2}{\omega_0} = \frac{1}{R_1 T_{v,max}}\sqrt{\frac{L}{C}}. \tag{3.44}$$

For a given inductor and capacitor the minimum bandwidth results if $R_1 >> R$. Show this is given by

$$\frac{\omega_1 - \omega_2}{\omega_0}(\min) = \frac{1}{R}\sqrt{\frac{L}{C}} \tag{3.45}$$

The reciprocal of $\frac{\omega_1 - \omega_2}{\omega_0}(\min)$ is known as the *quality factor* of the tuned circuit and is denoted by Q. Thus

$$Q = R\sqrt{\frac{C}{L}} \qquad \left(= \frac{R}{\omega_0 L} = R\omega_0 C\right). \tag{3.46}$$

Note finally that it follows from 3.42 and 3.43 and the expression for $\phi_{\text{out}} - \phi_{\text{in}}$ above that $\phi_{\text{out}} - \phi_{\text{in}} = +\frac{\pi}{4}$ at ω_1 and $-\frac{\pi}{4}$ at ω_2.

3.9.3 Perfectly coupled inductances (ideal transformer) with load

The labelled equivalent circuit is shown in Figure 3.11. The windings are in the same direction with the 'starts' indicated with black dots. When $M = \sqrt{L_p L_s}$ (perfect coupling) the voltage drop relations for coupled inductances

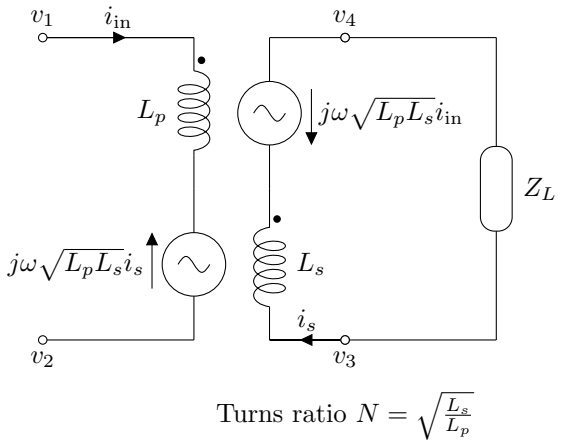


Figure 3.11: Perfectly coupled inductances with perfect conductors (ideal transformer)

are

$$v_1 - v_2 = j\omega L_p i_{\text{in}} + j\omega\sqrt{L_p L_s} i_s \tag{3.47}$$

$$v_3 - v_4 = j\omega L_s i_s + j\omega\sqrt{L_p L_s} i_{\text{in}}. \tag{3.48}$$

The load branch equation is:

$$v_4 - v_3 = i_s Z_L. \tag{3.49}$$

Show that:

$$v_3 - v_4 = N(v_1 - v_2) \tag{3.50}$$

which says that for an ideal transformer the voltage differences between the ends of the primary and secondary windings are related by N. This is a direct consequence of Faraday's law and is true whether there is a load connected or not.

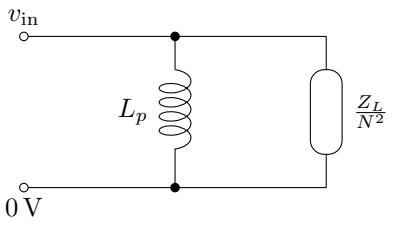


Figure 3.12: Input impedance of a loaded ideal transformer

Show that the input impedance of the loaded ideal transformer is given by:

$$Z_{\text{in}} = \frac{v_{\text{in}} - 0}{i_{\text{in}}} = \frac{1}{\frac{N^2}{Z_L} + \frac{1}{j\omega L_p}} \tag{3.51}$$

the impedance of the primary inductance in parallel with Z_L/N^2, as shown in Figure 3.12.

In practice, transformers are usually designed to make $\omega L_p >> Z_L/N^2$ so the input impedance reduces to Z_L/N^2.

Finally show that with Z_L removed the secondary side of the ideal transformer driven by a signal source of emf $\mathcal{E}_g$ and internal impedance Z_g can be represented as in Figure 3.13. (Hint; the first two equations are the same as before except that in the first one $\mathcal{E}_g$ replaces v_{in} and Z_g is added to $j\omega L_p$, the load branch equation does not apply.)

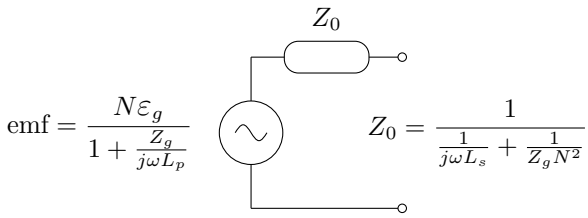


Figure 3.13: Representation of the output of an ideal transformer

4 Introduction To Logic Circuits

4.1 Representation of numbers in electronic systems

4.1.1 Analogue representation

In an analogue representation a range of numbers is represented by a range of voltages (or currents). Calculations are performed using an analogue circuit containing amplifiers, adders, integrators and multipliers which has been wired up specifically for the calculation desired. Such circuits, called analogue computers or simulators, are discussed in chapter 16.

The amplifiers, adders etc. employed in an analogue computer suffer from zero drifts, gain drifts and departures from their intended functions caused by temperature changes, power supply voltage changes, manufacturing inaccuracies, and ageing of components. There are also fundamental noise and man-made interference to contend with. These errors accumulate and limit the accuracy that can be achieved to about 1%. This means that only numbers with two or less significant digits can be processed without error in this way.

4.1.2 Digital representation

The kinds of errors mentioned above will be present in any electronic system. The only way to improve matters is to increase the range of voltage assigned to a particular number up to the point where errors cannot carry it outside this range.

An important step towards achieving this is to *represent each digit of a number separately.* For instance, we could represent the number 2.37 by using three separate wires one at 2 volts with respect to a zero, one at 3 volts and one at 7 volts. (This would be described as a parallel decimal digital representation.) Then any voltage between 1.5 and 2.5 volts would be recognized as a 2. i.e. errors of up to ±0.5 volt could be tolerated. A definition of which wire carried the units, tenths and hundredths would of course have to be made.

Another representation might be one in which the voltage levels were sent one after another down the same wire with some timing signal provided to indicate each digit and a definition made of whether the units or hundredths came first. (This would be called a serial decimal digital representation.)

In practice it turns out that, even with the digits represented separately, a decimal representation divides the voltage range into too many segments to ensure freedom from errors. In fact it is desirable to employ the absolute minimum number of segments which allows numbers to be represented. Obviously the number of voltage segments needed to represent numbers digitally is equal to the base of the number system. If we choose the number system with the smallest base, 2 (the binary system) we need only two segments one for each of the binary digits 1 and 0. This choice not only provides the best immunity from errors but also means that the circuits can be made from devices operating as on/off switches. (The invention of the binary number system is attributed to Leibniz.)

The binary digits 1 and 0 are called bits[1], a contraction of binary digit, (they are also referred to as true and false logic states). In circuits, 1 is usually represented by a voltage range surrounding a positive voltage (typically 5 or 10 volts) and 0 by a range (not overlapping the first) that includes 0 volts. In the interval between the two voltage ranges (which is usually traversed rapidly) the logic state is not defined. Multi-bit numbers are represented either as a train of binary voltage levels on a single wire (*serial* representation) or as binary levels present at some instant on a number of wires (*parallel* representation).

For the long-term storage of bits a variety of techniques are used e.g. the direction of magnetization of small regions on a magnetic tape or disc, the presence or absence of craters in a (compact) disc.

A group of eight bits is called a *byte*, one or more bytes make a *word*. Serial numbers require less complex circuits to process them but processing takes longer. Modern digital processors handle binary numbers as parallel words.

[1] To be distinguished from a *bit of information*, the unit in which information is measured, which is defined as the specification of one of two *equally probable* outcomes. An answer of 'heads' after the toss of an unbiased coin is an issue of one bit of information.

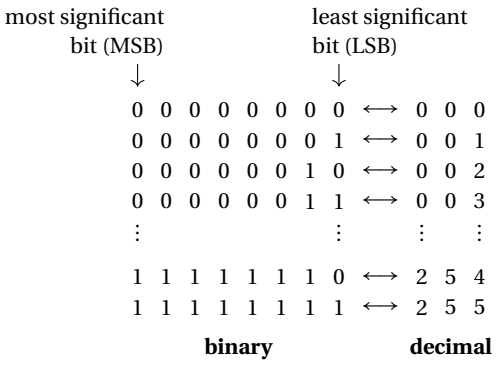


Figure 4.1: table
Simple allocation of bytes to decimal numbers 0 to 255.

Consider 1-byte words. The 256 different 1-byte words can represent 256 decimal (or any other) numbers. We naturally choose to map one set onto the other in a way that enables us to do arithmetic and to convert from one set to the other conveniently. One of the simplest ways of allocating them is shown in Table 4.1.

This is fine if we don't need negative numbers, if we do then we have to allocate some of our bytes to them, half to positive numbers and half to negative seems sensible. One way to do this, in which the negative numbers are called *twos complements*, is shown in Table 4.1. In this allocation the negative of a number b is obtained by inverting all its bits (1s to 0s and 0s to 1s) to form $b\sim$ and adding 1, i.e.

$$-b = b\sim +1 \tag{4.1}$$

This allocation is useful for two reasons. Firstly, it has the familiar property that the sum of a positive number and the negative number of equal magnitude is zero (provided the result is not outside the allocated range). Secondly, the most significant bit is 1 for all negative numbers and 0 for positive numbers (and zero) which makes it easy to distinguish between them. The MSB in the twos complement allocation is sometimes called the *sign bit*.

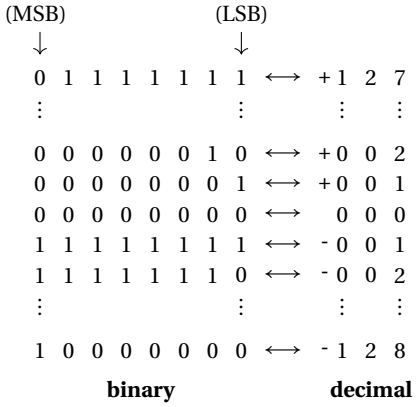


Table 4.1: Twos complement allocation of bytes to decimal numbers −128 to +127.

Words that are multiples of four bits long can be represented much more compactly in *hexadecimal* notation (*hex* for short) than in binary. The eight bits of a byte are represented by just two hex digits. The allocation of hex codes is shown in Table 4.2. The sixteen-bit (two-byte) binary word 1011 0001 1111 1000 becomes B1F8 in hex.

Of course numbers are not the only data, a lot of data processing is manipulation of text (word processing). In this case there is no need to do arithmetic and the allocation of binary words to text and formatting characters can be arbitrary. However it is clearly useful to have a standard allocation and one such is ASCII, the American Standard Code for Information Interchange. This 7 bit code is shown in Table 4.3.

4.2 Processing data in binary digital systems

Processing numbers and other data in a binary digital system involves moving words around between stores (registers) and carrying out arithmetic, logical or bitwise operations on them using 'logic circuits'. A logic circuit

binary	hex	decimal
0000	0	0
0001	1	1
0010	2	2
0011	3	3
0100	4	4
0101	5	5
0110	6	6
0111	7	7
1000	8	8
1001	9	9
1010	A	10
1011	B	11
1100	C	12
1101	D	13
1110	E	14
1111	F	15

Table 4.2: Binary and hex representations of the decimal numbers 0 to 15

most significant bits								least
000	001	010	011	100	101	110	111	sig. bits
non printable control codes			0	@	P	‘	p	0000
		!	1	A	Q	a	q	0001
		"	2	B	R	b	r	0010
		#	3	C	S	c	s	0011
		$	4	D	T	d	t	0100
		%	5	E	U	e	u	0101
		’	6	F	V	f	v	0110
		(	8	H	X	h	x	0111
		)	9	I	Y	i	y	1000
		*	:	J	Z	j	z	1001
		+	;	K	[	k	{	1010
		,	<	L	\	l	|	1011
		-	=	M	]	m	}	1101
		.	>	N	^	n	~	1110
		/	?	O	_	o	Del	1111

Table 4.3: ASCII printable characters (010 0000 is space)

whose output state (1 or 0) at any instant depends only on the currently existing input states is said to perform a *combinational logic* function while one which can store digits and whose output therefore can depend on past as well as present inputs, is said to perform a *sequential logic* function. We will consider both types of circuit beginning with combinational logic.

4.3 Combinational logic

4.3.1 Basic functions

As we shall see below, logic circuits which perform complicated functions are constructed from basic logic circuits which perform simple functions. These basic circuits are now made in *integrated circuit* (chip) form in vast numbers. They are collectively known as *gates*. The most used types are shown in Figure 4.2.

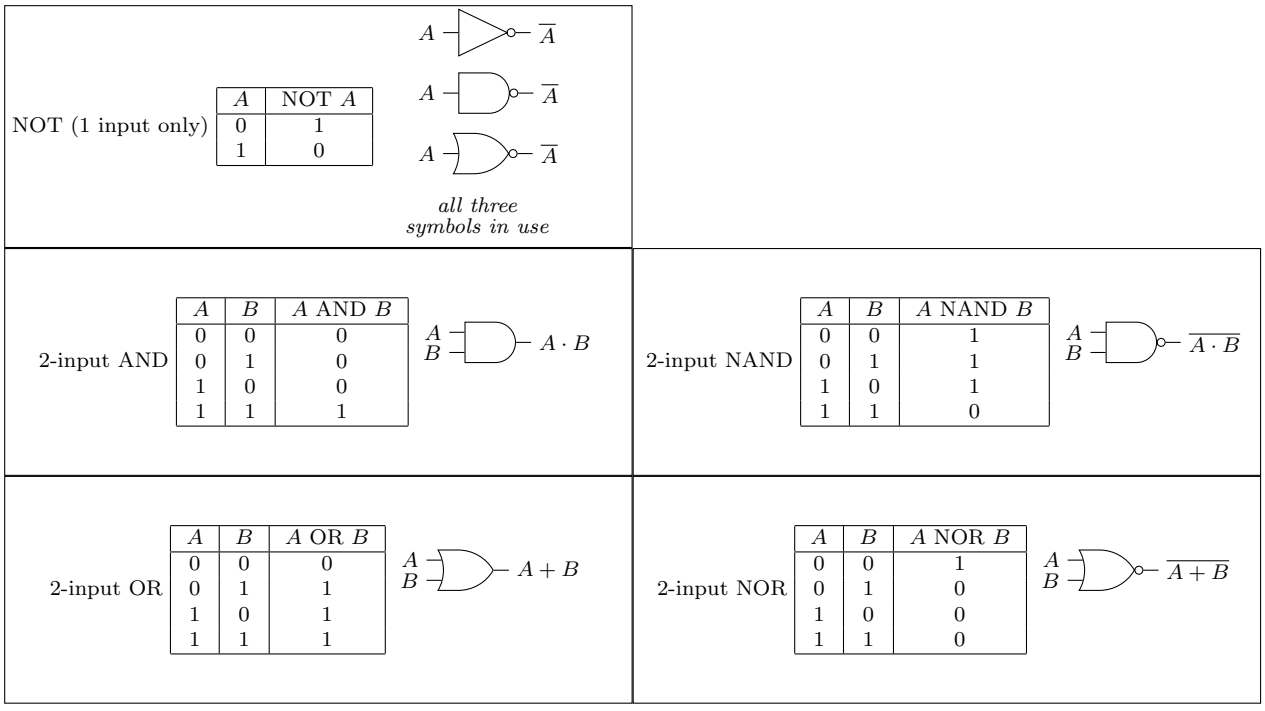


Figure 4.2: Basic logic functions: diagram symbols, boolean expressions, and truth tables.

A *truth table* is a way of defining a logic function by listing its value (output) for all possible combinations of inputs. A *Boolean*[1] *expression* is an algebraic description of a logic function. Note that the symbols + and . in the Boolean expressions have almost but not exactly the same effects as in ordinary algebra. NOT of course only has one input, the extension of the other functions to more inputs should be self evident.

4.3.2 Design of combinational logic functions

How do we devise a circuit to perform a given combinational logic function? The design process, which we shall illustrate below with an example, has several stages:

1. A truth table of the desired function is drawn up. (This completely defines it.)
2. From the truth table a Boolean expression representing the function is written down in terms of basic functions such as the ones listed above (this immediately enables a circuit to be drawn but usually it is not one we would wish to build).
3. Boolean algebra is then used to manipulate the function into the desired form, which usually means one involving only a particular chosen set of basic functions.
4. The circuit is then drawn.

Let us suppose by way of an example that we wish to construct a combinational logic circuit with three inputs (A, B, C) whose output is a 1 only when a majority of the three inputs are 1s. The truth table of the function (which we will call M) is shown in Table 4.4.

How do we express M as a Boolean expression? This is easy, refer to the table of basic functions above and consider the function:

$$\overline{A}.B.C + A.\overline{B}.C + A.B.\overline{C} + A.B.C \tag{4.2}$$

The first term is 1 only when $A = 0$, $B = 1$ and $C = 1$: the fourth line in the truth table for M. The second term gives a 1 in the sixth line etc. All the terms are then put together in an OR function to produce M.

Another way is to write down the function M is:

$$(A + B + C).(A + B + \overline{C}).(A + \overline{B} + C).(\overline{A} + B + C) \tag{4.3}$$

which picks out the rows of the table for which M is zero.

Clearly expressions like these could be written down for any logic function. The first form is simpler (has fewer terms) if the function has fewer 1s than 0s, the second is better if the function has fewer 0s than 1s.

As mentioned in stage (ii) of the design process, without doing any algebra, we could immediately draw a circuit to perform the function M using either of these two Boolean expressions e.g. the first form implies the circuit shown in Figure 4.3.

A	B	C	M
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1

Table 4.4: Truth table of function M

This circuit would certainly work but it might not be very attractive in practice, for instance it might call for a selection of basic circuits (gates) to be used which we do not want to use. Also it may not be the simplest circuit that will perform the function. Therefore we need a way of manipulating our expressions into forms which enable us to draw the simplest circuits using our chosen gates. The tool for this is Boolean algebra.

[1] In the mid nineteenth century George Boole and Augustus de Morgan developed a set of rules governing the relationships between the true and false statements of logic. This foundation was later developed into what we now call Boolean algebra.

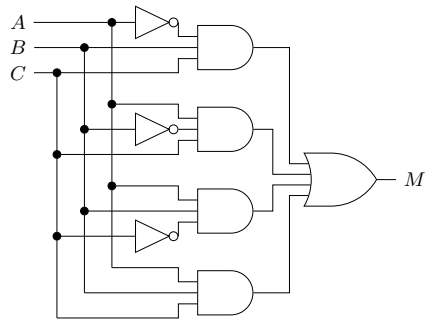


Figure 4.3: Naive implementation of M

4.3.3 Boolean algebra

There is only one non-trivial relation in Boolean Algebra: de Morgan's law, which we can write for three variables as

$$X + Y + Z = \overline{\overline{X}.\overline{Y}.\overline{Z}} \tag{4.4}$$

Alternatively, by taking the NOT of both sides and replacing variables by their inverses we have

$$X.Y.Z = \overline{\overline{X} + \overline{Y} + \overline{Z}} \tag{4.5}$$

These can be verified by comparing the truth tables of both sides. They show that the functions AND, OR and NOT are not independent: OR can be expressed in terms of NOT and NAND, while AND can be expressed in terms of NOT and NOR. Now we know from the example of M above that we can always describe any truth table in terms of an algebraic expression containing the three basic functions AND, OR and NOT. What de Morgan's law tells us is that we can in fact describe any truth table with just two basic functions as long as one of them is NOT. Technically this is very convenient because if we adopt say NOR as one of the basic functions the same design of integrated circuit chip can be used for the NOT (because this is just a one-input NOR circuit). The same is true for NAND.

The distributive, commutative and associative properties all operate in the algebra. There are also a number of useful identities, eight of which we list below. Their validity can be checked by comparing the truth tables of the left and right hand sides, however most are self evident e.g.:

$X + 0 = X$ $\quad X.0 = 0$ $\quad X + 1 = 1$ $\quad X.1 = X$
$X + \overline{X} = 1$ $\quad X.\overline{X} = 0$ $\quad X + X = X$ $\quad X.X = X$

4.3.4 Completion of the design of combinational logic

To show what can be done we return to the function M considered in section 4.3.2. The first expression for M is:

$$M = \overline{A}.B.C + A.\overline{B}.C + A.B.\overline{C} + A.B.C \tag{4.6}$$

We begin by complicating the expression by writing the last term three times using the identity $X = X + X$. Next we simplify it by combining each of the three terms $A.B.C$ with one of the first three terms using the identity $X + \overline{X} = 1$ which yields:

$$M = B.C + A.C + A.B \tag{4.7}$$

Finally using de Morgan's law we can manipulate it into a form that can be realized with NAND circuits only:

$$M = \overline{\overline{B.C}\,.\,\overline{A.C}\,.\,\overline{A.B}} \tag{4.8}$$

with the circuit diagram shown in Figure 4.4.

It is a good idea to label the gates on the diagram, as we have above, with the chip type numbers and their pin connections before starting to wire them up or lay out the design of a printed circuit board.

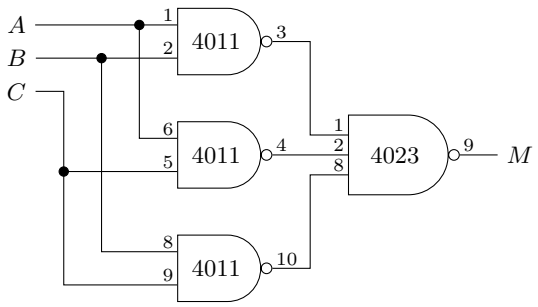


Figure 4.4: NAND implementation of M

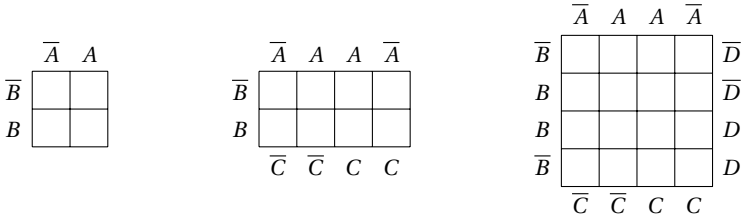


Figure 4.5: 2, 3 and 4 variable Karnaugh maps

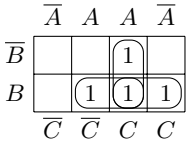


Figure 4.6: 3 variable Karnaugh map with M inserted

4.3.5 Karnaugh maps

So far we have used truth tables to display Boolean expressions. Another kind of table known as a *Karnaugh map* is also useful. It is more compact than a truth table and can be used as a tool for some manipulation of the function, particularly simplification. Examples of Karnaugh maps for functions of 2, 3 and 4 variables are shown in Figure 4.5. The usefulness of the maps is due to their labelling which is such that adjacent squares (not diagonals) differ only in the inversion of one of the variables.

Functions are first expressed as an ORed collection of terms and a 1 entered for each term into the appropriate box in the map. Our first expression for M is shown entered into the three variable map in Figure 4.6. We have combined the adjacent terms of M as indicated by the loops. The two terms in the bottom right hand corner of the map are $A.B.C$ and $\overline{A}.B.C$ can be combined into $B.C$ using $X + \overline{X} = 1$. The map makes it easy to spot such combinations as adjacent occupied squares, particularly cases like this one where it is useful to use the same term several times. In such a simple case we would not usually bother with a map but in complicated cases where it would be easy to make a slip in the Boolean algebra, maps are useful.

Combination loops must be full of 1s or 'don't cares' (*don't cares* are terms whose value doesn't matter) and must be rectangular or square. The edges of the map are functionally adjacent, a loop can leave the top (or side) and re-enter at the bottom (or other side).

4.4 Clocks

A circuit that produces a sequence of regularly spaced pulses is called a *clock*. Most digital systems contain one. Probably the simplest circuit is the one comprising a 2-input NAND Schmitt gate, one resistor and one capacitor shown in Figure 4.7. More precise clock circuits, like the one on your wrist, are controlled by the vibrations of a quartz crystal resonator.

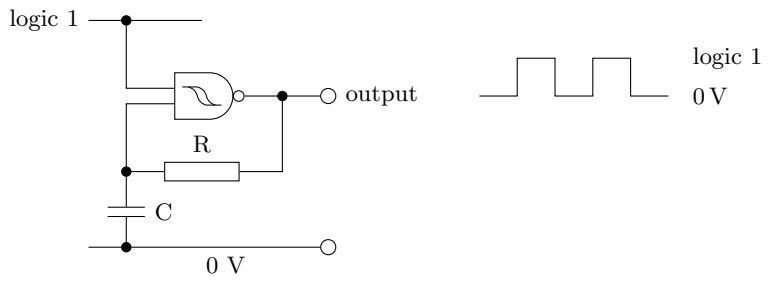


Figure 4.7: Simple clock

4.5 Sequential logic

4.5.1 Bistables ("Flip Flops")

Following the change of an input the output of a typical gate circuit remains at (remembers) its original state for only a few nanoseconds before it changes to a new state. A bistable is a basic logic circuit whose output persists in its original state until it is commanded to respond to its inputs by a signal on an additional (clock) input.

Bistables are given names according to the inputs provided and the way in which they respond to the clock. We shall describe two kinds; first, because it is the most general, the edge-triggered JK type. The trigger point is somewhere on the rising edge of the clock pulse. The truth table and the symbol are shown in Figure 4.8. Q_{new} is the state of the output Q after the clock edge, Q_{old} is the state of Q before the clock edge.

The symbol used to represent bistables on circuit diagrams is a rectangle showing the various inputs and outputs. $\overline{Q}$, the inverse of the output Q, is usually available also. A simpler type of bistable is the D type illustrated in Figure 4.9.

Some bistables have a pair of additional inputs which can be used to set up the Q output independently of the clock. These asynchronous inputs, called PRESET and CLEAR (or SET and RESET), set the Q output to 1 and 0 respectively.

Truth table

J	K	Q_{new}
0	0	Q_{old}
0	1	0
1	0	1
1	1	$\overline{Q_{\text{old}}}$

4cm

Symbol

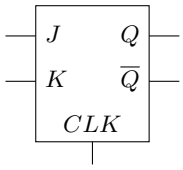


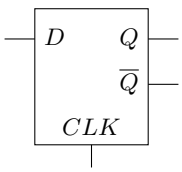
Figure 4.8: figure
JK flip flop

In all bistables there is a propagation delay (of the same order as a gate delay) between the instant the clock allows new data to be accepted (or an asynchronous input is activated) and its appearance at the output. Use is often made of this delay to allow old data to be used while new data is in the process of being loaded.

Truth table

D	Q_{new}
0	0
1	1

Symbol



5cm

Figure 4.9: D type flip flop

4.5.2 Registers

A bistable is a 1-bit store whose contents are indicated by the state of its Q output. An array of N bistables is called an N bit *register*. Registers are used for storing both parallel and serial data. Figure 4.10 shows a *register* for storing 1-byte parallel words: Figure 4.11 shows one for storing serial words.

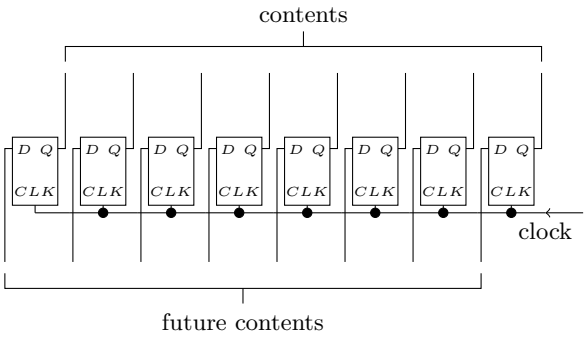


Figure 4.10: 1 byte parallel register

This latter arrangement in which a clock pulse moves the entire contents one place to the right is called a shift register. The design shown relies on the propagation delay mentioned in 4.5.1. To load it, a serial data byte is presented at the input with the control line at 'write' and eight clock pulses synchronous with the data are applied.

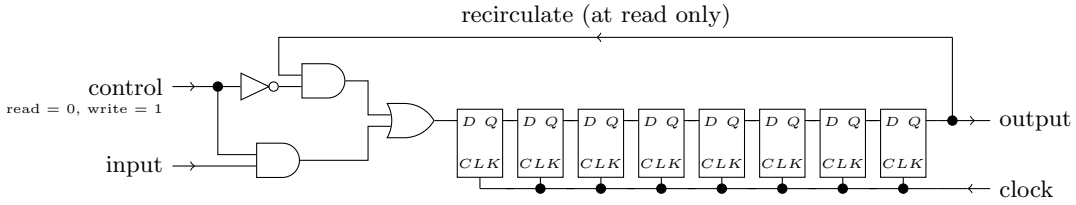


Figure 4.11: 1 byte serial register

If the control input is 0, i.e. 'read', and eight clock pulses are applied, a serial word is produced at the output. Note also that on read the data is fed back into the input (recirculated) to avoid losing it.

By using JK bistables with additional gating between them registers can be made to function as either parallel or shift registers on command. This allows such operations as serial to parallel and parallel to serial data conversion. Asynchronous loads can be performed if the bistables have PRESET and CLEAR inputs.

4.5.3 Clocks and dividers/counters

Pulse trains with lower frequencies can be derived from a clock input as shown in Figure 4.12, where each bistable produces clock pulses for the next in the chain.

The JK bistables are wired to change state (toggle) on each rising clock edge they receive. The Q output of each is then a train of square pulses at half the frequency of its clock input. This arrangement is called a 'ripple-through' binary divider or counter. We shall refer to the *state of the counter* at some instant of time, by this we mean the states of the Q outputs of all the bistables at that instant.

4.5.4 Generation of a particular repetitive timing/control waveform

One use of divider chains is in the generation of repetitive patterns of serial pulses. Suppose we wish to generate repeatedly an 8-bit serial word W = 01001000 and we have a binary divider chain of the type shown in 4.12 in which the first flip-flop has the required pulse width at its Q output. The waveform (A) at this output and at the outputs of the two following flip-flops (B and C) are shown in Figure 4.13 together with W. We have assumed that W is to start at t = 0. We require 1s in the second and fifth intervals (only). This is achieved by writing:

$$W = \overline{C}.B.A + C.B.\overline{A} \tag{4.9}$$

just as we did when describing a truth table algebraically. (You may have noticed that A, B, C are just like the inputs in a truth table.) The design of a circuit to realize W then proceeds as before (section 4.3).

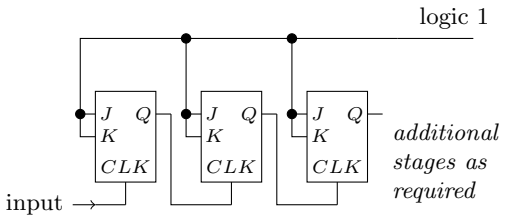


Figure 4.12: Binary divider

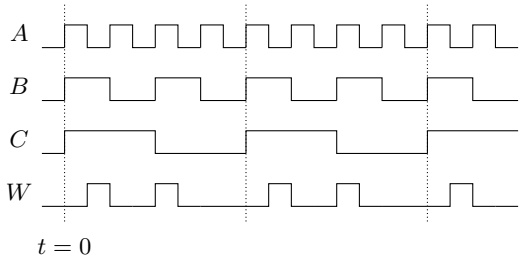


Figure 4.13: Binary divider waveforms and a desired serial word.

4.5.5 Effects of propagation delays

In the binary divider in Figure 4.12 the propagation delay between the edge of a clock pulse and the appearance of the new Q and $\overline{Q}$ outputs of a bistable causes the last in a chain of N bistables to change $(N-1)$ delays later than the first one. This means that there are short-lived states of the divider, while the clock edge is rippling through, which are not part of the desired sequence of states.

The short-lived out-of-sequence states in the ripple-through counter can give rise to short-lived unwanted states in our word which might cause it to be misinterpreted by circuits it is fed to. Unwanted states will not occur if:

1. all the flip-flops that change do so simultaneously, and
2. measures are taken to equalize delays in the decoding logic.

Condition 1 is satisfied if a *synchronous* counter is used. In a synchronous counter the clock input is fed to *all* the flip-flops simultaneously and the J and K inputs of each one are set up following each clock edge so that at the next edge all the flip-flops change to the next desired state. Such a counter can be designed to progress through *any* desired sequence of states and is sometimes called a *state machine*.

5 Mains Electricity Supply And Safety

5.1 Introduction

Mains power is supplied in the form of sinewave emfs with a frequency of 50 Hz. Alternating rather than direct voltage is used for the following reasons

(i) the voltage can be stepped up and down with very high efficiency (in transformers)

(ii) both the rotating machines which generate it and the rotating machines which it can drive are elegant and efficient

(iii) it is easy to derive direct voltages when they are needed (see section 15.4).

5.2 National and regional power distribution

The national grid is the backbone of the UK electric power distribution network to which generating stations and primary distribution stations are connected. To minimise I^2R losses (heating of the cables) the power is transmitted at the lowest practicable current and the highest practicable voltage, the factors limiting the latter being flash-over and leakage at insulators.

The highest voltage sections of the grid operate at 440 kVrms with respect to earth. Distribution branches operate at successively lower voltages the further from the backbone and nearer the users they are, typical voltages are 250 kV, 33 kV, and 11 kV with respect to earth. These parts of the network are invariably *three phase*, consisting of three wires carrying emfs with a phase difference of 120° between them, see Figure 5.1. (You should have noticed that pylons always carry multiples of three cables.) The advantages are that only three cables are needed for a supply and that the three phases used together (and reduced to an appropriate voltage) are ideally suited for powering rotating machinery. The structure of a three phase transformer is shown in Figure 5.2. Each primary winding is connected between two of the phases (Δ connected).

5.3 Local power distribution

Figure 5.3 shows the essentials of the local stage of distribution of power to a laboratory. The star point on the secondary of the transformer, called the neutral, is connected to a metal plate buried in the earth next to the transformer so at the transformer its voltage is the same as that of the earth. The earth is usually damp and is a reasonably good conductor. The order of occurrence of the positive voltage peaks in the three phases is indicated by blue, yellow and red colour coding of the wires. There is 240 volts rms between each phase and neutral and 415 volts rms between phases. Three phase supplies are used to drive equipment such as large motors and heavy duty ac to dc power converters.

For supplying low power fittings and apparatus the three phases are separated into three two wire supplies consisting of one live wire and neutral. Note that blue is used both for one of the three phases and for the neutral of the single phase outlets.

5.3.1 Voltage on the neutral line

When load current is being drawn there are small voltage drops along the live and neutral wires due to their resistance. The voltage on the neutral wire will therefore be slightly different in the building from its value (earth) at the substation, the difference tending to be greatest for the outlets furthest away from the transformer.

Many pieces of equipment and fittings are nonlinear and when a sine wave voltage is applied to them they draw currents with waveforms that are far from sinusoidal. (Fluorescent lights are a good example.) The waveform of the voltage with respect to earth of the neutral line may therefore be quite complicated.

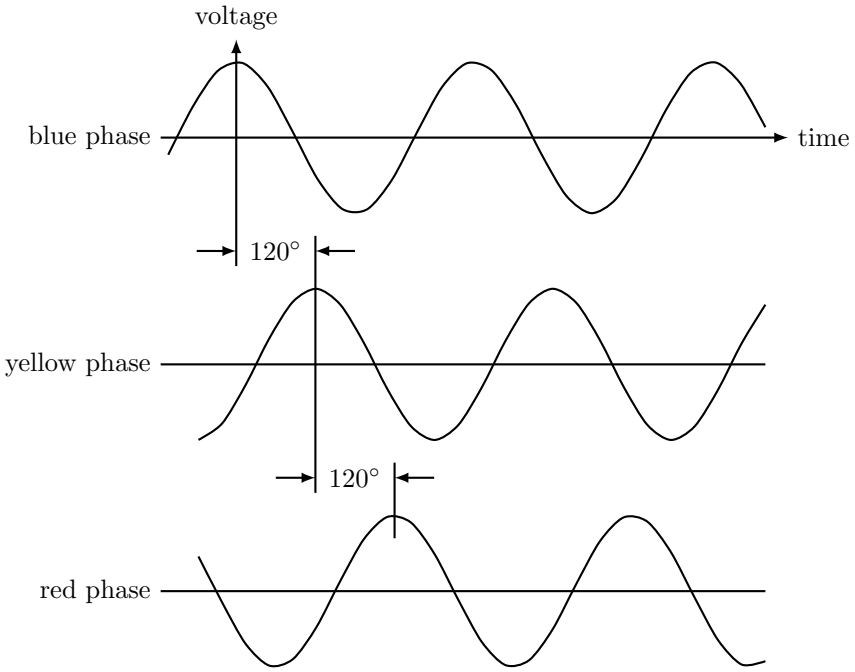


Figure 5.1: Emfs with respect to earth in a three-phase supply

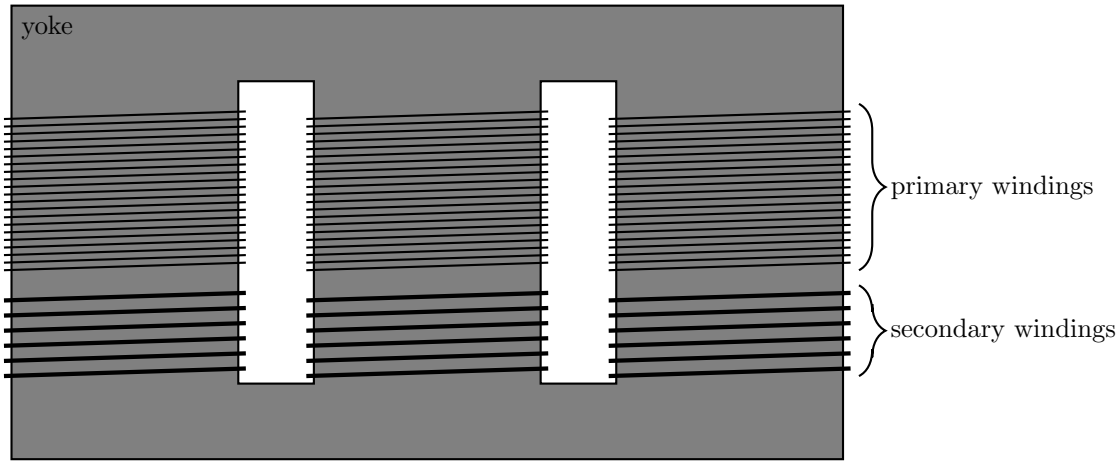
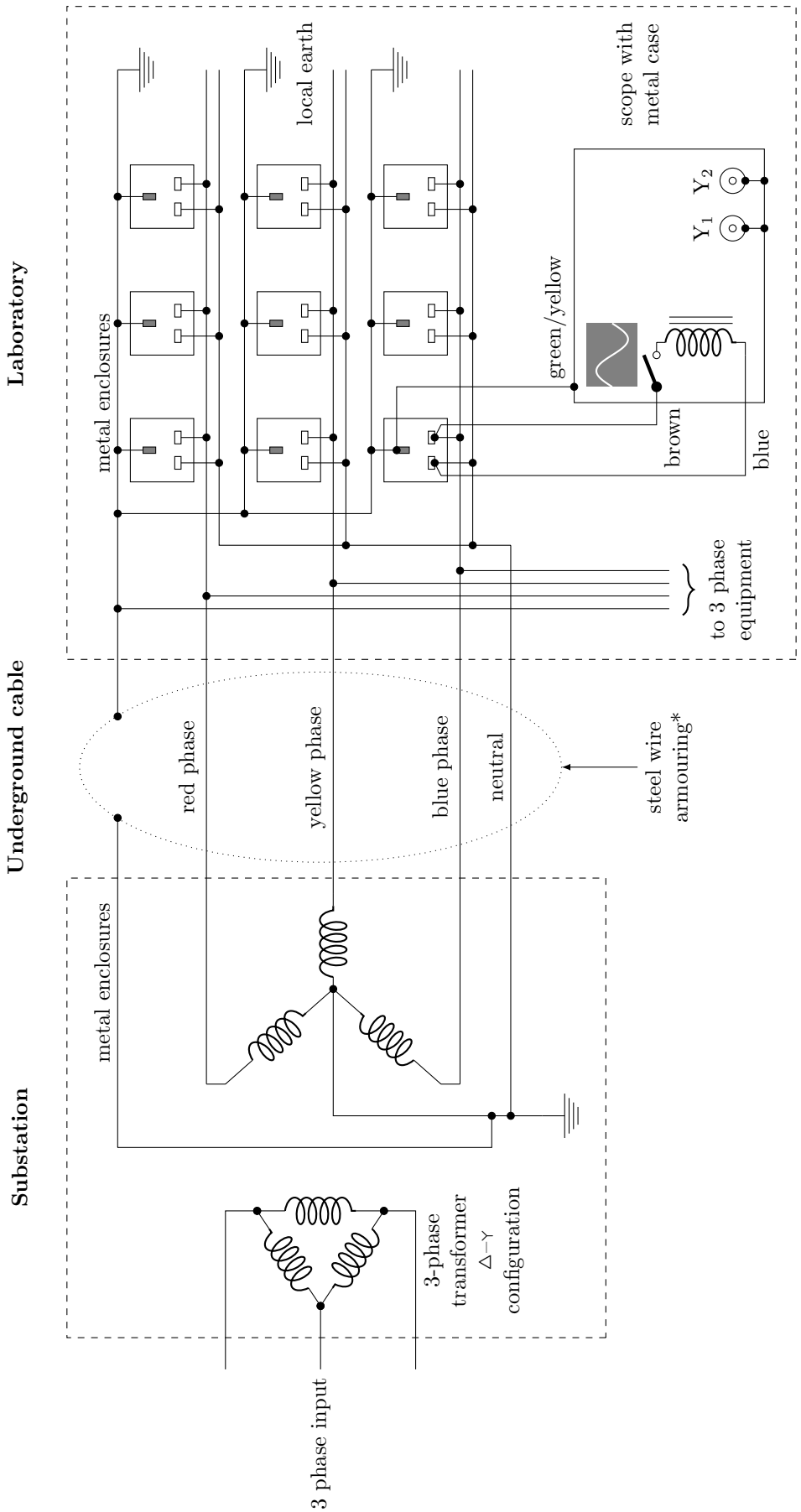


Figure 5.2: Three-phase transformer

5.3.2 Safety aspects

Most of the time you will be in some sort of contact with the earth via the structure of the building and if you touch the live wire you will complete another path back to the transformer via the earth (in parallel with the neutral wire) and a current will flow through you. 50 mA is usually fatal. For this reason live wires are carefully protected either with two separate insulating enclosures (*double insulated*) or by a metal case connected to earth.

In most houses all the earth wires (coloured green–yellow) are connected to the rising cold water main which is a metal pipe in good contact with the earth.



* There may also be an earth wire in the cable.

Figure 5.3: UK power distribution

6 Instruments

6.1 Signal Generators

Most analogue circuits are designed to modify an input signal in some way such as amplifying it. The instruments which supply input signals are called *signal generators* or *oscillators*.

Our signal generators, GW Instek type SFG2004, employ a technique called Direct Digital Synthesis (DDS) to produce an analogue output with a sine waveform. Square and triangle waveforms derived from the sine wave are also provided. The accuracy with which the frequency can be set is 0.002%. The circuit is quite complicated but as far as its output is concerned it behaves simply as an oscillating emf and an internal resistance (about 50 Ω) in series as shown in the figure below.

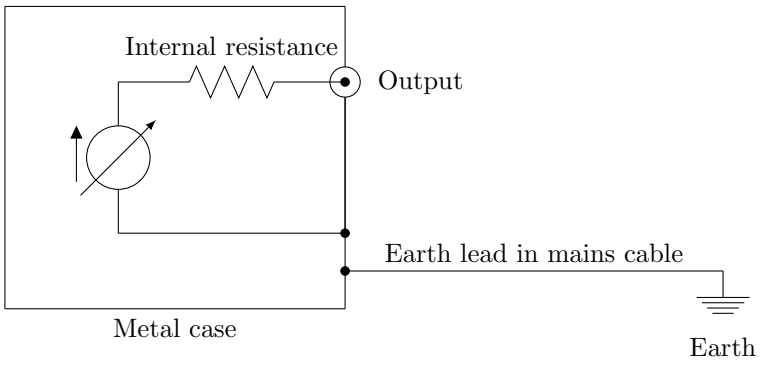


Figure 6.1: Signal generator equivalent circuit.

To explain the operation of a DDS system we consider the block diagram shown below. The numbers of bits shown are for illustration of the ideas only. In practice various design tricks are used to reduce the numbers.

The desired frequency is entered into register (i). The binary counter (iii) has 2^{28} states. The 28 bit word in the counter is fed to the 28 bit address input of the Read Only Memory (v). The memory contains 2^{28} 10 bit samples, equally spaced in phase, of one cycle of a sine wave. If the clock (ii) steps the counter through all its states it will take 2^{28} clock pulses to read out the whole sine wave from the ROM and the output frequency will be $10^7/2^{28}$Hz = 0.0372529...Hz. The digital samples clocked out from the memory are converted into analogue voltages in the digital to analogue converter (DAC), (vi). This produces a waveform with steps in it. The low pass filter (vii) smooths out these steps and delivers a good sine wave.

If the the counter were to jump M states on every clock pulse instead of one there would be M times fewer samples which would be clocked out in $1/M$ of the time making the output frequency M times 0.0372529...Hz. There would still be plenty of samples with which to construct the sine wave (Nyquist says we need only 2!).

Let us suppose that we want a frequency of 3156.0 Hz. We will get this if M = 84718.23..... but surely M must be a whole number? Yes, it must be, but suppose our jumps were not quite equal, say one jump of 84719 states for every three jumps of 84718 states or better 77 jumps of 84718 for every 23 jumps of 84719. The first seven figures of the average M would then be equal to the desired M and the idea can be extended if higher accuracy is required. The small jitter on the time axis due to the unequal jumps is removed by the filter. It should be stressed that this is just one way of doing DDS. Another way is to use an adder and a register rather than a counter and successively add the desired value of M, including sufficient figures after the decimal point, to the register contents and then dropping the numbers after the decimal point to produce the sequence of addresses for the ROM.

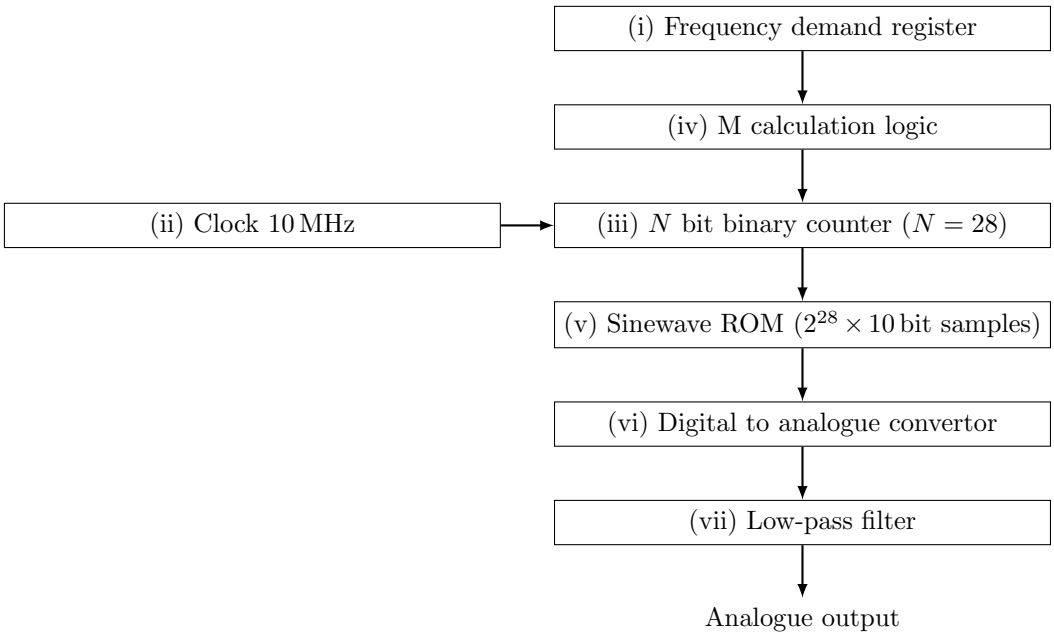


Figure 6.2: Block diagram of the Direct Digital Synthesis process.

Using the SFG2004

The desired frequency is entered using the push buttons on the front panel. There are 2 output sockets, one labelled OUTPUT 50 Ω (the internal resistance) delivers sine, square, or triangle waveforms (selected with the WAVE push-button) centred on earth voltage (which we define to be 0 volts). The other output socket labelled TTL/CMOS delivers positive rectangular pulses.

The outer parts of the sockets are connected to the metal case of the oscillator which (when the instrument is plugged in) is connected to a wire in the mains supply which finds its way (eventually) to a metal rod driven into the earth.

6.2 The analog oscilloscope

6.2.1 Cathode Ray Tube (CRT)

Look at Figure 6.3. Where the beam of cathode rays (electrons) hits the fluorescent coating a spot of light is produced. The brightness of the spot depends on the number of electrons in the beam which in turn depends on the voltage between the grid and the cathode which is set by the INTENSITY control. The FOCUS control works by varying the shape of the electric field in the anode region.

A voltage difference (and therefore an electric field) between the X plates deflects the beam in the horizontal direction. An electric field between the Y plates deflects the beam in the vertical direction.

6.2.2 Vertical deflection

The spot can be deflected vertically by the voltage applied to the CH1 (Channel 1) input, the voltage applied to the CH2 input, the sum of the two voltages, or the difference of the two voltages depending on the settings of the DISPLAY MODE and NORMAL/INVERT switches. Two other settings of the DISPLAY MODE switch (CHOP and ALT) allow the CH1 and CH2 inputs to be displayed alternately. The Y POSITION controls set the steady vertical deflections.

The input sockets accept *coaxial cables*, cables in which one conductor surrounds the other. When a cable is plugged in, its outer conductor is connected to the case of the oscilloscope which is connected to earth.

The input resistance of a channel is that of the chain of resistors associated with the VOLTS/DIV control and is almost invariably 1 MΩ. This is ten times smaller than the resistance of the DMM set to volts but is still high enough to have a negligible effect on most circuits it is connected to.

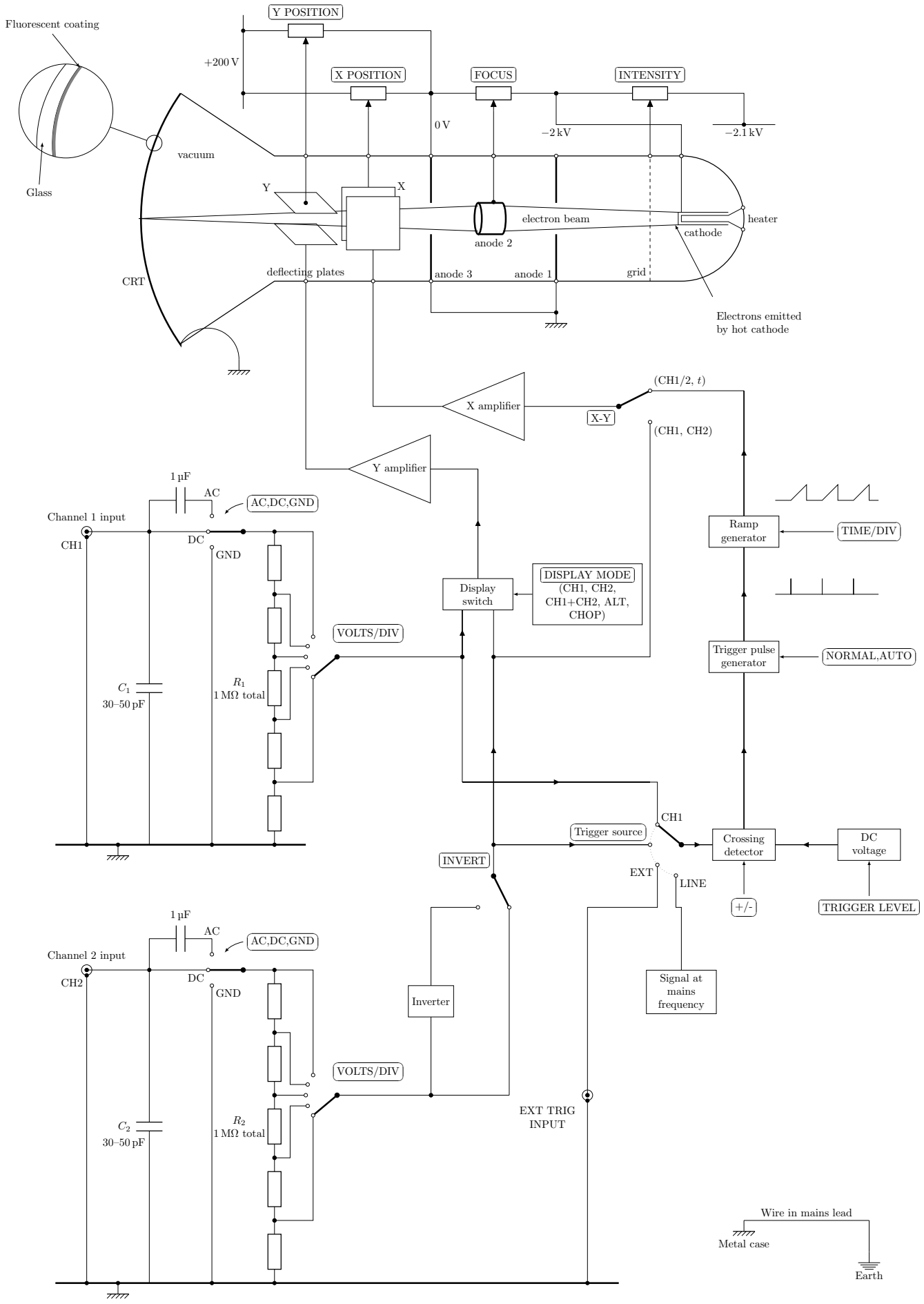


Figure 6.3: Typical 2-channel analog oscilloscope

6.2.3 Horizontal deflection

The voltage applied to one X plate comes from the X POSITION control. In normal operation the other plate has a voltage applied it called the *timebase*, which is generated by circuits inside the scope, see below. The X-Y switch allows the timebase to be replaced by the voltage in channel 2 when an X-Y display (CH1 against CH2) is required.

6.2.4 Timebase and triggering

The timebase comes from the ramp generator and has the sawtooth shaped waveform shown in Figure 6.4. It causes the spot to move horizontally from left to right at steady speed then flyback rapidly to the left hand side of the screen where it waits. Display of the voltage on the Y plates occurs during the ramp (or *sweep*), during the wait and the flyback periods the grid voltage in the CRT is changed to cut off the electron beam and nothing is visible. The duration of the ramp is set by the TIME/DIV control.

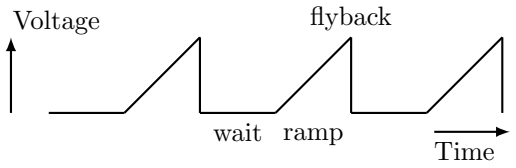


Figure 6.4: Oscilloscope timebase voltage showing the wait, ramp and flyback periods.

Because a spot dies away in about 1/10 second after the beam has been deflected to a new position a stationary display of a voltage on the Y plates is only possible if (i) the voltage repeats exactly so that successive traces can be overlaid, and (ii) it repeats often enough for the brightness to be maintained.

To overlay successive traces each ramp must start at the same point on the waveform being studied. This is achieved as follows. Suppose it is the waveform at the CH1 input we wish to observe; the first thing to do is to set the TRIGGER SOURCE to CH1. This feeds the signal in the CH1 channel to the crossing detector. A dc voltage, adjustable with the TRIGGER LEVEL control, is also fed to the crossing detector. When the CH1 voltage crosses this dc voltage the crossing detector generates a narrow pulse. This pulse is passed to the ramp generator causing it to start a ramp (see Figure 6.5).

If the CH1 voltage does cross the dc voltage it will actually do so twice per cycle, once when it is rising and once when it is falling, see the figure. The +/− switch selects which point produces a trigger pulse, + means the point where the CH1 voltage is rising (has a +ve slope).

If there are no crossings the screen will be blank which is unhelpful. Switching the NORMAL/AUTO switch to AUTO mode causes trigger pulses to be produced as above when there are crossings but causes pulses to be generated automatically (at about 30 per second) so something can be seen if there are no crossings. Setting the TRIGGER SOURCE switch to LINE causes trigger pulses to be generated at mains frequency.

6.3 The 2-channel digital oscilloscope

In the brief outline description given below any numbers quoted refer to our scopes. The control settings that need to be made on our digital scopes are mostly the same as those you would need to make for an analog scope (which some of you may have seen before) and many of the names are the same. The differences are that the digital scope has more options and they are selected using pop up menus rather than manual switches.

The channel 1 and channel 2 input circuits in the digital scope are the same as those shown in the bottom left hand corner of Figure 6.3 for the analog scope. The input of each channel can be set to AC, (a capacitor is inserted in series with the input), DC (the input is directly coupled), or GND (the input is disconnected from the signal source and connected to the metalwork of the scope which is connected to the earth wire in the power supply cable. Thereafter the treatment of the signals in a digital scope is different from that in an analog scope. On emerging from the input circuits of a digital scope the signal voltage is connected across a storage capacitor for a short time, probably of the order of 1 ns, using fast logic circuits as switches. This operation is called sample

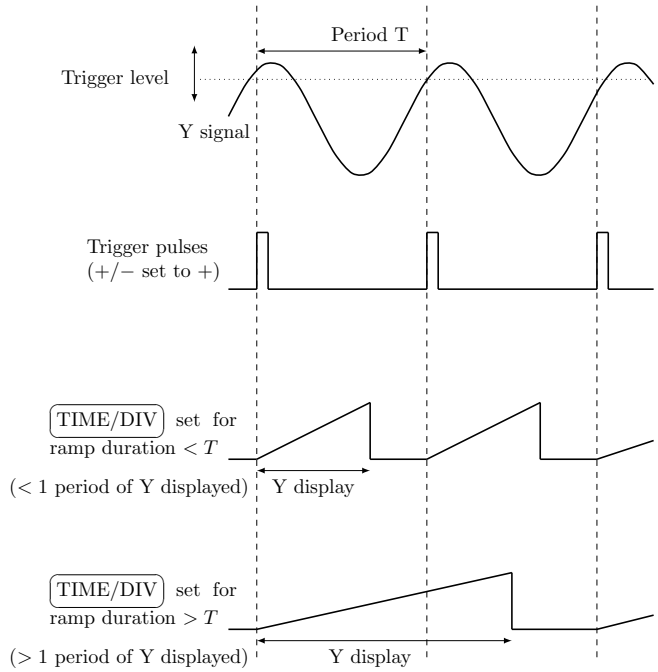


Figure 6.5: Oscilloscope triggering waveforms

and hold. Samples are taken at regular intervals appropriate for the signal being studied; the maximum rate of sampling is 10^9 per second.

The voltage on the capacitor at the end of the 1 ns sample window is applied to an analog to digital converter (ADC) that delivers an 8-bit word. At the maximum sample rate the digitisation has to be completed in not more than 1 ns. Only in the last decade or so has such performance been readily available.

For processing the data it is helpful to consider a particular number of consecutive samples and information about the time each sample was taken to constitute a record or data frame. Tektronix have defined a record as containing 2,500 samples. After a record has been acquired the system waits for the next trigger condition to be met before starting a new acquisition.

On the screen the x coordinate of an illuminated pixel is derived from the time the input was sampled and the y coordinate is derived from the digitised value of the input voltage at the time of the sample. Pixels are either illuminated or not, there is no control of pixel brightness.

The resolution of the screen is 320 pixels in the horizontal whereas a data frame has 2500 samples. The compression is realised by selecting a subset of the samples for display.

The above is provisional information. We are in contact with Tektronix to obtain more detail.

6.4 Digital Multimeters (DMMs)

Despite coming from a variety of manufacturers and looking somewhat different all our DMMs have very similar capabilities. Again the actual circuits are quite complicated but they behave as if they consist of an ideal voltmeter (one which draws no current) with extra components at the input depending on what is being measured. The ideal voltmeter consists of a range switch (which may be automatic) and a circuit which converts an analogue input voltage to a digital number which is then displayed, see the part of the figure below to the right of the dotted line.

Measuring voltage (Figure 6.6)

When the DMM is set to volts a high resistance is placed between the the input terminals of the ideal voltmeter. This resistance, the *input resistance* of the DMM set to volts, is made a standard value, almost invariably 10 MΩ. This is high enough for the act of connecting the DMM not to affect most circuits. Neither of the two input terminals is connected to earth.

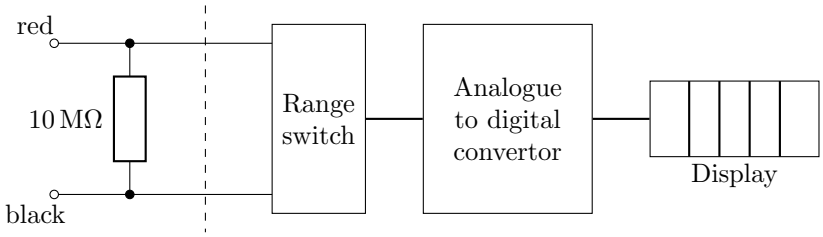


Figure 6.6: Digital multimeter measuring dc voltage.

Measuring current (Figure 6.7)

Setting the DMM to measure current places a low value resistance (typically 1 Ω) between the voltmeter terminals. The current to be measured I_x causes a voltage drop across this resistance which is measured by the ideal voltmeter.

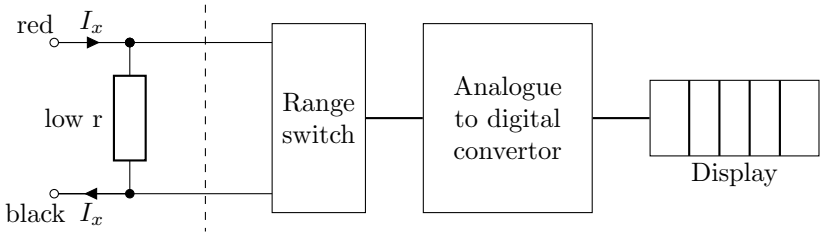


Figure 6.7: Digital multimeter measuring dc current.

Measuring resistance (Figure 6.8)

Setting the DMM to measure resistance places a source of current in parallel with the voltmeter. A direct current I_m flows out and through the resistance being measured and the resulting voltage drop is measured by the ideal voltmeter.

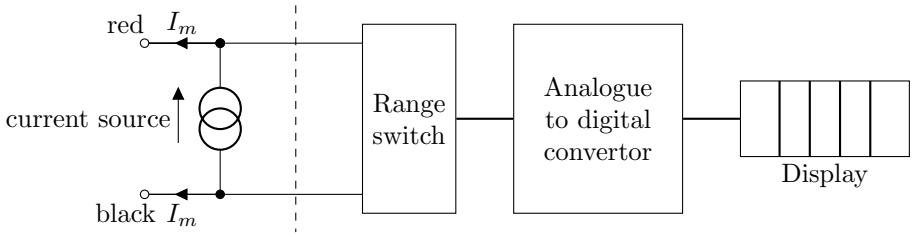


Figure 6.8: Digital multimeter measuring resistance.

6.5 More on digital multimeters

A block diagram of a typical DMM (the Hewlett Packard 3476A) is shown in Figure 6.9. We shall describe what each block does working backwards from the display.

6.5.1 The display

The digits in the display are of the 'seven segment' type thus: 8; the segments are bar shaped light emitting diodes (LEDs).

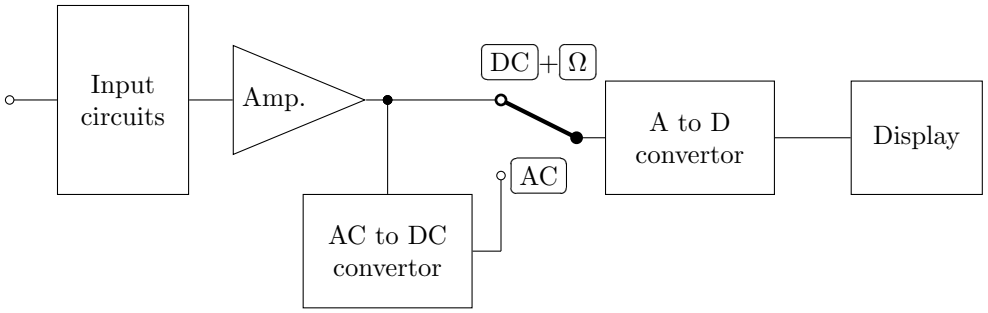


Figure 6.9: Block diagram of a typical digital multimeter (Hewlett Packard 3476A).

6.5.2 The analogue to digital converter (ADC)

Various kinds of ADC are used in digital multimeters. One common type is the dual-slope converter comprising a clock generating a square wave of period τ, a counter, and an integrator. A conversion cycle begins with the integrator output set to zero. V_{in} is then connected to the input of the integrator. After N_s counts (a time $N_s\tau$), V_{in} is disconnected and a negative reference voltage applied in its place. The integrator output then falls at a known and constant rate (see Figure 6.10) and the counter counts n clock pulses during the time t_f taken for the output to reach zero.

The interval t_f is proportional to the average value of V_{in} over $N_s\tau$. The time $n\tau$ represents t_f within the resolution of the ADC and by choosing suitable values for the variables within the converter, n represents the average value of V_{in}. At the end of the conversion cycle, n is presented at the output and the cycle then starts again.

In this country, $N_s\tau$ is chosen to be a multiple of 20 ms so that during dc measurements any interference from the 50 Hz mains averages to zero.

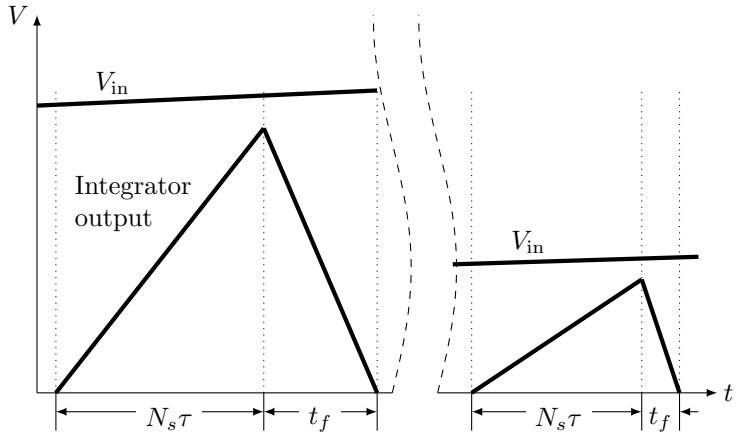


Figure 6.10: Operation of dual slope ADC

It should be noted that because t_f is represented by an integer number of clock pulses, a steady input voltage can lead to a display in which the last digit is flickering.

6.5.3 The ac to dc converter (rectifier)

The rectifier first subtracts any steady component in its input (like the AC setting of the AC/DC switch on the oscilloscope) and then inverts alternate half cycles of the ac waveform to produce a unidirectional output (unsmoothed dc).

6.5.4 Frequency responses of instruments

All voltmeters have a limited range of frequencies over which their calibration can be relied on. Typical frequency response curves for the instruments we have described are shown in Figure 6.11.

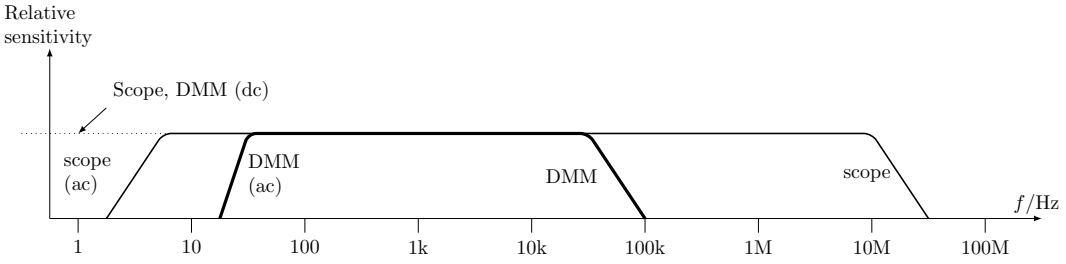


Figure 6.11: Typical frequency response curves for oscilloscopes and digital multimeters.

6.6 Bridges

For a network of components with four terminals (two input and two output), a quantity of interest is the ratio of the output voltage to the input voltage, the *voltage transmittance* T. A bridge is a four-terminal network with the property that the transmittance is zero when the components have particular values.

In a bridge used for measurements of component values one of the components in the network is the unknown whose value is being sought and one or two of the other components have adjustable values. Adjustments are made until the output is zero at which point the bridge is said to be *nulled* or *balanced*. From the settings of the adjustable components at balance the value of the unknown can be read off. Depending on the component being measured, the generator providing the input voltage will be dc or ac at some suitable frequency. The output voltage is indicated by a meter of the appropriate type, usually called a 'detector' in this application.

Examples of bridge networks are given below. It is a property of such networks that the positions of the generator and output meter can be interchanged. Until recently, component values were usually measured on bridges. Nowadays much more convenient instruments are available (see section 1.6) but bridge circuits are popular with examiners.

6.6.1 Resistance (Wheatstone) bridge

See Figure 6.12.

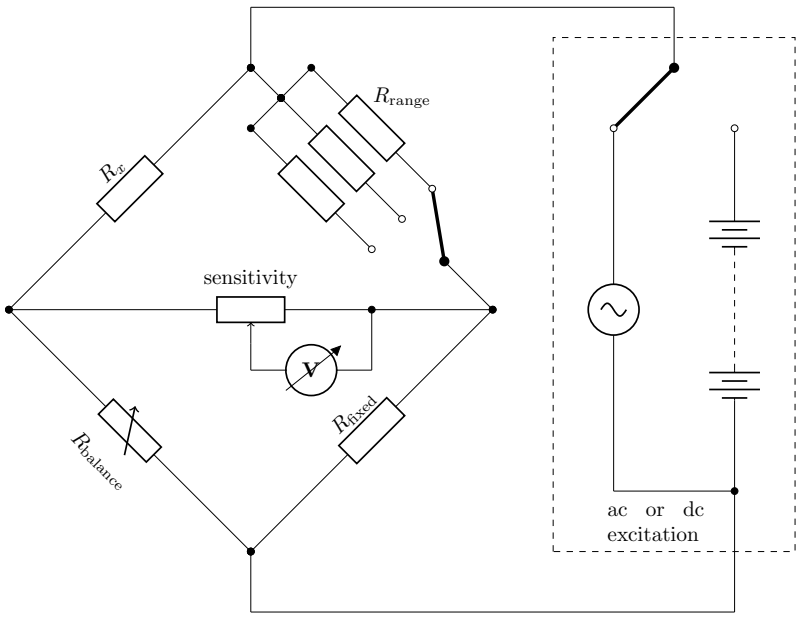


Figure 6.12: Resistance bridge

The balance condition is:

$$R_X = \frac{R_{\text{range}}}{R_{\text{fixed}}} R_{\text{balance}}$$

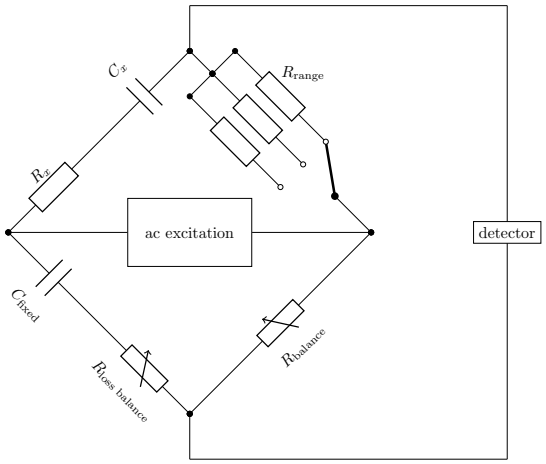


Figure 6.13: Capacitance bridge

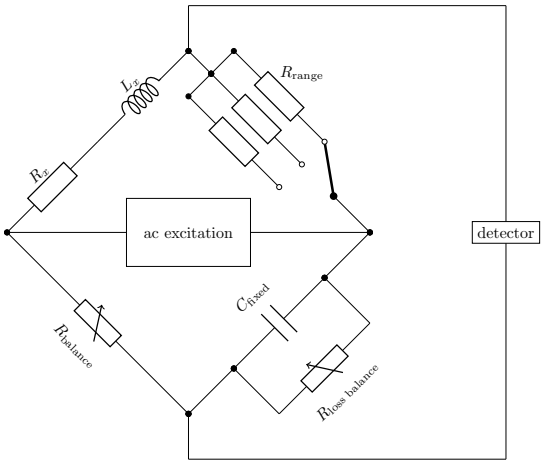


Figure 6.14: Inductance bridge

6.6.2 Capacitance bridge

See Figure 6.13. Both R_{balance} and $R_{\text{loss balance}}$ must be adjusted to achieve balance.

At balance,

$$C_X = \frac{R_{\text{balance}}}{R_{\text{range}}} C_{\text{fixed}}$$

and:

$$r_X = \frac{R_{\text{range}}}{R_{\text{balance}}} r_{\text{loss balance}}$$

Using this circuit the non-ideal behaviour of the capacitor at the frequency of the excitation is expressed as a stray series resistance. Other bridge circuits can express it as a stray parallel resistance.

6.6.3 Inductance bridge

See Figure 6.14.

The inductance of the inductor is given by:

$$L_X = R_{\text{balance}} R_{\text{range}} C_{\text{fixed}}$$

and its losses (at the operating frequency) expressed as a series resistance by

$$r_X = \frac{R_{\text{balance}}}{R_{\text{loss balance}}} R_{\text{range}}$$

6.7 RLC Meters

These instruments provide excellent examples of how microcomputers are making measurements easier for the user. We describe the *Topward RLC meter* whose system block diagram is shown in Figure 6.15. A sine wave current is passed through the device under test and a precision resistor and the relative amplitudes and the phase difference of the two voltage drops are measured. The calculations needed to turn this raw data into information about the elements in the equivalent circuit of the device under test are performed by the microcomputer which also selects the most appropriate value of precision resistor. The frequency at which a measurement is performed can be selected by the user; the default is 1 kHz.

The values of the strays in the test fixture or connecting leads are allowed for by measuring them before a test or whenever the test fixture or leads are altered. Measurements are made (i) with no device present ('open calibration') and (ii) with a shorting link between the test leads ('short calibration'), and the results are stored in the microcomputer memory.

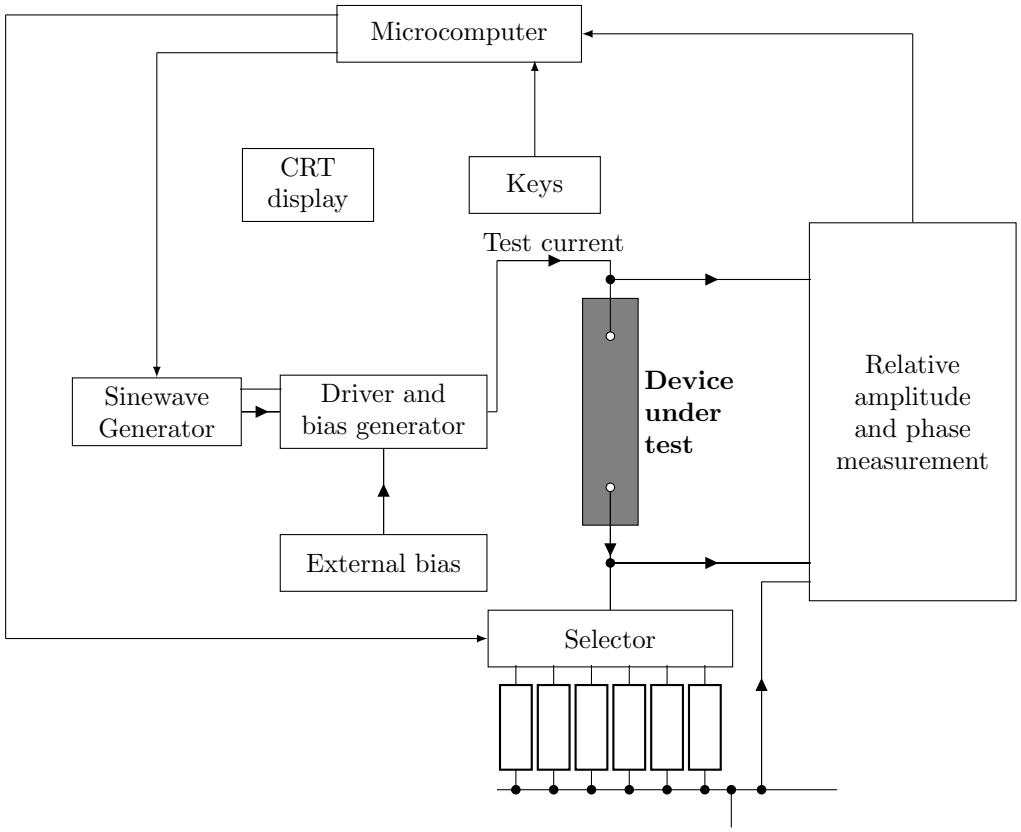


Figure 6.15: RLC Meter block diagram

6.8 Digital frequency and period meters and timers

6.8.1 Frequency meter

A block diagram of a frequency meter is shown in Figure 6.16. The input of unknown frequency is passed first through shaping circuits which convert it into a train of pulses. These pulses are then fed via a gate into a counter. The time for which the gate is open is derived by counting a number of pulses from a precise quartz crystal oscillator (typically operating at 1 MHz) which can be selected by a control on the front panel. (A mode of operation with automatic selection of a suitable gate time is often available also.) The position of the decimal point in the display is changed to reflect the gate time setting. The control logic produces not only the 'gate open' signal but also signals to reset the counter to zero and to transfer its contents to the display register at the end of the count.

6.8.2 Period meter

For low audio frequencies and below the gate time required for a measurement of the frequency of a signal to the desired accuracy may be inconveniently long. This problem is overcome if instead a period measurement is made. All that is required is to interchange the circuits in the two shaded boxes in Figure 6.16 so the counter is fed by the oscillator divider via the gate and the input signal, after suitable conditioning, feeds the control logic. A switch on the front panel (frequency/period mode) interchanges the circuits.

6.8.3 Timer

By using the circuit of Figure 6.16 in the period mode any process which can be made to produce suitable open and close signals for the gate can be timed.

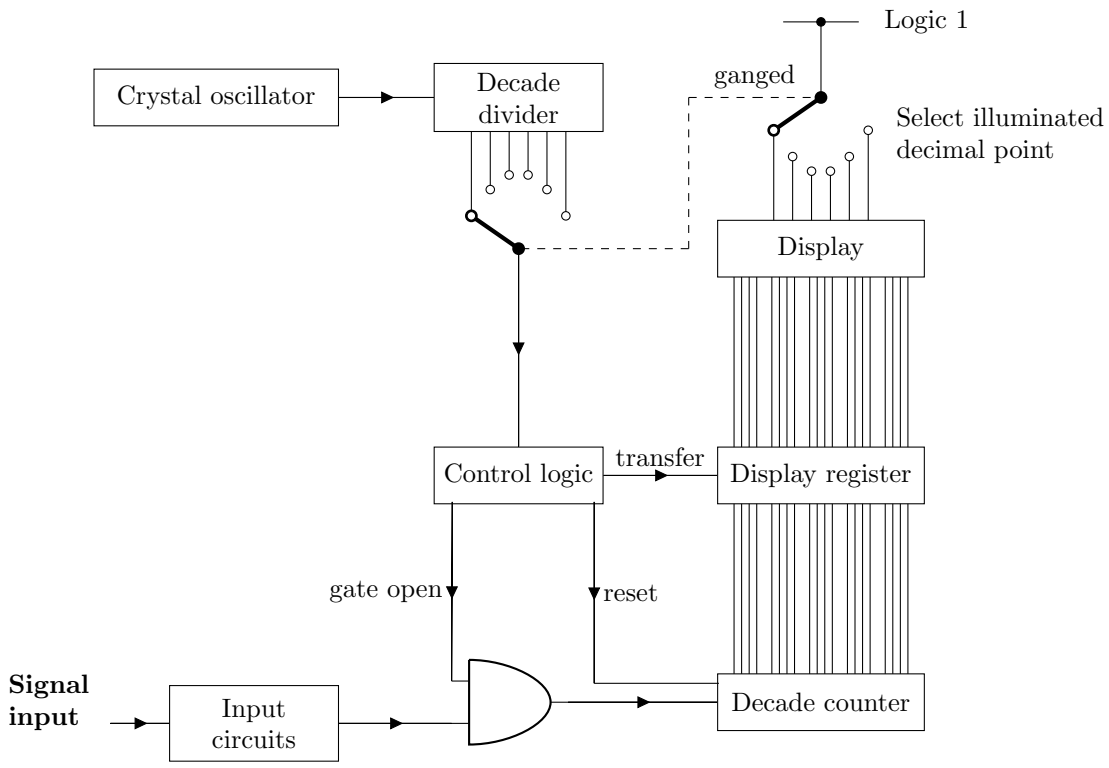


Figure 6.16: Digital timer block diagram

7 Observations, Errors and Tolerances

7.1 Introduction

In this chapter we consider observations, instrumental errors, and their simple treatment. The markings indicating the nominal values and tolerances of electronic components are also described.

7.2 Observations

We consider observations made using analogue and digital "instruments". By "instrument" we do not mean a particular oscilloscope. We mean a particular channel, range, and position setting on the oscilloscope.

All observations, both analogue and digital, are subject to *reading errors* and *calibration errors*. A typical observation comprises records of:

(i) the position of a pointer against a scale (analogue) or a displayed number (digital)

(ii) the range setting

(iii) a statement of the reading error (analogue) or resolution error (digital). These are random errors.

(iv) a statement of the calibration error (analogue and digital). These are systematic errors.

All of these records are required to make the observation complete. "Pointer" is rather a general term, it could for example be a point on a waveform trace on an oscilloscope.

Often we will be dealing with sets of observations made with the same instrument. Using the same instrument (which we remind you means not even adjusting the zero) has certain advantages as will be seen below.

7.2.1 Reading/resolution errors

The reading error (*re*) of an analogue observation is the error in reading the position of the pointer against the scale. Its maximum value is the smallest discernible difference in the position of the pointer and is usually much smaller than the finest marked scale divisions. The resolution error of a digital observation has a maximum value of half a 1 in the least significant figure.

7.2.2 Calibration errors

In a ideally calibrated analogue instrument the graduations on the scale correspond to truth to much better than the reading error, i.e. the reading error dominates. In a realistically calibrated instrument the calibration error may dominate. We take the calibration error of an analogue observation on a realistically calibrated instrument to be the departure from truth at the nearest scale division.

In an ideally calibrated digital instrument the inputs at which changes of the least significant digit occur correspond to truth to much better than the resolution error. We take the calibration error of a digital observation on a real instrument to be the departure from truth at the nearest least significant digit change.

We will restrict the discussion of the calibration errors of instruments to cases where they can be expressed as the combination of a zero error (*ze*) and a constant percentage of the reading, the latter implying that the intervals u between the inputs at which the least significant digit changes or the intervals v at which the pointer exactly overlays the divisions are all equal but differ from the nominal values by a scale error $1+\eta$ where $-\epsilon \le \eta \le \epsilon$ and ϵ is the value stated in the manufacturer's specification. The values of ϵ for our instruments are listed at the end of this chapter.

All instruments suffer drifts of their scale and zero errors over time often related to temperature changes. The drifts are usually worst immediately after switch on so turn on any instruments you plan to use as early as possible. Drifts in scale errors are usually less troublesome than zero error drifts.

7.3 processing of observations

Consider first just the two pairs of observations:-

(i) The position of an upper peak of a sinusoidal voltage measured using Channel 1 of an oscilloscope at a vertical scale of 2 volts/division:
9.35 above the bottom graticule line ±0.05 (re) ± (ze) ±5% (se) divisions,

The lower peaks of the sinusoidal voltage measured with the same instrument:
3.10 above the bottom graticule line ±0.05 (re) ± (ze) ±5% (se) divisions.

(ii) A dc voltage measured with a digital meter:
7.03 ± 0.005 (re) ± (ze) ±1% (se)

The voltage measured with no excitation:
0.03 ± 0.005 (re) ± (ze) ±1% (se)

The only processing we would probably want to do would be to take the differences of the two observations in each pair to eliminate the zero error. For the first pair this would yield the peak to peak voltage of the sinewave which would be, after taking into account the range setting, 12.5 ± 0.2 re ±5% se volts. For the second pair the difference would be 7.00 ± 0.01 (re) volts ±1% (se).

Note that in both cases the scaling error on the difference is the same as the scaling error on the individual observations, and that we have simply added the reading/resolution errors. Adding the reading errors is a reasonable thing to do with just two observations and in arriving at a combined error for the difference we would probably also add the scaling error to be safe because with just two observations there is a real chance that all the errors will add. (Note in passing that in our electronics laboratory signals are generally steady within the reading errors so there is no point in repeating measurements at exactly the the same level of excitation.)

Now consider a large set of observations on a system at different levels of excitation plus sufficiently frequent observations at zero excitation to track any zero drift. In the most favourable cases it will be sufficient just to take observations at zero excitation before and after a series of observations with the system excited.

The first step in processing would be to take differences between the excited observations and the most contemporary zero observations. With a large data set it would clearly be pessimistic to add the reading/resolution errors in each pair of observation so what would be a suitable way to treat these errors? We will examine this for the digital case and then show how the conclusions can be applied to the analogue case.

When a digital instrument displays a number N the value of the input can be expressed (using the definitions in section 7.2.2) as $N(1+\eta)u + \alpha u(1+\eta)/2 + ze$ where $-\epsilon \le \eta \le \epsilon$ and $-1 \le \alpha \le 1$ but we do not know their actual values.

A difference of two observations in a set taken with the same instrument is:

$$N_1(1+\eta)u + \frac{\alpha_1 u(1+\eta)}{2} - \left(N_2(1+\eta)u + \frac{\alpha_2 u(1+\eta)}{2}\right) = (N_1 - N_2)(1+\eta)u + \frac{(\alpha_1 - \alpha_2)u(1+\eta)}{2} \tag{7.1}$$

assuming the zero error has not changed significantly between the observations. The error for particular values of α_1, α_2 is δ, the difference between this expression and the difference $(N_1 - N_2)\left(u(1+\eta)\right)$ of the displayed values i.e.:

$$\delta = \frac{(\alpha_1 - \alpha_2)u(1+\eta)}{2} \tag{7.2}$$

We now assume that in our large set of observations the values of α_1, α_2 are uniformly distributed between −1 and +1 which allows us to form a mean square resolution error for differences of observations in the set as:

$$\frac{\int_{-1}^{1}\int_{-1}^{1} \frac{(\alpha_1-\alpha_2)^2 u^2}{4}\, d\alpha_1\, d\alpha_2}{\int_{-1}^{1}\int_{-1}^{1} d\alpha_1\, d\alpha_2} \tag{7.3}$$

where we have ignored the effect on this of the deviation of the scaling factor from unity as second order.

It is left as an exercise to show that the result is $\frac{1}{6}u^2$ so the root mean square resolution error in a difference (that should be used to calculate the resolution error when curve fitting in Mathematica) is $\sqrt{\frac{1}{6}}u$.

We can use the same analysis to work out the error to be ascribed to differences in analogue observations. Suppose we have an analogue reading R = 11.3 which we can see could equally well be 11.2 or 11.4 but we can see

could not be 11.1 or 11.5. We will then say that the value lies between 11.15 and 11.45, so the quantity equivalent to u, which we will call v, is 0.3. We could then describe the observation to be:

$$\frac{R}{v}(1+\eta)\,v+\alpha v\,(1+\eta)/2-zo \tag{7.4}$$

where $R = Nv$ and the rms error in a difference of analogue observations can be taken to be $\sqrt{\frac{1}{6}}v$.

So much for simple differences of observations. Frequently we are concerned with measurements of the transmittances or gains of circuits so we need to know how to calculate the error for a ratio of two differences e.g. in the digital case:

$$\frac{N_1 u(1+\epsilon)+\frac{\alpha_1 u}{2}-\left(N_2 u(1+\epsilon)+\frac{\alpha_2 u}{2}\right)}{N_3 u(1+\epsilon)+\frac{\alpha_3 u}{2}-\left(N_4 u(1+\epsilon)+\frac{\alpha_4 u}{2}\right)} \tag{7.5}$$

where $-1 < \alpha_1, \alpha_2, \alpha_3, \alpha_4 < 1$.

The error for particular values of $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ is the difference δ between this ratio and the ratio $(N_1 - N_2)/(N_3 - N_4)$ of the displayed values which is easily shown to be given by:

$$\delta = \frac{N_1 - N_2}{N_3 - N_4}\left(\frac{\alpha_1 - \alpha_2}{2(N_1 - N_2)} - \frac{\alpha_3 - \alpha_4}{2(N_3 - N_4)}\right) \tag{7.6}$$

if $\frac{(\alpha_1-\alpha_2)(\alpha_3-\alpha_4)}{4(N_1-N_2)(N_3-N_4)}$ can be ignored (which we will assume it can). Further assuming that the values of $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ are uniformly distributed between -1 and $+1$ the mean square error in the ratio of the two differences is:

$$\frac{\int_{-1}^{1}\int_{-1}^{1}\delta^2\,\mathrm{d}\alpha_1\,\mathrm{d}\alpha_2}{\int_{-1}^{1}\int_{-1}^{1}\mathrm{d}\alpha_1\,\mathrm{d}\alpha_2} \tag{7.7}$$

which it is left as an exercise to show is equal to:

$$\frac{1}{6}\left(\frac{N_1 - N_2}{N_3 - N_4}\right)^2\left(\frac{1}{(N_1 - N_2)^2}+\frac{1}{(N_3 - N_4)^2}\right) \tag{7.8}$$

so the rms error in the transmittances is:

$$\frac{1}{\sqrt{6}}\left(\frac{N_1 - N_2}{N_3 - N_4}\right)\sqrt{\frac{1}{(N_1 - N_2)^2}+\frac{1}{(N_3 - N_4)^2}} \tag{7.9}$$

This is the expression to be used to calculate the errors to be fed into Mathematica.

Setting $N = \frac{R}{v}$ shows that the rms error in the analogue case is:

$$\frac{1}{\sqrt{6}}\left(\frac{R_1 - R_2}{R_3 - R_4}\right)\sqrt{\frac{1}{(R_1 - R_2)^2}+\frac{1}{(R_3 - R_4)^2}} \tag{7.10}$$

The rms errors in squares of the transmittance are double these.

7.4 Calibration errors of instruments

(Values of ϵ in the maximum scaling factor $1+\epsilon$).

7.4.1 Oscillators or Arbitrary Function Generator (AFG)

Tektronix AFG-1022: (taken from datasheet available at https://www.tek.com/datasheet/arbitrary-function-generator)

Amplitude:

- Accuracy: (± 1 % of setting + 1 mV_{p-p}), (1 kHz sine waveform, 0 V offset)
- Resolution: 1 mV_{p-p}, 1 mV_{rms} or 4 digits

Frequency:

- Resolution: 1 µHz or 12 digits
- Internal Reference stability: ± 1 ppm at 0 - 40 C
- Internal reference aging: ± 1 ppm per year

DC offset:

- Range: ± (5 V_{pk} - Amplitude$_{p-p}$/2), 50 Ω load
- Range: ± (10 V_{pk} - Amplitude$_{p-p}$/2), open circuit of high Z load
- Accuracy: ± (1 % of [setting] + 1 mV + 0.5 % of amplitude (V_{p-p}))
- Resolution: 1 mV or 4 digits

7.4.2 Digital multimeters

Best accuracy:

Quantity	Tenma 72-8715
DC voltage (V)	$\pm(0.5\%+2)$
AC voltage (V)	$\pm(0.8\%+3)$
DC current (A)	$\pm(0.8\%+2)$
AC current (A)	$\pm(1.0\%+3)$
Resistance (Ω)	$\pm(0.8\%+3)$
Capacitance (F)	$\pm(4\%+3)$
Frequency (Hz)	$\pm(1.5\%+5)$
Temperature (C)	$\pm(1.2\%+3)$

Table 7.1

7.4.3 Oscilloscopes

The Tektronix TBS1052B is the main oscilloscope used in the electronics lab and has the following specifications (taken from the online datasheet: https://www.tek.com/oscilloscope/tbs1000b-edu-digital-storage-oscilloscope-manual/tbs1000b-and-tbs1000b-edu-series-oscil):

- DC gain accuracy: ± 3% from 10 mV/div to 5 V/div
- Time base accuracy: 50 ppm

The instrument listed below is intended for measurements on isolated devices i. e. devices that are not in a circuit. It may be damaged if it is connected to a live circuit or a charged capacitor.

7.4.4 Resistance, inductance and capacitance meter

Tenma 72-8155:

	Range	Best accuracy
Resistance (Ω)	200 Ω / 2 kΩ / 20 kΩ / 200 kΩ / 2 MΩ / 20 MΩ	± (0.8% + 1)
Capacitance (C)	2 nF / 20 nF / 200 nF / 2 µF / 20 µF / 200 µF / 600 µF	± (1% + 5)
Inductance (H)	2mH / 20 mH / 200 mH / 2H / 20 H	± (2% + 8)

Table 7.2

7.4.5 Semiconductor analyser

Peak Electronic Design Ltd, model DCA 55
Identifies the type and connections of almost any discrete semiconductor device.

7.5 Values markings and tolerances of passive components

7.5.1 Resistors

When resistors were first manufactured in quantity in the 1930s (using a carbon and clay technology) it was seldom necessary to know a value to better than ±20%. A geometric sequence with 6 values per decade (1, 1.5, 2.2, 3.3, 4.7, and 6.8) was adopted and values between 10 Ω and 10 MΩ were manufactured. Three coloured markings, interpreted as two digits and a decade multiplier, were used to indicate the value. Resistors measured as having values within 10% or 5% of the nominal value were given an additional silver or gold marking.

Today, resistors with values between 1 Ω and 10 MΩ are produced in very large quantities and values lying outside this range are readily available. "Ordinary" metal film resistors (see Figure M22.6) are available at 12 values per decade (1, 1.2, 1.5, 1.8, 2.2, 2.7, 3.3, 3.9, 4.7, 5.6, 6.8 and 8.2) and 1% tolerance (no longer determined by the value spacing) while the best quality mass produced resistors are available at 96 values per decade (~ 2.5% steps) and 0.1% tolerance. Five coloured bands are needed for the latter sequence, three digits, decade multiplier and tolerance. Some times a further band is used to indicate temperature coefficient. The four and five band (extra digit) colour codes in current use are shown in Table 7.3.

colour	digit	digit	(digit)	multiplier	tolerance
silver	-	-	-	×0.01 Ω	±10%
gold	-	-	-	×0.1 Ω	±5%
black	0	0	0	×1 Ω	-
brown	1	1	1	×10 Ω	±1%
red	2	2	2	×100 Ω	±2%
orange	3	3	3	×1 kΩ	-
yellow	4	4	4	×10 kΩ	-
green	5	5	5	×100 kΩ	-
blue	6	6	6	×1 MΩ	-
violet	7	7	7	-	-
grey	8	8	8	-	-
white	9	9	9	-	-

Table 7.3: Value shown: yellow, grey, violet, orange, brown is 487 kΩ ± 1%

The band closest to an end of the resistor is the first digit. Colour codes are not used on resistors designed to run at high temperatures (convection cooled high power types) because painted bands would be blackened and become unreadable. Heat resisting inks are used to mark the value and tolerance in plain text.

Note that on circuit diagrams 4.7 kΩ is often written 4k7, the multiplier letter replacing the decimal point and the Ω being omitted as redundant (given the presence of the resistor symbol on the diagram).

7.5.2 Capacitors

Capacitors are made with typically 6 values per decade and plainly marked with the value, tolerance and working voltage but smaller sizes might be colour coded or marked using the <*number*> <*number*> <*multiplier*> pF code (i.e. 104 represents 10×10^4pF or 100 nF). On circuit diagrams a capacitance of 2.2 nF might be labelled 2n2.

7.5.3 Inductors and transformers

In general, wound components are made specially for particular applications and may not be marked. However inductors with a few standard values are manufactured and these are usually plainly marked.

8 Measurement Techniques

8.1 Record keeping

Without duplicating material in the practical script, include in your working records:-

(a) the name of the practical and the starting date and time. Also record the time occasionally as the session proceeds.

(b) a list of the instruments used,

(c) the observations, see chapter7,

(d) a description of the data processing; and

(e) a discussion of the errors and their treatment.

If you are asked to observe some change in a parameter and within the resolution of your equipment you see no change, specify the limit of resolution, i.e. say 'the change observed was smaller than x', it is unhelpful to say 'no change was observed'.

8.2 Experimental considerations

8.2.1 Earths and earthing

Most mains powered electronic equipment has a metal case which is connected to earth and often one of the signal terminals is connected to the case. The scopes and oscillators we use are like this (see e.g. Figure 6.3) The existence of these permanent earth connections has important consequences for the way measurements can be done. The rule is that *all earthed wires that are connected to a circuit must go to the same point* — if you break this rule the parts of your circuit between the different points you have earthed will be short circuited because, out of sight, all earth wires are connected together (as shown in Figure 5.3).

8.2.2 Interference

Unless extraordinary precautions are taken the space in any laboratory will be threaded by significant electric and magnetic fields which can induce interfering signals in electronic apparatus.

The oscillating charge on the live wire in mains cables is a potent source of electric fields. Consider such a field terminating on a conductor which is connected to earth. As the field oscillates (at 50 Hz) the charge induced on the conductor varies in sympathy and a current flows to and fro between it and earth. If the path connecting the conductor to earth has significant resistance this current causes the voltage with respect to earth of the conductor to vary. The larger the resistance between the conductor and earth the larger the unwanted voltage on the conductor. The effect can be eliminated by placing conducting sheets above and below the sensitive conductor and connecting the sheets to earth via a low-resistance path so that as currents flow on and off the sheets their voltage does not change. Often a single earthed sheet placed under the sensitive conductor will reduce the interference to manageable levels by providing a large surface on which the interfering field lines can terminate.

An emf will be induced in a loop of circuit threaded by a changing magnetic field (Faraday's law). This effect can be reduced by repositioning the conductors to minimise the size of the loop, by changing the orientation of the loop with respect to the field, and/or by enclosing the loop in a box made of soft magnetic material which forms a bypass path for the field. Interfering magnetic fields are usually only significant near transformers, chokes and motors.

8.2.3 Cables

For transferring signals from terminals on one network to terminals on another, two wires are needed. In order of increasing immunity from pick-up of signals from external electric and magnetic fields the wires can be arranged as (i) separate wires (ii) twisted pair cable and (iii) coaxial cable. Where signal levels are small or impedances are high, coaxial cable should be used with the outer at the zero of voltage. If impedances are low, twisted pairs with one of the pair at the zero of voltage should be satisfactory. Separate wires should be used only for low impedances and large signals. This order of preference also applies to wires considered as sources of interference.

The immunity from pick-up of twisted-pair cable can be improved if both the wires have equal resistances to earth rather than one being connected to earth. This balanced arrangement is used for local distribution in the (copper) telephone network. The twisted pairs in multipair network cables are formed into further twisted bundles to minimise mutual interference or *crosstalk*.

8.2.4 Remarks on the use of voltmeters

An oscilloscope can be used as a voltmeter if an absolute accuracy of a few percent is acceptable. If more accurate voltage measurements are required and the frequency is within its specified range a digital voltmeter is the best choice. However if the waveform of the signal is not what it is thought to be or if some unsuspected interfering signal is present its readings may be seriously misleading. So a digital meter *should not be trusted unless there is an oscilloscope also connected to the circuit* to monitor what is happening.

You may choose to make measurements of, for example, the voltage amplification of an amplifier using different types of voltmeters to measure the input and the output. This is fine provided that before you start your measurements you first check the relative sensitivities of the instruments over the frequency range you intend to use. This is simply achieved by seeing how their readings compare when they are connected in parallel to the same signal source.

8.2.5 Some practical hints

(a) As explained in Chapter 7 even the simplest measurement requires two observations, one of which will usually be an observation after the excitation has been turned down to zero. It is usually sufficient to take these 'zero' observations before and after a series of measurements with the excitation applied rather than after each excited observation.

(b) Also in Chapter 7 the advantages of using the same voltmeter to measure the transmittance or gain of a circuit were spelled out. This is practical when the input and output signals are of comparable magnitude but in general they will not be and it will be necessary to change range. The calibration error will then no longer cancel out but the error introduced by changing range will probably be significantly less than the overall calibration error. Consult the manufacturer's specification.

(c) When measuring the transmittance or gain of a passive circuit, choose an input signal level that falls just below full scale to minimize the effects of reading/resolution errors. When investigating active circuits, excite them at a level just below the clipping value (see Section 13.6).

(d) Where possible, connect the COMMON (usually black) lead of your digital voltmeter to the zero volts (often earthed) and the live (usually red) lead to the point of interest. This reduces the loading on the circuit at high frequencies because the red lead has a lower capacitance to earth than the black lead.

(e) Keep mains leads away from input terminals and place an earthed metal plate under your circuit if you see interference from mains frequency electric fields. (These can be identified by the interference becoming stationary when the oscilloscope is triggered at mains frequency, the LINE position.

(f) Choose the best available waveform for triggering your scope. If you are investigating two erratic signals but there is a related good squarewave or sinewave signal present somewhere else in the circuit, use it via the EXTernal trigger input.

(g) We remind you that curve fitting in Mathematica takes account only of random errors.

8.3 Measurement of input and output resistances of networks

This section applies only in cases where any capacitances can be ignored.

8.3.1 Measurement of the output (internal) resistance of a source

We consider a Thevenin representation of the source with emf $\mathcal{E}$ and output resistance R_{out}. A voltmeter of resistance R_{M} connected to the source reads V_1, see Figure 8.1(a). Next a test resistor R_{T} is placed in parallel with the meter which then reads V_2, see Figure 8.1(b).

Using the potential divider formula, it is straightforward to show that

$$V_1 = \frac{\mathcal{E}}{1 + R_{\text{out}}/R_{\text{M}}} \quad \text{and} \quad V_2 = \frac{\mathcal{E}}{1 + R_{\text{out}}\left(\frac{1}{R_{\text{M}}} + \frac{1}{R_{\text{T}}}\right)}.$$

$\mathcal{E}$ is kept the same for both readings and it follows that

$$R_{\text{out}} = \frac{\frac{V_1}{V_2} - 1}{\left(\frac{1}{R_{\text{M}}} + \frac{1}{R_T}\right) - \frac{1}{R_{\text{M}}}\frac{V_1}{V_2}}.$$

When the resistance of the meter is large enough to be ignored, as is often the case in practice,

$$R_{\text{out}} \approx \left(\frac{V_1}{V_2} - 1\right) R_{\text{T}}.$$

Notice that:

(a) If both voltages are measured on the same range of the voltmeter their calibration errors will cancel out — which is good. For this reason you should resist the temptation to change to a more sensitive range for the smaller voltage (unless the improvement in reading error outweighs the worse calibration error).

(b) If your chosen value of R_{T} is much smaller than the unknown R_{out} then V_1 will be small and subject to large reading errors

(c) If your chosen value of R_{T} is much bigger than R_{out} then V_0/V_1 will be close to unity and the -1 in the bracket will greatly magnify the effect of the reading errors

(d) The optimum value of R_{T} is the one that drops the terminal voltage to 41.4% of its unloaded value (V_0) but one that drops it to anywhere between 30% and 60% will be satisfactory. Begin your measurement by trying different test resistors until you find one that meets this criterion. When you have found one, measure its value.

8.3.2 Measurement of the input resistance R_{in} of a network

A signal source $\mathcal{E}_g, R_g$ and a voltmeter of resistance R_{M} are connected to the input terminals of the network as shown in Figure 8.2(a). The voltmeter reads V_1. A test resistor R_{T} is then added as shown in Figure 8.2(b) and the voltmeter reads V_2.

Using the potential divider formula it is straightforward to show that

$$V_1 = \frac{R_\pi}{R_g + R_\pi}\mathcal{E}_g \quad \text{and} \quad V_3 = \frac{R_\pi}{R_g + R_{\text{T}} + R_\pi}\mathcal{E}_g,$$

where R_π is the resistance of R_{M} and R_{in} in parallel,

$$R_\pi = \frac{R_{\text{in}}R_{\text{M}}}{R_{\text{in}} + R_{\text{M}}}.$$

$\mathcal{E}_g$ is the same for both readings and can be eliminated, yielding

$$R_\pi = \frac{R_{\text{T}}}{V_1/V_2 - 1} - R_g.$$

Everything on the right hand side is known, so R_π can be calculated and then knowing R_{M}, R_{in} can be found.

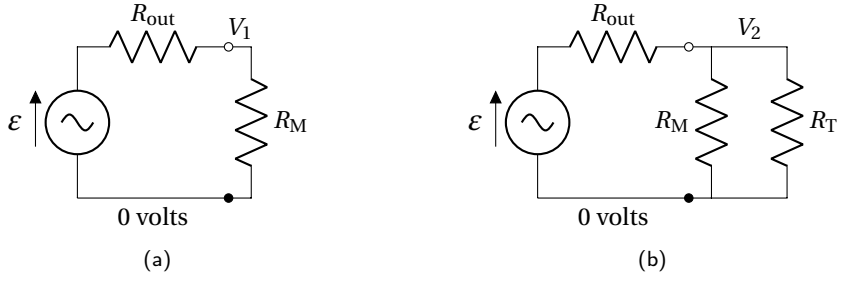


Figure 8.1: Measurement of output (internal) resistance

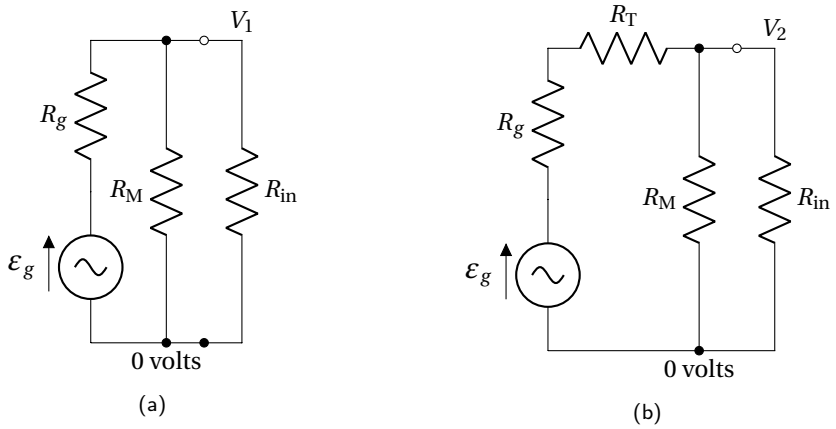


Figure 8.2: Measurement of input resistance

8.4 Phase shift measurement

Three methods using a scope:

(a) Adjust the vertical positions of the traces so that they are both accurately centred on a horizontal graticule line and measure the fraction of a period delay at the zero crossings,

(b) Use the maximum available Y shift on both beams and adjust the sensitivities so that the signals appear as hairpins. Measure the fraction of a period delay at the peaks. (Although the peaks are flat their sides enable the peak positions to be estimated quite well. This method eliminates the need to centre the traces.) If the phase shift is small and subject to large reading errors you should consider the costs/benefits of changing time base settings. You might do even better to get the period information from the oscillator and use the scope only for the delay measurement.

(c) If voltages with time variations $\sin(\omega t+\phi_1)$ and $\sin(\omega t+\phi_2)$ are applied to the X and Y inputs of an oscilloscope respectively an ellipse is seen. The value of $\phi_2-\phi_1$ can be extracted from the ellipse as described in Figure 8.3.

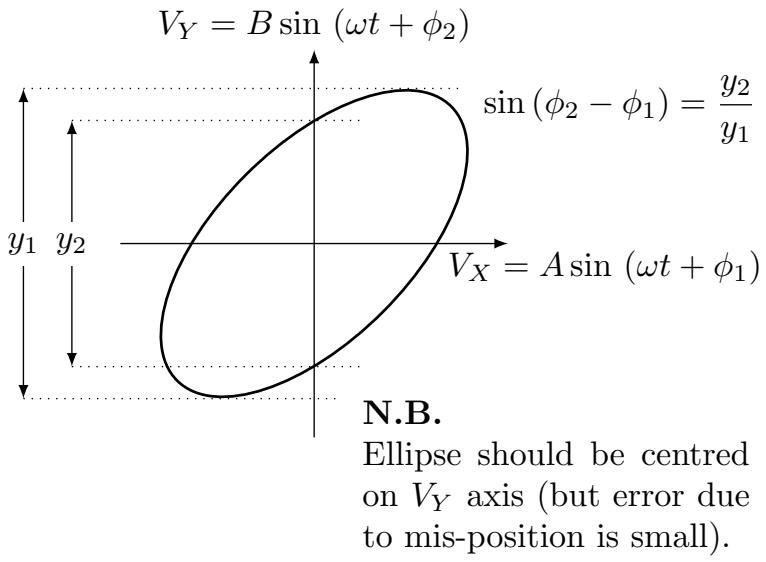


Figure 8.3: Phase shift measurement: plot of voltages with sinusoidal time variations.

9 Introduction To Semiconductors and the PN Junction Diode

9.1 Introduction

Semiconductor materials and devices are good examples of where the pictures we use to help us understand what is happening contain mixtures of classical and quantum ideas. We shall find it helpful to visualize current being carried by classical, positively and negatively charged free particles (holes and electrons) but we must bear in mind the underlying energy band structure. We give a brief discussion of semiconductor materials and follow this with descriptions of two devices, a uniform bar and a PN junction.

9.2 Bands, waves and particles

We picture a semiconductor as a uniform three dimensional lattice of atoms. The allowed energy states of the atomic electrons in the periodic potential of a perfect crystal form a discontinuous spectrum being grouped into *bands* with gaps between them in which there are no states. The corresponding wave-functions are travelling waves extending throughout the whole crystal. The lower bands of states are all full (equal numbers of waves travelling in opposite directions) and we need not concern ourselves with them except when we are considering polarization of the crystal. Our interest will centre on two of the upper bands known as the valence band and the conduction band. They are shown in Figure 9.1.

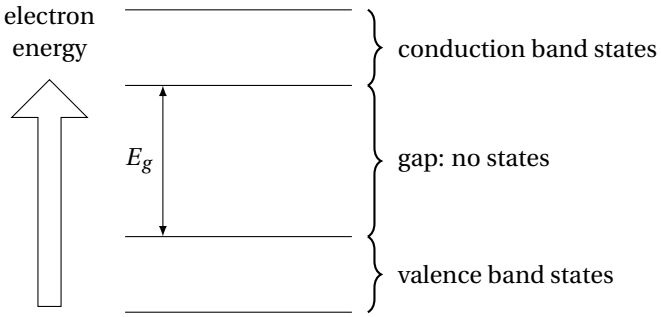


Figure 9.1: Electron states in a semiconductor

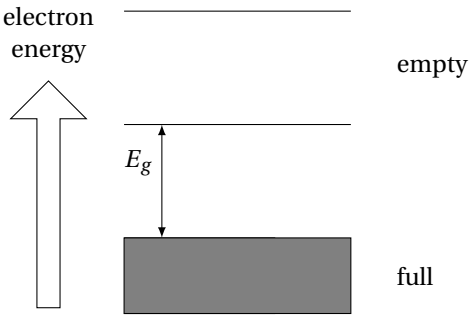
In thermodynamic equilibrium the probability that a state of energy E is occupied is described by the Fermi-Dirac distribution function:

$$\frac{1}{e^{\left(\frac{E-E_F}{kT}\right)}+1} \tag{9.1}$$

where E_F is the energy (the Fermi energy) at which the probability of occupation would be 50%.

In a pure perfect semiconductor crystal at absolute zero (Figure 9.2) the valence band is full and the conduction band is empty. The Fermi energy lies near the middle of the gap which is typically of the order of 1 eV wide. Full and empty bands cannot give rise to a current so the crystal is an insulator. At room temperature some electrons may be excited from the valence band states to occupy states at the bottom of the conduction band where an imbalance in the number of waves travelling in opposite directions can be maintained by an electric field i.e. a current can be made to flow. Moreover because the valence band is no longer full it can also carry current.

In a pure semiconductor at temperatures $\ll E_g/k$, the density of electrons in the conduction band is low and the exclusion principle has little effect. Wave packets constructed from states at the bottom of the conduction band move under the influence of electric and magnetic fields in the way expected of classical particles carrying a



Perfect, pure semiconductor at $T = 0\,\mathrm{K}$

Figure 9.2: Electron states occupied in a perfect pure semiconductor crystal at $T = 0\,\mathrm{K}$

single electron (negative) charge but having a different mass from electrons in free space. An argument involving unoccupied states at the top of the valence band and a number of negatives leads to a picture of positively charged particles called *holes*.

If we consider next the smoothing function used to derive the macroscopic approximation in electromagnetism (which may be ≈ 100 atom spacings wide) our wave packets/particles become smeared out to this extent. This however is still much smaller than typical device dimensions so we are left with a useful picture of electrons and holes in semiconductors as *fuzzy particles*.

Having justified the particle concepts one could (people do) write entirely particle based descriptions of semiconductors such as:

In a semiconductor crystal at absolute zero the atoms remain neutral, but at room temperature some electrons are liberated by thermal energy leaving some of the atoms ionized. The liberated electrons are able to wander throughout the whole crystal. A drift motion can be superimposed on the random motions (a current can be made to flow) by applying an electric field.

Current can also be carried in another way. A bound electron in a neutral atom can transfer to a neighbouring ionized atom without being excited into the free state. This can be visualized as a transfer of the state of positive ionization from one atom to the other in the opposite direction. The state of positive ionization, called a 'hole', is therefore also free to wander about the crystal and in fact behaves in many respects like a positively charged free particle.

We shall find the picture of conduction electrons and holes as free charged particles very helpful in describing the behaviour of semiconductors and of diodes and transistors made from them.

9.3 Carrier densities

9.3.1 Intrinsic semiconductors

The probability of occupation of the conduction band states is $\ll 1$ and is described by the Boltzmann limit of the Fermi-Dirac distribution:

$$e^{-\left(\frac{E-E_F}{kT}\right)} \tag{9.2}$$

In a pure semiconductor free electrons and holes are always created in pairs, so the number density p of holes is equal to the number density n of electrons. We refer to the electrons and holes collectively as '(charge) carriers' and to their densities in pure material as the *intrinsic* carrier density n_i. The minimum energy needed to generate an electron-hole pair is equal to the energy gap E_g between the valence and conduction bands of the semiconductor. Its value is about 1.1 eV for silicon and 0.7 eV for germanium. Some typical carrier densities are:

- densities of conduction electrons and holes at room temperature
  in pure Ge $n_i \approx 10^{19}\ \mathrm{m}^{-3}$ (≈ 1 per 10^{10} atoms)
  in pure Si $n_i \approx 10^{16}\ \mathrm{m}^{-3}$ (≈ 1 per 10^{13} atoms)
- density of conduction electrons in copper
  $n \approx 10^{29}\ \mathrm{m}^{-3}$ (≈ 1 per atom).

Band theory provides us with the equilibrium relation:

$$n_i^2 = \text{constant} \times T^3 e^{-\frac{E_g}{kT}} \tag{9.3}$$

9.3.2 Doped semiconductors

So much for pure perfect semiconductor crystals; things can be very different if the crystal contains impurities. Suppose we deliberately replace (a process known as 'doping') a small number (1 in 10^6 to 10^7) of the atoms in a Si crystal (valency 4) with phosphorus atoms (valency 5). What happens is that four of the electrons in the phosphorus atom form lattice bonds with the host atoms, but the fifth goes into a quite weakly bound orbit around the extra nuclear charge. The energy required to free this last electron, the *impurity ionization energy*, is much less than the energy gap and at room temperature practically all such impurities are ionized.

A semiconductor which has had its conductivity increased in this way is called N-type and the introduced impurity atoms are known as *electron donors*. In N-type material the free electrons (density n_N) are described as *majority carriers* and the holes (density p_N) as *minority carriers*. In typical transistor material the number density of the donors N_D is deliberately made much greater than the intrinsic carrier density n_i, and since nearly all the donors are ionized at room temperature the density of electrons n_N is closely equal to N_D.

If a trivalent impurity such as aluminium is introduced into Ge or Si, the conductivity is increased because of the extra holes formed. Such material is known as P-type and the impurities as *electron acceptors*. Again in typical transistor material the density of holes p_P is nearly equal to the density N_A of acceptors. In P-type material holes are the majority carriers and electrons (density n_P) the minority carriers.

Band theory provides us with the relation:

$$n_P p_P = n_N p_N = n_i^2 = \text{constant} \times T^3 e^{-\frac{E_g}{kT}} \tag{9.4}$$

9.3.3 Continuity equations

The hole density obeys the continuity equation:

$$\frac{\partial p}{\partial t} = -\nabla \cdot \frac{\mathbf{j}_h}{|e|} + g_h(\text{non-thermal}) + g_h(\text{thermal}) - \frac{p}{\tau_h} \tag{9.5}$$

which says that the rate of increase of hole density at some point in the crystal is the negative of the divergence of the (hole) current density plus the rate of generation of hole density per unit volume due to non-thermal causes (light, X-rays, injection etc.) plus the thermal equilibrium rate of hole density generation, minus the rate of reduction of hole density due to trapping, per unit volume. (Trapping is usually the first stage of recombination.) In thermal equilibrium this becomes

$$0 = 0 + 0 + g_h(\text{thermal}) - \frac{p_0}{\tau_h} \tag{9.6}$$

where p_0 is the equilibrium hole density. Using this result equation 9.5 can be written

$$\frac{\partial p}{\partial t} = -\nabla \frac{\mathbf{j}_h}{|e|} + g_h(\text{non-thermal}) - \frac{p - p_0}{\tau_h} \tag{9.7}$$

A similar equation holds for the electrons. In intrinsic material in equilibrium holes and electrons are generated (as pairs) at the same rate and therefore must both be also trapped at this rate.

9.4 Currents, mobility and conductivity

Our particle picture of a semiconductor crystal in equilibrium is one of electrons and holes flying about at random, being frequently deflected by the ionized impurities and the thermally agitated crystal lattice. If an electric field is applied, the carriers will be accelerated in or against the direction of the field but will tend to lose this extra drift velocity at each deflection. For normal strengths of field they acquire average (drift) velocities ($\mathbf{u}_e$, $\mathbf{u}_h$) proportional to the magnitude of the field so the material obeys Ohm's law. The constants of proportionality are called the *mobilities* of the carriers, the drift velocities acquired per unit field. They are defined by the relations:

$$\mu_h = \frac{u_h}{E} \qquad \text{and} \qquad \mu_e = -\frac{u_e}{E} \tag{9.8}$$

the minus sign making both mobilities positive numbers. The electric current density $\mathbf{j}$ due to drift of the carriers is given by the product of the carrier density, the charge they carry and their velocity, i.e.

$$\mathbf{j}_{\text{e,drift}} = -|e| n \mu_{\text{e}} \quad \text{for electrons, and} \quad \mathbf{j}_{\text{h,drift}} \quad \text{for the holes.}$$

Therefore in terms of the mobilities the total drift current density is given by

$$\mathbf{j}_{\text{drift}} = |e|\,(n\mu_e + p\mu_h)\,\mathbf{E} \tag{9.9}$$

The quantity $|e|\,(n\mu_e + p\mu_h)$ is called the *conductivity* of the material, it is usually denoted by σ.

In PN junctions and other more complicated structures (as distinct from uniform material), it is possible to set up and maintain non-uniform carrier densities. This leads to a new mechanism of current flow, diffusion, which does not require an electric field in the region where it is occurring. Diffusion arises simply from the tendency of the carriers to spread out due to their random motion. Consequently if we continuously feed in say holes at one place and remove them at another there will be a net current of holes between the two places even though there is no electric field and the motion of each hole is random.

It is found experimentally that particle current densities due to diffusion are generally described by Fick's law which states that the particle current density at a point is proportional to the gradient of the density of the particles at that point. A minus sign indicates that the flow of particles is down the gradient. The constant of proportionality D is called the diffusion coefficient. If we also multiply by $\pm|e|$ we obtain the electric current. For holes it is:

$$\mathbf{j}_{\text{h,diff}} = -|e|\,D_h \nabla p \tag{9.10}$$

and for electrons the electric current is

$$\mathbf{j}_{\text{e,diff}} = |e|\,D_h \nabla n \tag{9.11}$$

Using the foregoing we can assemble an expression for the current due to both drift and diffusion

$$\mathbf{j} = -|e|\,D_h \nabla p + |e|\,D_h \nabla n + |e|\,(n\mu_e + p\mu_h)\,\mathbf{E} \tag{9.12}$$

We also have

$$\mathbf{E} = -\nabla V \tag{9.13}$$

where V is the electric potential[1] and:

$$p = \text{constant} \times e^{-\frac{|e|V}{kT}} \tag{9.14}$$

which describes the relative probability of finding holes in particular regions according to their potential. (It reduces to the Boltzmann from the Fermi-Dirac form because the holes are assumed to be at low density.)

Now in equilibrium the hole and electron currents are both zero. Consider the hole current component in equation 9.12,

$$\mathbf{j}_h = -|e|\,D_h \nabla p + |e|\,p\mu_h \mathbf{E} = 0 \tag{9.15}$$

Using the above, we find

$$|e|\left(D_h \frac{|e|}{kT} - \mu_h\right) p \nabla V = 0 \tag{9.16}$$

which means that

$$\mu_h = D_h \frac{|e|}{kT} \tag{9.17}$$

a relationship deduced by Einstein. It shows that highly mobile holes have large diffusion coefficients. A similar relation holds for the electrons.

9.5 The uniform bar

9.5.1 Conductivity

From equation 9.12 we see that in order to measure the *conductivity*, σ (that is $\mathbf{j}_{\text{drift}}/\mathbf{E}$ or $|e|\,(n\mu_e + p\mu_h)$), of a material we must know that there are negligible variations of doping density *i.e.* negligible gradients of carrier density, in the specimen. The specimen is usually cut into the form of a uniform bar.

If we know that one type of carrier is in an overwhelming majority (an important simple case) measurement of σ tells us the product $n\mu_e$ (or $p\mu_h$). Note that measurements of $\mathbf{j}_{\text{drift}}$ and $\mathbf{E}$ made to determine σ will necessarily involve a *finite volume* of the material whereas the relation $\mathbf{j}_{\text{drift}} = \sigma\mathbf{E}$ applies at a *point*. This means that $\mathbf{j}_{\text{drift}}$ and $\mathbf{E}$ must be uniform over the volume in which they are measured if their ratio is to give a true value of the conductivity.

[1] It is usual to take V as the energy at the bottom of the conduction band but any level of energy that is fixed with respect to the band structure can be used

9.5.2 Hall coefficient

If, while a current is flowing along the bar (the $+x$ direction), a magnetic field B_z is applied in the $+z$ direction, the drifting holes and electrons experience an average force in the same $(-y)$ direction. The average force is $+|e|\mathbf{u}_h B_z$ for the holes and $-|e|\mathbf{u}_e B_z$ for the electrons. (These are in the same direction because $\mathbf{u}_h$ and $\mathbf{u}_e$ are in opposite directions.) When these forces are unequal the $\pm y$ surfaces of the sample become charged and a potential difference arises between them. The appearance of this voltage is the *Hall effect*.

Let us consider first P-type material which is sufficiently heavily doped that we can ignore the electrons. The transverse electric field $\mathbf{E}_y$ builds up until the transverse force due to it becomes equal and opposite to the force due to the magnetic field. The current then flows uniformly down the bar as it did before the magnetic field was applied. Then:

$$|e|\mathbf{E}_y = |e|\mathbf{u}_h B_z \tag{9.18}$$

On substituting for current density $\mathbf{j}_x$ in terms of hole drift velocity $\mathbf{u}_h$ we obtain:

$$\mathbf{E}_y = \frac{1}{|e|\,p_P}\mathbf{j}_x B_z \tag{9.19}$$

If, instead of considering P-type material, we go through the corresponding analysis for strongly N-type material this simply changes the sign of the charge on the carrier in all of the above equations and leads to:

$$\mathbf{E}_y = -\frac{1}{|e|\,n_N}\mathbf{j}_x B_z \tag{9.20}$$

We define the *Hall coefficient* R_H (for right handed axes) by the relation:

$$\mathbf{E} = R_H \mathbf{j}_x B_z \tag{9.21}$$

Thus R_H is given by:

$$R_H = \frac{1}{|e|\,p_P} \text{ or } R_H = -\frac{1}{|e|\,n_N} \tag{9.22}$$

In other words when one carrier is in the overwhelming majority, measurement of the magnitude and sign of the Hall coefficient tells us the density and type of the carriers from which their mobility can be obtained via σ. This is very useful. If one carrier is not in an overwhelming majority, measurements of conductivity and Hall coefficient over a range of temperatures can be made to yield the densities and mobilities of both.

9.6 The PN junction diode

9.6.1 Introduction

The PN junction diode is a two terminal semiconductor device, belonging to the family known as rectifiers, through which current flows much more easily in one direction than in the other. It consists of a piece of semiconductor crystal which has been doped so that there is a change from P-type to N-type as shown in Figure **??**. The crystal structure is continuous across the junction, only the doping changes.

The forward, or easy, direction of conduction is with the voltage V applied so that the P-type side supplies holes and the N-type side electrons, i.e. with the current I flowing in the direction of the 'arrowhead' in the circuit diagram symbol for the diode shown in Figure **??**.

In a diode with a very abrupt junction (change of doping) I is related to the applied voltage V by the Shockley law:

$$I = I_0\left(e^{\left(\frac{eV}{kT}\right)} - 1\right) \tag{9.23}$$

in which I_0 is a small constant current, to which I tends when V is large and negative, the hard or reverse direction of conduction. I_0 does not depend upon V, up to a breakdown value, but it does depend very strongly upon temperature.

A graph of the relation between V and I for a device is known as its *characteristic*. If this is a straight line through the origin, the device is said to be *linear*, or to *obey Ohm's law*. Clearly the PN junction does not obey Ohm's law.

An abrupt junction results from the alloy method of construction.

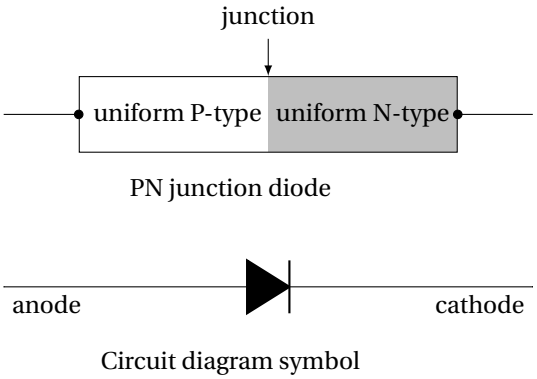


Figure 9.3: PN junction diode structure and circuit diagram symbol

9.6.2 The PN junction in equilibrium (no voltage applied)

Although PN junctions cannot be made in this way, we will consider what happens when a piece of electrically neutral P-type semiconductor is brought into contact with a piece of neutral N-type semiconductor so that the crystal lattices join perfectly. Consider the holes first. Since the P-type side contains a much higher density of holes than the N-type side, the P-type side loses holes and becomes increasingly negatively charged with respect to the N-type side as the holes spread out into the enlarged crystal. A potential difference builds up between the two ends of the device. Shockley has shown that the potential gradient (electric field) is not uniform but is confined to a layer in the neighbourhood of the junction (Figure 9.4(a)). Carriers that are thermally generated in the layer are rapidly swept away by the field, and the density in the layer is very low. This region, in which the fixed impurity ions are largely un-neutralized by free electrons and holes is called the *depletion layer* (see Figure 9.4(c)).

The field in the depletion layer opposes the flow of holes from the P-type side, but does not affect the flow of holes in the opposite direction as this depends only on the rate at which the minority holes on the N-side arrive at the edge of the layer. The potential difference across the layer builds up to the value V_b at which the net hole flow becomes zero, which is when the fraction of the holes on the P-type side of the layer with energy greater than the potential step is equal to the entire density of holes on the N-type side of the layer.

The same thing happens to the electrons with the net flow becoming zero at the same magnitude and sign of V_b.

This picture of the behaviour in terms of four components of current (the hole and electron flows in both directions) is a useful one, and we give the components names as shown in Figure 9.4(b). You will remember from section 9.3 that electron-hole pairs are being generated thermally throughout the material. The minority carriers generated near the depletion layer have a chance of reaching its edge and falling down/up the potential step before recombining. These carrier flows comprise the *generation* currents. The *annihilation* currents are so called because those majority carriers with sufficient thermal energy to surmount the potential step, and which succeed in doing so, will become minority carriers on the other side eventually to recombine. In equilibrium, the hole annihilation and generation currents are equal and opposite and so are the electron currents.

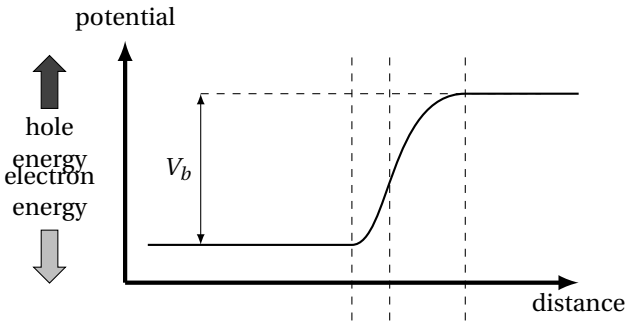
9.6.3 PN junction with large reverse voltage applied

An external voltage applied to the diode appears at the junction, very little of it is dropped in the bulk material. Reverse voltage adds to the barrier height. A particularly simple situation exists when a reverse voltage strong enough to suppress the annihilation currents completely is applied. The current then reduces to I_0, the sum of the two generation components. Figure 9.5 shows the densities of the minority carriers near the depletion layer under these conditions.

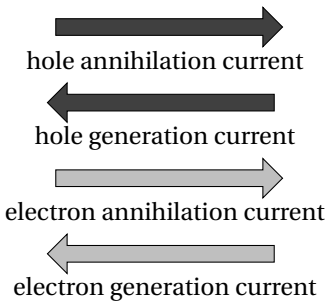
The densities are close to zero in the depletion layer, where the carriers are swept away by the large electric field, and they recover to their equilibrium densities n_p and p_n back from the edges of the layer with characteristic lengths L_e and L_h called *diffusion lengths*. The hole density at x in the N-type material is given by

$$p_N(x) = p_N\left(1 - e^{-x/L_h}\right) \tag{9.24}$$

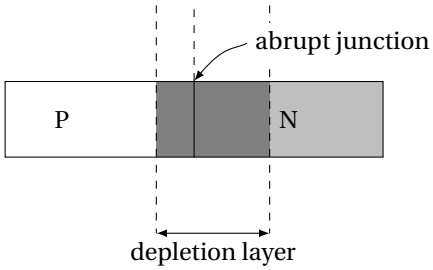
so the hole density gradient at plane **B** ($x = 0$) is p_N/L_h. Similarly for the electrons the density gradient at plane **A** is $-n_P/L_e$. We can use these ideas to evaluate the current crossing the junction. When the holes flowing from right to left emerge from the depletion layer on the P-side (plane **A**) they become majority carriers and it is not



(a) PN junction potential as a function of position (no applied voltage)



(b) Charge carrier currents



(c) PN junction diode structure showing the depletion region

Figure 9.4: PN junction potential as a function of position (no applied voltage)

easy to say what the current will be. However at plane **B** they are minority carriers with a known density gradient so we can say that there is at least a current density given by Fick's law:

$$+|e|\left(\frac{-D_h p_N}{L_h}\right) \tag{9.25}$$

Similarly, we can say there is at least an electron diffusion current density at plane **A** of:

$$-|e|\left(\frac{D_e n_P}{L_e}\right) \tag{9.26}$$

These currents are not augmented significantly by a drift component because the electric field outside the depletion layer is weak. (This statement is justified below).

If we assume that there is no recombination in the depletion layer, the total current (I_0) is the sum of the two diffusion components i.e.:

$$I_0 = -|e|\,A\left(\frac{D_h p_N}{L_h} + \frac{D_e n_P}{L_e}\right) \tag{9.27}$$

where A is the area of the junction. The minus sign simply indicates that the current flows from right to left.

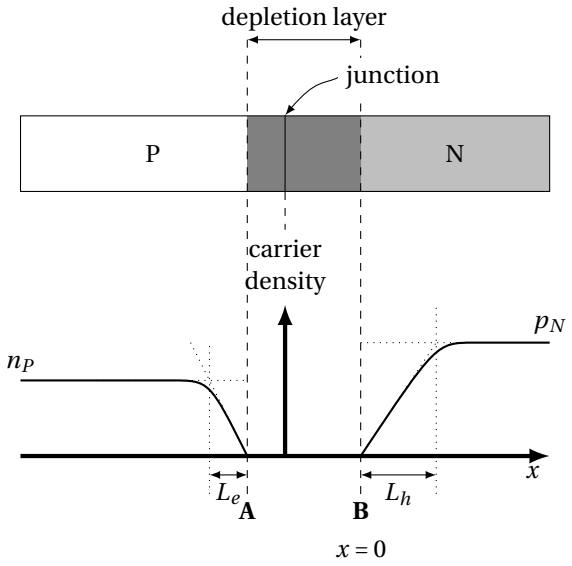


Figure 9.5: PN junction potential as a function of position (applied reverse voltage)

The current crossing any plane parallel to the junction is of course constant. However many diffusion lengths away from the junction there are no gradients of density of either carrier on either side so there can be no current due to diffusion. What happens is that the minority diffusion components which are completely dominant at planes **A** and **B** give way to majority drift components due to the weak electric field many diffusion lengths from the junction. The fact that the electric field is only strong enough to produce a *majority* drift current equal to a *minority* diffusion current justifies our ignoring it in evaluating the current above.

In this model the reverse current I_0 is independent of the voltage applied. In practice there is usually some voltage dependence due to leakage, also the diode will 'breakdown' and pass a large current if its rated reverse voltage is exceeded.

9.6.4 The Shockley law

As remarked above any external voltage V applied to the diode appears almost entirely across the depletion layer and alters the height of the potential step, the electric field in the rest of the material remaining small. The step becomes $V_b - V$ if V is in the forward direction.

The probability of a hole falling down the potential step does not depend on its height. Therefore the hole generation current I_{gh}, which is proportional to the rate at which holes arrive at the depletion layer by diffusion from the points at which they were generated by thermal excitation, does not depend upon V. The hole annihilation current I_{ah}, however, is proportional to the number of holes on the P-type side with energy exceeding the height of the step, and statistical mechanics tells us this number is proportional to:

$$\int_{V_b - V}^{\infty} e^{-\frac{|e|V}{kT}}\,\mathrm{d}V \tag{9.28}$$

which is proportional to $e^{-\frac{|e|(V_b - V)}{kT}}$ at constant temperature.

The net hole current I_h is thus given by

$$I_h = Ke^{-\frac{|e|(V_b - V)}{kT}} - I_{\text{gh}} \tag{9.29}$$

where K is some constant. But in equilibrium $(V = 0)$, we know that I_h is zero, so:

$$0 = Ke^{-\frac{|e|V_b}{kT}} - I_{\text{gh}} \tag{9.30}$$

We can use this to eliminate K and we then find:

$$I_h = I_{\text{gh}}\left(e^{-\frac{|e|V}{kT}} - 1\right) \tag{9.31}$$

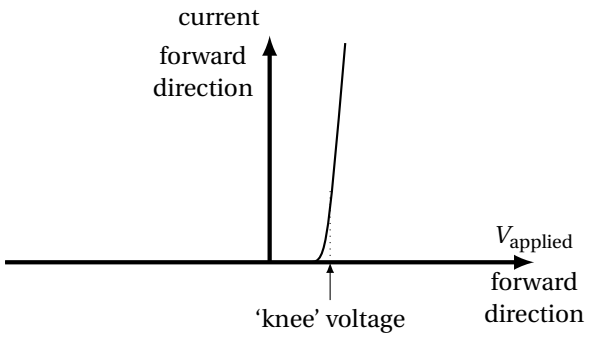
Similarly, being careful about signs and directions we obtain the net electron current, I_e,

$$I_e = I_{\text{ge}}\left(e^{\frac{|e|V}{kT}} - 1\right) \tag{9.32}$$

The total current $I = I_h + I_e$ is then given by:

$$I = I_0\left(e^{\frac{|e|V}{kT}} - 1\right) \tag{9.33}$$

writing I_0 for the sum of the hole and electron generation currents. This celebrated law due to Shockley gives an accurate description of abrupt junctions in germanium. It is illustrated in Figure 9.6.



The forward direction is the same as that of the arrow-head on the symbol.

Figure 9.6: Shockley law V-I diagram

9.6.5 The PN junction with large forward voltage applied

Under strong forward bias the electron density in the P-region is shown in Figure 9.7.

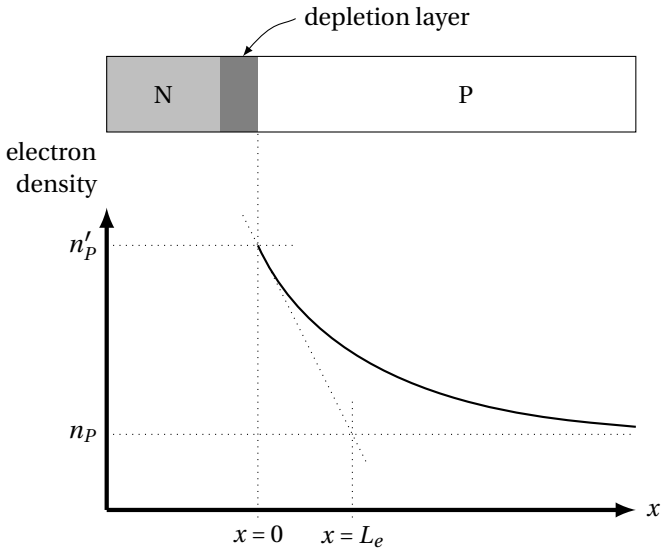


Figure 9.7: PN junction potential as a function of position (applied forward voltage)

The applied voltage maintains an electron density of n'_P just outside the depletion layer on the P-side. The density falls exponentially with distance to the equilibrium value n_p as given by:

$$n_P(x) - n_P = \left(n'_P - n_P\right)e^{-x/L_e} \tag{9.34}$$

The density gradient at $x = 0$ is therefore:

$$-\frac{n'_P - n_P}{L_e} \tag{9.35}$$

which in the case of strong forward bias, $n'_P \gg n_P$, reduces to $-n'_P - n_P/L_e$. As in the case of strong reverse bias, the current can be evaluated by summing the diffusion currents at each edge of the depletion layer.

9.6.6 Temperature dependence of I_0

Using the relations given in section 9.4 and assuming that $n_N = N_D$ and $p_P = N_A$, the expression for I_0 given in section 9.6.3 becomes:

$$I_0 = |e| A \left(\frac{D_h}{L_h N_D} + \frac{D_e}{L_e N_A} \right) \tag{9.36}$$

The term in the square bracket has a negligible temperature dependence compared with the exponential.

9.6.7 Note on silicon junctions

Silicon junctions are not as well described by the Shockley law as those made in germanium. The reverse current I_0 both increases with voltage and is much larger than expected, although still much smaller than in a germanium junction of similar area. These effects are due to thermal generation of electron-hole pairs in the depletion layer, which is an insignificant process in germanium junctions. The voltage dependence arises because the total generation rate depends on the volume of the depletion layer and this grows in thickness with increasing reverse voltage.

The knee voltage is about 0.6 V compared with about 0.2 V for germanium.

9.6.8 Capacitances of junctions

Reverse biased junctions with small signals[1] applied exhibit capacitances which depend on the type of junction.

For junctions made by the alloying process the N- to P-type transition is abrupt and the capacitance is given by:

$$C_\mathrm{a} = K_\mathrm{a} \left(V_\mathrm{a} + V \right)^{-\frac{1}{2}} \tag{9.37}$$

where K_a, V_a are constants and V is the reverse bias voltage.

For junctions made by diffusion of impurities the transition from N-type to P-type is gradual and the capacitance is given by:

$$C_\mathrm{g} = K_\mathrm{g} \left(V_\mathrm{g} + V \right)^{-\frac{1}{3}} \tag{9.38}$$

where K_g, V_g are constants and V is the reverse bias voltage. The depletion layer is usually much *thinner* than the region in which the doping density changes.

Note: the depletion layer relates to the position of the junction as shown when the P-type side is more heavily doped than the N-side.

[1] See end of section 10.4.5

10 The Diffusion Transistor

10.1 Introduction

Diffusion or bipolar transistors come in two varieties PNP and NPN. We shall describe the NPN-type in which the more important charge carriers are minority electrons. (In PNP devices the important carriers are minority holes and applied voltages have the opposite polarity.) We use the name *diffusion* transistor because this is the mechanism of current flow at the heart of the device, *not* because they are made (as are almost all semiconductor devices) by diffusing donor and acceptor atoms into regions of the semiconductor crystal.

10.2 Manufacture of NPN transistors in silicon

10.2.1 Semiconductor processing

A slice is cut from a single crystal of N-type silicon and is polished. Finished slices are typically 100 mm diameter and 0.1 mm thick. The slice is exposed to steam at 1000 °C which converts a surface layer to SiO_2, and one face is then coated with a photo-resist material[1]. An image of a rectangular grid, whose mesh delineates the area to be occupied by each transistor, is then formed on the photo-resist using ultra-violet light. The area required for a single transistor ranges from less than 1 mm^2 to more than 20 mm^2 depending on the power it is designed to handle. Thousands of small devices may be made from one slice.

A developer is used to remove the unexposed photo-resist leaving an array of rectangular holes in which the oxide layer is exposed. An etchant then removes the oxide layer in the holes using the developed photo-resist grid as a mask. The rest of the photo-resist is then removed. The slice is then exposed to a hot vapour containing P-type impurities which diffuse into the silicon where it is not masked by the oxide grid, the density of the introduced acceptors being great enough to change the type from N-type to (weakly) P-type. The whole of the oxide layer is then removed leaving the slice section shown in Fig 10.1. Next, following further steps involving a photo-resist mask with smaller holes, an oxide mask and exposure to a hot vapour of donor atoms, a preponderance of donors is introduced to form a heavily N-type region N+ as shown below. The result is an N^+PN structure.

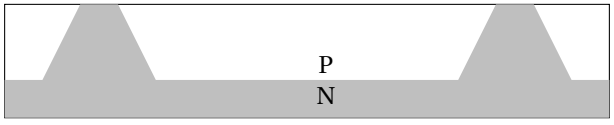


Figure 10.1: Section of a slice after first diffusion

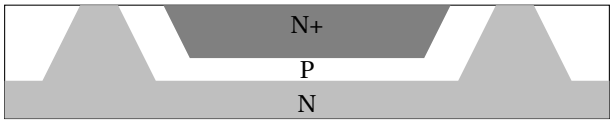


Figure 10.2: Section of slice after section diffusion

Tests are performed on a few of the transistors on the slice using pin probes to make contact and if the performance is satisfactory the slice is cut up into individual transistor chips.

[1] A material that can be hardened by ultravioled light.

10.2.2 Encapsulation

It remains to attach leads to the relevant regions of the chip, to provide a path for it to dissipate heat and to protect it. The two basic approaches to encapsulation are the metal case and the block of cast plastic but many variations are employed.

Metal case

The chip is bonded N-side down to a metal plate or 'header' which carries two glass bead seals through which leads pass. Two connections, *base* and *emitter*, are made with aluminium or gold bond wires (the third, *collector*, being to the header) and a metal cover is then cold welded over the assembly as shown in Figure 10.3. This encapsulation has good hermeticity and allows good heat dissipation (when the header is clamped to a 'heat-sink') but is expensive.

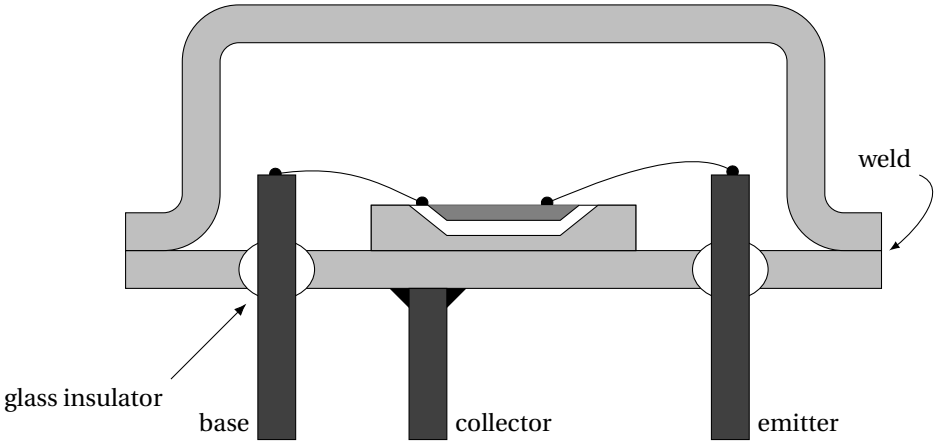


Figure 10.3: NPN transistor encapsulated in a metal casing.

Plastic block

The header in this case is a three tined sheet metal stamped shape as shown in Figure 10.4. The wafer is mounted on the centre tine and bond wires are taken to the other two. A plastic block is then cast around the assembly and finally the support strip is cropped off. This encapsulation is cheap because it is simple, little material is needed and strips of devices can be handled together. Heat dissipation is not very good as the only metal conducting path is down the collector lead so this encapsulation is only used for low power devices. Also over long periods of time, water may diffuse into the package.

10.3 Physical picture of the dc operation of an NPN transistor

The NPN diffusion transistor consists of NP and PN junctions much less than a diffusion length apart in a single crystal of semiconductor. The physics of PN junctions with and without bias has been examined in chapter 9.

10.3.1 Dependence of I_c on V_{be}

In normal operation the emitter-base junction of an NPN transistor is strongly forward biassed and the base collector junction is strongly reverse biassed. The forward current I_e crossing the depletion layer at the emitter junction is made up of two annihilation currents: electrons flowing into the base and holes flowing out. ('Strong' forward bias means that the generation currents are small enough to be ignored.) The voltage dependence of this current is described well by the Shockley law (even for silicon transistors where the magnitude of I_0 is not well accounted for) so we write, ignoring the -1 (the generation currents):

$$I_e = I_{e0} e^{\left(\frac{eV_{be}}{kT}\right)} \tag{10.1}$$

where $I_{e0} = C(T)\, e^{-E_g/kT}$ and the factor $C(T)$ varies weakly with temperature compared with the exponential. The electron component of this current is:

$$I_e(e) = \bar{\gamma} I_{e0} e^{\left(\frac{eV_{be}}{kT}\right)} \tag{10.2}$$

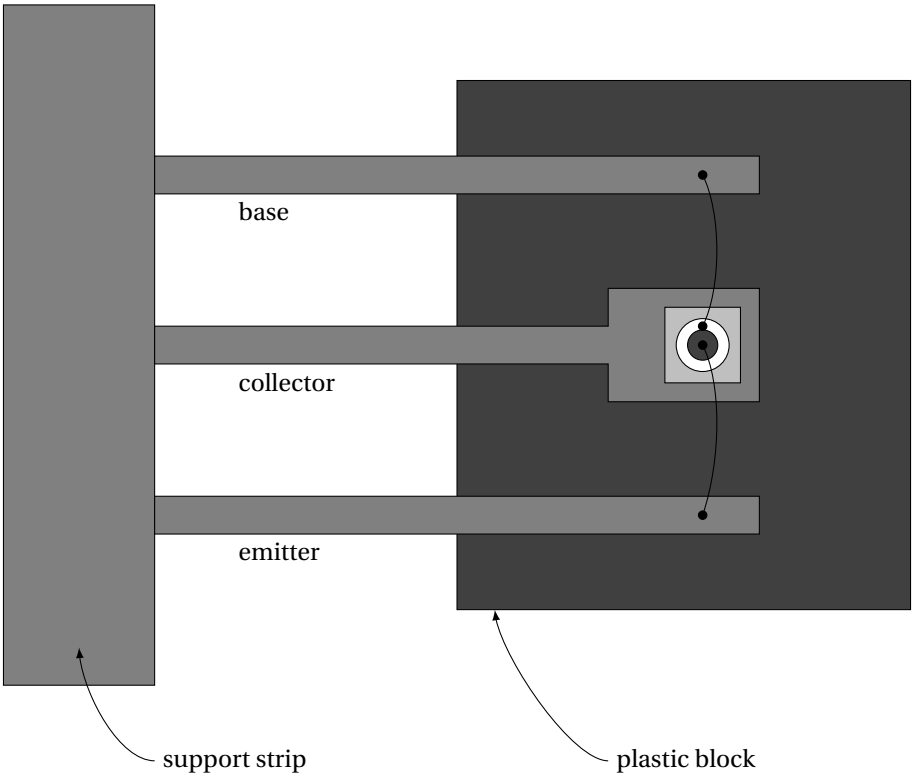


Figure 10.4: NPN transistor encapsulated in a plastic casing.

where the *emitter efficiency* $\bar{\gamma}$, the fraction of the emitter current carried by electrons entering the base, is typically > 0.99. The hole component of I_e is very much less than the electron component because the N-type emitter is much more heavily doped than the P-type base. The hole component will be discussed in section 10.3.4.

On emerging from the depletion layer at the emitter junction the electrons find themselves in a region of the base which is essentially free of electric field and they simply wander about at random i.e. *diffuse*. Electrons reaching the edge of the depletion layer at the collector junction are drawn into it and hence into the collector by the electric field. The width of the field-free region of the base between the edges of the two depletion layers is very much less than a diffusion length and the *base transport factor* $\bar{\delta}$, the probability of an electron reaching the collector depletion layer before it is lost by recombination, is, like $\bar{\gamma}$, typically > 0.99.

The electron current reaching the edge of the collector depletion layer (the collector current) may therefore be written:

$$I_c = \bar{\gamma}\bar{\delta} I_{e0} e^{\left(\frac{eV_{be}}{kT}\right)} \tag{10.3}$$

$\bar{\delta}$ is slightly dependent on V_{ce}, see section 10.3.2. It is usual to lump the two constants $\bar{\gamma}\bar{\delta}$ into a single constant $\bar{\alpha}$ and write

$$I_c = \bar{\alpha} I_{e0} e^{\left(\frac{eV_{be}}{kT}\right)} \tag{10.4}$$

We shall refer to this relation between the collector current and the voltage between emitter and base as the *transistor law*.

The current is greater in a transistor than it would be in a diode of the same area with the same doping as the emitter base junction. The reason is the following. The electron component of I_e, $I_{eb}(e)$ maintains an electron density n'_p in the base just outside the emitter depletion layer which is the same regardless of whether or not there is a collector junction nearby. When there is a collector junction, the electron density in the base material is forced to zero at the edge of the collector depletion layer as shown in Figure 10.5 below. When there is no collector junction (a diode) the density of the injected electrons falls off as the diffusion length as shown in Figure 9.7. When, as is usual, recombination is small the density gradient $\frac{dn}{dx}$ in the transistor is nearly constant and is given by:

$$\frac{dn}{dx} = \frac{\text{electron density } n'_p \text{ at edge of base-emitter diffusion layer}}{\text{distance } w^* \text{ between edges of the two depletion layers}} \tag{10.5}$$

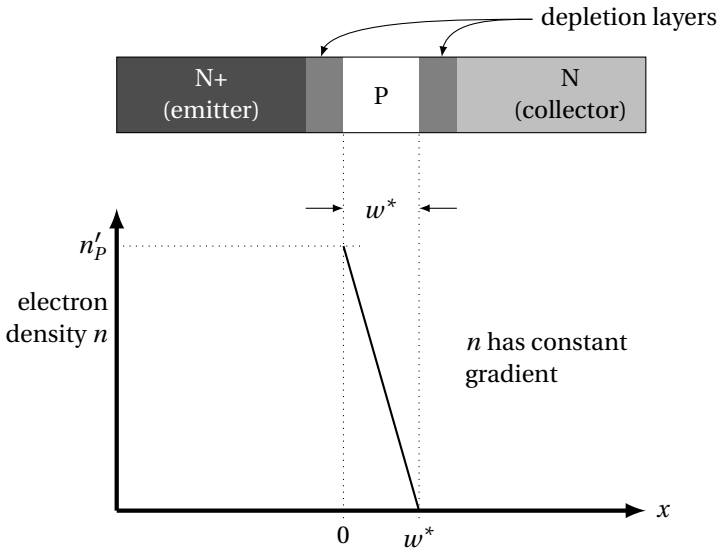


Figure 10.5: NPN transistor electron density versus position

The electron diffusion current crossing the base is therefore given by

$$I_{\mathrm{bc}}(\mathrm{e}) = \frac{|e| D_N n'_p}{w^*} \tag{10.6}$$

When there is no collector junction the electron diffusion current just outside the emitter depletion layer is

$$I_{\mathrm{diode}}(\mathrm{e}) = \frac{|e| D_N n'_p}{L_{\mathrm{e}}} \tag{10.7}$$

which is smaller by a factor L_{e}/w^*. (It will be recalled from the discussion of the PN junction that n'_p is equal to the density of electrons in the emitter with energy exceeding the barrier height and is proportional to $\exp\left(eV_{\mathrm{be}}/kT\right)$.)

10.3.2 Dependence of I_{c} on V_{ce}

The effectiveness of the potential 'cliff' at the collector junction depletion layer in gathering in the electrons that wander to its edge is almost perfect and independent of V_{ce} so long as the junction is reverse biased. An increase of V_{ce} simply increases the height of the cliff over which the electrons fall and this has little effect. However, an increase of V_{ce} also widens the collector depletion layer and pushes its edge further into the base. This reduces slightly the width of the field-free region of the base in which the electrons travel by diffusion, thus reducing recombination and increasing the density gradient. This means that there is a slight dependence of $\bar{\delta}$ on the collector voltage.

10.3.3 Dependence of I_{c} on temperature

Temperature enters the transistor law in two ways: explicitly via $\exp\left(eV_{\mathrm{be}}/kT\right)$ and implicitly via I_{e0} $\left(= I_{\mathrm{e0}}/\gamma\delta\right)$. The latter dependence (which is in the opposite direction to the explicit one) is dominant and the collector current increases with increasing temperature.

The simple theory on which the Shockley law is based predicts:

$$\frac{1}{I_{\mathrm{c}}}\frac{\partial I_{\mathrm{c}}}{\partial T} V_{\mathrm{be}} = \frac{E_{\mathrm{g}} - eV_{\mathrm{be}}}{kT} \approx +6\%\ \text{per}\ ^\circ\mathrm{C} \tag{10.8}$$

for $E_{\mathrm{g}} = 1.1$ eV, and $V_{\mathrm{be}} = 0.6$ volts.

If we had held I_{c} constant in some way, the temperature coefficient of interest would have been $\left(\partial V_{\mathrm{be}}/\partial T\right) I_{\mathrm{c}}$, predicted by the simple theory to be:

$$\frac{\partial V_{\mathrm{be}}}{\partial T} I_{\mathrm{c}} = \frac{V_{\mathrm{be}} - E/e}{T} = -1.7\,\mathrm{mV}/^\circ\mathrm{C} \tag{10.9}$$

with the values assumed above. The experimental value is usually nearer −2.3 mV/°C.

As mentioned in Chapter 9 the magnitude and temperature dependence of I_0 are described well by the Shockley law for germanium junction diodes. For silicon, diodes although the predicted value of I_0 is orders of magnitude too low, the temperature coefficients are usually within a factor of two of the Shockley values.

10.3.4 The base current I_b

The small fraction $1-\bar{\gamma}$ of the emitter current which is carried (in an NPN transistor) by holes leaving (rather than electrons entering) the base and the small fraction $1-\bar{\delta}$ of the electrons diffusing across the base which recombine contribute to a current in the base lead given by:

$$I_b = (1-\bar{\gamma})\,I_e + \left(1-\bar{\delta}\right)\bar{\gamma}\,I_e \tag{10.10}$$

$$= \left(1-\bar{\gamma}\bar{\delta}\right)I_e \tag{10.11}$$

$$= \frac{\left(1-\bar{\gamma}\bar{\delta}\right)I_c}{\bar{\gamma}\bar{\delta}} \tag{10.12}$$

$$\therefore \frac{I_c}{I_b} = \frac{\bar{\gamma}\bar{\delta}}{\left(1-\bar{\gamma}\bar{\delta}\right)} = \frac{\bar{\alpha}}{(1-\bar{\alpha})} \tag{10.13}$$

This is an important quantity denoted by $\bar{\beta}$ and called the *dc current gain*. As $\gamma\delta$ is typically 0.99 or greater, $\bar{\beta}$ is typically 100 or more. The hole contribution to the current crossing the emitter-base junction is a fixed fraction of the total current because, as may be seen by examining the expression for I_0 in Chapter 9, it depends on the doping. This flow of holes out of the base must be made up by injection of holes at the base contact.

The rate of recombination of electrons in the base is proportional to the number of excess electrons above the equilibrium value. With the triangular distribution shown in Figure 10.5 this number is proportional to the density gradient, and hence also to the current. For each electron that recombines a hole has to be injected at the base contact.

Thus *both components of the current in the base lead are proportional to the emitter current*, and we expect to find that the current gain is constant, independent of the value of I_c.

10.4 The diffusion transistor as a circuit element

10.4.1 Circuit symbols

In Figure 10.6 are shown the major current flows in and typical voltages applied to NPN and PNP diffusion transistors together with the symbols[1] generally used to depict them on circuit diagrams.

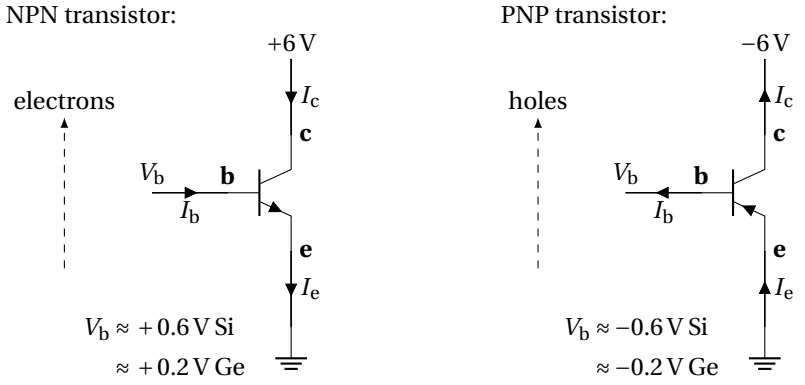


Figure 10.6: NPN and PNP transistor symbols and typical base voltages

[1] The unrepresentative appearance of these symbols is due to the fact that they were originally designed to represent (rather well) devices of a totally different construction: the 'catswhisker' type transistors, now obsolete, forwhose invention Bardeen, Brattain and Shockley were awarded the Nobel Prize in 1956.

10.4.2 DC characteristics of an NPN transistor

The voltages applied to diffusion transistors in analogue circuits are usually such that at all times the emitter-base junction is forward biased and the collector-base junction is reverse biased. In an NPN transistor under these conditions, electrons flow from the emitter into the base, diffuse across the base, and are captured by the collector. (In a PNP transistor holes follow the same path.) The current in the collector lead depends strongly on the voltage between the base and the emitter and on temperature. The dependence on the voltage between collector and base is weak and is usually negligible. The physical model discussed in the previous sections predicts, in reasonable agreement with observation, that the current in the collector lead, ignoring the dependence on the voltage between collector and base, is given by:

$$I_c = C(T)\, e^{-\frac{E_g}{kT}} e^{\frac{e(V_b - V_e)}{kT}} \tag{10.14}$$

The model predicts also that the current in the base lead is a small constant fraction of the collector current given by:

$$I_b = \frac{I_c}{\beta} \tag{10.15}$$

where β is typically > 100.

10.4.3 Biasing

In circuits transistors are made active by applying dc voltages to them which cause steady currents to flow, a process known as *biasing*. Writing the dc voltage $V_b - V_e$ as $V_{be:bias}$ gives a collector current $I_{c:bias}$ of

$$I_{c:bias} = A(T)\, e^{-\frac{E_g}{kT}} e^{\frac{eV_{be:bias}}{kT}} \tag{10.16}$$

It turns out that because of the exponential dependence on $V_{be:bias}$ and the fact that larger transistors tend to have larger bias currents, the bias voltage between base and emitter found in practice is always close to 0.6 volts for silicon transistors (0.2 volts for germanium).

10.4.4 Effects of temperature on biasing

Temperature appears in the transistor law, Equation 10.14, in three places. The dependence on temperature of $C(T)$ can be ignored compared with that of the exponentials. At constant V_{be} the change in bias current due to a change of temperature δT is given by

$$\frac{1}{I_c}\frac{\delta I_{c:bias}}{T} V_{be} = \frac{E_g - eV_{be:bias}}{kT^2} \tag{10.17}$$

This is a large effect, typically 6% K^{-1}, the current approximately doubling for a 10 K rise in temperature. This dependence is a very important consideration in the design of amplifiers.

10.4.5 Low frequency, small signals

Once the bias or working point has been set up, signal voltages v_{be} (v_{be} means $v_b - v_e$) are also applied between emitter and base making the total voltage applied $V_{be:bias} + v_{be}$. Then, as long as the highest frequency in v_{be} is lower than that at which there is any dependence on frequency in the transistor characteristics (a low frequency signal) the collector current is given by:

$$\begin{aligned}
I_c &= Ce^{-\frac{E_g}{kT}} e^{\frac{e(V_{be:bias}+v_{be})}{kT}} && (10.18)\\
&= Ce^{-\frac{E_g}{kT}} e^{\frac{eV_{be:bias}}{kT}} e^{\frac{ev_{be}}{kT}} && (10.19)\\
&= Ce^{-\frac{E_g}{kT}} e^{\frac{eV_{be:bias}}{kT}} + \left(\frac{ev_{be}}{kT} + \frac{1}{2}\left(\frac{ev_{be}}{kT}\right)^2 + \ldots\right) && (10.20)\\
&= I_{c:bias} + I_{c:bias}\left(\frac{ev_{be}}{kT} + \frac{1}{2}\left(\frac{ev_{be}}{kT}\right)^2 + \ldots\right) && (10.21)
\end{aligned}$$

The first term is the collector bias current. The second term is the signal current, the part of the collector current due to the signal voltage, which we usually want to be undistorted i.e. to have the same waveform as the signal voltage v_{be}. This will only be the case if we can ignore the higher terms, i.e. $\left(ev_{be}/kT\right) \ll 1$ (or $v_{be} \ll 25\,\text{mV}$).

Looked at another way, what the signal voltage does is to cause small excursions about the working point along the $I_c - V_{be}$ characteristic (equation 10.14 above). The condition for linearity (that the signal current has the same waveform as v_{be}) is that the excursions are small enough for the portion of the characteristic covered to be considered straight. We take this as a definition of a *small signal*. Note that it depends on "how linear" we want the system to be.

10.4.6 Low frequency, small signal parameters

When the signal is small as defined above the signal current is given by:

$$i_c = \frac{eI_{c:bias}}{kT} v_{be} \tag{10.22}$$

$$= g_m v_{be} \tag{10.23}$$

where we have written the constant of proportionality $eI_{c:bias}/kT$, called the *small signal mutual conductance,* as g_m. It is the slope of the $I_c - V_{be}$ characteristic at the working point which, because of the exponential, is proportional to $I_{c:bias}$. At room temperature and an $I_{c:bias}$ of 1 mA, $g_m \approx 40\,\text{mA}\text{V}^{-1}$. All bipolar transistors, regardless of their size, have the same g_m at the same temperature and $I_{c:bias}$.

In terms of the bias and signal currents in the base lead the collector current is given by:

$$I_c = I_{c:bias} + i_c = \bar{\beta} I_{b:bias} + \beta_0 i_b \tag{10.24}$$

β_0 is called the *small signal current gain.* (In our simple model it is the same as the ratio of the collector and base bias currents $\bar{\beta}$.) Its value varies from transistor to transistor.

Another small signal parameter of interest is the resistance r_e between base and emitter, this is given by:

$$r_e = \frac{v_{be}}{i_b} = \frac{v_{be}}{i_c}\frac{i_c}{i_b} = \frac{\beta}{g_m} \tag{10.25}$$

At room temperature, with $I_{c:bias} = 1\,\text{mA}$ and $\beta_0 = 100$, $r_e \approx 2.5\,\text{k}\Omega$.

Summarising, the low frequency small signal parameters of our model are a mutual conductance proportional to the bias current, a current gain independent of the bias current, and a resistance between base and emitter inversely proportional to the bias current. We have also assumed that the collector current does not depend on the voltage on the collector, in terms of small signal parameters this amounts to assuming that the output resistance is infinite.

10.4.7 Low frequency, small signal equivalent circuit

We need equivalent circuits of transistors if we are to analyse circuits containing them. We can intuitively construct one following the discussion of the small signal parameters above. We fix our attention on the mutual conductance as the key parameter and taking into account the infinite output resistance make the first element in our equivalent circuit a current generator $g_m v_{be}$ or $\beta_0 i_b$ between collector and emitter, see section **??**. The equivalent circuit is completed by adding a resistance β_0/g_m between base and emitter (Figure 10.7). The entire small signal behaviour of our basic low frequency model of a bipolar transistor is contained in this two element equivalent circuit.

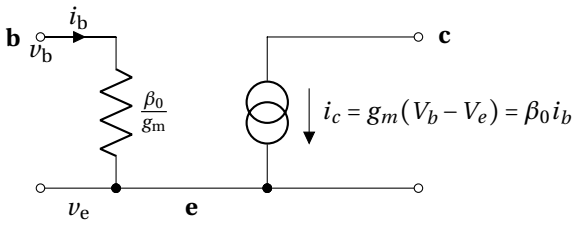


Figure 10.7: Low frequency, small signal, equivalent circuit

10.4.8 A more complete small signal equivalent circuit

We discuss the elements added to Figure 10.4.7 to make Figure 10.7. They account for the following effects.

(i) The current in the base lead flows transversely through the P tpe base material to reach the planar N^+PN region, see e.g. Figure 10.3. This path has significant resistance (typically a few hundred ohms). It shows as $r_{bb'}$ in Figure 10.7.

(ii) It is observed that the collector current increases slighltly when the collector coltage is increased. Part of the increase is due to the increase in width of the collector depletion layer which narrows the field free region of the base and slightly increases the diffusion current, see Figure 10.5. Another cause of the increase is a slight increase of the base voltage due to the increase of collector voltage. These effects are accounted for by the resistance in parallel with the current generator and the resistance between collcetor and base. The value of $1/\mu_0$ is of order 10^4.

(iii) The reverse biased collector base junction has a capacitnce, see paragraph 9.6.8. It shows as C_c in Figure 10.7. In a typical small device its value is of order 10 pF.

(iv) The current entering the base from the emitter is in phase with the voltage V_{be} between base and emitter. The current at the collector side of the base is delayed relative to V_{be} due to the diffusion process. This effect is accounted for by the capacitance

$$\frac{g_m}{\beta_0 \omega_\beta} \approx \pi \tau_e \left(\frac{w^*}{L_e} \right)^2 ,$$

where τ_e is the electron lifetime in the P type base, and the other quantities are defined in chapter 9.

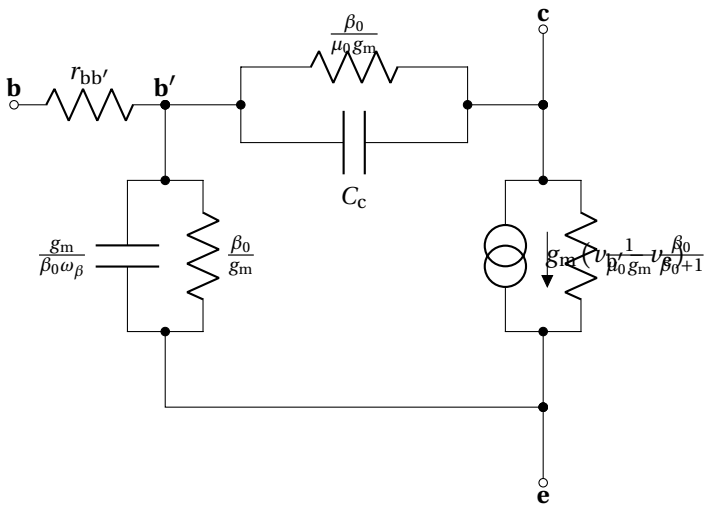


Figure 10.8: Small signal equivalent

10.5 Use of the diffusion transistor as an amplifier

10.5.1 An amplifier circuit

The circuit diagram of a simple amplifier is shown in Figure 10.9. Typically the transistor is biased to a working point with I_c such that $V_{out} \approx 0.5 V_{supply}$.

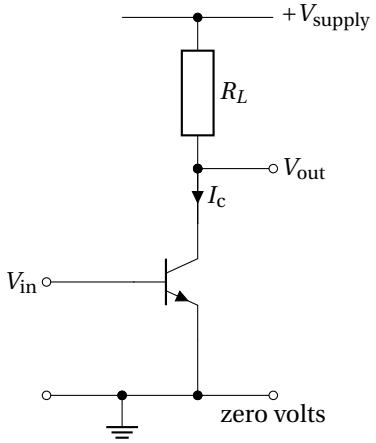


Figure 10.9: Circuit diagram of a simple amplifier

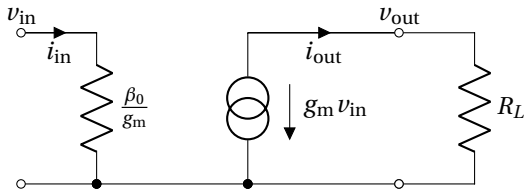


Figure 10.10: Circuit diagram of a simple amplifier

The operation of the circuit is as follows; when V_{in} changes (i.e. an input signal is applied), I_{c} changes and therefore V_{out} changes. The change in V_{out} (the output signal) can be much greater than the change in V_{in} and amplification is obtained.

10.5.2 Small signal low frequency amplifier

When the input signal is small we can use the small signal equivalent circuit in an equivalent circuit of the amplifier. When the input frequency is low enough for the capacitances to be ignored and if we also ignore $r_{\text{bb}'}$ and the resistance involving $1/\mu_0$, the small signal equivalent circuit of the amplifier reduces to that shown in Figure 10.10.

From this equivalent circuit we see that:

$$v_{\text{out}} = i_{\text{out}} R_L = -g_{\text{m}} v_{\text{in}} R_L \tag{10.26}$$

Therefore:

$$\frac{v_{\text{out}}}{v_{\text{in}}} = -g_{\text{m}} R_L \tag{10.27}$$

10.6 Problems with using diffusion transistors in amplifiers

10.6.1 Large signals

As the signal level is increased there is more and more characteristic distortion due to the non-linear relationship between I_{c} and V_{be}. For a sine wave signal input voltage the result is a lopsided output as one half cycle is amplified more than the other. The new frequencies occurring in the output are predominantly even harmonics of the input.

10.6.2 Temperature effects

If a transistor is amplifying a very slowly varying ('quasi dc') signal, any drift in its working point causes a problem as the two effects on the output cannot be distinguished.

For ac signals (at other than extremely low frequencies) we can easily distinguish signals from drifts in bias levels and we can also use blocking capacitors to subtract steady or slowly varying mean levels. Small drifts are therefore not in themselves a problem as they are in dc amplifiers. However, unless precautions are taken drifts might be large, an increase in temperature of a mere 10 K will double I_{c} in the circuit of Figure 10.9 and carry the steady voltage at the collector to zero volts, the bottom end of the working range, making it useless as an amplifier.

The solutions to these problems are explored in chapters 13, 14, and 15. Negative feedback and a circuit using two identical transistors in a balanced arrangement (a long-tailed pair) feature prominently.

10.7 Use of the diffusion transistor as a switch

If an input sinewave signal voltage is large enough the transistor will spend half its time passing no current (completely *cut-off*) and the other half conducting so well (*saturated*) that all the supply voltage is dropped across R_L. In an amplifier this is undesirable and would be described as severe *limiting* (or *clipping*) *distortion*, the sinewave input having become an approximate square wave at the output. However for the transistors to spend most of their time either cut-off or saturated is exactly what we want in binary logic circuits and in these circuits the input signals are deliberately made so large that this is what happens. Under these conditions the transistors are being turned on and off like switches.

11 The Junction Field Effect Transistor

11.1 Introduction

There are two main families of field effect transistors (FET) the *junction gate* family and the *insulated gate* family. Within these families there are two varieties, *P-channel* and *N-channel*, and also, in the insulated gate family, the channels may be permanent or created by biasing. Our main discussion will relate to an N-channel, junction gate type (JFET) in which the important charge carriers are majority electrons. With obvious changes of voltage polarities the account could equally well apply to a P-channel, junction gate device. Insulated gate FETs (MOSFETs) are discussed in Chapter 18.

11.2 Manufacture of N-channel JFETs

As in the manufacture of diffusion transistors, many FETs are made simultaneously on a slice of silicon crystal. The same kinds of masking and diffusion processes are used and will not be described again.

The starting material is a slice of P-type silicon. One face of the slice is exposed to a gas of H_2 and $SiCl_4$ molecules and donor atoms at 1200°C causing an N-type layer to grow on the surface, see Figure 11.1.

A P-type pattern, a small part of which is shown in Figure 11.2, is then diffused into the surface, penetrating down to the original P-type slice. (The dashed lines will be the edges of the individual transistor chips.)

In a second P-type diffusion some of each N-type region is ocnverted back to P-type. The structures formed, one of which is shown in Figure 11.3, consist of an N-type conducting path of length L and width W surrounded by P-type material. The conducting path is called tha *channel*, the ends of the channel the *source* and the *drain* and the P-type material is called the *gate*.

'Channel' is rather a misleading name as it is typically not long and thin but short and wide. In fact L may be a few microns while W may be many mm in a power device. The channel thickness is typically 0.1 μm.

To make best use of the area of the wafer, the W dimension is usually not straight but follows a meander pattern so a plan view of the channel might be as shown in Figure 11.4.

A metal encapsulation often used for low power devices is shown in Figure 11.5. A single glass bead seal in the header carries three lead wires. One of the wires has a nail-head formed on its end on which the FET wafer is mounted. This wire is the gate lead. Bond wires connect the source and the drain to the other two leads.

11.3 JFET characteristics

If the source is taken to be at zero volts and the P-type gate is held at various negative voltages V_g with respect to the source, the I_d versus V_d characteristics shown in Figure 11.6 are found. The curves are symmetrical about the origin, a consequence of the symmetry of the device.

The behaviour observed depends on whether V_d is less than or greater than the "knee" voltage indicated in Figure 11.6.

Qualitatively what is happening is this. Electrons flow along the channel from source to drain due to the influence of V_d. At any point in the channel the cross-section left available for conduction depends on how much has been made insulating by penetration of the depletion layer the surrounding gate junction. This depends on the reverse bias on the junction at the point in question which in turn depends on the voltages on the gate and in the channel. The latter, of course, varies from zero at the source to V_d at the drain.

When V_d is less than the knee voltage, the shape of the depletion layer depends only on V_g and the channel behaves as a voltage controlled resistor. This is seen as the family of straight lines through the origin in Figure 11.6. This property of FETs can be used to make voltage controlled potential dividers.

When V_d is greater than the knee voltage it has an important influence on the shape of the depletion layer causing it to penetrate the channel more deeply at the drain end. An increase of V_d further throttles the channel and it turns out that this effect almost exactly cancels the increase in current that would have resulted if no additional throttling had occurred. The drain current becomes almost independent of V_d and the device behaves

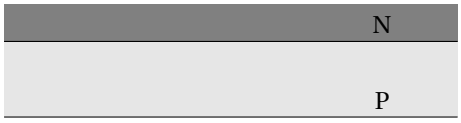


Figure 11.1: Section of slice with grown N-type layer

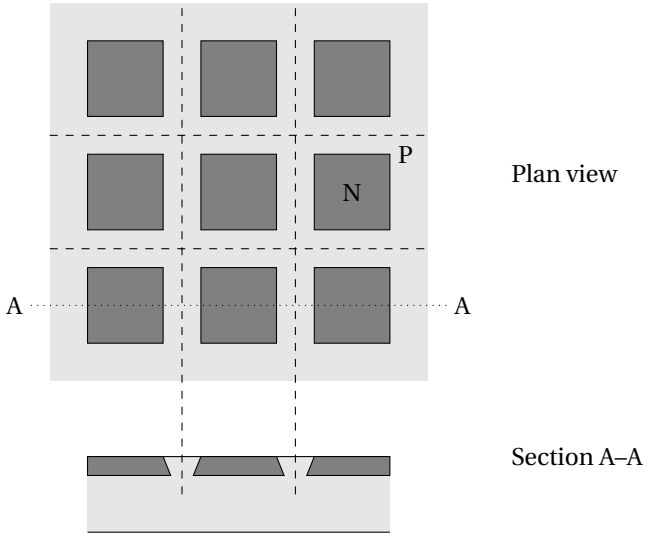


Figure 11.2: Plan view and section of slice after the first P-type diffusion.

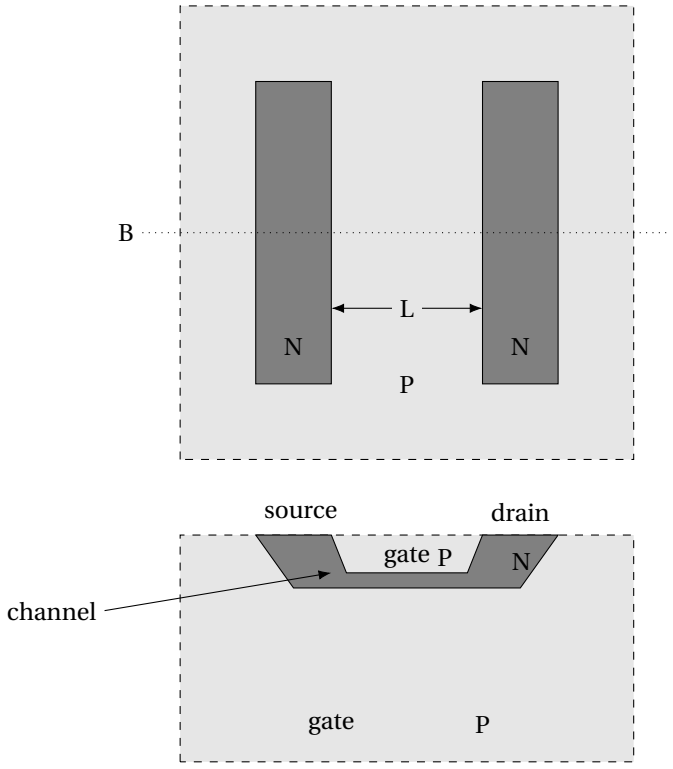


Figure 11.3: Enlarged view of one FET chip after the second P-type diffusion.

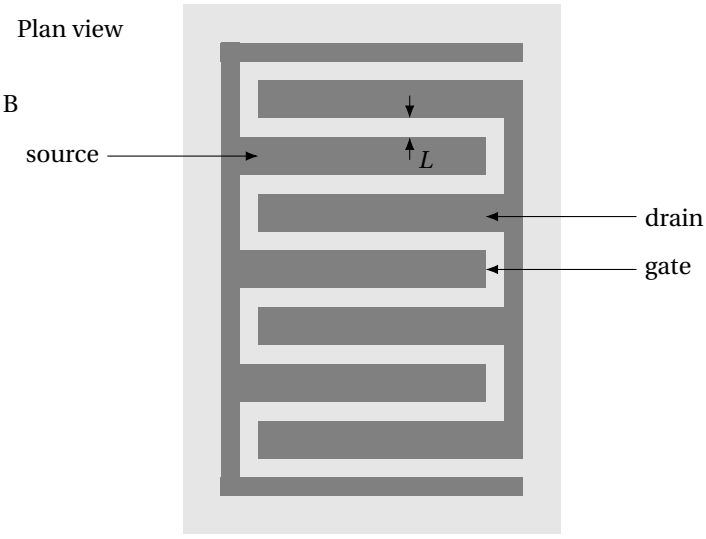


Figure 11.4: Plan view of a meandering channel layout

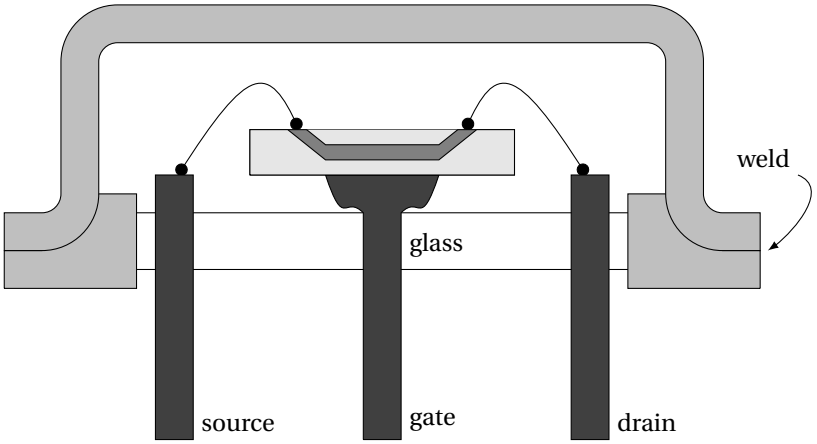


Figure 11.5: N-channel JFET metal encapsulation.

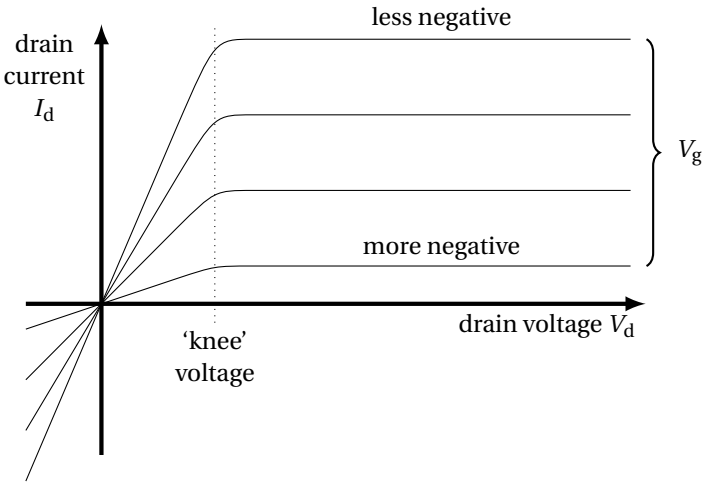


Figure 11.6: N-channel JFET I_d–V_d characteristics, source at 0 volts.

as a voltage controlled current source. This is displayed as the family of horizontal lines in Figure 11.6 (c.f. the diffusion transistor).

In a batch of FETs of the same type considerable variations in the knee voltage are found. A range of 0.5 to 5 V is not unusual.

The derivation of the shapes of the characteristics from the device geometry and material properties will not be attempted here. It is difficult for the high V_d case. Fortunately, although difficult to derive, the empirical dependence of I_d on V_g in the V_d large regime is given quite well for many devices by the simple quadratic relation (in which the effect of V_d is ignored):

$$I_d = I_{dss}\left(1 - V_g/V_p\right)^2 \tag{11.1}$$

where I_{dss} and V_p are constants for the device.

The gate current I_g, being simply the reverse current of a small area silicon PN-junction, is very small, typically ≈ 1 pA.

An increase in temperature affects the drain current through several mechanisms. Two of the most important are an increase in the resistance of the channel and a decrease in the penetration of the gate depletion layer into the channel. These are of opposite sign and since they are of similar magnitude the resulting temperature coefficient can be small and of unpredictable sign. In general, the temperature dependence of the characteristics of FETs is much weaker than that of diffusion transistors.

11.4 JFET circuit symbols

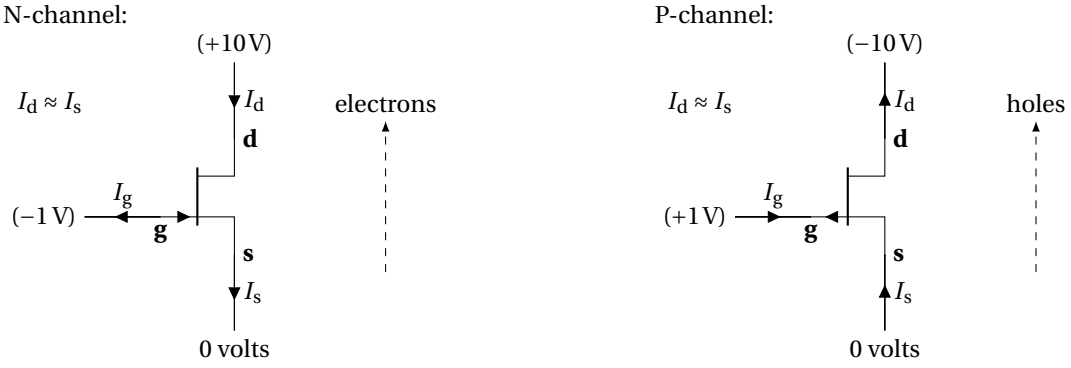


Figure 11.7: JFET circuit symbols

Typical voltages with respect to source in normal operation are shown in brackets.

11.5 Use of the JFET as a small signal amplifier

Before any input signal is applied, the transistor is made active by biasing it to a particular working point, i.e. setting it up with particular steady values of V_d and I_d. V_d is invariably greater than the knee voltage. The comments about small signals made in the previous chapter apply also to the FET although its quadratic characteristic is not so dramatically nonlinear as the exponential of the diffusion transistor.

An FET operated in the high V_d regime is an even better example of a voltage controlled current source than a diffusion transistor. A small signal equivalent circuit is shown in Figure 11.8 below. The delay in the drain current is very short, of the order of 10 ps, and its effects are negligible compared with those of the time constants due to the capacitances. (The FET is essentially a much faster device than the diffusion transistor because it relies on conduction current rather than diffusion current.)

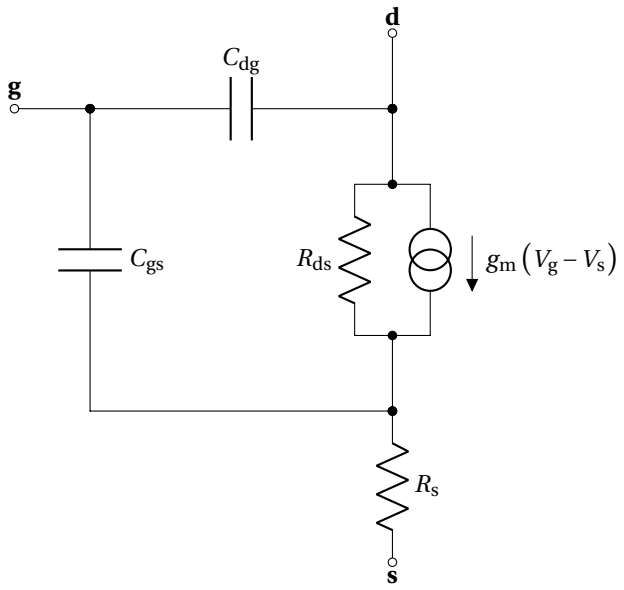


Figure 11.8: Junction FET signal, small amplifier equivalent circuit.

The small signal mutual conductance of the device, g_m, is the slope of the I_d versus V_g characteristic at the working point. We can easily find an expression for g_m since:

$$I_d = I_{dss}\left(1 - \frac{V_g}{V_p}\right)^2 \tag{11.2}$$

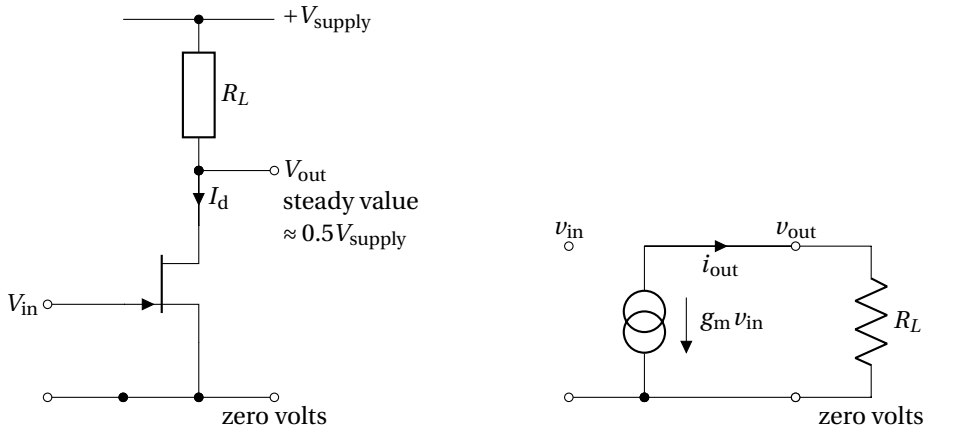
so:

$$g_m = \frac{2\sqrt{I_{dss}}}{V_p}\sqrt{I_d} \tag{11.3}$$

Note particularly that g_m is proportional to $\sqrt{I_d}$ and unlike the g_m of diffusion transistors depends on the size of the device through I_{dss}. A typical value of g_m in a small device passing a drain current of 1 mA is 1 mAV^{-1} much smaller than the 40 mAV^{-1} value of a diffusion transistor at 1 mA collector current.

The capacitances C_{dg} and C_{ds} are larger and R_{ds} is smaller the greater the area of the device. In a small signal JFET the capacitances are typically of order 10 pF and R_{ds} is of order 100 kΩ. The gate current is so small that it is usually ignored. This means that the input resistance and the dc current gain of the device are usually taken as infinite. The resistance R_s can usually be ignored except in high current devices.

The circuit diagram of an amplifier and its small signal low frequency equivalent circuit are shown in Figure 11.9 (R_{ds} and R_s have been ignored).



(a) Amplifier circuit.

(b) Small signal low frequency equivalent circuit.

Figure 11.9: Junction FET signal amplifier circuit.

From the equivalent circuit we see that:

$$V_{out} = i_{out} R_L = -g_m V_{in} R_L \tag{11.4}$$

so:

$$\frac{V_{out}}{V_{in}} = -g_m R_L \tag{11.5}$$

There is no input current.

11.6 Problems in using JFETs

11.6.1 Large signals

As the signal level is increased there is more and more *characteristic distortion* due to the non-linear relationship between I_d and V_g. For a sine wave input voltage the result is a lopsided output as one half-cycle is amplified more than the other. If the characteristic is of precisely quadratic form the only new components in the output are at dc and the second harmonic frequency. Distortion can be reduced by the use of circuits such as the long-tailed pair or the application of negative feedback (see Chapter 12 and Chapter 14).

11.6.2 Effects of temperature changes

The comments made in Chapter 10 concerning the effects of temperature changes on amplifiers using diffusion transistors apply also to amplifiers using FETs. However, as mentioned above, the temperature dependence of the characteristics of FETs is generally much weaker than that of diffusion transistors and the problems are much less severe.

As with diffusion transistors, the use of two identical FETs in a balanced arrangement (a long-tailed pair) can go a long way towards solving this problem (see next chapter).

11.7 Use of the JFET as a switch

JFETs are rarely used as switches but their cousins, the MOSFETs have a starring role in today's digital circuits. Some hint of this is given in Chapter 18.

12 Transistor Stages

12.1 Introduction

In this chapter we begin to describe how circuits are built up using diffusion and field effect transistors. The ideas apply to both integrated circuits (chips) and circuits made with separate ("discrete") components. Opamps and voltage regulators are examples of circuits which may contain fifty or more components. Designers create them by assembling a selection of smaller circuits, typically containing less than half a dozen components, whose properties they are familiar with. It is hardly surprising therefore that the key to understanding complicated circuits is to be able to recognise these smaller circuits, or *stages* as we will call them.

12.2 Small signals and large signals

Describing a signal applied to a non-linear device such as a transistor as "small" is a concise way of saying things like:- 'the excursions about the working point along the device characteristics due to the signal are small enough for the curvature of the characteristics to be ignored', 'the device behaves linearly for small signals', 'signal wave-shapes are independent of signal level and amplitudes of outputs are proportional to amplitudes of inputs'.

When signals are small, the analysis can include all the high frequency and minor effects represented in the small signal equivalent circuits of the transistors.

We define large signals as those for which the curvature of the transistor's characteristics cannot be ignored. This enormously complicates analysis. For instance, the idea of a working point becomes less clear as bias levels change with signal amplitude. Accordingly, for large signals we will limit ourselves to studying only the effects of the non-linearity at low frequencies and not attempt to describe any high frequency or second order effects. The (time domain) models used will be the low frequency characteristic curves:

$$I_c = I_{c0} e^{e\left(V_{b'} - V_e\right)/kT} \tag{12.1}$$

where $I_b = I_c/\bar{\beta}$ and we have written I_{c0} for $C\left(T\right) e^{-E_g/kT}$. The level of approximation is equivalent to considering only the resistance between b′ and e and the current generator in the small signal equivalent circuit.

For field effect transistors:

$$I_d = I_{dss}\left(1 - \frac{V_d}{V_p}\right)^2 \tag{12.2}$$

an approximation equivalent to retaining only the generator in the small signal case.

It might be thought that the non-linearity of transistor characteristics would inevitably give rise to serious distortion of large signals. It turns out that the situation is not as bad as might be feared.

12.3 Biasing, signal zeros, and terminals

In the next chapter (Opamps) we will see that trains of stages are usually coupled together so that each biases the transistors in the next. Accordingly we will not pay much attention to how the working points of the transistors are maintained in this chapter.

In all circuits there will be some points whose voltages do not vary when there is a signal present. Examples are points connected either directly, or via large value capacitors, to power supplies or to the zero of voltage. On small signal equivalent circuits such points are shown as connected directly to the zero of voltage (which usually means earth). We may refer to such points as *signal zero* points.

Input and output terminals will be represented in circuit diagrams and equivalent circuits by small circles.

12.4 One transistor stages

Given that a signal source has two terminals, a transistor has three terminals, and any load to which the amplified signal is delivered has two terminals, it should be clear that however the interconnections are made the transistor will have one terminal in common with the input and output (signal source and load) circuits. This provides a convenient way of referring to the arrangement e.g. common emitter. The amplifying arrangements[1] are shown in Figure 12.1 in which the input terminals are on the left and the output terminals on right. (PNP and P-channel devices could equally well have been used for illustration.)

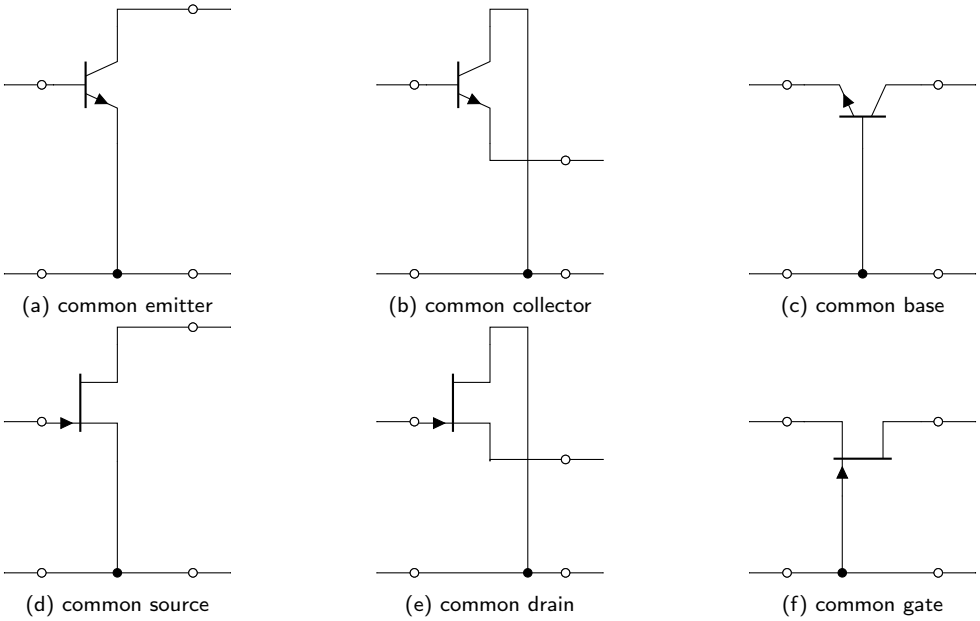


Figure 12.1: One-transistor amplifying arrangements.

We now turn to detailed descriptions of the three one-transistor amplifying arrangements. We use the small signal equivalent circuit of the bipolar transistor to construct small signal equivalent circuits of the three stages. The expressions are easily generalised to take account of the load not being resistive, R_L is simply replaced by Z_L. If the source is non-resistive R_S is replaced by Z_S.

The corresponding (simpler) expressions for the gains and impedances of field effect transistor stages can usually be obtained from the bipolar ones by omitting elements, see Figures 10.8 and 11.9.

The complete expressions are rather complicated. Simplifications can be made for signal frequencies < 10 kHz by omitting the capacitances from the equivalent circuits of the transistors. Also, the errors involved in ignoring the resistances in parallel with C_c and in parallel with the current generator are usually acceptable. At higher frequencies it is usually not possible to see if it is permissible to omit any capacitors before starting the analysis. For instance, in the common emitter circuit, it turns out after doing the algebra that it is usually C_c not the diffusion capacitance (the one in Z_2) which causes the voltage gain of the stage to first begin to fall with frequency so in retrospect the diffusion capacitance could have been ignored. This would probably not have been apparent at the start.

A table of the greatly simplified expressions resulting from simplifying the transistor equivalent circuit to just two elements (Table 12.1) is given in section 12.6.

12.4.1 Common emitter/common source stages

Figure 12.2(a) is a circuit diagram of a common emitter stage using an NPN bipolar transistor. Figure 12.2(b) is an equivalent circuit of the common emitter stage connected to a signal source and load. (It would be the same for a PNP transistor). The voltage of the common wire has been taken as zero. With figure (b) as labelled the node and branch equations in the frequency domain are:

[1] There are three non-amplifying configurations in which the signal sources and loads are interchanged.

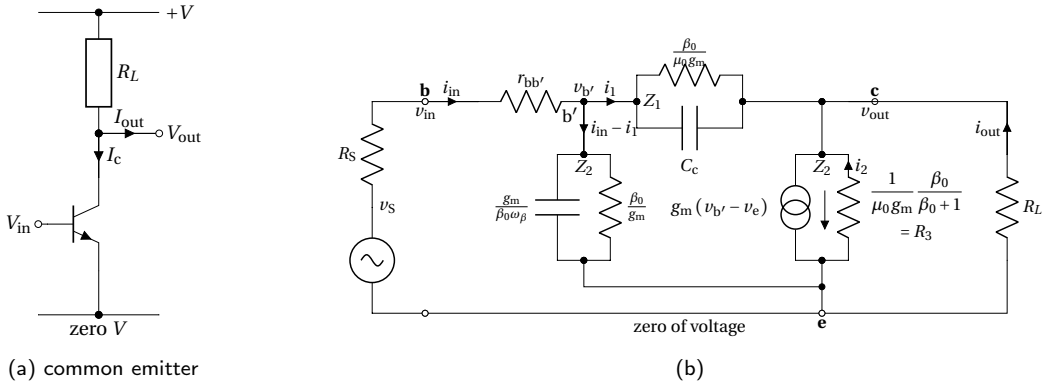


Figure 12.2: Common emitter circuit diagram (a) and equivalent circuit (b)

$$
\begin{aligned}
i_1 &= g_m(v_{b'} - v_e) - i_{out} - i_2 && (12.3)\\
v_e &= 0 && (12.4)\\
v_{in} - v_{b'} &= i_{in} r_{bb'} && (12.5)\\
v_{b'} - 0 &= (i_{in} - i_1) Z_2 && (12.6)\\
v_{out} - 0 &= -i_{out} R_L && (12.7)\\
v_{out} - 0 &= -i2 \frac{1}{\mu_0 g_m} \frac{\beta_0}{\beta_0 + 1} = -i_2 R_3 && (12.8)\\
v_{b'} - v_{out} &= i_1 Z_1 && (12.9)\\
v_S - v_{b'} &= i_{in}(R_S + r_{bb'}) && (12.10)
\end{aligned}
$$

where:

$$Z_1 = \frac{\beta_0}{\mu_0 g_m + j\omega C_c \beta_0} \tag{12.11}$$

and:

$$Z_2 = \frac{\beta_0}{g_m\left(1 + j\frac{\omega}{\omega_\beta}\right)} \tag{12.12}$$

from which the following expressions may be obtained:

voltage gain

$$\frac{v_{in}}{v_{out}} = \frac{g_m R_L - \frac{R_L}{Z_1}}{\left(1 + \frac{R_L}{R_3}\right)\left(1 + r_{bb'}\left(\frac{1}{Z_1} + \frac{1}{Z_2}\right) + \frac{R_L}{Z_1}\left(1 + r_{bb'}\left(g_m + \frac{1}{Z_2}\right)\right)\right)} \tag{12.13}$$

current gain

$$\frac{i_{out}}{i_{in}} = \frac{g_m - \frac{1}{Z_1}}{\left(1 + \frac{R_L}{R_3}\right)\left(\frac{1}{Z_1} + \frac{1}{Z_2}\right) + \frac{R_L}{Z_1}\left(g_m + \frac{1}{Z_2}\right)} \tag{12.14}$$

input impedance

$$\frac{v_{in}}{i_{in}} = Z_{in} = r_{bb'} + \frac{1 + \frac{R_L}{R_3} + \frac{R_L}{Z_1}}{\left(1 + \frac{R_L}{R_3}\right)\left(\frac{1}{Z_1} + \frac{1}{Z_2}\right) + \frac{R_L}{Z_1}\left(g_m + \frac{1}{Z_2}\right)} \tag{12.15}$$

output impedance

$$Z_{\text{out}} = \frac{1 + (R_S + r_{\text{bb}'})\left(\frac{1}{Z_1} + \frac{1}{Z_2}\right)}{\frac{1}{R_3} + (R_S + r_{\text{bb}'})\left(\frac{1}{R_3}\left(\frac{1}{Z_1} + \frac{1}{Z_2}\right) + \frac{1}{Z_1}\left(\frac{1}{Z_2} + g_{\text{m}}\right)\right) + \frac{1}{Z_1}} \tag{12.16}$$

When typical values are inserted in these expressions it is found that the voltage gain is roughly proportional to g_{m}, a parameter that depends on working point, whereas the current gain is roughly proportional to β and only weakly dependent on g_{m}. The consequence is that the large signal voltages suffer severe distortion but large current signals do not (provided that the assumption of constant β is justified).

If we consider only cases in which R_S can be ignored, it is clear from comparing the equivalent circuits of the diffusion and field effect transistors that the corresponding expressions for the common source stage can be obtained by setting $\beta_0 = \infty$, $Z_2 = 1/j\omega C_{\text{gs}}$, $Z_1 = 1/j\omega C_{\text{dg}}$, $\left(1/\mu_0 g_{\text{m}}\right) = R_{\text{ds}}$ and $r_{\text{bb}'} = 0$.

Turning now to the large signal case, bearing in mind the comments in section 12.2, and referring to Figure 12.2(a). When there is no signal:

$$V_{\text{out}} = V_{\text{out:bias}} \tag{12.17}$$

$$I_{\text{c}} = I_{\text{c:bias}} \tag{12.18}$$

$$+V - V_{\text{out:bias}} = I_{\text{c:bias}} R_L \tag{12.19}$$

assuming $I_{\text{out}} = 0$.

When there is a signal,

$$+V - V_{\text{out:sig}} = I_{\text{c:sig}} R_L \tag{12.20}$$

So:

$$V_{\text{out:sig}} - V_{\text{out:bias}} = -\left(I_{\text{c:s}} - I_{\text{c:bias}}\right) R_L \tag{12.21}$$

$$= -I_{\text{c:bias}} R_L \left(e^{e\left(V_{\text{b}'\text{:sig}} - V_{\text{b}'\text{:bias}}\right)/kT} - 1 \right) \tag{12.22}$$

The change in output voltage is related to the change of input voltage by an exponential. Deriving the corresponding expression for the common source stage is left as an exercise.

12.4.2 Common collector/common drain stages (also called emitter followers/source followers)

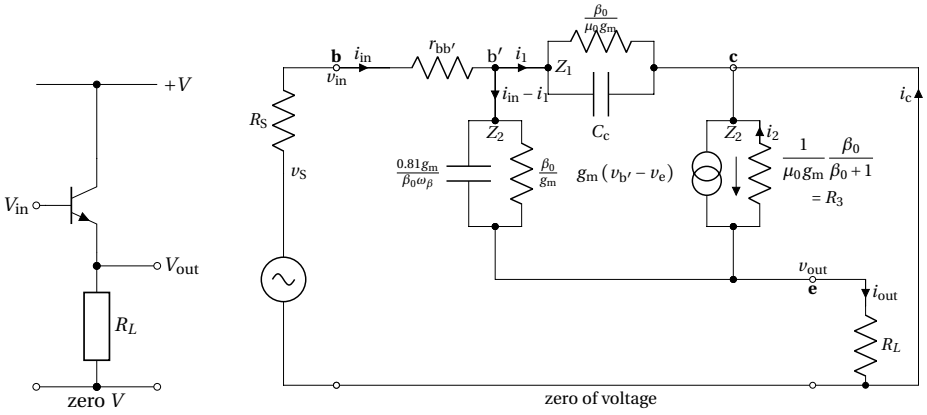


Figure 12.3: Common collector (emitter follower) circuit diagram and equivalent circuit

Figures 12.3(a) is a circuit diagram of an emitter follower stage. Figure 12.3 (b) is the equivalent circuit. Refer to the beginning of section 12.4.1 for other relevant remarks. The load is connected between the emitter and the zero volts line and all the signal voltage developed across it is in series with the input. (This as we shall learn in chapter 15 is a case of 100% series voltage negative feedback.)

The node and branch equations in the frequency domain are:

$$\begin{aligned}
i_{\text{out}} &= i_{\text{in}} - i_1 + g_{\text{m}}\left(v_{\text{b}'} - v_{\text{e}}\right) - i_2 && (12.23)\\
i_2 + i_1 + i_{\text{c}} &= g_{\text{m}}\left(v_{\text{b}'} - v_{\text{e}}\right) && (12.24)\\
v_{\text{e}} &= v_{\text{out}} && (12.25)\\
v_{\text{in}} - v_{\text{b}'} &= i_{\text{in}} r_{\text{bb}'} && (12.26)\\
v_{\text{b}'} - v_{\text{out}} &= \left(i_{\text{in}} - i_1\right) Z_2 && (12.27)\\
v_{\text{out}} - 0 &= i_{\text{out}} R_L && (12.28)\\
v_{\text{out}} - 0 &= -i_2 \frac{1}{\mu_0 g_{\text{m}}} \frac{\beta_0}{\beta_0 + 1} = i_2 R_3 && (12.29)\\
v_{\text{b}'} - 0 &= i_1 Z_1 && (12.30)\\
v_{\text{s}} - v_{\text{b}'} &= i_{\text{in}}\left(R_S + r_{\text{bb}'}\right) && (12.31)
\end{aligned}$$

with Z_1 and Z_2 as in section 12.4.1.

We are not interested in i_{c} so the third equation is redundant. From the rest of the equations the following expressions may be derived:

voltage gain

$$\frac{v_{\text{out}}}{v_{\text{in}}} = \frac{g_{\text{m}} + \frac{1}{Z_2}}{\left(1 + \frac{r_{\text{bb}'}}{Z_1}\right)\left(g_{\text{m}} + \frac{1}{Z_2}\right) + \left(\frac{1}{R_L} + \frac{1}{R_3}\right)\left(1 + r_{\text{bb}'}\left(\frac{1}{Z_1} + \frac{1}{Z_2}\right)\right)} \tag{12.32}$$

current gain

$$\frac{i_{\text{out}}}{i_{\text{in}}} = \frac{g_{\text{m}} + \frac{1}{Z_2}}{\frac{R_L}{Z_1}\left(g_{\text{m}} + \frac{1}{Z_2}\right) + \left(\frac{1}{Z_1} + \frac{1}{Z_2}\right)\left(1 + \frac{R_L}{R_3}\right)} \tag{12.33}$$

input impedance

$$\frac{v_{\text{in}}}{i_{\text{in}}} = Z_{\text{in}} = r_{\text{bb}'} + \frac{g_{\text{m}} + \frac{1}{Z_2} + \frac{1}{R_L} + \frac{1}{R_3}}{\frac{1}{Z_1}\left(g_{\text{m}} + \frac{1}{Z_2}\right) + \left(\frac{1}{R_L} + \frac{1}{R_3}\right)\left(\frac{1}{Z_1} + \frac{1}{Z_2}\right)} \tag{12.34}$$

output impedance

$$Z_{\text{out}} = \frac{1 + \left(R_S + r_{\text{bb}'}\right)\left(\frac{1}{Z_1} + \frac{1}{Z_2}\right)}{\left(1 + \frac{R_S + r_{\text{bb}'}}{Z_1}\right)\left(g_{\text{m}} + \frac{1}{Z_2}\right) + \frac{1}{R_3}\left(1 + \left(R_S + r_{\text{bb}'}\right)\left(\frac{1}{Z_1} + \frac{1}{Z_2}\right)\right)} \tag{12.35}$$

With typical values inserted the voltage gain is found to be close to (but just below) +1 and is only weakly dependent on the working point. In other words the voltage at the output (the emitter) closely follows the voltage at the input and the distortion is low. The expressions for the common drain/source follower stage are easily derived by the method used in section 12.4.1.

Given an R_S which is not too high the higher g_{m} of bipolar transistors tends to make the output impedance of emitter followers lower than that of source followers.

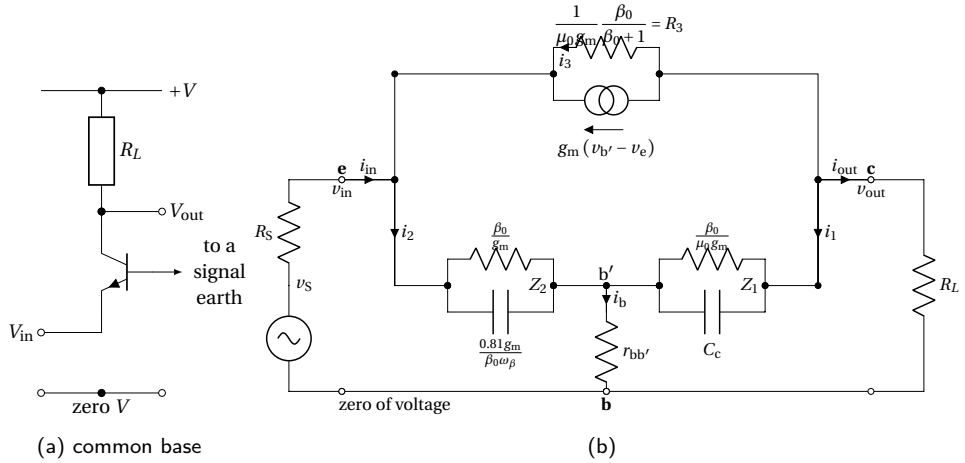


Figure 12.4: Common base/common gate stage circuits and equivalent circuit

12.4.3 Common base/common gate stages

Figures 12.4(a) and (b) are a circuit diagram and an equivalent circuit of a common base stage. Refer to the beginning of section 12.4.1 for other relevant remarks.

The node and branch equations in the frequency domain are:

$$\begin{aligned}
i_2 &= i_{in} + i_3 + g_m(v_{b'} - v_{in}) && (12.36)\\
i_1 + i_2 &= i_b && (12.37)\\
0 &= i_{out} + i_1 + g_m(v_{b'} - v_{in}) + i_3 && (12.38)\\
v_{b'} &= i_b r_{bb'} && (12.39)\\
v_{in} - v_{b'} &= i_2 Z_2 && (12.40)\\
v_{out} - v_{b'} &= i_1 Z_1 && (12.41)\\
v_{out} - v_{in} &= i_3 \frac{1}{\mu_0 g_m}\frac{\beta_0}{\beta_0+1} = i_3 R_3 && (12.42)
\end{aligned}$$

with Z_1 and Z_2 as in section 12.4.1.

From these equations the following expressions may be derived:

voltage gain

$$\frac{v_{out}}{v_{in}} = \frac{\frac{Z_1 Z_2}{R_3}(1 + g_m R_3) + r_{bb'}\left(1 + \frac{Z_1+Z_2}{R_3} + g_m Z_2\right)}{\left(\frac{1}{R_3} + \frac{1}{R_L}\right)\left((Z_1+Z_2)r_{bb'} + Z_1 Z_2\right) + Z_2 + r_{bb'}(1 + g_m Z_2)} \qquad (12.43)$$

current gain

$$\frac{i_{out}}{i_{in}} = \frac{\frac{Z_1 Z_2}{R_3}(1 + g_m R_3) + r_{bb'}\left(1 + \frac{Z_1+Z_2}{R_3} + g_m Z_2\right)}{Z_1\left(1 + \frac{Z_1+Z_2}{R_3} + g_m Z_2\right) + (R_L + r_{bb'})\left(1 + \frac{Z_1+Z_2}{R_3} + g_m Z_2\right)} \qquad (12.44)$$

input impedance

$$Z_{in} = R_L \frac{\left(\frac{1}{R_3} + \frac{1}{R_L}\right)\left((Z_1+Z_2)r_{bb'} + Z_1 Z_2\right) + Z_2 + r_{bb'}(1 + g_m Z_2)}{Z_1\left(1 + \frac{Z_1+Z_2}{R_3} + g_m Z_2\right) + (R_L + r_{bb'})\left(1 + \frac{Z_1+Z_2}{R_3} + g_m Z_2\right)} \qquad (12.45)$$

output impedance

$$Z_{\text{out}} = \frac{(Z_1 + r_{\text{bb}'})(1 + g_{\text{m}} Z_2) R_S + \left(1 + \frac{R_S}{R_3}\right)\left((Z_1 + Z_2) r_{\text{bb}'} + Z_1 Z_2\right)}{Z_2 + (R_S + r_{\text{bb}'})\left(1 + \frac{Z_1 + Z_2}{R_3} + g_{\text{m}} Z_2\right)} \tag{12.46}$$

The common base stage has near unity current gain and compared with the common emitter stage, a low input impedance, a high output impedance, and a greater bandwidth. The expressions for the common gate stage are easily derived by the method used in section 12.4.1.

12.5 Two transistor stages

These are combinations of the one transistor stages and we will concentrate on the features resulting from the combinations rather than giving full treatments.

12.5.1 The long tailed pair stage

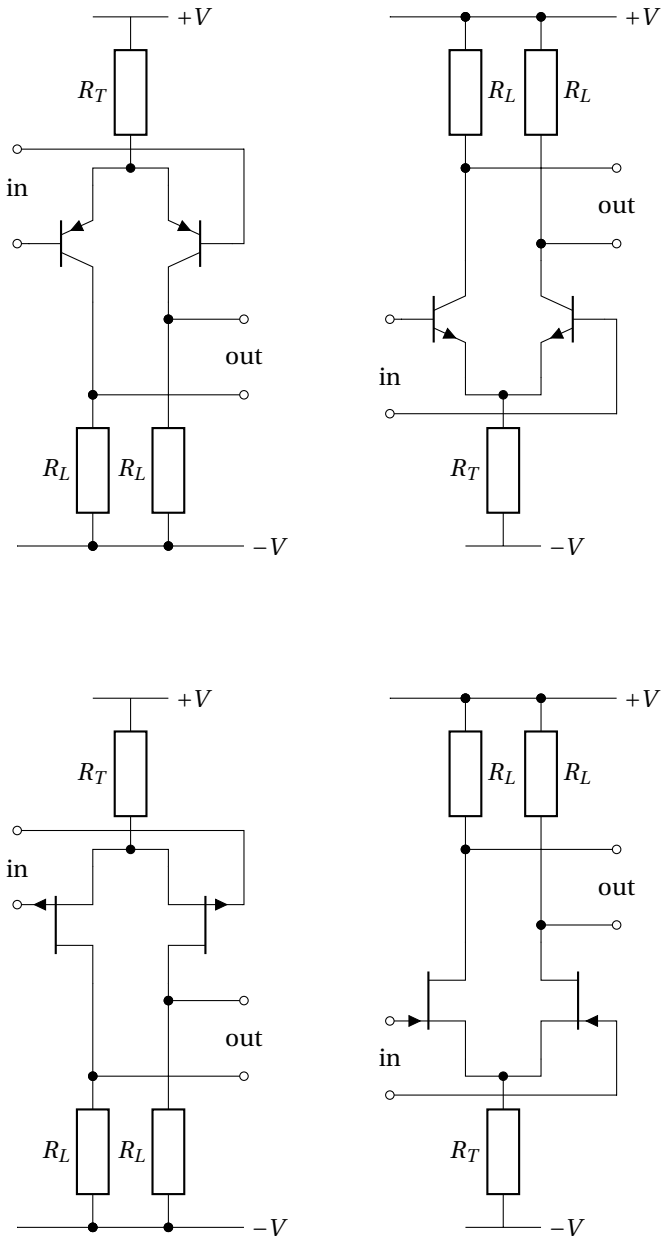
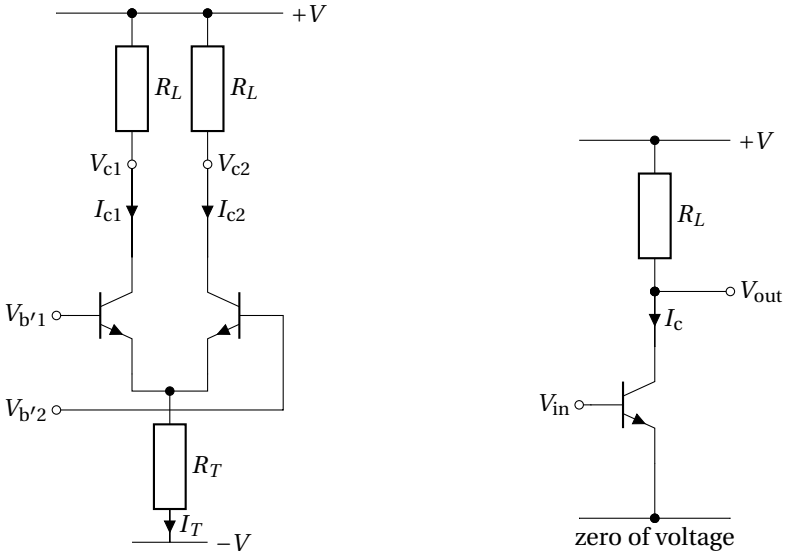


Figure 12.5: Circuit diagrams of Long Tailed Pair (LTP) stages made using various transistor types

The long tailed pair[1] (LTP) is a combination of two common emitter (or common source) stages in which each transistor amplifies half the signal. Circuit diagrams of LTPs made with various types of transistor are shown in Figure 12.5. For definiteness the operation of the NPN circuit shown in Figure 12.6(a) will be described. Assuming the two transistors are identical and set up with equal bias voltages (often near zero) on their bases, the emitter currents which flow through the "long tail" resistor R_T to the negative supply are equal and the collector voltages are also equal.



(a) Long-tailed pair (b) Common emitter

Figure 12.6: Long Tailed Pair (LTP) stage using NPN transistors and a common-emitter configuration

Used as an amplifier stage the input signal is applied between the two bases and controls the way the tail current I_T divides between the two transistors. The output is taken between the two collectors. The bipolar LTP has important advantages over a single common emitter stage. In particular, the linearity is much better as will now be demonstrated.

Referring to Figure 12.6(a) and bearing in mind the comments in section 12.2 we can write:

$$I_{c1} = I_{c0} e^{e\left(V_{b'_1} - V_e\right)/kT} \tag{12.47}$$

and

$$I_{c2} = I_{c0} e^{e\left(V_{b'_2} - V_e\right)/kT} \tag{12.48}$$

Also we have

$$+V - V_{c1} = I_1 R_L \tag{12.49}$$

and:

$$+V - V_{c2} = I_2 R_L \tag{12.50}$$

From these it is straightforward to derive:

$$V_{c1} - V_{c2} = -\frac{V_e - -V}{R_T} \tanh\left(\frac{e\left(V_{b'_1} - V_{b'_2}\right)}{2kT}\right) \tag{12.51}$$

The leading fraction is I_T which because changes of V_e with signal are invariably small compared with $-V$ is essentially constant. The output signal voltage (the voltage between the two collectors) is related to the input signal voltage (the voltage between the bases) by the tanh function rather than the exponential found for the common emitter (CE) stage in section 12.4.1. The tanh function has a point of inflection at the working point so for the same size of input signal voltage the distortion is less for the LTP. This is not too surprising: the exponential in the CE expression produces harmonics of an input sinewave of which the largest is the second. The LTP being

[1] The first use of this most valuable circuit is attributed to Alan Blumlein, the chief engineer of EMI killed in an air crash during the second world war. (He of course used thermionic valves.)

symmetrical cannot produce any even harmonics so the first to appear is the third. This is an example of where the use of non linear devices does not produce as much distortion as might have been expected. (In fact the odd harmonics are also less in the LTP because the transistor that is conducting harder tends to pull up V_e slightly.) Note that it makes little difference to the output voltage whether half the input is fed to each device (balanced signal source with centre tap at signal zero) or all the input is fed to one side (single ended signal source). There is a change in V_e but this is small compared with $-V$.

When the signal is small we have:

$$V_{c1} - V_{c2} = -\frac{eI_T}{2kT} R_L \left(V_{b'_1} - V_{b'_2} \right) \tag{12.52}$$

$$= -g_m R_L \left(V_{b'_1} - V_{b'_2} \right) \tag{12.53}$$

where g_m is the mutual conductance of one of the transistors at a bias current of $I_T/2$, so the voltage gain is just what we would expect for two cooperating common emitter stages. The input resistance of the LTP is twice that of the CE stage.

Effect of temperature changes on the LTP

It was mentioned in Chapter 10 that unless precautions are taken the bias current in a common emitter stage will have a large dependence on temperature. One of the most important benefits of the LTP is the insensitivity of the bias currents to temperature changes.

Consider what happens when the (equal) base voltages are fixed and the temperature increases by a few Kelvin. The result is an increase of a few mV in the voltages of the emitters of the two transistors. This change is a very small fraction of the voltage across R_T which determines the tail current so the temperature dependence of this current is small. The next thing required is that the tail current continues to divide equally between the two transistors. This will occur if both devices follow the temperature change in step. In practice this is achieved by forming the transistors side by side on the same chip or, less satisfactorily, mounting discrete transistors in thermal contact.

Summary of the advantages of the long-tailed pair stage

(a) working points insensitive to temperature changes

(b) lower distortion: the lop-sidedness of large signals seen in common emitter/source stages is eliminated.

(c) output insensitive to fluctuations in the power supply voltages: both sides are affected equally and the input and output signals are voltage differences between the two halves.

(d) can be coupled easily to other long-tailed pairs without capacitors to make circuits which work down to 0 Hz (dc).

These advantages make the long-tailed pair stage (made either with diffusion or field effect devices) the ideal building block for the early stages of opamps. See for example, Figure M13.1.

12.5.2 The long tailed pair with current mirror

A circuit known as a 'current mirror' (CM) is shown acting as a load for a LTP in Figure 12.7. The two transistors in the mirror are identical and at the same temperature. From the diagram:

$$\frac{I_T}{2} - \delta I = I_c \left(1 + \frac{2}{\beta} \right) \tag{12.54}$$

and

$$\frac{I_T}{2} + \delta I = I_{out} + I_c \tag{12.55}$$

from which it follows that

$$I_{out} = 2\frac{\beta + 1}{\beta + 2} \delta I + \frac{I_T}{\beta + 2} \tag{12.56}$$

When the signal input is zero ($\delta I = 0$):

$$I_{out:no\ sig} = \frac{I_T}{\beta + 2} \tag{12.57}$$

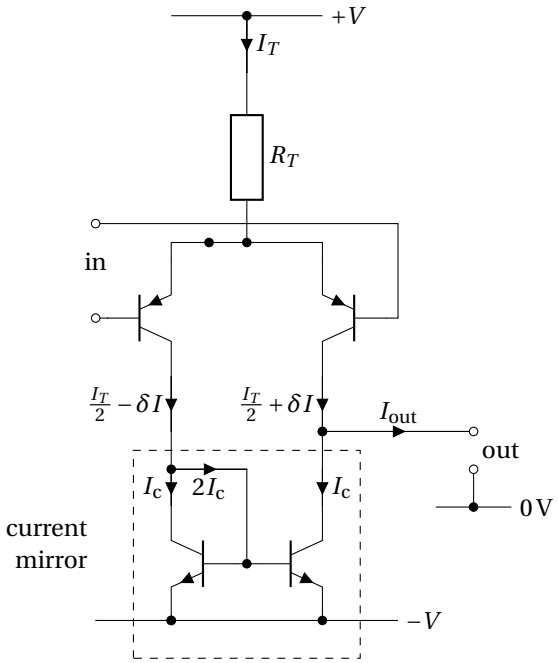


Figure 12.7: Long Tailed Pair (LTP) stage with current mirror as load

So:

$$I_{out} - I_{out:no\ sig} = 2\frac{\beta+1}{\beta+2}\delta I \tag{12.58}$$

$$\approx 2\delta I \tag{12.59}$$

The high output impedance of the LTP + CM tends to define the input signal *current* of the following stage. Where, as it usually is, this is a common emitter stage its linearity is improved (see section 12.4.1). What happens is that the large signal input voltage of the CE is distorted in just the right way to improve its output.

12.5.3 The cascode stage

The impedance Z_1 between the collector and base in the equivalent circuit of the CE stage (or the drain and gate of the common source stage) has a voltage $v_{out} + v_{b'e} \approx (G_v + 1)\, v_{b'e}$ across it where G_v is the voltage gain. The amount of current flowing from the signal source into Z_1 is the same as if the input voltage $(\approx v_{b'e})$ had been applied to an impedance $Z_1/(G_v + 1)$. The important part of Z_1 is the C_c (or the C_{ds}) so its contribution to the input impedance of the stage is a shunt capacitance of $\approx (G_v + 1)\, C_c$ which can be embarrassingly large for a stage to be used at high frequencies. (This magnification of the input capacitance is called the *Miller effect.*) One solution is to couple the common emitter or common source stage not directly into the load but via a common base or common gate stage which have low input resistances. All the signal current still flows in the load so the voltage gain is the same as before but the voltage gain of the common emitter or common source part of the stage is now low so the Miller effect is greatly reduced. This is the cascode circuit illustrated in Figure 12.8.

12.6 Approximate properties of the small-signal stages

The properties of the small-signal stages for $r_{bb'} = \mu_0 = C_{diff} = C_c = 0$, $Z_2 = \beta_0/g_m$, and $Z_1 = \infty$ (the transistor equivalent circuit shown in Figure 10.7) are given in the table below.

12.7 Power output stages

A significant fraction of the electrical power fed to an output stage is spent in heating the transistors. This heat must be able to escape rapidly enough for the temperature of the semiconductor to be kept down to a safe level.

To enable the heat generated to be conducted away, power devices are spread over much larger areas of silicon than small signal devices. This leads to larger capacitances. Also diffusion power transistors are made

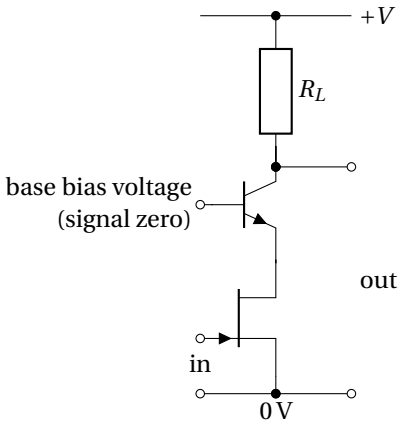


Figure 12.8: Cascode stage

with thicker base regions in order to withstand high collector voltages with the result that their current gains are lower and transit times longer. These effects tend to reduce the frequency response of circuits containing them. For these reasons it is not a good idea to choose a much larger device than is really needed. This involves knowing the limits on the maximum current and voltage as well as their product (the maximum power) devices can withstand.

The extent of this permissible working region (called the *safe operating region* or SOR) of a device is specified by the manufacturer after exhaustive tests. A typical SOR is shown in Figure 12.9.

12.7.1 A common emitter/common source power stage

An NPN common emitter power output stage with a transformer coupled load is shown in Figure 12.10. It is biased (by circuits not shown) to a current I and voltage V =(+V assuming the transformer primary winding has zero resistance). The biasing is such that collector current flows throughout a whole cycle of the maximum signal input voltage, a type of operation known as *Class A*.

If the transistor was linear (signal collector current proportional to V_{in}) the maximum output without clipping would be obtained when the current is swung from 0 to 2I and the collector voltage is swung from twice +V to 0 at the negative and positive peaks of v_{in} respectively. Assuming an ideal transformer this would occur when $V/I = R_L/n^2$, a value of resistance known as the *optimum load* on the transistor. The use of a transformer of suitable turns ration n enables any R_L to present the optimum load to the transistor, a technique known as *matching*. Note that this optimum load, which allows the working ranges in both voltage and current to be filled simultaneously, depends only on the bias levels. It has nothing to do with the output resistance of the stage (which on our simplest model is infinite) or the maximum power transfer theorem (which does not apply when voltage swings are limited).

Continuing to assume linearity it is easy to show that the efficiency η, given by

$$\eta = \frac{\text{signal power dissipated in load}}{\text{power taken from supply}} \tag{12.60}$$

stage	**voltage gain**	**current gain**	**input resistance**	**output resistance**
CE	$-g_{\text{m}}R_L$	β_0	$\frac{\beta_0}{g_{\text{m}}}$	∞
CC	$\frac{1}{1+\frac{\beta_0}{\beta_0+1}\frac{1}{g_{\text{m}}R_L}}$	β_0+1	$\frac{\beta_0}{g_{\text{m}}}+(\beta_0+1)R_L$	$\frac{1}{g_{\text{m}}}\frac{\beta_0}{\beta_0+1}+\frac{R_S}{\beta_0+1}$
CB	$g_{\text{m}}R_L$	$\frac{\beta_0}{\beta_0+1}$	$\frac{1}{g_{\text{m}}}$	∞
LTP	$-g_{\text{m}}R_L$	β_0	$\frac{2\beta_0}{g_{\text{m}}}$	∞
LTP+CM	$-g_{\text{m}}R_L$* *R_L, the input resistance of the next stage, can be large.	$2\beta_0$	$\frac{2\beta_0}{g_{\text{m}}}$	∞

Table 12.1: Approximate properties of small-signal stages

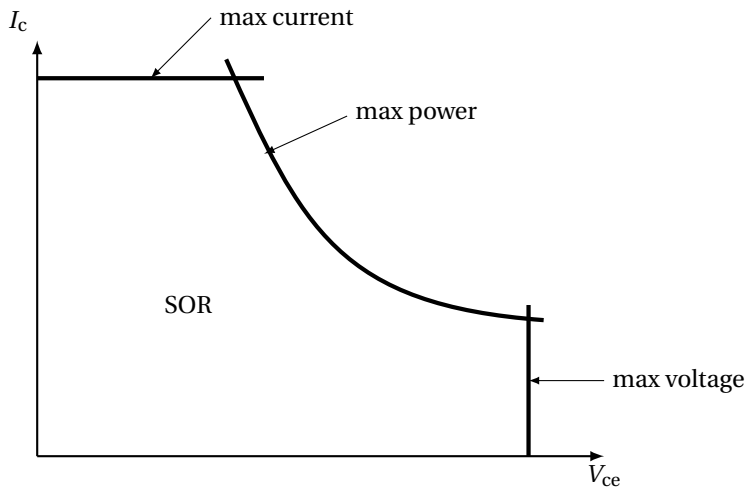


Figure 12.9: Safe operating region $V-I$ plot for a power output stage

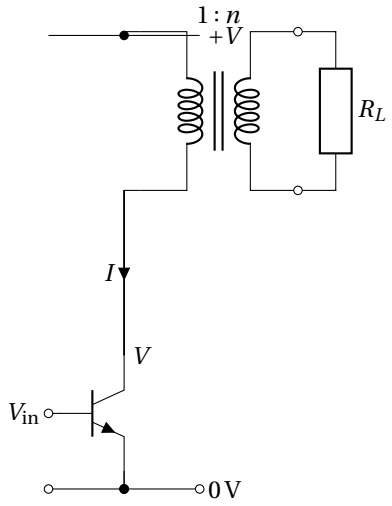


Figure 12.10: NPN common emitter power output stage

has for a sinewave input signal a *maximum* value of 50% (when the stage is fully driven). The efficiency is much lower for spiky signals like speech and music with peaks just reaching the ends of the working range and falls to zero for small signals because of the continuous dissipation of bias power. Real transistors are of course non-linear so the choice of a suitable bias point is more difficult, also considerable distortion occurs before the output approaches the ends of the working range and signal levels are usually restricted to reduce this distortion. This leads to even lower efficiency.

This circuit was often used as the last stage of the audio amplifier in car radios where plenty of power was available from the battery.

12.7.2 The compound emitter follower power stage

Where high efficiency is required the circuit shown in Figure 12.11, known as a *compound, Class B push-pull*, emitter follower output stage, is used. By compound is meant that a pair of NPN and PNP transistors are used to feed the same load, by Class B is meant that the transistors are biased to the point of just conducting and that each conducts on the alternate half cycle of the input that turns it on and remains non conducting for the other half cycle. "Push-pull" indicates that one device sources the current through the load and the other sinks it.

In the absence of signal, the output terminal **C** is at zero volts, and so is the point **B**. Bias current flowing through the diodes ensures that point **A** is at about +0.6 V and that point **A′** is at about −0.6 V, just the voltages needed to make the output transistors slightly conducting. The resulting small current that flows through the transistors from the $+V$ supply to the $-V$ supply does not have a very well defined magnitude. To control it the resistors R_{2a} and R_{2b} are included. The voltage drop across these increases if the current increases (due for

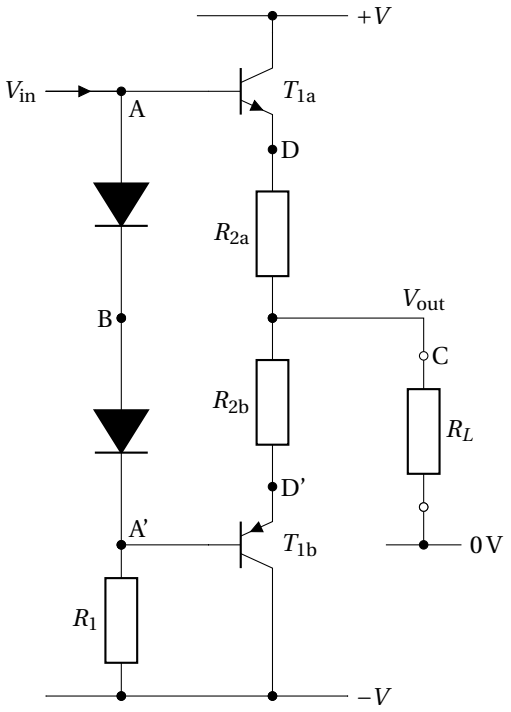


Figure 12.11: Circuit diagram of a compound emitter follower output stage

instance to a temperature increase) and tends to reduce the V_{be} applied to the transistors thereby reducing the increase in the current

When there is an input signal, points **A**, **B**, **A′** vary in voltage with very nearly the same waveform, but with **A** about 0.6 V higher and **A′** about 0.6 V lower in voltage than point **B** at all times. When the voltage at **A** is above the steady bias value, T_{1a} conducts and a current flows in R_L; when the voltage at **A′** is below its steady bias value, T_{1b} conducts and a current flows in R_L in the opposite direction. T_{1a} and T_{1b} therefore conduct on alternate half cycles of the signal, and the full signal waveform appears across R_L. The advantage of this arrangement is that a large current can be taken from the supply and delivered to R_L at peaks of signal (with an efficiency of 78.5% at maximum sine wave output) but unlike class A *the current taken drops to a low value if there is no signal.* This has obvious advantages if batteries are used as the supply. The output stage of the audio amplifier feeding the loudspeaker in a portable radio is invariably of this type.

The compound Class B output stage has advantages in linearity as well as efficiency. If the NPN and PNP device characteristics are identical (except for the reversals of sign) the circuit is symmetrical and like the long tailed pair cannot produce even harmonic distortion. This would be true with the complementary transistors in either the common emitter or emitter follower configurations. In the emitter follower configuration we have been discussing there is a further benefit of reduced distortion as mentioned in section 12.4.1. A disadvantage is that an input voltage of $+V-(-V)$ peak to peak is required to drive the stage to full output.

Note that if the diodes were not present, and the bases of T_{1a} and T_{1b} were simply connected together, point B would have to move at least 0.6 V up or down in voltage before either T_{1a} or T_{1b} would conduct significantly, and severe *cross-over distortion* would result.

If the current gains of the two transistors are the same, the input and output resistances of the stage may be considered to be the same as either half alone (since only one half is conducting at a time).

13 Opamps

13.1 Introduction

An opamp is a chain of several amplifying stages, usually three. Generally we require that the chain has a high input impedance, inverting and non-inverting (but otherwise equivalent) inputs, high voltage gain, adequate frequency response, and the ability to supply power to low resistance loads. Typically the first (input) stage in the chain is a long tailed pair providing the inverting and non-inverting inputs, the second stage is a common emitter providing most of the voltage gain, and the third (power output) stage is a complementary emitter follower.

Opamps used to be made from discrete components with the stages being biased individually and coupled together with dc blocking capacitors. Current practice in both integrated and discrete opamp designs is to couple stages directly to each other so that each biases the next. All that is then required is to set the working point of the whole chain. (This is one of the functions of a *feedback network* which will be dealt with in the next chapter.) The reasons for coupling the stages directly are that the unloaded voltage gain (A) does not fall to zero at 0 Hz (dc), large value capacitors cannot be made on silicon chips anyway, and in the discrete component case the coupling and bypass capacitors and extra resistors needed make the old approach very expensive in component cost and assembly time.

It is not realistic to attempt to study all the possible directly coupled multistage chains, nor is it necessary. The ideas can be revealed by examining a few typical designs.

13.2 Typical opamp configurations

13.2.1 A basic discrete component opamp

A basic opamp with three stages is shown in Figure 13.1.

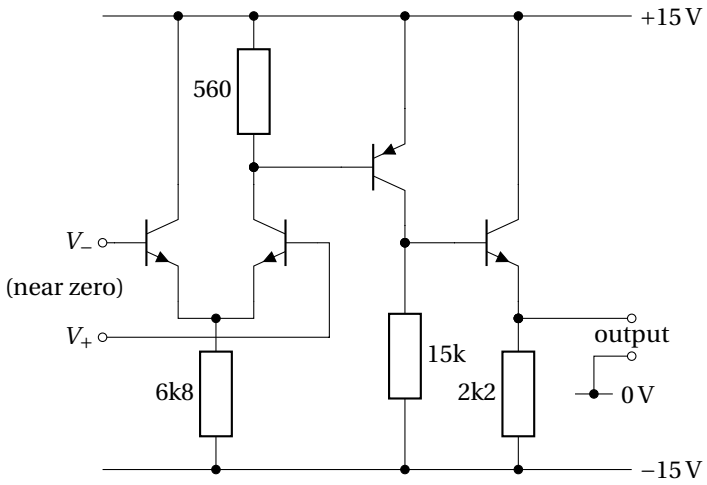


Figure 13.1: Basic discrete component opamp.

It comprises an NPN bipolar long tailed pair, a PNP common emitter stage, and a class A emitter follower output stage. Note that no capacitors and only four resistors are used. Most of the voltage gain comes from the common emitter stage, the gain of the long-tailed pair is only ~ 10 due to the low load resistance and the fact that only one half of its output is used. The live output terminal can be pulled up to near +15 V but the negative half of the working range is limited by the load resistance. E. g. for a load resistance of 2.2 kΩ the live output terminal cannot go below −7.5 V.

13.2.2 An improved design

A more advanced discrete component opamp is shown in Figure 13.2.

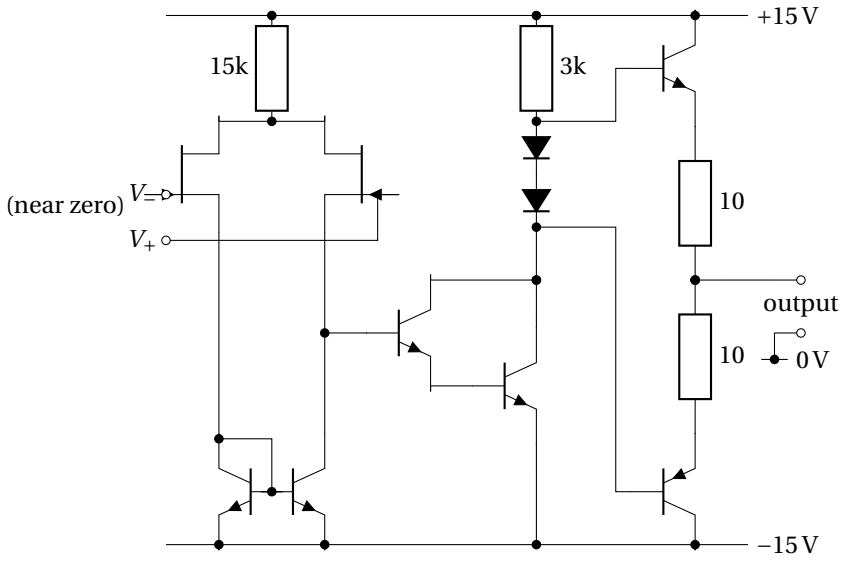


Figure 13.2: Improved opamp design.

The improvements over the previous design are:

(a) The FET long tailed pair input stage with current mirror load gives much higher input resistance and gain.

(b) The Emitter follower added to the second stage inreases its current gain. (This combination, known as a Darlington pair, is effectively an NPN transistor with very high current gain.)

(c) The compound, class B, emitter follower output stage gives greater efficiency and a working range of nearly −15 V to 15 V.

13.2.3 An integrated circuit opamp

The circuit of a Fairchild type 741 chip opamp is shown in Figure 13.3.

It has essentially the same three stages as the circuit in Figure 13.2 but they are much more complicated. The input stage amounts to a PNP long tailed pair with a current mirror load (Q_5, Q_6) giving it a high gain. The emitter follower-common base arrangement (Q_1, Q_3, Q_2, Q_4) is used to obtain a high common mode input voltage range. An emitter follower (Q_{16}) is placed in front of the middle common emitter stage (Q_{17}). The load on Q_{17} is the double collector Q_{13} part of a current mirror system used for bias current control in the preceding stages. (Fairly complex bias control circuits are needed to give the chip a wide operating temperature range.) The class B compound emitter follower output stage is also complex and includes protection against short circuited loads. Dominant lag frequency compensation is imposed by the capacitor C_1. (See section 13.7.)

13.3 Equivalent circuits and circuit diagram symbols

To use an opamp we need an equivalent circuit which shows what happens at its terminals and we need to know the frequency dependence of its unloaded voltage gain A. It is also useful to have a circuit diagram symbol. Typical diagrams, which were introduced in Table 13.1, are repeated in Figure 13.4 and Figure 13.5 below. The input terminals will usually be the bases or gates of an input long-tailed pair.

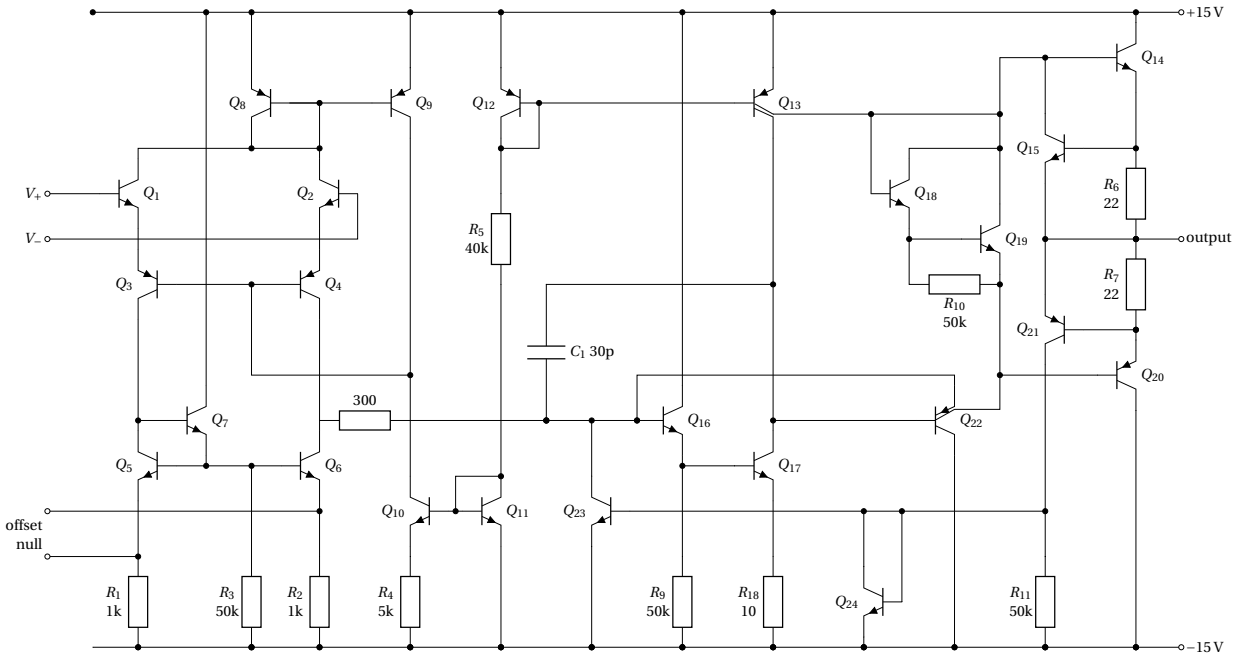


Figure 13.3: Circuit of a Fairchild type 741 chip opamp.

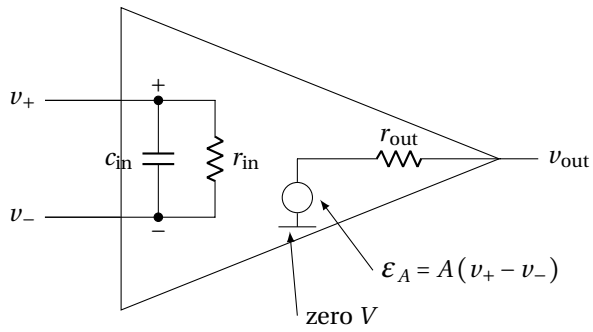


Figure 13.4: Opamp equivalent circuit

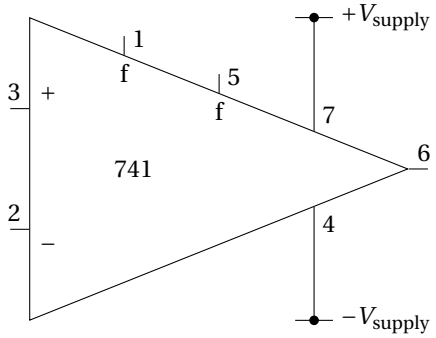


Figure 13.5: Circuit diagram symbol of a type 741 opamp in an 8 pin dual in line package.

The dc power supply connections are usually omitted from equivalent circuits as they rarely feature in the analysis of signal behaviour.

Chip opamps are usually packaged so that we have access only to the input terminals of the first stage, the output terminal of the output stage, the power supply connections and (sometimes) some pins (labelled f f in Figure 13.5) to which we can attach resistors and capacitors to tailor its frequency response (see section 1.7). In the circuit diagram symbol the features inside the triangular outline are omitted but the manufacturers type number, the power supply connections and package pin numbering are often shown.

13.4 An equivalent circuit containing an opamp

Consider the equivalent circuit (of an amplifier) shown in Figure 13.6 (a version of Figure 13.2 with a non-ideal opamp).

The node and branch equations for excitation at angular frequency ω are

$$\begin{aligned}
i_{\text{in}} &= -i_1 + i_2 && (13.1)\\
v_S - v_{\text{in}} &= i_{\text{in}} R_S && (13.2)\\
v_{\text{in}} - v_- &= i_{\text{in}} R_1 && (13.3)\\
v_- - 0 &= -i_1 z_{\text{in}} && (13.4)\\
v_- - v_{\text{out}} &= i_2 z_2 && (13.5)\\
\mathcal{E}_A - v_{\text{out}} &= i_3 r_{\text{out}} && (13.6)\\
v_{\text{out}} &= i_{\text{out}} R_L && (13.7)\\
i_{\text{out}} &= i_2 + i_3 && (13.8)
\end{aligned}$$

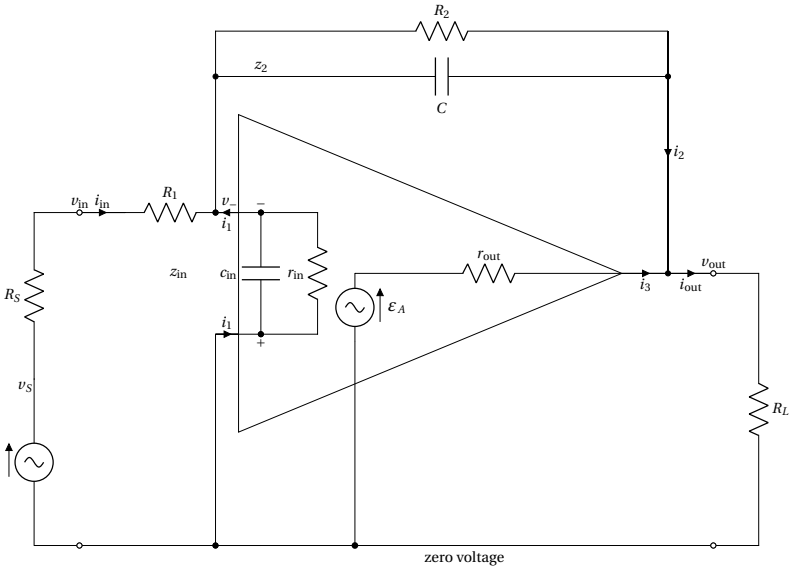


Figure 13.6: Equivalent circuit containing an opamp

and the opamp relation is:

$$\varepsilon_A = A(v_+ - v_-) \tag{13.9}$$

From these equations any ratio of currents or voltages can be derived. For instance, the voltage gain,

$$\frac{v_{\text{out}}}{v_{\text{in}}} = \frac{-\frac{z_2}{R_1} - \frac{v_{\text{out}}}{AR_1}}{1 + \frac{v_{\text{out}}}{AR_1}\left(\left(1 + \frac{R_1}{z_{\text{in}}}\right) + z_2\left(\frac{1}{R_L} + \frac{1}{r_{\text{out}}}\right)\left(1 + \frac{R_1}{z_{\text{in}}} + \frac{R_1}{z_2}\right)\right)} \tag{13.10}$$

where:

$$\frac{1}{z_2} = \frac{1}{R_2} + j\omega C \tag{13.11}$$

and:

$$\frac{1}{z_{\text{in}}} = \frac{1}{r_{\text{in}}} + j\omega C_{\text{in}} \tag{13.12}$$

The special case of $R_1 = R_2$, $C = 0$, and A large gives $v_{\text{out}}/v_{\text{in}}$ close to -1. The circuit is then called an *inverting buffer*.

The derivation of the input impedance $v_{\text{in}}/i_{\text{in}}$ is left as an exercise; if A is large, it tends to R_1.

13.5 Ideal opamps

In the example given above we see that the larger r_{in} and A and the smaller c_{in} and r_{out} the smaller their effect on $v_{\text{out}}/v_{\text{in}}$ (and, it turns out, on any of the ratios). Now a competent designer will choose an opamp whose guaranteed worst values of r_{in} and A etc. (see section 13.9 for examples) are such that the performance of the circuit is not affected by more than is allowed by the specification it must meet. (Equivalent to making the second terms in the numerator and denominator of $v_{\text{out}}/v_{\text{in}}$ negligible.) Therefore, when we work out for ourselves the behaviour of a circuit which we believe has been well designed *we may, initially at least, consider the opamp as having* r_{in} *and* A *infinite and* c_{in} *and* r_{out} *zero* as this reduces the amount of algebra considerably. Perhaps of more interest is the fact that examiners often ask for opamps to be considered in this way. These properties define the concept of an *ideal opamp* introduced in Chapter 1. Note that it depends on the circumstances; a type 741 which can be considered ideal in one circuit with one specification may well be non-ideal in another.

13.6 Power supplies and working range

By definition an amplifying device is one capable of delivering more signal power to a load than is absorbed by its input. To avoid violating a well trusted law an amount of power equal to at least this difference must be provided by another source. An opamp amplifies by converting power from a dc supply into signal power.

The dc power usually comes from a well smoothed dual supply, typically ±15 volts (see section 15.5) with such low impedance outputs that any signal currents flowing into them cause negligible voltage variations. This is the reason why power supplies rarely feature in the analysis of signal behaviour.

When applying a signal to an entirely passive circuit there is seldom any concern over it being large enough to cause non linear behaviour (except perhaps if the circuit contains inductors with ferromagnetic cores). In contrast, the voltage at the output of an opamp is restricted to a *working range* slightly narrower than that bounded by the power supply voltages (except in some cases where the load is inductive). Too large a signal at the input terminals of an opamp will produce an output that is limited or clipped, see the typical opamp transfer characteristic shown in Figure 13.7.

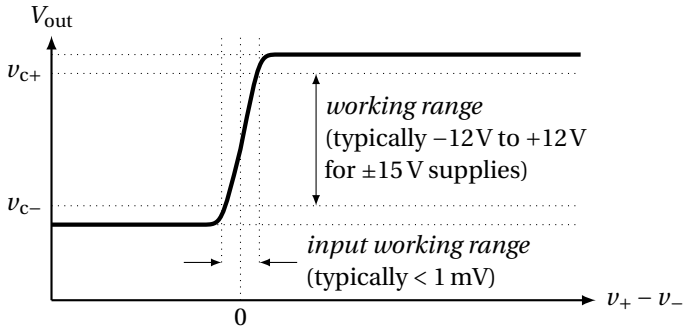


Figure 13.7: Opamp transfer characteristic

13.7 Discussion of the frequency dependence of A

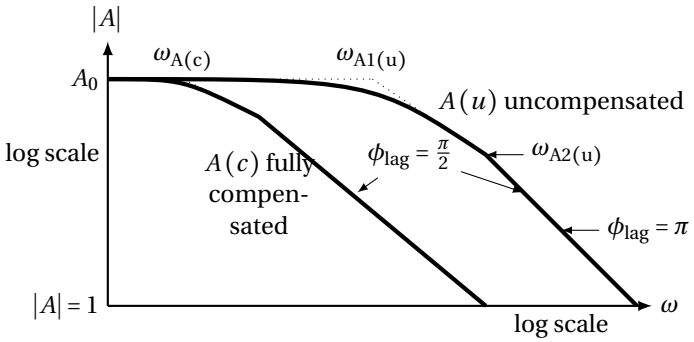


Figure 13.8: Dependence of gain voltage A on frequency ω

If the open loop voltage gain A of an opamp has not been deliberately reduced at high frequencies it will look something like the outer curve $A_{(u)}$ in Figure 13.8. By adding appropriate RC networks to the opamp circuit a frequency response like the inner curve A(c), described as 'fully compensated' or having a 'dominant lag', can be achieved. $A_{(c)}$ is given by:

$$\frac{A_0}{1 + j\frac{\omega}{\omega_{A_{(c)}}}} \tag{13.13}$$

Above $\omega_{A_{(c)}}$ the product of $|A|$ and the frequency is constant and is called the *gain–bandwidth* product of the opamp. Many opamps are manufactured with this full compensation built in (and not accessible for modification). An example of a fully compensated opamp is the type 741, $|A|$ falls by a factor of 10 for each decade increase in frequency above 10 Hz $\left(\omega_{A_{(c)}}/2\pi\right)$. Its gain–bandwidth product is 10^6 Hz.

On the outer curve the segments of steepening slope are due to high frequency effects in the transistors in the output stage and capacitances in earlier stages.

There are phase shifts associated with the falling segments of the curves. A segment falling by a factor of ten per ten times increase in frequency (the expression for $A_{(c)}$ above when $\omega \gg \omega_{A_{(c)}}$) is associated with a phase shift of 90°, a segment falling by a factor of 100 per decade in frequency e.g. $A_{(u)}$ above $\omega_{A_{2(u)}}$ (due to two time constants) is associated with a phase shift of 180° i.e. a complete reversal of phase.

13.8 DC considerations

13.8.1 Biasing

The stages in an opamp need dc bias currents flowing through them to make them active. The first step towards providing these is to connect dc voltages (typically ±15 V) to the supply pins which feed all the stages in the chip. To complete the definition of the quiescent (no signal) state of the opamp it is necessary to define the dc voltages of the input pins (which are the input terminals of the first stage) and feed some current to them (or sink some from them depending on the type of transistors in the input stage).

13.8.2 Input offsets

If the input pins of a real chip opamp are connected together we do not find zero voltage on the output pin. In fact the output voltage is likely to be found hard against one end of the working range. This effect, which is due to the high value of A and the difficulty of achieving precise balance in the input stage in manufacture, can be represented by a small dc voltage in series with one of the input terminals and is known as the *input offset voltage* of the opamp. Depending on the opamp design and technology its value may lie in the tenths of μV or mV ranges. Manufacturers specify an upper limit for the magnitude of the input offset voltage but do not specify its sign.

The bias currents taken by the two inputs will not be precisely the same, the difference, the *input offset current*, again having its maximum magnitude but not its sign specified. (Input bias currents are usually small, lying in the μA or pA range depending on the type of transistors used in the input stage of the opamp but the offset can be a significant fraction of the mean value.)

These offsets (δV_{in} and δI_{in}) and the input bias current I_{in} are shown in Figure 13.9. This is the equivalent circuit to use in calculations of dc offsets.

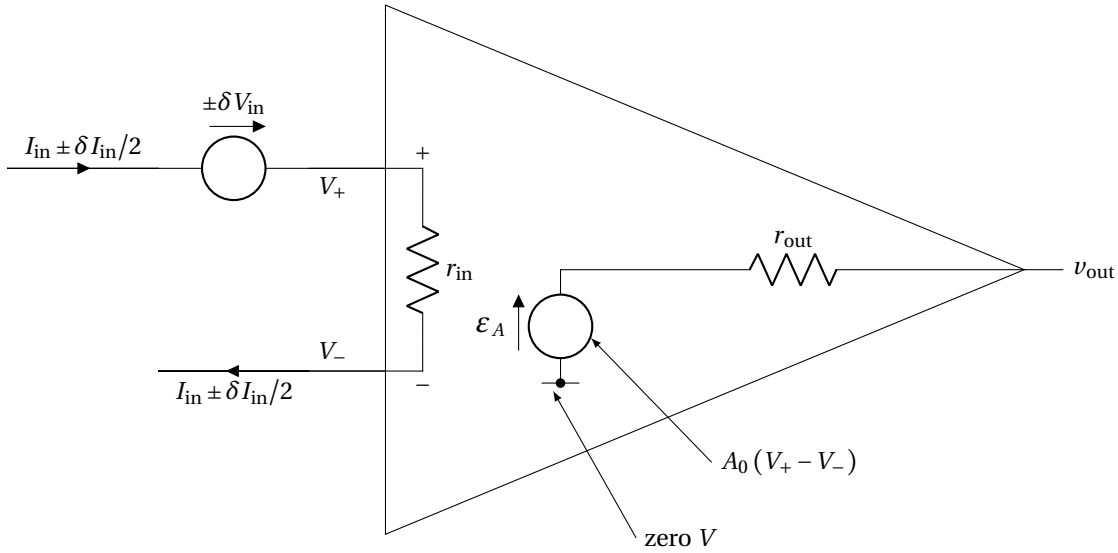


Figure 13.9: Equivalent circuit for calculations of input offsets

13.9 Properties of two opamps

In order to design circuits that will perform within specification a designer needs to know the guaranteed worst case properties of the components he/she plans to use. Data for two opamps, operating at room temperature and with ±15 volt power supplies, are given in the tables below. For some properties the worst case is a maximum, for some it is a minimum.

	min	**max**	**units**	**notes**
$\|A\|$ (l.f.)	2×10^3			(1)
gain-bandwidth product	50		MHz	(2)
r_{in}	0.3		MΩ	
r_{out}		200	Ω	
input offset voltage		1	mV	
input bias current		5	μA	
input offset current		300	nA	

(1) and (2) indicate fall in $|A|$ starts at 25 kHz.

Table 13.1: Type AD847 (bipolar, fully compensated)

	min	**max**	**units**	**notes**
$\|A\|$ (l.f.)	5×10^4			(1)
gain-bandwidth product	3		MHz	(2)
r_{in}	10^5		MΩ	
r_{out}		150	Ω	
input offset voltage		6	mV	
input bias current		200	pA	
input offset current		100	pA	

(1) and (2) indicate fall in $|A|$ starts at 60 Hz.

Table 13.2: Type TL081 (FET input, fully compensated)

14 Amplifier Design and Negative Feedback

14.1 Introduction

In this chapter we begin to describe how linear circuits are designed. Designing circuits is a more interesting task than simply analysing those produced by other people. We focus on linear amplifiers using commercial opamps, not only because they provide good examples of the design procedure, but because they are of fundamental importance in analogue electronics.

14.2 The designer's task

An amplifier takes a signal from a source and delivers a related, higher power signal to a load. Amplifiers can generally be classified according to which ratio of output and input quantities is of most interest. There are four such ratios:

(a) v_{out}/v_S, the voltage gain

(b) i_{out}/v_S, the transconductance gain

(c) v_{out}/i_S, the transresistance gain

(d) i_{out}/i_S, the current gain

The performance required of an amplifier will have been set down in a specification. This will state the gain required and its precision and the range of source and load resistances and the range of frequencies (or alternatively the permissible rise time and overshoot for a step input) over which this precision must be maintained. The designer's task is to ensure that the specified quantities remain within their specified limits even when the properties of the components used all happen to take their worst case values.

14.3 Negative feedback and the basic amplifiers

In all the amplifier circuits we have examined (and in fact in all other practical linear amplifier circuits) there is a passive component network which connects the output of the amplifier back to the input of the opamp in some way. We say that the circuits employ *feedback*. The combination of the forward signal path through the opamp and the return path through the network is called a *feedback loop*.

Within the range of frequencies over which the performance of the amplifier is specified, the feedback is usually such as to reduce the voltage or current at the opamp input pins below the levels that would obtain if the feedback were not present. Such feedback is said to be *negative*. Negative feedback is a central concept in the design of linear amplifiers. It enables them to be designed with predictable, precise and stable performance despite the fact that opamp properties are specified rather loosely (see section 13.9) and are dependent on supply voltage, temperature and signal level.

There are four ways in which negative feedback can be applied corresponding to the four types of amplifier mentioned in section 14.2 above. (In some situations combinations of the four methods are employed but this does not involve any new ideas and we shall not discuss them here.) *The feedback signal can be proportional to either the output voltage or the output current of the amplifier and can be applied either as a voltage in series with or a current in parallel with its input.* We have assembled the basic information about the four feedback configurations and a bare opamp in the columns of Table 14.1. A bare opamp is included because it enables the effects of the feedback to be clearly distinguished from the separate effects of the finite input and output resistances and finite A of the opamp. (We emphasize that a bare opamp is introduced for comparative purposes only: an opamp by itself does not constitute a practical linear amplifier.)

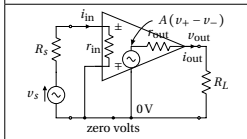
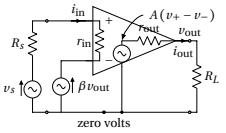
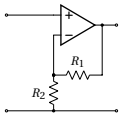
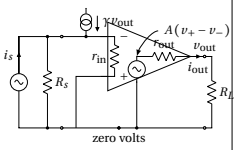
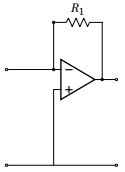
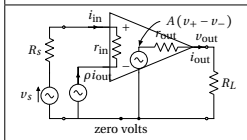
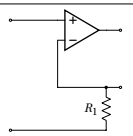
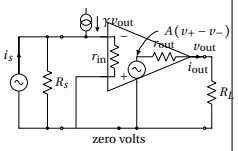
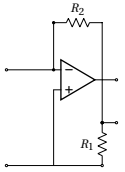
Essential configuration and type of feedback	**Voltage gain G_V** $\frac{v_{\text{out}}}{v_s}$	**Transresistance gain G_R** $\frac{v_{\text{out}}}{i_s}$	**Transconductance gain G_G** $\frac{i_{\text{out}}}{v_s}$	**Current gain G_I** $\frac{i_{\text{out}}}{i_s}$	**Principal gain and name of amplifier** [inverting or non-inverting]	**Input resistance R_{in}** $\frac{v_{\text{in}}}{i_{\text{in}}}$	**Output resistance R_{out}** $\frac{v_{\text{out}}}{i_{\text{out}}}$	**Practical version**	**Approximate feedback factor**
 No feedback	$\pm\frac{AR_L}{R_L+r_{\text{out}}}\frac{1}{1+\frac{R_s}{r_{\text{in}}}}$	$\pm\frac{AR_L}{R_L+r_{\text{out}}}\frac{R_s}{1+\frac{R_s}{r_{\text{in}}}}$	$\pm\frac{A}{R_L+r_{\text{out}}}\frac{1}{1+\frac{R_s}{r_{\text{in}}}}$	$\pm\frac{A}{R_L+r_{\text{out}}}\frac{R_s}{1+\frac{R_s}{r_{\text{in}}}}$	– [either]	r_{in}	r_{out}	*This is not a practical circuit*	–
 Voltage-series feedback feedback factor β	$\frac{AR_L}{R_L+r_{\text{out}}}\frac{1}{1+\frac{\beta AR_L}{R_L+r_{\text{out}}}+\frac{R_s}{r_{\text{in}}}}$ $\left(\frac{1}{\beta}\right)$	$R_s\frac{v_{\text{out}}}{v_s}$	$\frac{1}{R_L}\frac{v_{\text{out}}}{v_s}$	$\frac{R_s}{R_L}\frac{v_{\text{out}}}{v_s}$	Voltage [non-inverting]	$r_{\text{in}}\left(1+\frac{\beta AR_L}{R_L+r_{\text{out}}}\right)$ ($\gg r_{\text{in}}$)	$\frac{r_{\text{out}}}{1+\frac{\beta Ar_{\text{in}}}{R_s+r_{\text{in}}}}$ ($\ll r_{\text{out}}$)		$\beta=\frac{R_2}{R_1+R_2}$
 Voltage-shunt feedback feedback factor γ	$\frac{1}{R_s}\frac{v_{\text{out}}}{i_s}$	$\frac{-AR_L}{R_L+r_{\text{out}}}\frac{R_s}{1+\frac{\gamma AR_sR_L}{R_L+r_{\text{out}}}+\frac{R_s}{r_{\text{in}}}}$ $\left(-\frac{1}{\gamma}\right)$	$\frac{1}{R_LR_s}\frac{v_{\text{out}}}{i_s}$	$\frac{1}{R_L}\frac{v_{\text{out}}}{i_s}$	Transresistance [inverting]	$\frac{r_{\text{in}}}{1+\frac{\gamma Ar_{\text{in}}R_L}{R_L+r_{\text{out}}}}$ ($\ll r_{\text{in}}$)	$\frac{r_{\text{out}}}{1+\frac{\gamma Ar_{\text{in}}R_s}{R_s+r_{\text{in}}}}$ ($\gg r_{\text{out}}$)		$\gamma=\frac{1}{R_1}$
 Current-series feedback feedback factor ρ	$\frac{AR_L}{R_L+r_{\text{out}}}\frac{1}{1+\frac{\beta AR_L}{R_L+r_{\text{out}}}+\frac{R_s}{r_{\text{in}}}}$	$R_s\frac{v_{\text{out}}}{v_s}$	$\frac{1}{R_L}\frac{v_{\text{out}}}{v_s}$ $\left(\frac{1}{\rho}\right)$	$\frac{R_s}{R_L}\frac{v_{\text{out}}}{v_s}$	Transconductance [non-inverting]	$r_{\text{in}}\left(1+\frac{\rho A}{R_L+r_{\text{out}}}\right)$ ($\gg r_{\text{in}}$)	$r_{\text{out}}\left(1+\frac{\rho Ar_{\text{in}}}{r_{\text{out}}(R_s+r_{\text{in}})}\right)$ ($\ll r_{\text{out}}$)		$\rho=R_1$
 Current-shunt feedback feedback factor α	$\frac{1}{R_s}\frac{v_{\text{out}}}{i_s}$	$\frac{-AR_L}{R_L+r_{\text{out}}}\frac{R_s}{1+\frac{\gamma AR_sR_L}{R_L+r_{\text{out}}}+\frac{R_s}{r_{\text{in}}}}$	$\frac{1}{R_LR_s}\frac{v_{\text{out}}}{i_s}$	$\frac{1}{R_L}\frac{v_{\text{out}}}{i_s}$ $\left(-\frac{1}{\alpha}\right)$	Current [inverting]	$\frac{r_{\text{in}}}{1+\frac{\alpha Ar_{\text{in}}}{R_L+r_{\text{out}}}}$ ($\ll r_{\text{in}}$)	$r_{\text{out}}\left(1+\frac{\alpha AR_sr_{\text{in}}}{r_{\text{out}}(R_s+r_{\text{in}})}\right)$ ($\gg r_{\text{out}}$)		$\alpha=\frac{R_1}{R_2}$

Table 14.1: Feedback configurations

In the table the amplifiers are shown in their essentials with the feedback represented by generators. The input and output terminals of the amplifiers are represented as usual by small circles making it clear that the sources and loads are not parts of the amplifiers. The quantities α, β, γ and ρ are called *feedback factors*, the first two are dimensionless the second two have dimensions siemens (ohms^{-1}) and ohms respectively. In practice they are realized (approximately) by the networks shown in the ninth column of the table.

We have shown voltage or current generators in the signal sources as appropriate for the type of amplifier. These are of course related by:

$$v_S = i_S R_S \tag{14.1}$$

(the *generator equivalence relation*). We also have the load relation:

$$v_{\text{out}} = i_{\text{out}} R_L. \tag{14.2}$$

These relations make it easy to write down the remaining three gain expressions for an amplifier once one has been derived from the node and branch equations.

The input resistance is the resistance seen by removing the source and looking into the input terminals with the load connected. The output resistance is that seen looking into the output terminals with the load removed and the source connected. The expressions for the input and output resistances for the four configurations are shown in the right-hand columns of the table.

The table exhibits some interesting symmetries and repays some study. When, as is usual in practice, the amount of feedback applied is large, (by which we mean that the *loop gains* βA, $\gamma A r_{\text{in}}$, $\rho A/R_L$ and αA are all $\gg 1$) all the expressions simplify but *one of the gain expressions for each configuration simplifies right down to the reciprocal of its feedback factor*. We call this gain which is least affected by other things the principal gain and use it as the name of the amplifier. We have put boxes around the four principal gains in the table. Values for large loop gain are shown in brackets.

Notice that the feedback makes the output resistance of each amplifier dependent on R_S, the resistance of the source, and the input resistance dependent on the load resistance R_L. Notice the pattern of raising and lowering of these resistances by the feedback compared with the bare opamp values.

The gains in the table relate the output quantities to the source generators. The gains referred to the input voltages and currents at the amplifier terminals may be derived from these by setting $v_S = v_{\text{in}}$ when $R_S = 0$ and $i_S = i_{\text{in}}$ when $R_S = \infty$.

14.4 Principal gains and resistances of the standard amplifiers

We give below for reference the results of analysis of the practical versions of the four basic amplifiers shown in Table 14.1. If the input signal is at a single frequency the expressions may be generalized by substituting impedances for resistances. The expressions are quite complicated but with practical values of the variables inserted many of the terms they contain, particularly those involving $1/A$, will be small.

14.4.1 Voltage amplifier

The equivalent circuit is shown in Figure 14.1. The feedback signal is the voltage developed across R_2 by i_1 applied in series with the input. The branch equations are:

$$\begin{aligned}
v_{\text{in}} - 0 &= i_{\text{in}}\left(r_{\text{in}} + R_2\right) + i_1 R_2 && (14.3)\\
v_{\text{out}} - 0 &= i_1 R_1 + \left(i_1 + i_{\text{in}}\right) R_2 && (14.4)\\
v_{\text{out}} - 0 &= i_{\text{out}} R_L && (14.5)\\
A\left(v_{\text{in}} - \left(i_1 + i_{\text{in}}\right) R_2\right) - v_{\text{out}} &= \left(i_1 + i_{\text{out}}\right) r_{\text{out}} && (14.6)\\
v_S - v_{\text{in}} &= i_{\text{in}} R_S && (14.7)
\end{aligned}$$

We have taken care of the node equations by the current labelling on the diagram. Eliminating i_1, i_{out}, i_{in} and v_{in} we find (after some algebra) the voltage gain:

$$\frac{v_{\text{out}}}{v_{\text{S}}} = \frac{\frac{R_1+R_2}{R_2} + \frac{r_{\text{out}}}{A r_{\text{in}}}}{1 + \frac{r_{\text{out}} R_2}{(R_1+R_2) A r_{\text{in}}} + \frac{1}{A r_{\text{in}}}\left(\frac{(R_1+R_2)(r_{\text{in}}+R_S)}{R_2} + R_1\right)\left(1 + \frac{r_{\text{out}}}{R_L} + \frac{r_{\text{out}}}{R_1+R_2}\right)} \tag{14.8}$$

The special case of $R_1 = 0$, $R_2 = \infty$ gives $v_{\text{out}}/v_s \approx +1$ (unity gain buffer or voltage follower). We find R_{in} by manipulating the gain expression to put the denominator into the form $M\left\{R_s + [\ ]\right\}$ where M does not contain

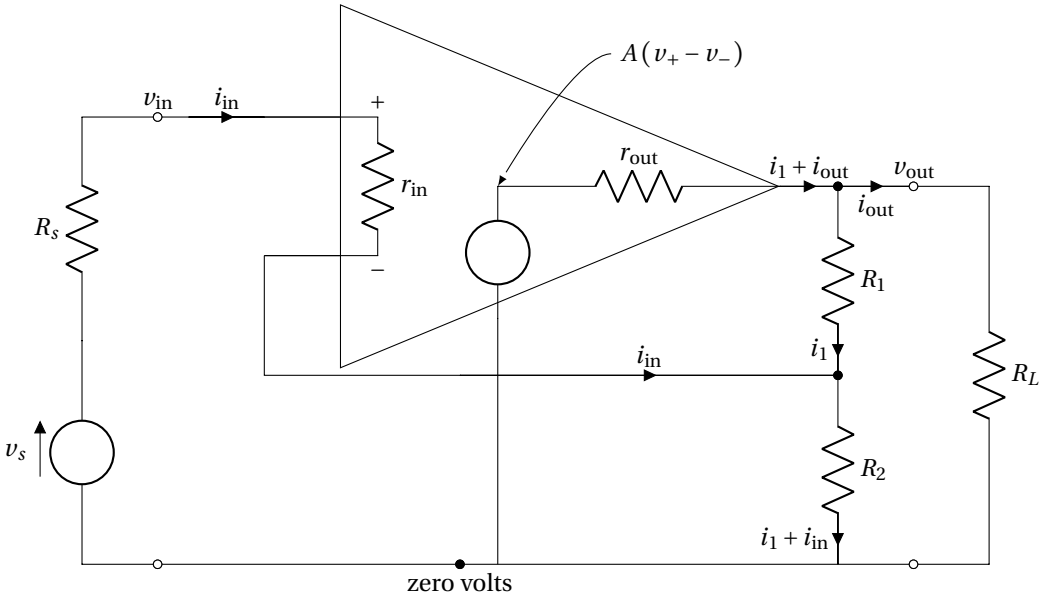


Figure 14.1: Voltage amplifier equivalent circuit

R_s. The quantity in the square bracket is then R_{in}. We find R_{out} in the same way by expressing the denominator as $N\{R_L + [\ ]\}$. The results are:

$$\frac{v_{\text{out}}}{i_{\text{in}}} = \frac{r_{\text{in}}\left(A\frac{R_2}{R_1+R_2}\frac{R_L}{R_L+r_{\text{out}}} + \frac{r_{\text{out}}R_L}{(R_1+R_2)(R_L+r_{\text{out}})} + 1\right) + \frac{R_1R_2}{R_1+R_2}}{1 + \frac{r_{\text{out}}}{R_L+r_{\text{out}}}\frac{R_2}{R_1+R_2}} \tag{14.9}$$

$$R_{\text{out}} = r_{\text{out}}\frac{1 + \frac{R_1}{R_S}\frac{R_2}{R_1+R_2}\frac{R_S}{R_S+r_{\text{in}}}}{1 + \frac{R_S}{R_S+r_{\text{in}}}\frac{R_2}{R_1+R_2}\left(\frac{Ar_{\text{in}}}{R_S} + \frac{r_{\text{out}}}{R_2} + \frac{r_{\text{out}}}{R_S}\left(\frac{r_{\text{in}}}{R_2}+1\right)\right) + \frac{R_1}{R_S}\frac{R_2}{R_1+R_2}\frac{R_S}{R_S+r_{\text{in}}}} \tag{14.10}$$

14.4.2 Transresistance amplifier

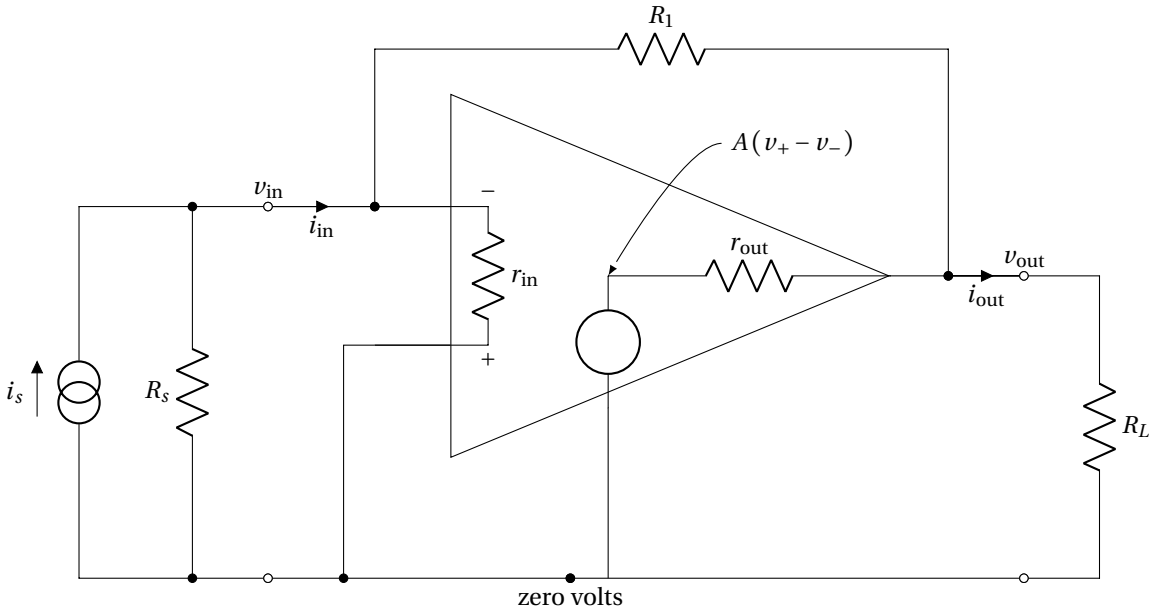


Figure 14.2: Transresistance amplifier equivalent circuit

$$\frac{v_{\text{out}}}{i_S} = \frac{-R_1 + \frac{r_{\text{out}}}{A}}{1 + \frac{1}{A}\left(\left(1 + \frac{r_{\text{out}}}{R_L}\right)\left(1 + R_1\left(\frac{1}{r_{\text{in}}} + \frac{1}{R_S}\right)\right) + r_{\text{out}}\left(\frac{1}{r_{\text{in}}} + \frac{1}{R_S}\right)\right)} \tag{14.11}$$

$$R_{\text{out}} = \frac{r_{\text{out}}}{1 + \frac{AR_S r_{\text{in}} + r_{\text{out}}(R_S + r_{\text{in}})}{R_1(R_S + r_{\text{in}}) + R_S r_{\text{in}}}} \tag{14.12}$$

$$R_{\text{in}} = \frac{r_{\text{in}}}{1 + \frac{\frac{1}{R_1}\left((A+1)R_L + r_{\text{out}}\right)r_{\text{in}}}{R_L + r_{\text{out}} + \frac{r_{\text{out}} R_L}{R_1}}} \tag{14.13}$$

14.4.3 Transconductance amplifier

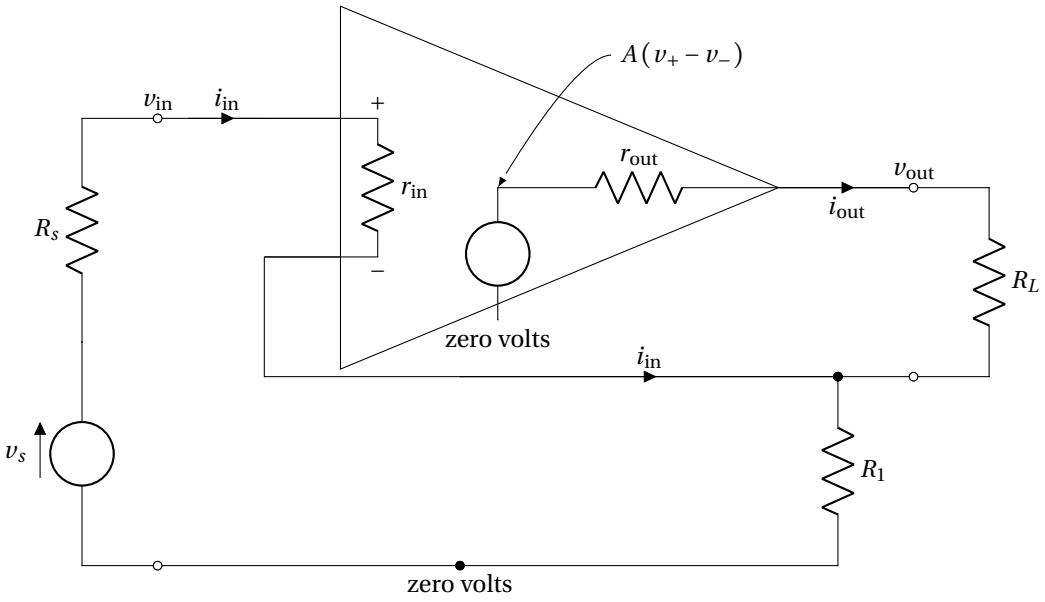


Figure 14.3: Transconductance amplifier circuit

$$\frac{i_{\text{out}}}{v_S} = \frac{\frac{1}{R_1} - \frac{1}{Ar_{\text{in}}}}{1 + \frac{1}{A}\left(\frac{r_{\text{out}} + R_L}{R_1}\left(1 + \frac{R_S}{r_{\text{in}}}\right) + \frac{r_{\text{out}} + R_L}{r_{\text{in}}} + \left(1 + \frac{R_S}{r_{\text{in}}}\right)\right)} \tag{14.14}$$

$$R_{\text{out}} = r_{\text{out}}\left(1 + \frac{R_1\left((A+1)\,r_{\text{in}} + R_S\right)}{R_S + r_{\text{in}} + R_1}\right) \tag{14.15}$$

$$R_{\text{in}} = r_{\text{in}}\left(1 + \frac{R_1\left(Ar_{\text{in}} + R_L + r_{\text{out}}\right)}{r_{\text{in}}\left(R_L + r_{\text{out}} + R_1\right)}\right) \tag{14.16}$$

14.4.4 Current amplifier

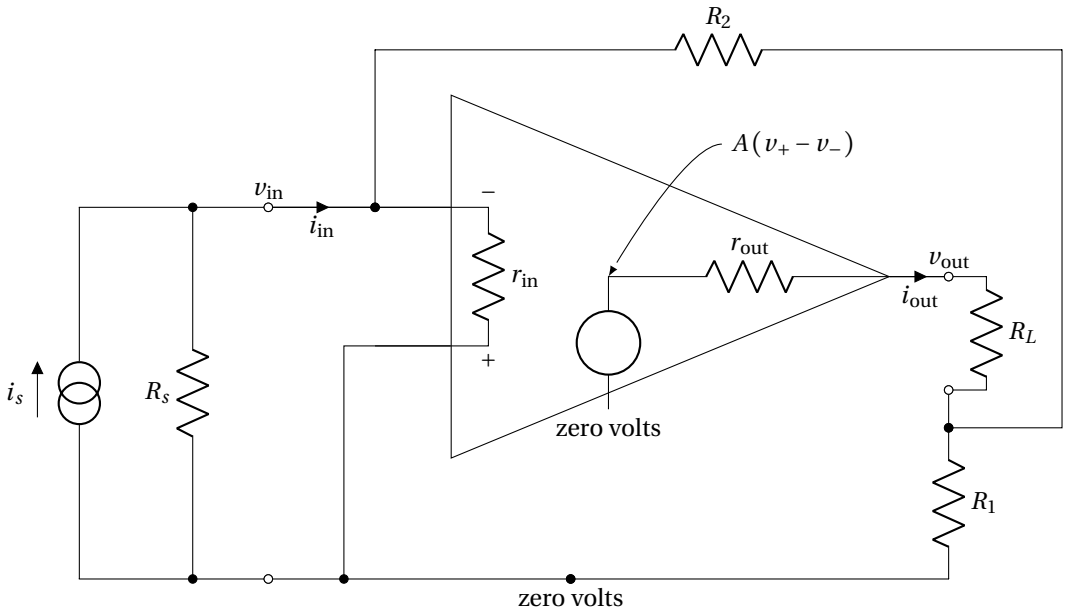


Figure 14.4: Current amplifier circuit

$$\frac{i_{out}}{i_{in}} = \frac{-AR_S + \frac{R_1}{R_1+R_2}}{(R_L + r_{out})\left(1 + \frac{\frac{R_1}{R_2}(1+A)R_S}{\left(1+\frac{R_1}{R_2}\right)(R_L+r_{out})} + R_S\left(\frac{1}{R_1+R_2} + \frac{1}{r_{in}}\right)\right)} \tag{14.17}$$

$$R_{out} = r_{out}\left(1 + \frac{R_S r_{in}(1+A)\frac{R_1}{R_2} + R_1(R_S + r_{in})}{r_{out}\left((R_S + r_{in})\left(1+\frac{R_1}{R_2}\right) + \frac{R_S r_{in}}{R_2}\right)}\right) \tag{14.18}$$

$$R_{in} = \frac{r_{in}}{1 + \frac{\frac{R_1}{R_2}(1+A)r_{in}\frac{1}{R_L+r_{out}} + \frac{r_{in}}{R_2}}{1+\frac{R_1}{R_2} + \frac{R_1}{R_2+r_{out}}}} \tag{14.19}$$

14.5 Non-standard configurations

Specifications which cannot be met by employing one of the standard amplifiers can usually be satisfied by using slightly modified versions. For example, suppose we need an inverting amplifier with a high input resistance and a low output resistance. None of the standard amplifiers has this combination of properties but such an amplifier can be built by using a standard transresistance configuration to obtain the inversion and the low output resistance and adding a resistance in series with the input to give the desired input resistance. (An opamp with low input current will usually be needed.) The result is the so called 'see saw' amplifier. We already have a figure and the analysis of this as it was used as an example in section 13.4.

14.6 Defining the passband or the pulse response required

What we have said so far has involved resistances only and, if we ignore any frequency dependence of the magnitude or phase shift of A, all the amplifier properties derived apply whatever the time dependence of the input signal.

14.6.1 Frequency response

Amplifier specifications written in frequency domain language will include requirements on the variation of the magnitude and possibly also the phase of the gain with frequency. Tailoring the frequency response is straightforward if it simply involves changing some or all of the single resistance branches in the feedback network into branches containing combinations of resistances and capacitances (inductances are generally avoided) because we can use the resistance expression as a starting point. If a more complicated feedback network is required a new gain expression will need to be worked out.

14.6.2 Pulse responses

Linear amplifiers used for amplifying pulses are usually specified in terms of the permissible rise-time and overshoot in the response to a step input. A desired time domain response may be obtained in the following way:

(a) construct it from the time domain functions listed in Figure 4.1

(b) write down its transform

(c) multiply the transform by $1/s$

(d) replace s by $j\omega$

(e) find a configuration with this frequency domain response

The design of linear pulse amplifiers is a rather specialized business. It is of particular interest to nuclear experimenters for conditioning the signals from particle detectors.

14.7 Stability of amplifiers

We choose the voltage amplifier to illustrate the ideas. Taking the gain expression from the second box in the second row of Table 14.1, modifying it to be $v_{\text{out}}/v_{\text{in}}$ (see end of section 14.3), and generalising to impedances (see section 14.4) we have:

$$\frac{v_{\text{out}}}{v_{\text{in}}} = \frac{AZ_L}{Z_L + z_{\text{out}}} \frac{1}{1 + \frac{AZ_L\beta}{Z_L + z_{\text{out}}}} \tag{14.20}$$

The quantity $\frac{AZ_L\beta}{Z_L+z_{\text{out}}}$, called teh *loop gain*, is the product of the gain of the forward paths through the opamp $\frac{AZ_L}{Z_L+z_{\text{out}}}$ and the transmittance of the feedback path β.

We consider first the case when Z_L and z_{out} are resistances and β is the transmittance of a potential divider constructed from resistances so the only phase in the loop gain is due to A. At frequencies below the first corner in the A of the opamp (see Figure 13.8) the loop gain will be real, usually large and positive. If the opamp is fully compensated the phase lag in the loop approaches but cannot exceed 90°. Therefore at high frequencies the falling loop gain is pure imaginary, it cannot approach -1 and the amplifier is stable. (This certainty of stability is the reason why fully compensated opamps are manufactured despite the drastic lowering of the high frequency gain such compensation entails.)

If the opamp is uncompensated and has an A described by say the outer curve in Figure 13.8 the phase lag in the loop at frequencies above the second corner approaches 180°. This is dangerous because the loop gain then becomes real and negative. If it reaches -1 the gain of the apmplifier becomes infinite and the circuit becomes an oscillator. One might then ask, why do manufacturers make uncompensated opamps? The reason is to give designers more flexibility. The magnitude of the loop gain depends on β as well as A, the highest value of β used (corresponding to a voltage amplifier gain of unity) requiring the severest compensation. For smaller values of β (higher amplifier gains) less than full compensation may be required to achieve satisfactory stability and by preserving more of the high frequency gain of the opamp better performance can be achieved. A typical stability criterion is that the loop phase lag at the frequency at which the magnitude of the loop gain has to fallen to unity should be less than 135°.

Thus the designer needs to control the magnitude and phase of the loop gain all the way up to the frequency (usually well above the high frequency edge of the amplifiers passband) at which its magnitude has fallen below unity. He/she does this by tailoring the feedback network e.g. by adding capacitors to give phase leads, and applying the minimum acceptable compensation to the opamp. Uncompensated opamps have additional connections to their circuits brought out to pins for this purpose (see Figure 13.5).

The situation is more complicated when Z_L, z_{out} and β are complex. However no new ideas are involved so we will not take the discussion further except to remark that shunt capacitance in Z_L can be a problem.

14.8 Significance of the gain–bandwidth product

Suppose that a specification calls for a voltage amplifier with a gain of $10 \pm 2\%$ from a few Hz to 20 kHz. This frequency range is called the passband (or bandpass!) of the amplifier. Now we know from Table 14.1 that to make the expression for the voltage gain of a voltage amplifier simplify to $1/\beta$, i.e. become least dependent on other factors, we must have $A\beta$ large, 100 might be needed to achieve the precision required in the specification. In our example $\beta = \frac{1}{10}$ which means that $|A|$ must be at least 10^3 up to 20 kHz. In other words the opamp must have a gain bandwidth product of 2×10^7. Clearly the 741 is not good enough for this, its gain bandwidth product is only 10^6 ($|A|$ has fallen to 50 at 20 kHz). We must choose an opamp with a larger gain bandwidth product, perhaps an uncompensated one.

14.9 Consequences of biasing requirements and input offsets of opamps for amplifier design

The provision of the + and − power supply connections to the opamp chip does not affect the amplifier circuit design but the need to define the quiescent voltages of the input pins of the opamp and supply some current to them does place some constraints on the design of the feedback network. The input currents and the offsets mentioned in section 13.8 cause a voltage offset at the output of an amplifier which depends on the network and the component values.

Let us first consider the effect of the input offset voltage. In an amplifier whose specification allows the gain to fall to unity or below at dc (often referred to as an ac amplifier) the feedback network can be designed to ensure that the amplification of the offset voltage is small (more feedback at dc) which reduces its importance. In a dc amplifier (one whose gain is maintained down to 0 Hz) there is no alternative to using an opamp whose offset voltage is suitably low (or can be trimmed).

Now consider the input bias currents, if they are equal and there are equal resistance paths back to a common dc voltage, then provided the voltage drops in the paths are not so great as to carry the opamp outside its allowed common mode input voltage range all will be well. If the resistances are different and/or there is an offset current a voltage difference will arise at the input terminals which will have an effect similar to that of the input offset voltage.

14.10 Power output

Suppose that using an amplifier fed from ±15 volt supplies we wish to deliver a sine wave signal with a power of 30 mW to a 15 Ω load. This corresponds to a voltage of 0.67 V rms across the load and a current of 45 mA rms. Currents of this order would be taken from the power supplies and the voltage working range would be nothing like filled.

If we were concerned about the current drain on the supply (as we would be if batteries were being used) we would prefer the amplifier to produce the power required with the maximum voltage swing (i.e. just within the working range) and the corresponding minimum current swing.

We might aim for values such as 24 V p-p (which is 8.5 V rms) and 3.5 mA rms which correspond to 30 mW being dissipated in a 2.4 kΩ load. Use of a transformer (as we saw in section 3.9.3) enables the 15 Ω load to be converted into a 2.4 kΩ load on the amplifier.

15 Nonlinear Circuits

15.1 Introduction

15.1.1 Linear circuits

The fact that the waveform of a signal is distorted after it has passed through a circuit is not a reliable indicator that the circuit is nonlinear. For instance, a square wave becomes distorted if it is passed through a *linear* circuit in which the magnitude and/or phase shift of the transmittance are frequency dependent. Distortion arises because the different harmonic components in the squarewave are amplified and phase shifted by different amounts so that their sum is no longer a squarewave.

Expressed in time domain language a linear circuit is characterized by saying that the shape of the output waveform is the same whatever the amplitude of the input. In frequency domain language we may characterize a linear circuit by saying that only frequency components present in the input appear in the output. Another general way of defining a linear circuit is to say that superposition applies; the output arising from two inputs is the sum of the outputs arising from each separately.

15.1.2 Nonlinear circuits

Characteristic features of a nonlinear circuit are that frequencies different from those present in the input are found in the output and the shape of the output depends on the amplitude of the input. Superposition does not apply. Nonlinear effects generally become dominant only if signals have substantial amplitudes, (~volts); most circuits behave linearly when excited at very low levels. Exceptions are multipliers and circuits with positive feedback.

Nonlinear circuits can be roughly divided into two categories, those in which we concentrate most on the waveshapes i.e. their behaviour in the time domain, and those in which the frequency content of the signals i.e. their behaviour in the frequency domain is of most interest. As you are now aware, we have powerful analytical ways of deriving the responses of linear circuits in the frequency and time domains. With nonlinear circuits quantitative analysis is much more difficult and we usually settle for more qualitative descriptions.

15.2 Rectifier / harmonic generator

The simple circuit shown in Figure M1.1 comprising a semiconductor diode and a resistor might be described as either a rectifier or a harmonic generator depending on whether we were thinking about the waveshape or the harmonic content.

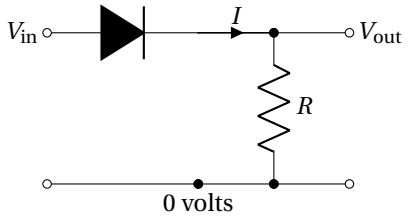


Figure 15.1: Rectifier or harmonic generator

For a silicon junction diode the relation between the current and the voltage drop is well described by the nonlinear expression (the Shockley law, see 9.6):

$$I = I_0 \left(e^{\frac{eV}{kT}} - 1 \right) \tag{15.1}$$

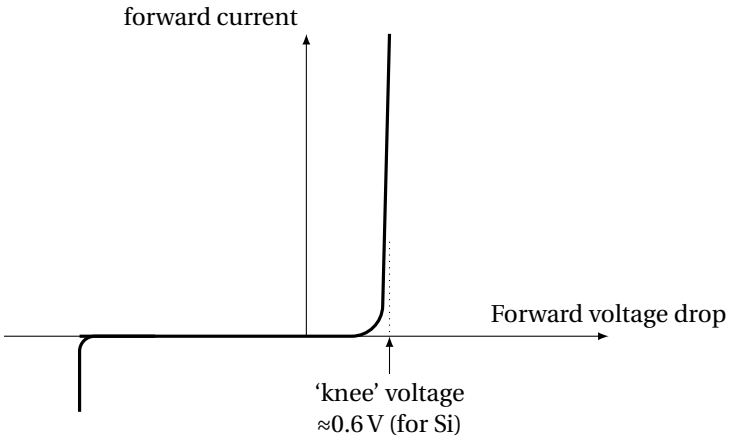


Figure 15.2: I-V characteristic of a silicon junction diode.

where I_0 is a constant at constant temperature. The diode conducts strongly in the forward direction once the "knee" voltage (about 0.6 volts) is reached and hardly conducts at all in the reverse direction (unless the breakdown voltage is exceeded) see Figure 15.2. V_{out} in the circuit above is then given by:

$$\frac{V_{\text{out}}}{R} = I_0\left(e^{\frac{e(V_{\text{in}}-V_{\text{out}})}{kT}} - 1\right) \tag{15.2}$$

This is completely intractable, emphasizing the point made at the end of section 15.1.2.

If V_{in} is a pure sine wave of (angular) frequency ω, and we imagine the exponential written as a power series, we can see that harmonically related frequency components at 0, 2ω, etc. appear in the output. [For example, a 0 Hz (dc) component and a 2ω component arise from the term in ${V_{\text{in}}}^2$ since $\sin^2\omega t$ is equal to $0.5\,(1-\cos 2\omega t)$.] The circuit is clearly behaving as a *harmonic generator*.

As far as the waveshape of V_{out} is concerned we can get a good picture of what is happening by imagining that the diode has a small voltage drop for current flowing in the forward direction and allows no current to flow in the opposite direction. It is then clear that the waveform of V_{out}, if V_{in} is much greater than 0.6 V peak, is just the positive half of the input sine wave. The circuit is behaving as a *half-wave rectifier*. Knowing this enables us to look up the harmonic content in Table 3.1 and so bypasses the problem of trying to find it from the expression above. Note that only *even* harmonics appear.

15.3 An amplitude limiter/odd harmonic generator

The circuit diagram below (Figure 15.3) using two silicon diodes and a resistor may be called an amplitude limiter or an odd harmonic generator depending on whether we are thinking about it in the time or frequency domains. We take a lesson from 15.2 above and consider the waveshape first. Clearly the output will approximate to a squarewave of about 1.2 Volts p-p for large sinusoidal inputs. This is worth realizing in itself but it also enables us simply to look up the harmonic content in Table 3.1 again. Note that only *odd* harmonics are generated. (You may care to ponder on the relation between the symmetry of the waveforms in the table and their harmonic content.)

15.4 Analog multipliers

Multipliers are based on the Jones/Gilbert cell shown in Figure 15.4.

D_1 and D_2 are transistors connected as diodes. $I_x + \Delta I_x$ and $I_x - \Delta I_x$ are inputs, $I_c + \Delta I_c$ and $I_c - \Delta I_c$ are outputs. We assume the transistors are identical and their base currents can be neglected. Then for T_1 and T_2

$$I_c + \Delta I_c = I_0 \mathrm{e}^{\frac{e(V_{\text{b1}}-V_{\text{e}})}{kT}}$$

$$I_c - \Delta I_c = I_0 \mathrm{e}^{\frac{e(V_{\text{b2}}-V_{\text{e}})}{kT}}$$

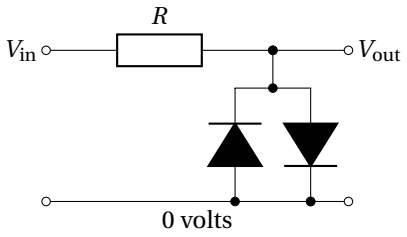


Figure 15.3: Amplitude limiter or odd harmonic generator circuit

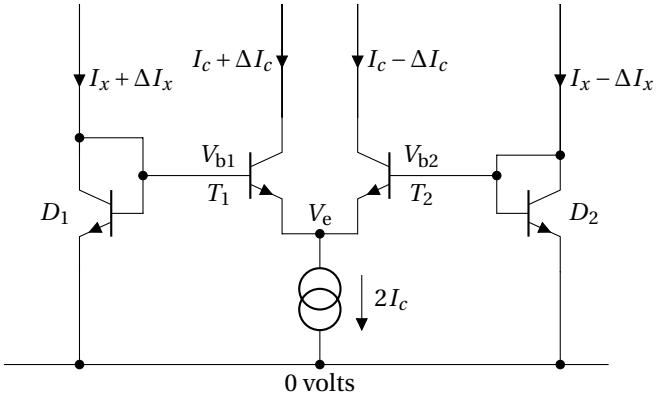


Figure 15.4: Jones/Gilbert cell

where I_0 is a constant. Dividing leads to

$$V_{b1} - V_{b2} = \frac{kT}{e} \ln \frac{I_c + \Delta I_c}{I_c - \Delta I_c}$$

For the diode connected transistors we have

$$I_x + \Delta I_x = I_0 e^{\frac{e(V_{b1}-0)}{kT}}$$

$$\text{and} \quad I_x - \Delta I_x = I_0 e^{\frac{e(V_{b2}-0)}{kT}}$$

leading to

$$V_{b1} - V_{b2} = \frac{kT}{e} \ln \frac{I_x + \Delta I_x}{I_x - \Delta I_x}.$$

Equating the two expressions for $V_{b1} - V_{b2}$ gives us

$$\Delta I_c = \frac{I_c}{I_x} \Delta I x;$$

ΔI_c is proportional to the product of I_c and ΔI_x. We have multiplication but these are inconvenient variables, we would prefer the inputs and output to be voltages. An arrangement that achieves this is shown in Figure15.5. Two cells $Q_{1\,A,B}$ and $Q_{2\,A,B}$ are used with their outputs cross coupled and fed to a difference amplifier (opamp A1 and four resistors R) giving an output with respect to 0 volts. The inputs are the voltages $V_{x1} - V_{x2}$ and $V_{y1} - V_{y2}$ and the transfer function is

$$V_{out} = k(V_{x1} - V_{x2})(V_{y1} - V_{y2}),$$

where k is typically 0.1 volts^{-1}. One of the x input terminals and one of the y input terminals may be connected to 0 volts if desired.

If $V_{x1} - V_{x2}$ and $V_{y1} - V_{y2}$ and V_{out} can all be of either polarity with respect to the zero of voltage then the device is said to be a "four quadrant multiplier". Note that the device is linear in $V_{x1} - V_{x2}$ if $V_{y1} - V_{y2}$ is kept constant and vice versa. Currently available devices have accuracies better than 1%. See for example the data for the AD534 in the Analogue computing script EL14.

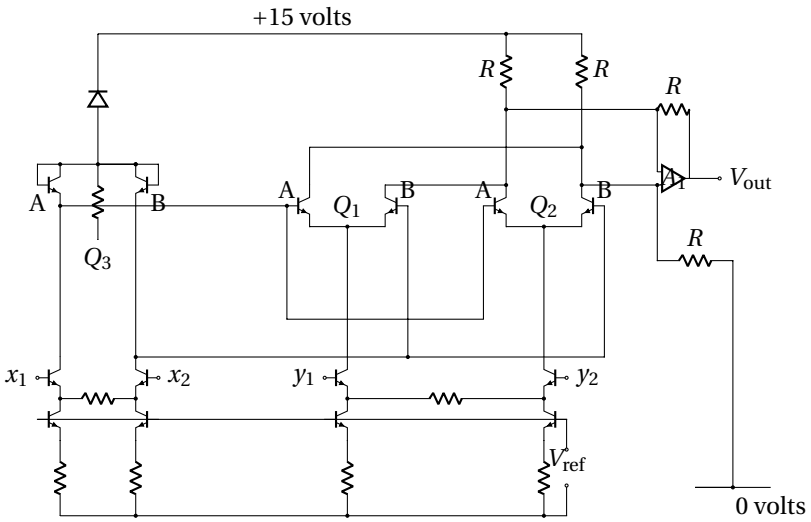


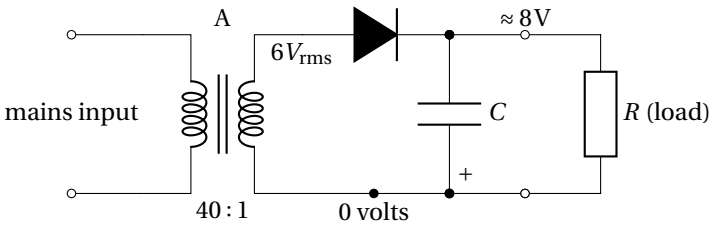
Figure 15.5: Voltage multiplier

15.5 Power supplies

Electronic circuits generally need to be fed with power from low voltage, dc supplies. Some such as watches and portable radios are designed for minimum power demand so that batteries or solar cells can be used. Circuits requiring a significant amount of power are normally run from the mains. The circuits that convert the high alternating voltage of the mains to smooth, unidirectional, low voltages are usually called "power supplies" — they are really power *converters* of course.

15.5.1 Half wave power converter

A simple circuit providing approximately +8 V is shown below.



(See Table 1.5 for meanings of symbols)

Figure 15.6: Half wave power converter circuit

Consider the waveforms in the steady state (a long time after switch on). The transformer produces a 6 V rms sinewave output from the 240 V rms, 50 Hz mains sinewave input. The capacitor C charges through the diode for the small fraction of a cycle during which the voltage at A exceeds the voltage across the capacitor. The capacitor then discharges through the load R for the rest of the cycle. The current through the diode consists of pulses 20 ms apart. It is easy to see what the shape of the output voltage is. It follows the input sine wave while the diode is conducting (neglecting the forward voltage drop) and then follows an exponential decay until the next positive peak of the input (Figure 15.7).

The peak to peak value of the *ripple* is easily calculated as part of an exponential decay with time constant RC. The ripple voltage causes a ripple current with the same waveform to flow through R.

Taking the frequency domain view of the steady state, the single frequency (50 Hz) output of the transformer causes a current to flow through the diode which contains a dc component and harmonic components with frequencies which are multiples of 50 Hz. The dc component flows through the load. The high frequency components flow through the capacitor giving rise to negligible voltage drops across it. The lower frequency components give rise to significant voltage drops whose waveforms when combined make up the ripple voltage.

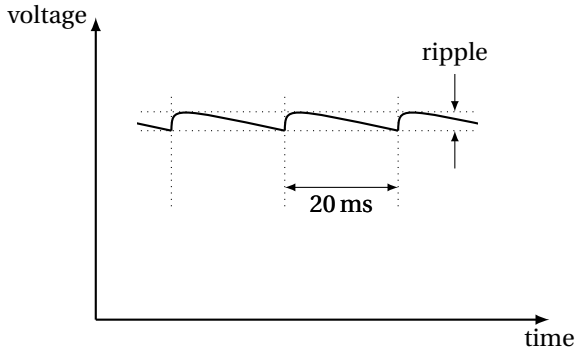


Figure 15.7: Ripple voltage on the output of a typical half-wave power converter

15.5.2 Full wave converters

(a) By using a transformer with an additional (identical) secondary winding and adding another diode as shown in Figure 15.8(a) a full wave circuit results in which there is a pulse of current through each diode for each cycle of the mains. The ripple has the same waveform as that shown in Figure 15.6 but is at twice the frequency and has a lower amplitude making it easier to smooth.

(b) For most circuits using opamps a "double sided" power supply with three terminals (typically +15, 0 and −15 volts) is required. The sign of V_{out} in the circuits above would be changed if the diodes were reversed. Therefore by adding another capacitor and another two diodes as shown below we can make a double sided full wave power supply (Figure 15.8(b)). The 0 volt line is usually connected to earth.

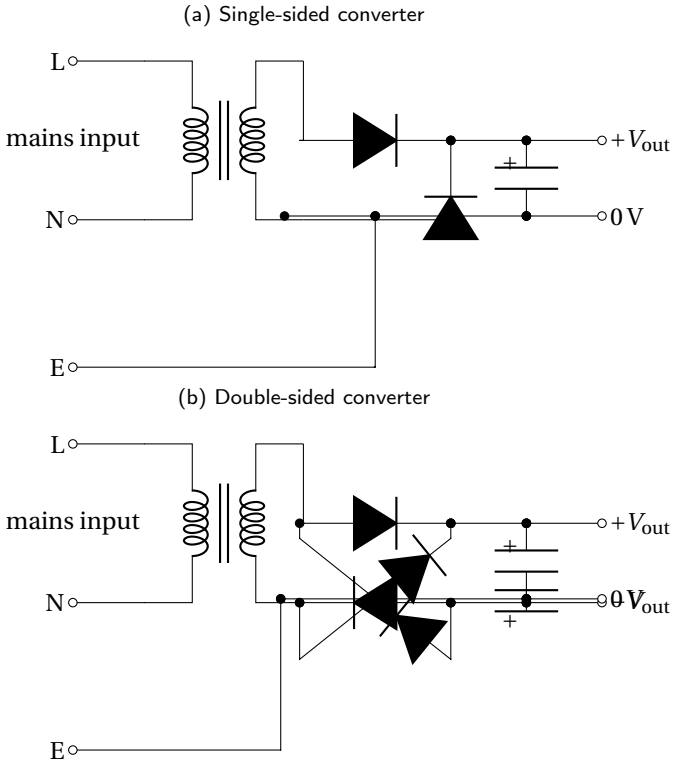


Figure 15.8: Full-wave converters

15.5.3 Stabilisation of the output voltage of power supplies

The amount of ripple voltage and the magnitude of the output resistance which can be tolerated in a power supply will depend on the application. The simple circuits we have examined may be acceptable as they stand but often much lower values of ripple and output resistance are required and it is sometimes important that the

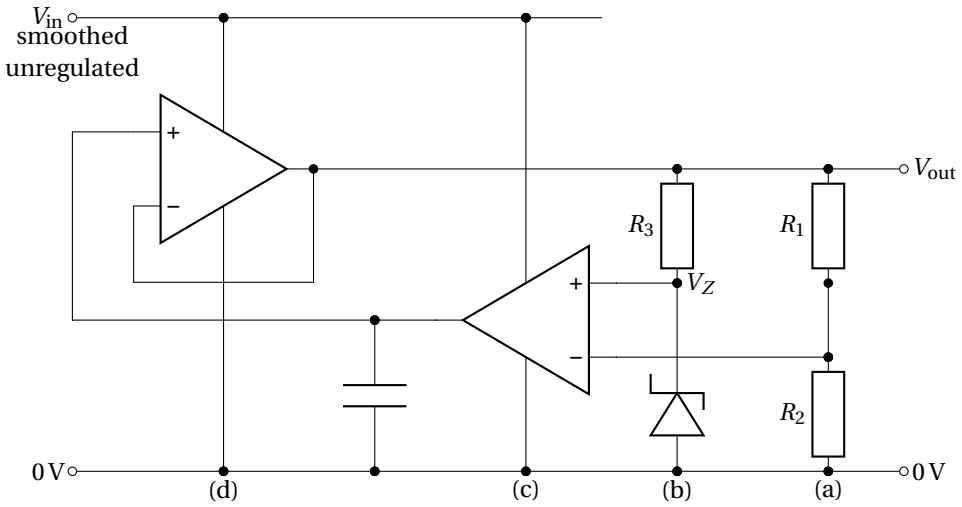


Figure 15.9: Voltage regulator

output voltage has a precise value. Improvements in these characteristics can be obtained by adding *regulators* to the simple circuits.

The circuit of a simple voltage regulator is shown in Figure 15.9. It comprises:

(a) a potential divider which delivers a fraction of the output voltage to the inverting input of the amplifier (c)

(b) a zener diode Z fed with current via R_3, (a zener diode is a nonlinear semiconductor device which maintains a nearly constant voltage drop V_Z for a wide range of currents). A stable low-ripple voltage V_Z is delivered to the non-inverting input of amplifier (c)

(c) an amplifier whose output depends on the difference between (a) and (b)

(d) a second (higher power) amplifier connected as a unity gain buffer.

The circuit is a negative feedback loop which maintains the output voltage in a fixed relation to V_Z.

$$V_{out} = \frac{R_1 + R_2}{R_2} V_2$$

15.6 Nonlinear amplifiers

15.6.1 Non-limiting

We are concerned here with amplifiers driven so that their outputs lie comfortably within their working ranges (see Figure 13.7). The non-linearities in the transfer characteristics of the opamps (slight curvature and perhaps small kinks due to crossover distortion in class B output stages) are not usually large enough or well enough defined for any use to be made of them. (Not really surprising since opamps are generally designed for use in linear circuits.)

Non-linear amplifiers are constructed by including nonlinear elements in the negative feedback network. The example most likely to be encountered is the logarithmic transresistance amplifier shown in Figure 15.10 in which a junction diode is used.

Assuming that the amplifier is ideal, the diode obeys the Shockley law, and $i_{in} \gg I_0$, the output voltage and input current are related by

$$V_{out} = \frac{kT}{e} \ln\left(\frac{i_{in}}{I_0}\right) \tag{15.3}$$

The circuit is often turned into a pseudo logarithmic voltage amplifier with the addition of a resistor (R) in series with the input (see section 14.5). The output voltage is then related to the input voltage by

$$v_{out} = \frac{kT}{e} \ln\left(\frac{v_{in}}{I_0 R}\right) \tag{15.4}$$

The sensitivity of these simple circuits to temperature changes of the diode means that more complicated circuits are used in practice.

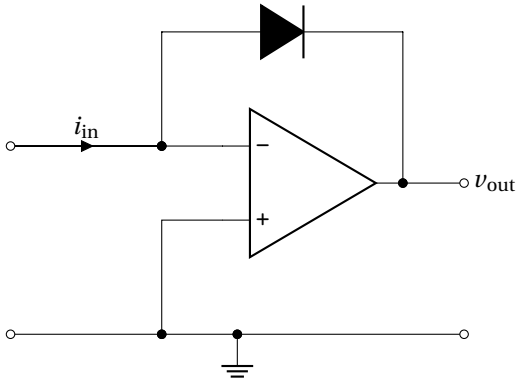


Figure 15.10: Logarithmic transresistance amplifier

15.6.2 Limiting

Any amplifier becomes a limiter if driven beyond the point where its working range is just filled. An example is an amplifier consisting simply of an opamp without feedback used to perform sinewave to squarewave conversion. Other circuits in which limiting is important are described below.

15.7 Switching circuits with positive feedback

When the feedback in an amplifying circuit is positive small signals are reinforced and increase to levels at which a nonlinear effect (usually limiting) controls what happens. Positive feedback speeds up transitions between limiting levels.

15.7.1 Schmitt trigger

If the input to a logic gate changes too slowly there is a danger of a burst of high frequency oscillation occurring at the switching point (due to the large high frequency gain of logic gates while they are active). The effect is to feed a burst of extra logic transitions into the system, clearly very undesirable. The solution is to feed the slow edge to the logic circuit via a Schmitt trigger circuit. This is shown in Figure 15.11.

The output of this circuit spends most of its time hard against one or other of the clipping levels, V_{c+} and V_{c-}, (the ends of the working range) of the opamp, any transitions being very rapid (due to the feedback) even if the input is slowly varying. The input voltage at which the output makes a transition depends on which clipping level it is setting out from since the potential divider providing the feedback also sets the voltage of the non-inverting input of the opamp. The difference between the two input voltages at which switching occurs is called the *hysteresis* and is given by:

$$\frac{R_2}{R_1 + R_2}\left(V_{c+} - V_{c-}\right) \tag{15.5}$$

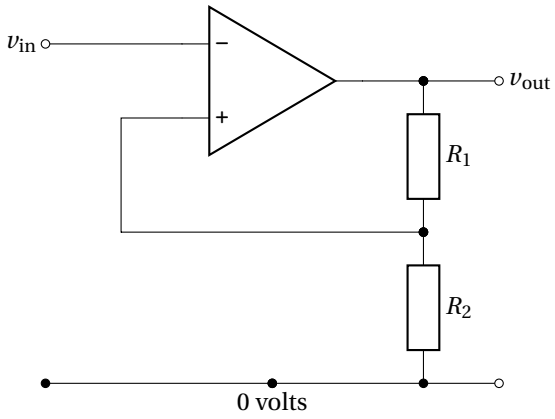


Figure 15.11: Schmitt trigger

15.7.2 Monostable

The monostable circuit, shown in Figure 15.12, produces a pulse of defined width following triggering by an arbitrary pulse of sufficient amplitude (and the correct sign). The circuit rests with the output of the left hand amplifier at the lower saturation level and the output of the right hand amplifier at the upper saturation level. A positive trigger pulse applied as indicated reverses this situation until the capacitor has had time to change its charge sufficiently whereupon the original state is regained. The positive feedback speeds up the transitions. Single transistors make satisfactory amplifiers in this circuit and the one following.

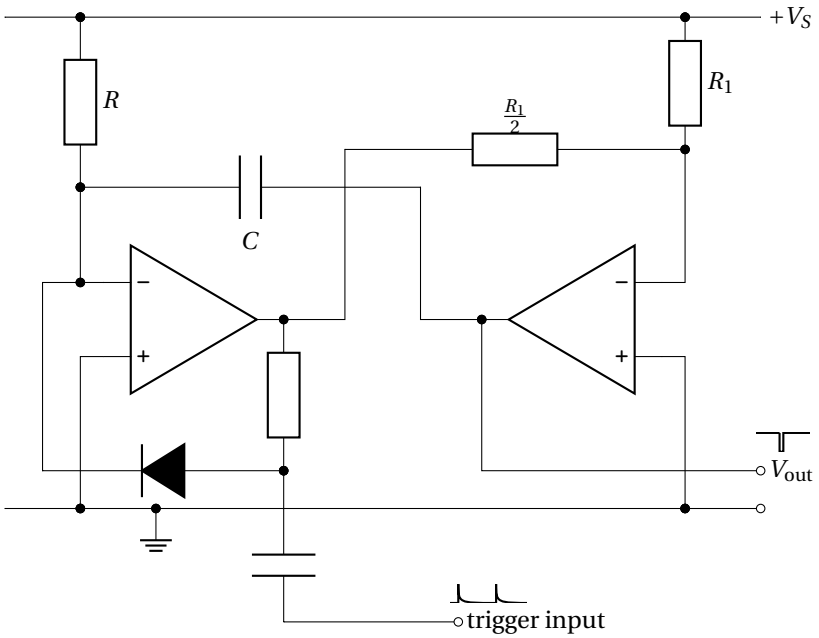


Figure 15.12: Monostable

15.7.3 Bistable

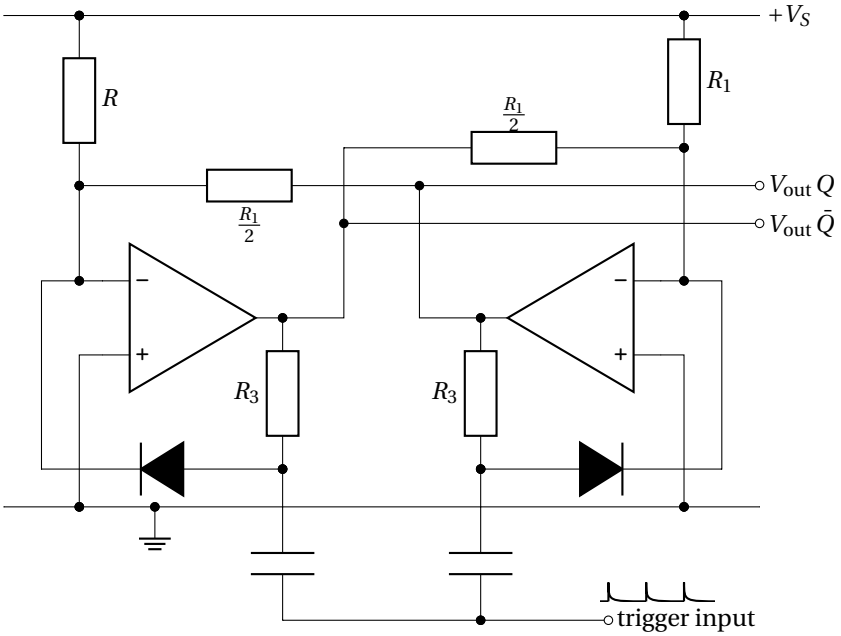


Figure 15.13: Bistable

Bistables are of course used as 1 bit stores and should need no further introduction here. A circuit is shown in Figure 15.13. Its operation should be clear from the foregoing except perhaps for the function of the diodes (which is to steer the trigger pulses to the correct amplifier).

(It was the monostable which was originally and aptly called a flip-flop (it is flipped one way and then flops back to its resting state in its own time). The name has come to be applied, rather inappropriately, to bistables also.)

15.8 Mixers

A nonlinear circuit with two inputs is shown in Figure 15.14. (A, ϕ, ω) represents a sinewave of amplitude A, phase ϕ and frequency ω; $m = n = 0$ allows for a possible dc component in the output.

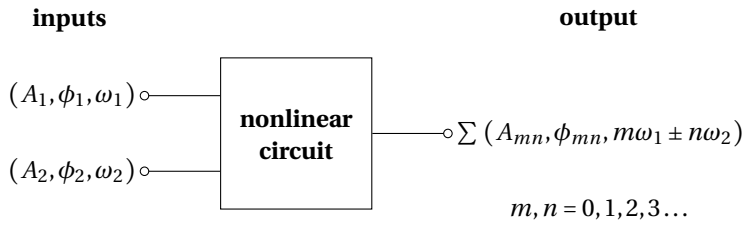


Figure 15.14: Mixer circuit

In general the output contains components at harmonics of the input frequencies and at sum and difference frequencies. The generation of sum and difference frequency components from two input signals is an important operation in many signal processing systems and is called *mixing*. (Sound studio "mixers" are analogue adders.) Any nonlinear device can in principle be used as a mixer but only a few are in practice.

Diodes are used as mixers at microwave frequencies and a circuit consisting of two transformers and a ring of four diodes is often used at lower frequencies.

Another device which is occasionally used as a mixer and which is easy to analyse is the field effect transistor. In a suitable circuit its transfer function has the form:

$$V_{\text{out}} = C\left(V_p - V_{\text{in}}\right)^2 \tag{15.6}$$

where C and V_p are constants.

If $V_{\text{in}} = A_1 \sin\omega_1 t + A_2 \sin\omega_2 t$ (ignoring any phases which could easily be put in):

$$V_{\text{out}} = C\left[V_p^2 - (A_1 \sin\omega_1 t + A_2 \sin\omega_2 t)\right]^2 \tag{15.7}$$

$$= C\left(V_p^2 - \frac{A_1^2 + A_2^2}{2}\right) \tag{15.8}$$

$$- C\left(2V_p\left(A_1 \sin\omega_1 t + A_2 \sin\omega_2 t\right)\right) \tag{15.9}$$

$$- C\left(\frac{A_1^2}{2}\cos 2\omega_1 t + \frac{A_2^2}{2}\cos 2\omega_2 t\right) \tag{15.10}$$

$$+ C\left(A_1 A_2 \cos\left(\omega_1 - \omega_2\right) t - A_1 A_2 \cos\left(\omega_1 - \omega_2\right) t\right) \tag{15.11}$$

so even with this simple quadratic non-linearity we have five new components (including the one at dc).

Another device which can be used as a mixer and is also easy to analyse is the analog multiplier. It is particularly convenient to use because it has two separate input ports. See Section 15.4.

15.9 Communications

15.9.1 Introduction

One of the most important characteristics of an analogue signal is the range of frequencies it contains. For "communications quality" speech the range is typically 300 Hz–2.5 kHz, for "hi-fi" music 20 Hz–16 kHz, and for broadcast quality video the range extends to 10 MHz. The upper limit for high speed data is currently in the 100 MHz region but is being rapidly increased. These bands of frequencies are called the *basebands* of the signals. The width of a signals baseband defines the minimum bandwidth of the channel needed to communicate the signal.

If more than one signal of the same type is sent down the same cable simultaneously it is not possible to separate them at the receiving end if their basebands overlap. However, if each of them were to be shifted up in frequency by a different amount so that they then occupied non-overlapping bands of frequency it would be possible to separate them (with filters) at the receiving end. Then provided they could be shifted back to their baseband the original signals could be recovered.

Another reason for wanting to shift the position of a signal in the spectrum is to make it possible to broadcast it from an aerial. In this case a radio or television receiver performs the separation and down-shifting to the baseband.

(Up-shifting can be achieved not only by passing signals through a mixer but also by using the baseband signals to vary the *frequency* of a carrier signal. This is the technique used in VHF radio broadcasting. We will not pursue it here.)

The processes of up-shifting and down-shifting are called *modulation* and *demodulation*.

15.9.2 Point-to-point communications

Long distance telephone and most other point-to-point communications systems employ the method known as single sideband suppressed carrier (SSSC) transmission. Let $V_b \sin \omega_b t$ be one component of a signal occupying a baseband. At the transmitting end of a communications channel this is multiplied by a high frequency sinewave of constant amplitude $V_c \sin \omega_c t$, known as the carrier. The result is (using the expression for a product of sines):

$$\frac{1}{2} V_b V_c \left[\sin \left(\omega_c + \omega_b \right) t + \sin \left(\omega_c - \omega_b \right) t \right] \tag{15.12}$$

The first term shows that the baseband has been shifted to lie just above ω_c the carrier frequency. The second term shows that there is an inverted version of the baseband lying just below the carrier frequency. These two shifted basebands are called *sidebands*. There is no signal at the carrier frequency ω_c. The sidebands extend out from the carrier frequency on either side by an amount equal to the highest frequency in the baseband signal.

Both sidebands carry the same information so we can halve the bandwidth needed (back to the width of the signal's baseband) by transmitting only one of them (removing the other with a filter). Let's assume we remove the inverted one. Having transmitted the remaining sideband we can easily return it to baseband (demodulate it) at the receiving end; we simply multiply it by a locally generated carrier $V_c' \sin \omega_c t$. The result is:

$$\frac{V_c V_c'}{4} V_b \left[\sin \omega_b t + \sin \left(2\omega_c + \omega_b \right) t \right] \tag{15.13}$$

It is easy to remove the second term with a filter leaving the original baseband signal component (apart from a constant factor).

If there is an error in the frequency of the local oscillator in the receiver, the recovered baseband signal will be shifted in frequency by the same amount. Provided the shift is small, it is usually of no consequence.

15.9.3 AM broadcasting

If before multiplication by the carrier at the transmitting end, the quantity $V_b \left(1 + m \sin \omega_b t\right)$ is formed from the signal component $V_b \sin \omega_b t$ (where m is a constant < 1 called the *modulation depth*), the product becomes

$$V_b \left(1 + m \sin \omega_b t\right) V_c \sin \omega_c t \tag{15.14}$$

which is a carrier wave with a slowly varying amplitude (amplitude modulated (AM) wave). This is identical to:

$$V_b V_c \left[\sin \omega_c t + \frac{m}{2} \sin \left(\omega_c + \omega_b \right) t + \frac{m}{2} \sin \left(\omega_c - \omega_b \right) t \right] \tag{15.15}$$

showing that there is now a signal at the carrier frequency as well as two sidebands. The spectrum is shown in Figure 15.15 for $m = 0.5$.

Again the sidebands extend out from the carrier frequency on either side by an amount equal to the highest frequency in the baseband signal and again in the lower frequency sideband the distribution of frequencies is reversed compared with the baseband.

This is the kind of radio signal which is broadcast on the short, medium and long wavebands. The disadvantages of the technique, that it uses twice the minimum necessary bandwidth and wastes power transmitting a carrier, are currently outweighed by the simplicity of the receivers which do not need to be highly stable or contain demodulation oscillators. (See below for the method of demodulation used.)

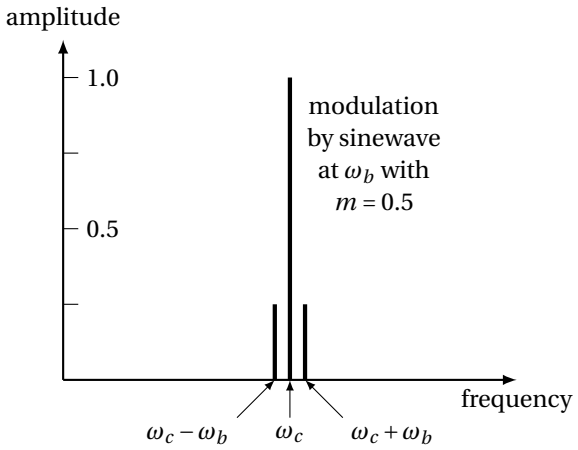


Figure 15.15: RF spectrum of an AM signal

In order to be able to pack a lot of stations into the crowded spectrum in the medium waveband the highest modulation frequency the engineers transmit is usually limited to 4.5 kHz which entails some loss of fidelity of music signals. Care is taken to ensure that the modulation depth is such that the downward peaks of the baseband signal never cut off the carrier as this produces signals both at multiples of the carrier frequencies and outside the intended sidebands.

In principle, the modulation can be recovered from an amplitude modulated carrier using a mixer in which the second input is from an oscillator in the receiver tuned to the carrier frequency. However there is a snag. Both sidebands (and the carrier) now mix with the locally generated input and if there is an error in the frequency of the local oscillator, there is no component in the mixer output which is simply the modulation. The local oscillator must be *phase locked* to the carrier for it to work.

The method of demodulation used in AM radio receivers in practice does not require a local oscillator input at all. Consider the smoothed half wave rectifier shown in Figure 15.16 (compare with Figure 15.6).

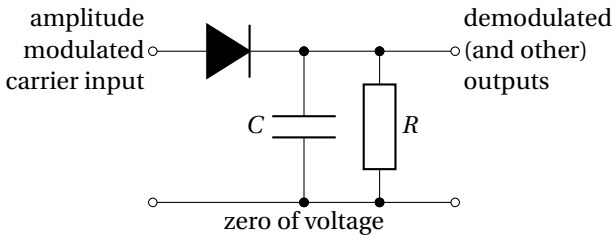


Figure 15.16: Half-wave rectifier used for AM demodulation

The smoothing time constant RC is chosen to be long compared with the period of the carrier and short compared with the shortest period of the modulation so that the output follows the peak value of the input as it varies with the modulation.

There are actually three components in the output voltage of this demodulator:

(a) the wanted audio signal (originally the amplitude modulation)

(b) a mean dc level

(c) a residual ripple at the carrier frequency and harmonics of it.

Another way to look at the operation of this demodulator is to say that the nonlinear characteristic of the diode is mixing the carrier with the sidebands to produce, among other things, the modulation (the carrier of course has exactly the right phase for this to work).

15.9.4 The superheterodyne AM radio receiver

A radio receiver must be capable of being tuned to the correct carrier frequency and must have a spectral response suitable for the signals being received. Generally we would like the response to be uniform for the full width of the interval desired (9 kHz for medium wave AM radio) and zero outside so that signals at adjacent and distant frequencies are rejected. With currently available techniques it is not possible to make filters of ideal shape which are tunable over wide frequency ranges. This was even more true in 1925 when Edwin Armstrong invented the *superheterodyne* receiver circuit to overcome this problem. In this circuit the frequencies of incoming carriers are converted to a fixed frequency called the *intermediate frequency* or i.f. (at which a good filter *can* be made) using a mixer fed with a second input from a local oscillator tuned to be always a fixed frequency above (or below) the frequency of the wanted incoming signal.

The signal is then amplified and filtered at the intermediate frequency, demodulated by a peak rectifying diode (Figure 15.16), further amplified at baseband frequencies (audio) and passed to a loudspeaker.

15.10 Use of mixers in analogue instruments

We return to the mixer demodulator we dismissed for AM demodulation in section 15.9.3. In systems in which the baseband signal is impressed onto the carrier locally rather than at a distant transmitter the carrier is clearly also available to be used in a mixer demodulator. A good example of such a system is an infrared radiometer in which the incoming radiation beam is interrupted at a carrier frequency so the output from the radiation detector is an amplitude modulated signal at this frequency the amplitude being proportional to the strength of the incoming beam.

After filtering and amplification the signal is demodulated by a mixer using a local input phase locked to the carrier obtained from the circuit driving the beam interrupter mechanism. This kind of signal processing has advantages in reducing the effect of random (noise) voltages accompanying the wanted signal. In such systems different names from the ones used in communications tend to be given to the parts, modulator becomes “chopper”, demodulator becomes “phase sensitive rectifier”, the second input to the demodulator is called the “reference” signal and the whole signal processing chain is often called a “lock-in amplifier”. See section 20.6.3.

16 Analogue Computing

16.1 Introduction

Differential equations describing physical systems are now solved routinely by digital computers and where a precise investigation of one or two solutions is required we need use no other tool. However, there are some equations (particularly non-linear ones exhibiting chaotic behaviour) where it is of interest to look first for regularities in the solutions as parameters in the equation are varied. This can often be most conveniently achieved by constructing an electronic circuit which is an analogue of the system and examining the waveforms generated. Machines called analogue simulators or computers are used for this purpose. Although they cannot compete in accuracy with digital computers the effect of varying parameters can be seen instantly and a panoramic view of the nature of the solutions quickly obtained.

16.2 Ideal building blocks

The building blocks used to construct analogue computers are summing amplifiers and integrators (which invert), multipliers, and potential dividers. Differentiators can be used but are not favoured because of the degradation in signal-to-noise ratio that often results.

16.2.1 Inverting summing amplifier (Figure 16.1)

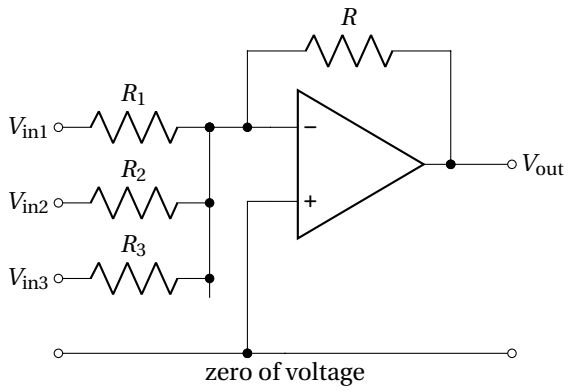


Figure 16.1: Inverting summing amplifier with 3 inputs

Assuming the opamp is ideal, its inverting input is at the zero of voltage and its output is given by

$$V_{\text{out}} = -\left(\frac{V_{\text{in1}}}{R_1} + \frac{V_{\text{in2}}}{R_2} + \frac{V_{\text{in3}}}{R_3}\right) R \tag{16.1}$$

16.2.2 Inverting summing integrator

The basic integrator circuit is shown in Figure 16.2.

Assuming the opamp is ideal its inverting input is at the zero of potential and all the current flowing through the input resistor R charges the capacitor C. The output voltage at time t is then given by

$$V_{\text{out}}(t) = \frac{Q_0}{C} - \frac{1}{RC}\int_0^t V_{\text{in}}\,\mathrm{d}t \tag{16.2}$$

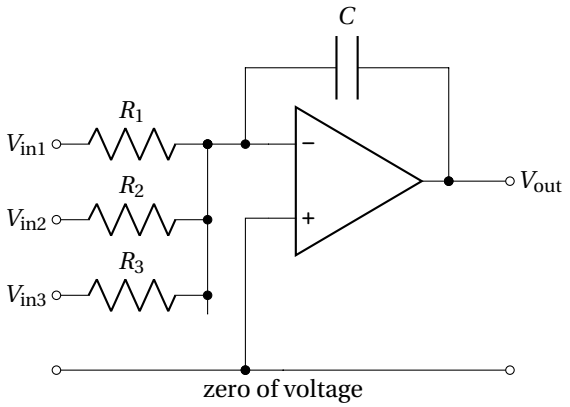


Figure 16.2: Inverting summing integrator with 3 inputs

Assumong the opamp is ideal its inverting input is at the zero of voltage and the output voltage at time tis

$$V_{\text{out}}(t) = \frac{Q_0}{C} - \frac{1}{R_1 C}\int_0^t V_{\text{in1}}\,\mathrm{d}t - \frac{1}{R_2 C}\int_0^t V_{\text{in2}}\,\mathrm{d}t - \frac{1}{R_3 C}\int_0^t V_{\text{in3}}\,\mathrm{d}t, \tag{16.3}$$

where Q_0 is the charge on C at $t = 0$.

Setting up the initial conditions before starting to run a simulation will include setting the initial value of Q_0 on the capacitor of each integrator. Switches must be added to the integrator circuit to enable it to be changed from set mode to run (compute) mode. These are shown in the circuit of Figure 16.3. When the switches are in the set position, $V_{\text{out}} = Q_0/C$.

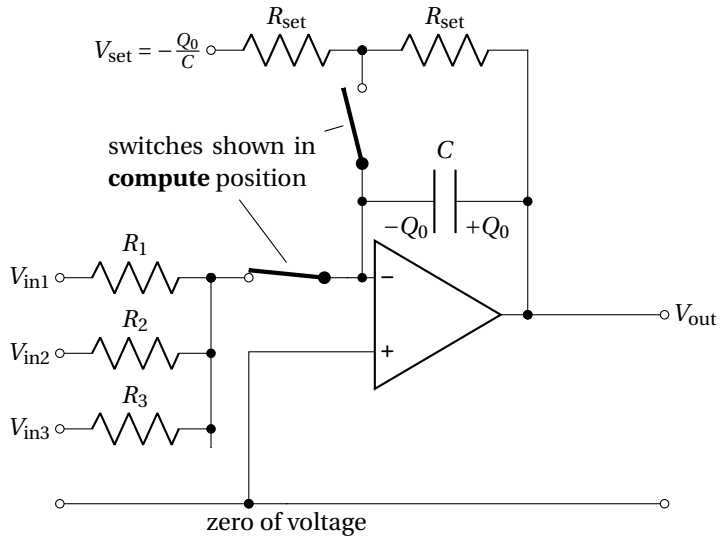


Figure 16.3: Inverting summing integrator with switches to select 'set' and 'run/compute' modes.

16.2.3 Multipliers

Analog multipliers were described in chapter 15.

For an ideal device with two inputs the output voltage at any instant is given by:

$$V_{\text{out}} = k V_{\text{in1}} V_{\text{in2}} \tag{16.4}$$

where k is a constant.

16.2.4 Voltage dividers

When any load is negligible the output voltage at any instant is given by:

$$V_{\text{out}} = \alpha V_{\text{in}} \tag{16.5}$$

where α is a (manually) adjustable constant less than unity.

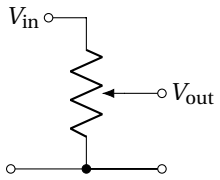


Figure 16.4: Voltage divider

16.3 Simulation of differential equations

Consider as an example the equation:

$$\ddot{y} + A\dot{y} + By + Cy^n = F\sin\omega t \tag{16.6}$$

where A, B, C and F are constants. We can formally integrate this twice to yield

$$y + A\int y\,\mathrm{d}t + B\int\left[\int y\,\mathrm{d}t\right]\mathrm{d}t + C\int\left[\int y^n\,\mathrm{d}t\right]\mathrm{d}t + Dt + E = -\frac{F}{\omega^2}\sin\omega t \tag{16.7}$$

where D and E are constants of integration (initial conditions). This is of course no help in providing a paper solution but it is of just the form we need to see how to devise an electronic simulation using the building blocks and an oscillator. We simulate not the differential equation but the integral equation derived from it.

16.4 An example: disk dynamo loaded with a motor

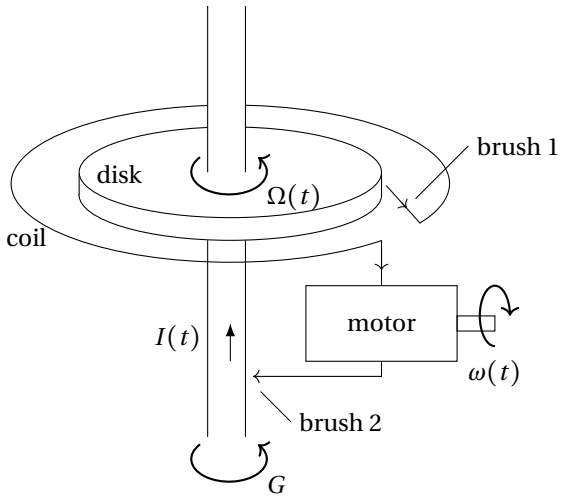


Figure 16.5: A system comprising a Faraday disk dynamo driving a motor

Figure 16.5 shows a system comprising a Faraday disk dynamo driving a motor. One of the reasons for being interested in this system is that the current, and therefore the magnetic field, can exhibit chaotic behaviour similar to that seen in the geological record of the Earth's magnetic field.

In the dynamo, the stationary coil surrounding the disk is of inductance L and one end is connected by a brush (sliding contact) to the rim of the disk. The radially conducting disk is fixed to the axle and contact is made to the axle via another brush. The motor completes the circuit. The disk is made to rotate at angular speed $\Omega(t)$ by a steady external torque G applied to the axle. If any axial magnetic field is present a radial emf is developed in the disk sending a current I through the coil and the motor.

The torque produced by the motor is assumed proportional to the current. If ω is the speed of rotation of the motor armature, B the moment of inertia of the armature, HI the torque on the armature due to the current, and $D\omega$ the decelerating torque on it due to friction, we get:

$$B\dot{\omega} = HI - D\omega \tag{16.8}$$

Denoting by $2\pi M$ the mutual inductance between the coil and the rim of the disk, assuming that the coil is wound in the sense shown in the figure, that Ω is positive when in the same sense as that of the applied couple, and that the magnetic field acting on the disk is produced by the current I in the system, the emf generated by the motion of the disk is $MI\Omega$. This emf is balanced by the sum of the voltage drops RI, $L\dot{I}$, and $H\omega$ where R is the total electrical resistance of the system, i.e.:

$$L\dot{I} + RI + H\omega = MI\Omega \tag{16.9}$$

A third equation comes from considering the torque on the disk which if A is the moment of inertia of the disk and K the coefficient of friction, is given by:

$$G - MI^2 - K\Omega = A\dot{\Omega} \tag{16.10}$$

It is convenient to replace the dimensional dependent variables $(I(t), \Omega(t), \omega(t))$ and independent variable t by the dimensionless variables $(x(\tau), y(\tau), z(\tau))$ and τ defined by

$$\tau \equiv \frac{R}{L}t, \qquad x \equiv \left(\frac{M}{G}\right)^{\frac{1}{2}} I, \qquad y \equiv \frac{M}{R}\Omega, \qquad z \equiv \frac{R}{L}\frac{B}{H}\left(\frac{M}{G}\right)^{\frac{1}{2}} \omega \tag{16.11}$$

The set of equations 16.8–16.10 then take the dimensionless forms:

$$\dot{x} = x(y-1) - \beta z \tag{16.12}$$

$$\dot{y} = \alpha\left(1 - x^2\right) - \kappa y \tag{16.13}$$

$$\dot{z} = x - \lambda z \tag{16.14}$$

where:

$$\alpha \equiv \frac{GLM}{R^2 A} \tag{16.15}$$

$$\kappa \equiv \frac{KL}{RA} \tag{16.16}$$

$$\beta \equiv \frac{H^2 L}{R^2 B} \tag{16.17}$$

$$\lambda \equiv \frac{DL}{RB} \tag{16.18}$$

Accordingly we simulate:

$$x = \int_0^t \left\{x(y-1) - \beta z\right\} \mathrm{d}t + e_1 \tag{16.19}$$

$$y = \int_0^t \left\{\alpha\left(1 - x^2\right) - \kappa y\right\} \mathrm{d}t + e_2 \tag{16.20}$$

$$z = \int_0^t \left\{x - \lambda z\right\} \mathrm{d}t + e_3 \tag{16.21}$$

$$\tag{16.22}$$

and represent x, y, z by voltages.

An example of the circuitry needed to realise the first integral equation is shown in Figure 16.6 below. The integrators are unclamped and the simulation begins to run at $t = 0$. The initial condition e_1 is the integrator output voltage set up before $t = 0$ as described in 16.2.2.

16.5 Practical considerations

16.5.1 Signal levels

The number of volts per unit for each variable must be chosen so that signal levels remain comfortably within the working ranges of the opamps and multipliers and sufficiently above the noise level at all times during a simulation.

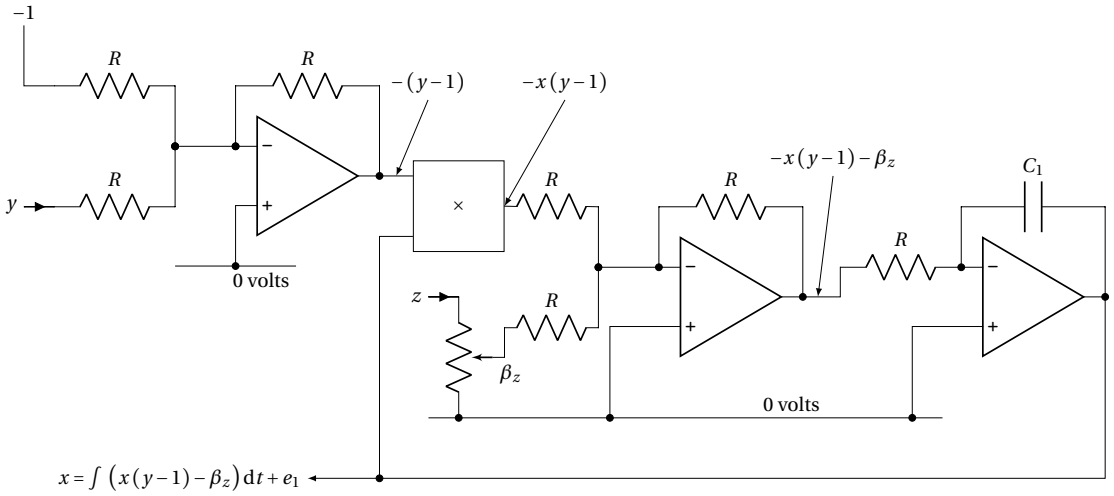


Figure 16.6: Simulation circuit

16.5.2 Time scaling

Time scaling is chosen for convenience of display and so that frequencies are below those at which the performance of the active devices starts to deteriorate. Typically this means that characteristic times are scaled to be in the range 1–10 ms.

16.5.3 Offsets

The outputs of the active devices will not be precisely zero when their inputs are zero. Improvements can usually be obtained by trimming. As a minimum:

(a) In summing amplifiers, low input current opamps should be employed and should have their input offset voltage (see 13.8) trimmed.

(b) In integrators, very low input current opamps should be employed and should first have their input offset voltage trimmed and then their input offset current trimmed for minimum drift (by supplying or sinking a small current at the inverting input).

(c) Multipliers cannot usually be trimmed by the user but in some a voltage can be added to the output and used to trim the offset.

(d) The reference conductor on which the active devices are mounted should be a continuous metal sheet to minimise voltage drops across it.

16.5.4 Drifts

Temperature changes will affect supply voltages, voltages used as parameters, and offsets. Adequate time must be allowed for temperatures to stabilise after any changes, particularly after soldering. The most useful basic precaution is to protect the circuit from draughts.

17 Analog Oscillators

17.1 Introduction

If an amplifier has sufficient positive feedback to make the loop gain greater than unity for small signals, an output, beginning as random noise or a switch-on transient, will build up until some non linear process intervenes. If the range of frequencies over which the loop gain is greater than unity extends down to 0 Hz (dc) a persistent saturated state results. If it does not a repetitive output waveform is produced whose shape and period depend on the feedback network and the type of nonlinearity. We will consider examples of two broad classes of such circuits: (i) van der Pol, Robinson and multiplier controlled 'sinewave' oscillators, and (ii) relaxation oscillators with square wave, triangle wave, and pulse outputs. We assume we have chosen opamps whose defects can be ignored in the circuits presented, i.e. they can be treated as ideal.

17.2 Oscillators with 'sinewave' outputs

No real oscillator can produce a perfect sinewave. (The best commercial audio oscillators have distortion levels of less than 0.001%, more ordinary ones have distortions of a few percent.) The components of a sinewave oscillator circuit are an opamp, a positive feedback network containing frequency determining elements, and some arrangement to control the amplitude.

17.2.1 Classification of 'sinewave' oscillators

We classify the circuits according to the type of frequency determining network and the method of amplitude control.

Frequency determining networks

Oscillation occurs at the frequency at which the total phase shift in the positive feedback loop is zero. The phase shift in the amplifier will usually be close to either 0 or π depending on whether it is connected as non-inverting or inverting. Some frequency determining networks are shown in Figure 17.1.

Networks f(i) and f(ii) have a transmission peak at the operating frequency and f(iii) does not (it is not essential that there is a peak). Network f(iv), known as a 'twin tee', provides a transmission notch and is used in a negative feedback loop around the opamp within the overall positive feedback loop. A rapid rate of change of phase shift with frequency is desirable.

Methods of amplitude control

a(i) *Soft limiting* (known as van der Pol operation). The circuit has weak positive feedback and a non-linear amplifier whose gain falls gently as the amplitude of the oscillation builds up. The amplitude settles at some value well within the clipping levels.

a(ii) *Hard limiting* (known as Robinson operation). The circuit has strong positive feedback and drives itself to an amplitude set by a limiter or the clipping levels of the amplifier.

a(iii) *Gain control.* The positive feedback network contains a device such as a thermistor or a multiplier which enables its transmission to be controlled by the oscillation amplitude.

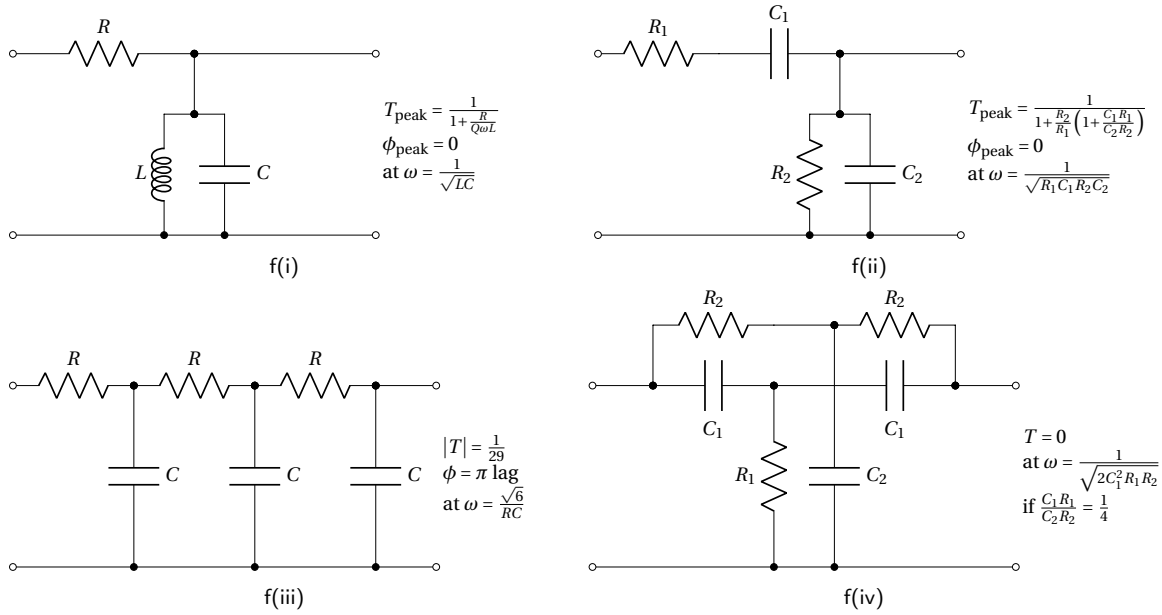


Figure 17.1: Frequency determining networks

17.2.2 Type f(i) a(i) [van der Pol] oscillator

We discuss the circuit shown in Figure 17.2. The (gentle) nonlinearity is represented by the device NL.

The positive feedback is via the narrow bandpass filter network R_3, L, C, where $R_3 \gg Q\omega L$, the resonant impedance of the tuned circuit. We assume that the current I in R_3 is related to V, the voltage at the non inverting input of the amplifier, by the power series:

$$I = H_0 V + H_1 V^2 + H_2 V^3 + \ldots \tag{17.1}$$

and that $V = V_0 \sin\omega t$ where ω is the resonant frequency of the tuned circuit. In the steady state the power converted into heat in the tuned circuit must equal the power delivered to the tuned circuit by the amplifier, i.e.:

$$\frac{V_0^2}{2Q\omega L} = \frac{\omega}{2\pi}\int_0^{2\pi/\omega} I V_0 \sin\omega t \,\mathrm{d}t \tag{17.2}$$

$$= \frac{\omega}{2\pi}\int_0^{2\pi/\omega} \left(H_0 V_0^2 \sin^2\omega t + H_1 V_0^3 \sin^3\omega t + H_2 V_0^4 \sin^4\omega t + \ldots\right)\mathrm{d}t \tag{17.3}$$

$$= \frac{1}{2}H_0 V_0^2 + \frac{3}{8}H_2 V_0^4 + \frac{5}{16}H_4 V_0^6 + \ldots \tag{17.4}$$

(The integrals in the odd powers of V are zero.)

H_0 must be positive for oscillations to start so for this equation to have a solution other than $V_0 = 0$, one of the H_2, H_4, etc. must be negative. For most non linear elements the higher the power of V_0 the smaller the coefficient. If we take H_4 and higher coefficients to be negligible we find:

$$V_0^2 = \frac{4}{3}\frac{H_0 - \frac{1}{Q\omega L}}{-H_2} \tag{17.5}$$

We see that H_0 and the losses in the tuned circuit must be such that $H_0 > 1/Q\omega L$ and that H_2 must be negative.

Assuming a progressive fall in the size of the coefficients we have:

$$H_0 \approx \frac{\partial I}{\partial V} \qquad \text{must be positive and} \qquad H_2 \approx \frac{1}{6}\frac{\partial^3 I}{\partial V^3} \tag{17.6}$$

for stable operation. It is interesting to test a few device characteristics for this condition.

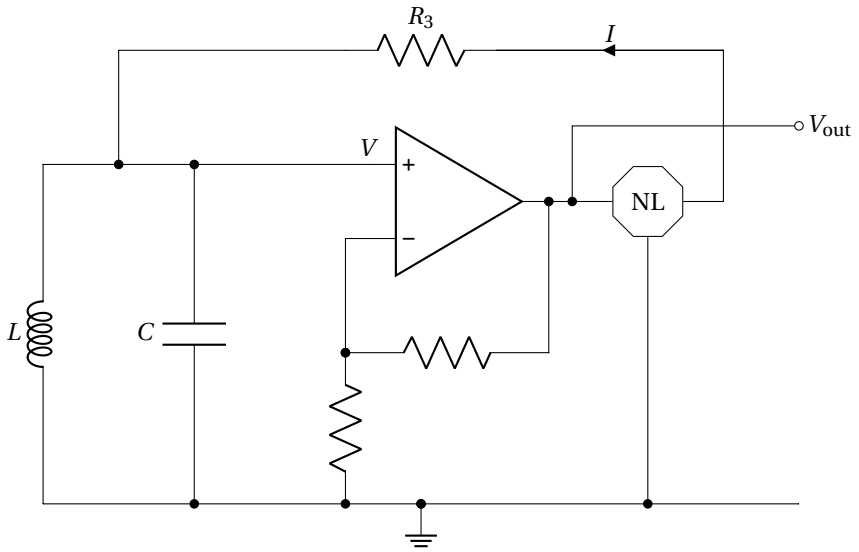


Figure 17.2: van de Pol oscillator

Pentode valve

The transfer characteristic is $I = K\left(V + A\right)^{3/2}$ where K, A are constants (Child's law):

$$\frac{\partial I}{\partial V} = \frac{3}{2}K\left(V + A\right)^{1/2} \qquad \text{and} \qquad \frac{1}{6}\frac{\partial^3 I}{\partial V^3} = -\frac{K}{16}\left(V + A\right)^{-3/2} \tag{17.7}$$

so the pentode is suitable for van der Pol operation.

Field Effect Transistor

The transfer characteristic is $I = I_0\left(1 + V/V_p\right)^2$:

$$\frac{\partial I}{\partial V} = \frac{2I_0}{V_p}\left(1 + \frac{V}{V_p}\right) \qquad \text{and} \qquad \frac{1}{6}\frac{\partial^3 I}{\partial V^3} = 0 \tag{17.8}$$

so stable van der Pol operation is not possible for a true square law device.

Bipolar (diffusion) transistor

The transfer characteristic is $I = I_0 \exp\left(eV/kT\right)$ and as all the derivatives have the same sign, van der Pol operation is not possible.

Van der Pol developed his treatment of oscillators in 1927 during the era of triode valves which obey Child's law. As our brief survey shows single semiconductor devices are unsuitable for this kind of operation so its relevance has declined.

The soft limiting in van der Pol oscillators can mean that the amplitude is strongly influenced by component tolerances and that the harmonic content of the output is difficult to predict.

17.2.3 Type f(i) a(ii) [Robinson] oscillator

We discuss the circuit shown in Figure 17.3.

Assuming that $R_3 \gg Q\omega L$, the resonant impedance of the tuned circuit, and that V_{out} is large enough to cause hard limiting by R_4, D_1, D_2, the current through R_3 is a squarewave of amplitude $I_0 \approx V_{\text{lim}}/R_3$ (zero to peak) where V_{lim} is the maximum limiter output. The voltage across the tuned circuit will be close to a sine wave provided the tuned circuit has a Q > 5. The power balance equation becomes:

$$\frac{V_0^2}{2Q\omega L} = 2\frac{\omega}{2\pi}\int_0^{\pi/\omega} I V_0 \sin\omega t \,\mathrm{d}t = \frac{2V_{\text{lim}}V_0}{\pi R_3} \tag{17.9}$$

$$\therefore V_0 = \frac{4V_{\text{lim}}Q\omega L}{\pi R_3} \text{ (zero to peak)} \tag{17.10}$$

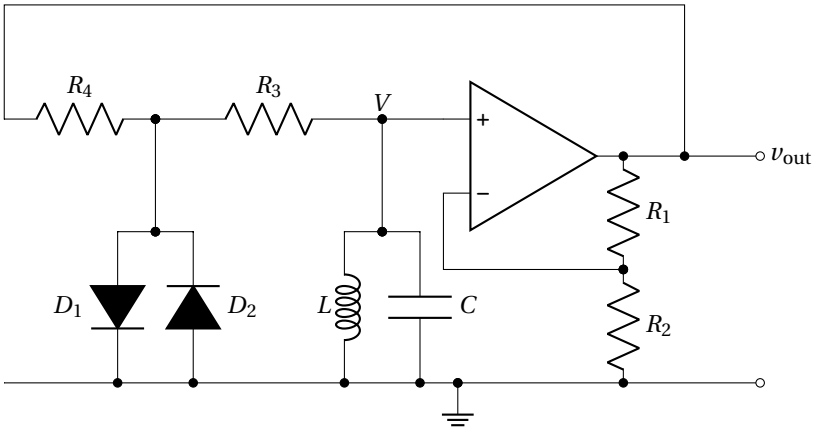


Figure 17.3: Robinson oscillator

just what we would expect for a fundamental component of $I = 4I_0/\pi$ (zero to peak). The third harmonic component of I is $I = 4I_0/3\pi$ (zero to peak), using this it is straightforward to show that the third harmonic content of the output is $V_0/9Q$ (zero to peak). With a decent Q, say 50, we see that a very respectably pure sine wave output is obtained even though the circuit involves hard limiting.

The amplitude of a Robinson oscillator is less dependent on component tolerances than a van der Pol oscillator and as we have seen the purity of its waveform is readily calculated also.

17.2.4 Type f(i) a(iii) oscillator

We discuss the circuit shown in Figure 17.4. R_1 is a temperature-dependent resistor (thermistor) whose temperature depends on the amplitude of v_{out}.

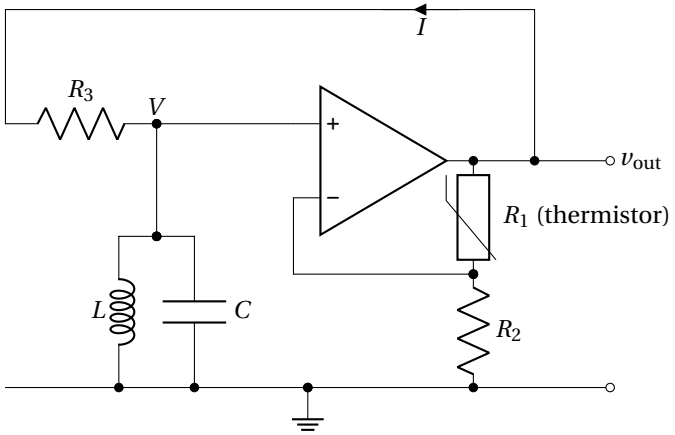


Figure 17.4: Type f(i) a(iii) oscillator

The positive feedback is via the narrow bandpass filter network R_3, L, C, where $R_3 \gg Q\omega L$, the resonant impedance of the tuned circuit. The voltage amplifier has a gain which falls smoothly with oscillation amplitude due to the (negative temperature coefficient) thermistor R_1 in its negative feedback network. Typically $R_1 = R_0 e^{-\alpha(T-T_0)}$ where R_0 is the resistance at T_0 (room temperature) and $\alpha \approx 0.04\,\mathrm{K}^{-1}$. If Newton's law of cooling applies to the thermistor $T - T_0$ will be given by:

$$T - T_0 = kV_{out}^2$$

Oscillation occurs at an amplitude such that the loop gain is unity i.e. (assuming an ideal opamp):

$$\frac{1 + \frac{R_0 e^{-\alpha k V_{out}^2}}{R_2}}{1 + \frac{R_3}{Q\omega L}} = 1$$

The amplitude of oscillation in this circuit is sensitive to ambient temperature changes. At very low frequencies the thermistor causes distortion.

17.2.5 Type f(iii) a(iii) oscillator

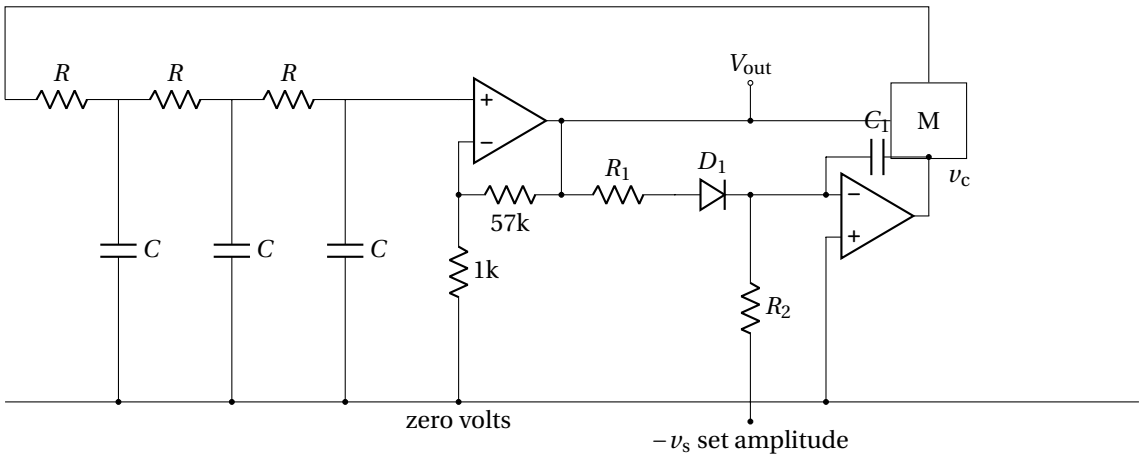


Figure 17.5: Type f(iii) a(iii) oscillator

(See Figure 17.5.) At the frequency of oscillation the RC "ladder" network has a transmittance of $-1/29$. The amplifier has a gain of 58 and the multiplier (set to invert) has a gain of $-kv_c$. The loop gain is therefore $0.2v_c$ (for $k = 0.1\,\mathrm{V}^{-1}$). Since the loop gain is unity at the oscillation frequency the control voltage v_c will settle at 5 V.

The amplitude control works as follows:—

Current with the waveform of a half-wave rectified sine wave flows onto the left hand plate of C_1 via D_1 and R_1 and a steady current flows off the plate to the negative set amplitude supply $-v_s$ via R_2. v_{out} varies until the average value of the sum of these two currents is zero.

17.2.6 Low distortion oscillator

A problem with the circuit in section 17.2.5 is that the amplitude control voltage v_c inevitably has some residual ripple on it at the oscillation frequency. This ripple modulates the loop gain and causes distortion.

The integrator time constant $\frac{R_1R_2}{R_1+R_2}C_1$ would have to be infinite to eliminate the ripple completely which woudl mean waiting for ever for the amplitude to stabilise. Ideally, we would like a control voltage that responded instantly to changes of amplitude but carried negligible ripple. A way to achieve this is outplined below.

Consider an oscillator in which the positive feedback loop is comprised of a negating multiplier and two circuits like the one shown in Figure 17.6.

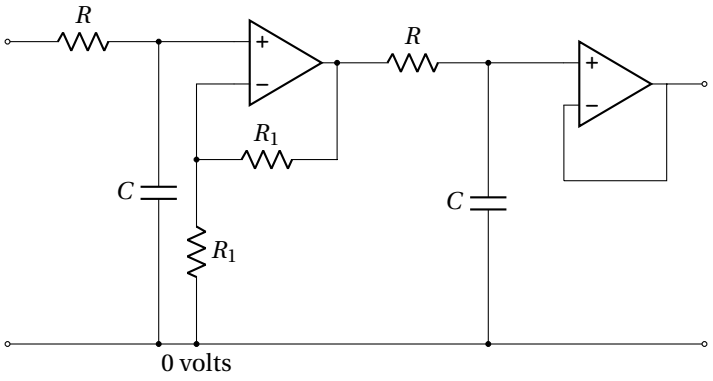


Figure 17.6: Section of positive feedback loop

Using four RC sections to achieve a π phase lag and placing voltage amplifiers between them has a unmber of advantages. Each RC section produces a $\pi/4$ phase lag and a maximum rate of change of phase with frequency, each amplifier output is a low impedance source, and most important for us we have outputs available with equal amplitude and $\pi/2$ phase difference i.e. $v_0\cos\omega t$ and $v_0\sin\omega t$. We use these outputs as shown in Figure 17.7.

The current flowing onto the left hand plate of C_1 from the multipliers is $\frac{kV_0^2}{R_3}$ free of ripple. The amplitude control then works as in section 17.2.5.

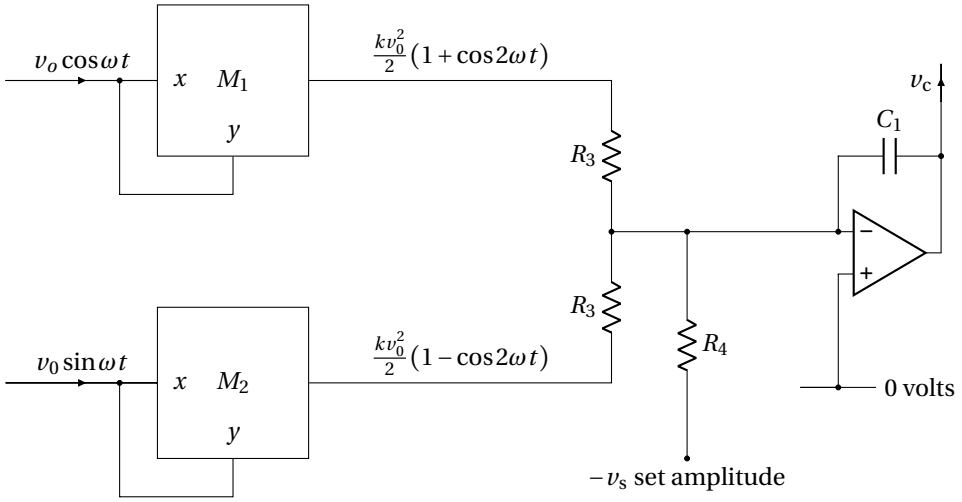


Figure 17.7: Generation of amplitude-control voltage

17.3 Squarewave, triangle, pulse, and sawtooth oscillators

These waveforms are usually generated in circuits which switch between limiting or clipping states. The fundamental frequency often depends on the clipping levels so the functions of amplitude and frequency control are not in general separable as they are for most near sinewave oscillators. A typical cycle of operation is for a circuit to switch to one of its saturated states and then wait for the charge on a capacitor to build up or decay until a level of voltage is reached at which the circuit switches back to the first state. It then waits again before returning to the second state. The name *relaxation oscillator* is sometimes used for such circuits.

17.3.1 Square and triangle waves

Square and triangle waves may be generated by a Schmitt trigger with an additional feedback loop as shown in Figure 17.8.

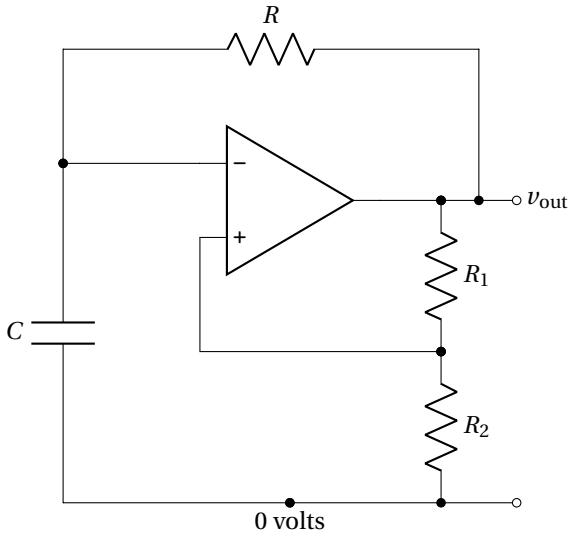


Figure 17.8: Square and triangle wave oscillator using a Schmitt trigger with feedback loop

Assume that v_{out} makes an upward transition from V_{c-} to V_{c+} at some instant. This means that the voltage of the non-inverting input of the opamp rises from βV_{c-} to βV_{c+} where:

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

For the transition to have occurred, the inverting input must have been at βV_{c-}. After the transition the potential of the inverting input rises as C charges through R and it eventually reaches βV_{c+} at which a transition of the

output back to V_{c-} occurs. This cycle repeats and thus a square wave appears at the output. The frequency of the square wave is determined by the time it takes for the voltage across the capacitor to relax between βV_{c-} and βV_{c+}.

When $R_2 \ll R_1$, a good triangle wave exists at the inverting input of the opamp. If this is wanted it should be taken via a buffer amplifier with high input impedance.

17.3.2 Unequal mark-to-space ratio

Circuits with two time constants may be used to generate waveforms with unequal mark-to-space ratios. An example is shown in Figure 17.9.

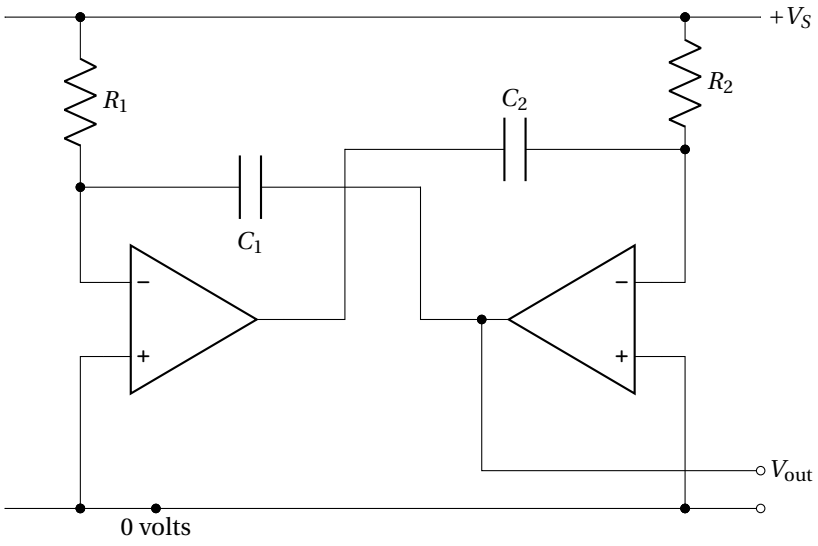


Figure 17.9: Circuit for generating waveforms with unequal mark-to-space ratios

Single bipolar transistors in the common emitter configuration make satisfactory amplifiers in this circuit.

18 MOSFETs and Logic Gates

18.1 Introduction

In junction FETs (Chapter 11) the current flowing in the channel under the influence of the drain-source voltage is controlled by the voltage applied to the gate-channel junction. In MOSFETs (Metal-Oxide-Semiconductor-FETs) control of the current is exercised by a metal gate electrode deposited on a thin layer of silicon dioxide grown on the silicon surface. Both P and N channel MOSFETs can be made in two forms, one in which there is a channel at $V_g = 0$ (c.f. the JFET), described as depletion mode devices and one in which a gate-source voltage must be applied to create a channel, enhancement mode devices. The circuit diagram symbols of the four varieties are shown in Figure 18.1. They are rather neat in the way they indicate the mode of operation.

18.2 The N channel enhancement mode MOSFET

MOSFETs are manufactured using the same photolithography, etching, and diffusion processes that are used to make junction devices. An N channel enhancement mode MOSFET has the structure shown in Figure 18.2.

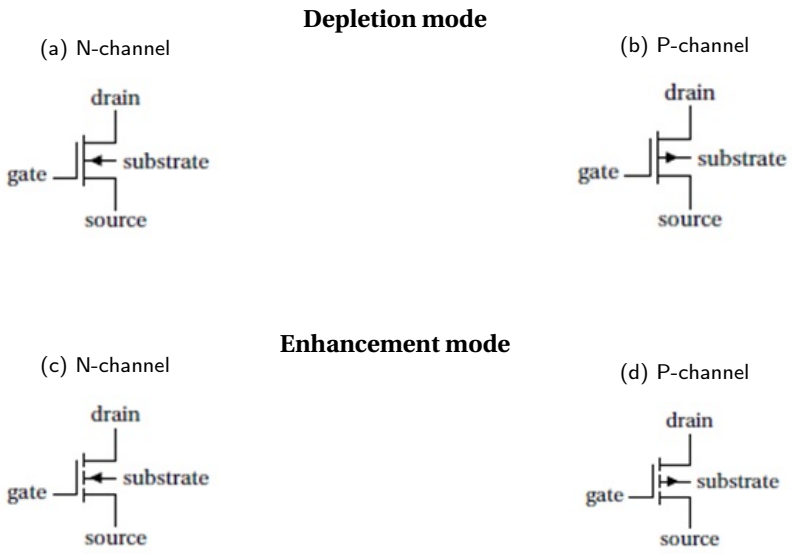


Figure 18.1: MOSFET varieties

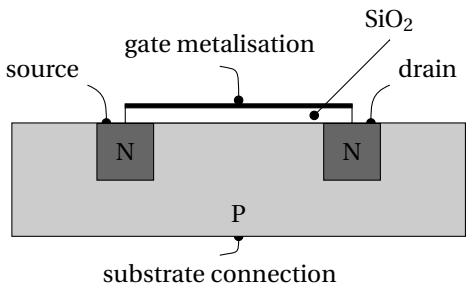


Figure 18.2: Construction of an N channel enhancement mode MOSFET

With no voltage applied between gate and channel (no charge on the gate-channel capacitance) there is no conduction between the source and the drain because the two back to back PN junctions are more than a diffusion length apart. When the gate is charged positively the surface of the P type silicon under the oxide layer becomes negatively charged. This occurs through the attraction of conduction electrons and the expulsion of holes. At a high enough voltage on the gate (the threshold voltage V_t) an N type layer (called an inversion layer) forms on the surface of the silicon making it possible for majority electrons to flow between the source and drain (which are N type regions). The N type channel formed gets thicker (more conducting) if the gate voltage is further increased. So for gate-channel voltages greater than V_t current will flow when a drain-source voltage V_d is applied. As is seen in JFETs the characteristics show two behaviours, that of a voltage controlled resistance for low V_d and that of a voltage controlled current source for high V_d. The drain current is given approximately by:

$$I_\mathrm{d} = I_\mathrm{dss}\left(\frac{V_\mathrm{g}}{V_t} - 1\right)^2 \text{ for } \left|V_\mathrm{g}\right| > |V_t| \tag{18.1}$$

$$= 0 \text{ for } \left|V_\mathrm{g}\right| < |V_t| \tag{18.2}$$

V_t is the threshold voltage at which the channel first forms. The gate current, being due only to leakages, is very small, even less than in JFETs. The substrate is usually connected to the negative supply rail ensuring that the P/N junctions at the ends of the channel are revese biased.

The maximum charge that can be put onto the silicon surface by charging the gate metal/silicon dioxide/silicon capacitor is $\epsilon_0\epsilon_r E_\mathrm{max}$ where E_max is the breakdown field of the SiO_2 dielectric and ϵ_r its relative permittivity. Taking ϵ_r as 4 and E_max as $10^8\,\mathrm{V\,m^{-1}}$ yields $3.5\times10^{-3}\,\mathrm{C\,m^{-2}}$ for the maximum surface charge ($\sim 2.2\times10^{16}$ electron charges $\mathrm{m^{-2}}$). For the device to work, the original density of holes in the P-type silicon must be low enough for this to create an N-type layer at the surface.

18.3 CMOS logic circuits

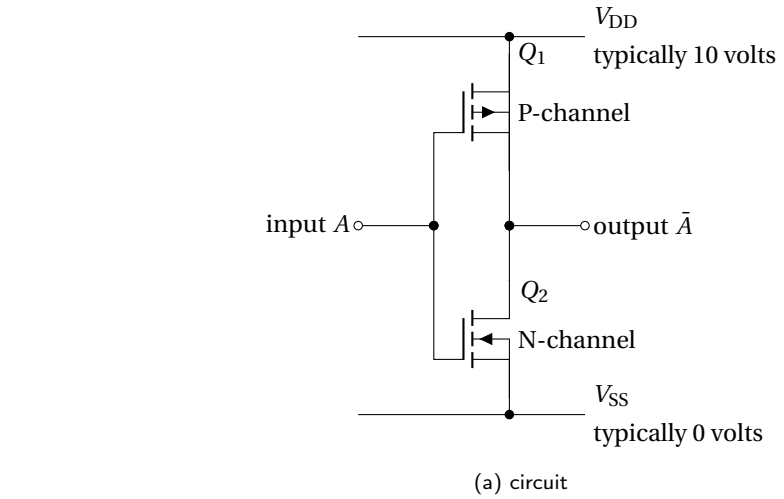
Most logic circuits are now made entirely from enhancement mode MOSFETs. This currently dominant technology is referred to as complementary MOS or CMOS, complementary simply meaning that both N- and P-channel devices are used. The ideas involved are best illustrated by considering some gate circuits.

18.3.1 NOT gate (inverter)

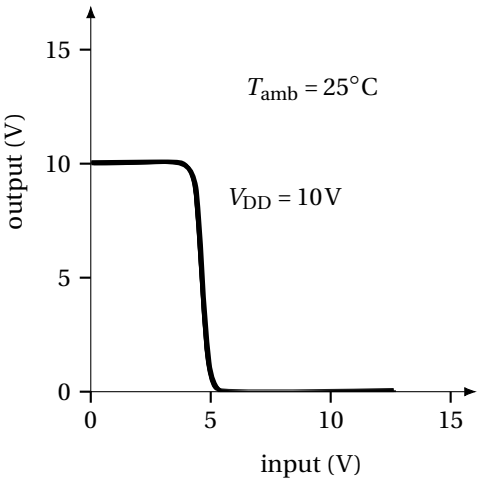
The circuit diagram of a CMOS NOT gate, its transfer characteristic, and the current it draws from the power supply (with no load connected to its output) are shown in Figure 18.3(a), (b), (c) respectively. Note that the N channel device is at the bottom and the P channel device is at the top in the circuit diagram. (This is always the case in CMOS circuits drawn with the +ve supply at the top.)

When the input voltage is zero, there is a thick P type channel in the upper device Q_1 connecting the output to the +10 volts supply. There is no channel in the lower device Q_2 which is therefore not conducting and no current is taken from the supply. When the input is increased to about 2 volts an N type channel starts to be formed in Q_2 and as Q_1 is still conducting strongly current starts to be drawn from the supply. The output voltage remains near 10 volts. At an input voltage near 5 volts both transistors are conducting equally, the output has fallen to 5 volts, and the maximum current is being drawn from the supply. As the input voltage is increased further the P type channel in Q_1 becomes thinner and the N type channel in Q_2 connecting the output to 0 volts becomes thicker. The output voltage approaches 0 volts and the current drawn from the supply falls. When the input reaches about 8 volts the P channel in Q_1 disappears completely, the current falls to zero, and the output is close 0 volts where it remains as the input is raised to 10 volts.

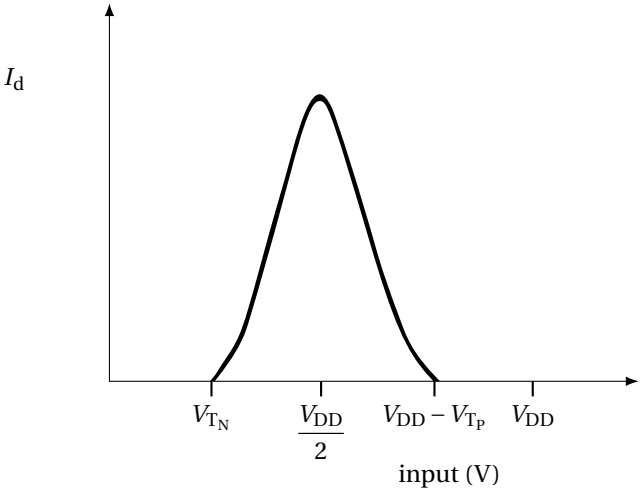
In using this circuit as a logic inverter (NOT gate) logic 0 is restricted to the range 0–2 volts and logic 1 to 8–10 volts to avoid the conditions under which both transistors are conducting. Power is dissipated in the channel resistances when the output changes states because of the current just mentioned and also because any capacitance loading the output is charged through a P channel when the output goes to 1 and discharges through an N channel when the output goes to 0, transferring charge from V_dd to 0 and dissipating $\frac{1}{2}CV_\mathrm{dd}^2$. Rapid transition between logic levels can minimise the power dissipation due to both devices conducting but for both mechanisms the dissipation is proportional to the frequency of the transitions.



(a) circuit



(b) Output voltage as a function of input voltage



(c) Drain current I_d as a function of input voltage

Figure 18.3: CMOS NOT gate

18.3.2 AND gate

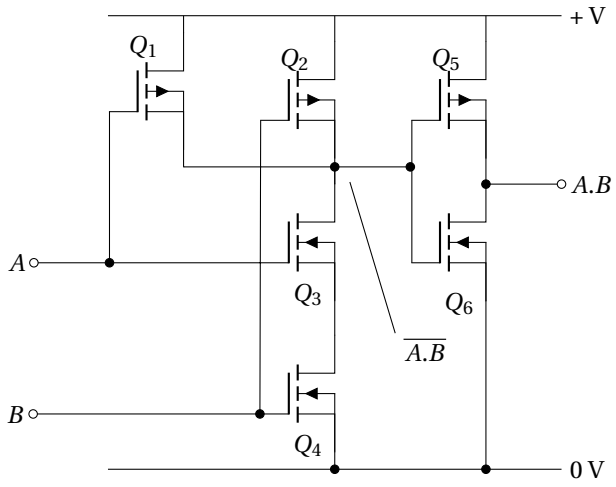


Figure 18.4: CMOS AND gate

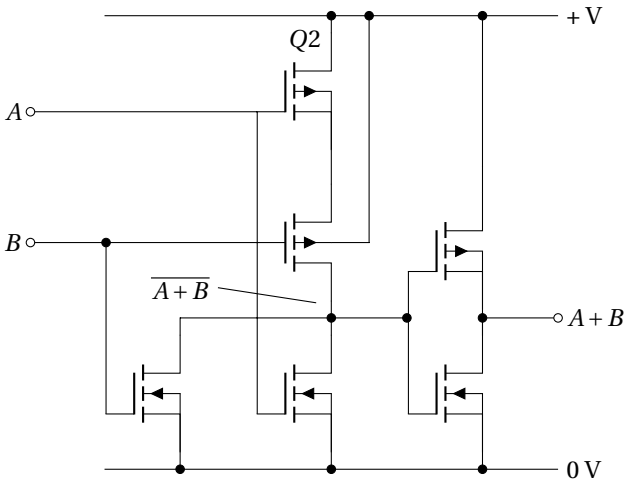


Figure 18.5: CMOS OR gate

The circuit diagram of a two input CMOS AND gate is shown in Figure 18.4.

The right hand pair of transistors are connected as a NOT gate so the four devices on the left must perform the NAND function for the overall function to be AND. In this gate there are two N channel devices in series at the bottom of the circuit and two P channel devices in parallel at the top. A logic 0 on one of the inputs will turn

one of the P channel devices on and one of the N channel devices off so the input to the inverter will be a logic 1. A logic 1 on both inputs will turn both the P channel devices off and both N channel devices on making the input to the inverter a logic 0. To provide another input another pair of devices is added to the circuit, an N channel device in series with the others at the bottom and a P channel device in parallel with the others at the top.

18.3.3 OR gate

The circuit diagram of a two input OR gate is shown in Figure 18.5.

Again there is an inverter on the right hand side so the left hand four transistors must perform the NOR function. There are two N channel devices in parallel at the bottom and two P channel devices in series at the top. The operation and the way of extending of the number of inputs should be obvious from the previous discussion.

18.3.4 Transmission gate

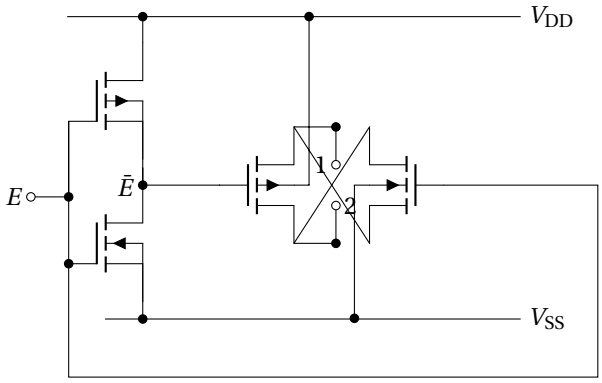


Figure 18.6: CMOS transmission gate (1 = input or output, 2 = output or input)

The circuit diagram of a transmission gate, a logic controlled switch, is shown in Figure 18.6. Terminals 1 and 2 are connected when E is at logic 1 and disconnected when E is at logic 0. It consists of P and N channel devices in parallel and an inverter. The signal being switched must remain within the voltage range 0 to V_{dd}. Assuming initially that terminals 1 and 2 are near $V_{dd}/2$, E at logic 1 turns on the N channel device and the inverter produces a 0 at the gate of the P channel device which also turns it on. When terminals 1 and 2 are near V_{dd} or V_{ss} one of the devices may cease to be conducting but the other will remain on. When E is at logic 0 both devices remain off whatever the voltages on terminals 1 and 2 (provided they remain between 0 and V_{dd}).

Transmission gates can be used to switch analogue and digital signals.

18.3.5 Flip flops

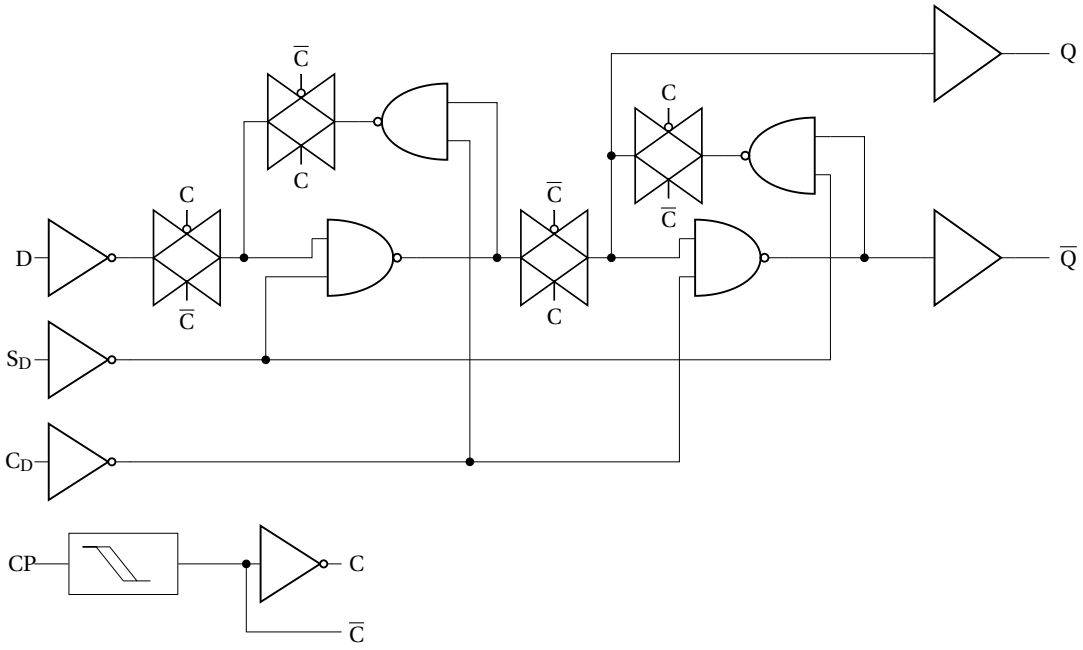


Figure 18.7: CMOS D type flip-flop

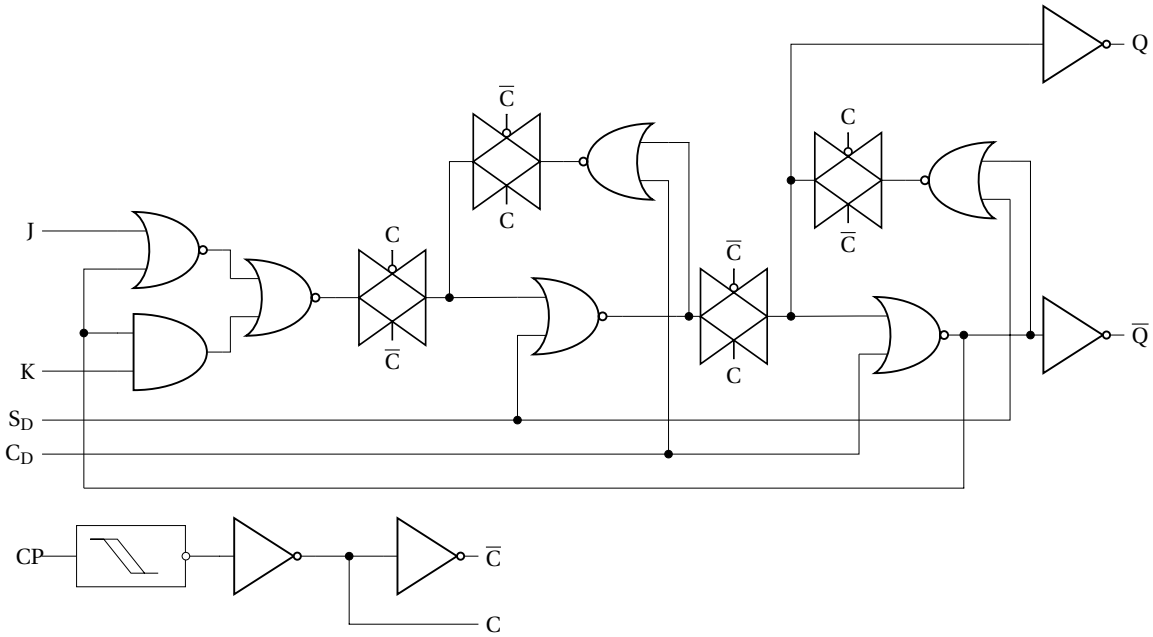


Figure 18.8: CMOS JK type flip-flop

Block diagrams of D type and JK type flip flops are shown in Figures 18.7 and 18.8 respectively. The symbol with four triangles represents the parallel pair of devices in the transmission gate shown in Figure 18.6. It is left as an exercise for the reader to reconcile these diagrams with the truth tables given in Chapter 4.

18.3.6 Random access memory

Random access memory (RAM) chips carry rectangular arrays of flip flops which can be individually addressed. Because their interfaces are defined the flip flops do not have to be general purpose types like the D and JK described above and can be of simplified design to economise on silicon area. A typical circuit is shown in Figure 18.9. The cross-coupled N channel transistors T_3 and T_4 with the p channel T_1 and T_2 as drain loads make up the basic bistable. The N channel transistors T_5 and T_6 provide access to the bistable for reading or writing when the row and column lines are simultaneously made high.

The flip flop described above is the simplest circuit which can retain the bit value it is loaded with for as long as the power supply is maintained. It is a *static* memory cell. However, with the voracious demand for memory by present day programs, much less silicon-hungry memory cell circuits are needed. Simpler cells can be used if we drop the requirement that they be static and arrange for periodic *refreshing* of the cell contents. Such a cell is described as *dynamic* and typically has the circuit shown in Figure 18.10. When its row and column lines are simultaneously addressed (made high) the access transistor is turned on and either some charge is transferred into the capacitor if it was originally uncharged (logic 0) or a little charge is transferred if it was at logic 1. The amount of charge supplied by the column select lines is sensed and used to identify the original contents i.e. the cell is read. Following the read operation the contents are then restored to their original state by sourcing or sinking current from the column select lines.

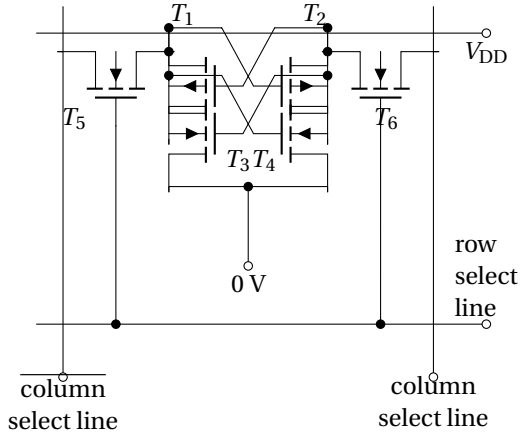


Figure 18.9: Typical static RAM circuit

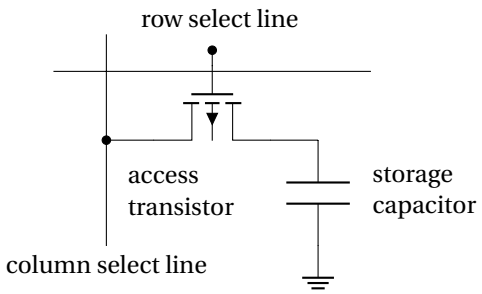


Figure 18.10: Dynamic RAM circuit

It is difficult (and therefore expensive) to fabricate large arrays of memory cells without defects and the strategy is adopted of providing spare rows and columns which can be used to replace faulty ones. This replacement is carried out by opening fusible links in the memory addressing and decoding circuits with a laser.

18.4 Analogue uses of MOSFETs

The fact that circuits made entirely of MOSFETs can be used as amplifiers should be evident from considering a CMOS inverter biased to give an output of half the power supply voltage. CMOS inverters do not make very practical opamps but circuits with the usual inverting and non-inverting inputs are manufactured using the CMOS technology and can have useful properties such as very high input resistance, low power dissipation, and sometimes odd things like a common mode input range which extends below the negative power supply voltage. See Chapter 13.

18.5 Mixed analogue and digital uses of MOSFETs

CMOS transmission gates can be used to switch analogue voltages.

18.5.1 4 channel multiplexer

A circuit to connect one of four analogue signals in succession to an analogue to digital converter (ADC) is shown in Figure 18.11. In the Johnson counter a single logic 1 circulates around a ring of four bistables.

18.5.2 Synchronous rectifier

See section 20.6.3 and section 15.10.

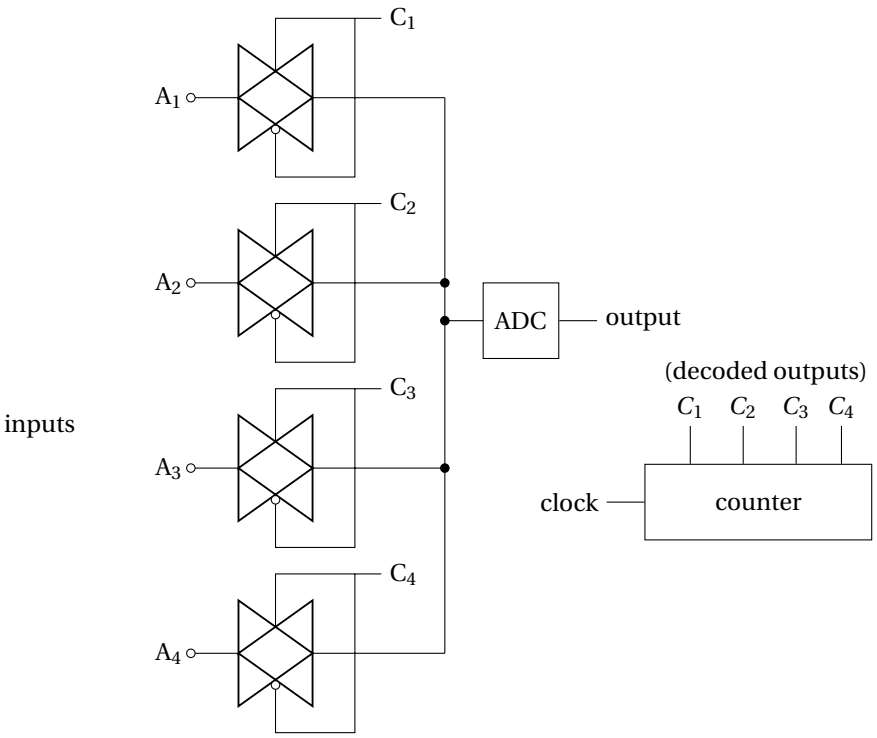


Figure 18.11: 4 channel multiplexer

19 Introduction to the Computer

19.1 Introduction

By “computer” we mean a machine which accepts and stores information (programs and data) in the form of digital words, processes the data in accordance with the programs, and transmits the results to the outside world. Even the simplest mass produced machines are too complicated to be understood in detail in two days so we have built a special machine, the EC1, to illustrate the concepts. Although it has only 13 instructions and a tiny memory, 256 bytes, it is still capable of performing useful tasks. This chapter is devoted to a description of the machine. If you need to refresh your memory of logic circuits or hexadecimal notation, see Chapter 4.

19.2 Architecture of the EC1

The machine is composed largely of 8-bit (1-byte) registers, integrated circuits (chips) containing an array of 8 bistables. The bistables in each register are connected to an 8-track pathway called the bus, see Figure 19.1. The nth bistable in a register is always associated with the nth track of the bus and can only receive a bit from or present a bit to, that track.

The registers **I**nstruction, **X**, and **O**utput can only accept a byte from the bus, **E**xternal can only transmit a byte to the bus, **A**ccumulator can receive a byte from the Arithmetic and Logic Unit **ALU** and can transmit a byte to the bus, **B**uffer has a bi-directional connection to the bus, **Y** can take its input from **A**. The memory **M** is an array of 256 1-byte registers sharing a single bi-directional connection to the bus. The program counter **PC** can receive a byte and also count up, the return address counter **RAC** can present a byte to the bus, receive a byte directly from the **PC**, and also count up. The memory address register **MA** can receive a byte either directly from the **PC** or from the bus. To provide these differing capabilities a number of different types of register chip are used. When a byte is copied from a register or from the bus to a register all the bits are transferred simultaneously (parallel operation).

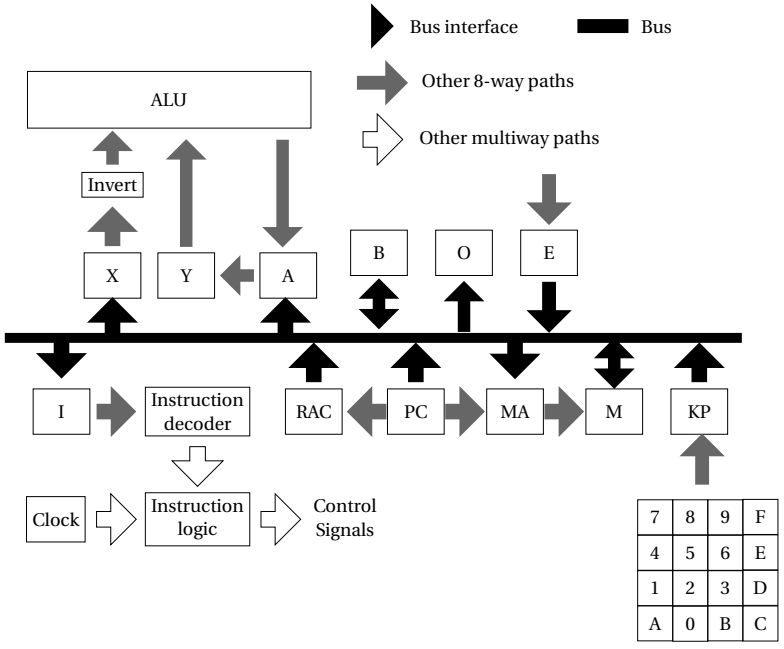


Figure 19.1: EC1 block diagram

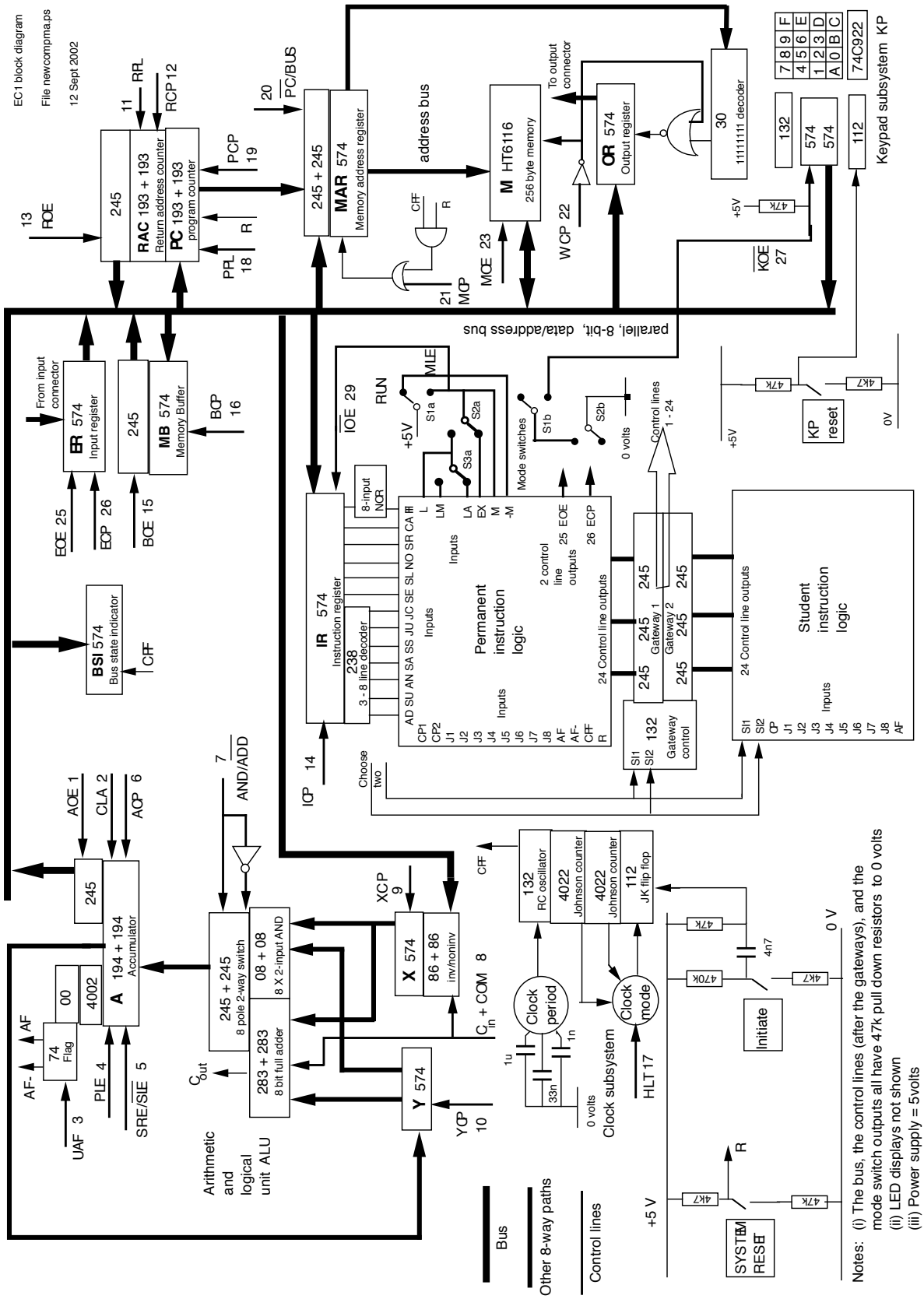


Figure 19.2: EC1 architecture diagram

Line	Block	Name	Description
1	A	AOE	Copies the contents of A (the accumulator) onto the bus when 1.
2		CLA	Sets all bits in A to zero when 1.
3		UAF	Updates A flag on ↑.
4		PLE	Enables the loading of A from the ALU output when 1, shift when 0.
5		SRE / $\overline{\text{SLE}}$	Enables shift right (to lsb end) of contents of A when 1, left when 0. (Msb or lsb respectively become 0.)
6		ACP	Loads A or shifts contents of A on ↑.
7	ALU	AND / $\overline{\text{ADD}}$	Selects AND path in ALU when 1, ADD path when 0.
8		C_{in} + COM	Sets adder carry-in to 1 and inverts all bits of X input from bus when 1.
9	X	XCP	Loads X register from bus via inverter/non inverter on. (Inverter bloack is set to invert all bits or not according to state of line 8.)
10	Y	YCP	Loads the Y register with accumulator contents on ↑.
11	RAC	RPL	Loads the RAC (return address counter) from the PC when 1.
12		RCP	Increments the contents of the RAC by 1 on ↑.
13		ROE	Copies the contents of the RAC onto the bus when 1.
14	IR	ICP	Loads IR (the instruction register) from the bus on ↑.
15	MB	BOE	Copies the contents of MB (the memory buffer) onto the bus when 1.
16		BCP	Loads the MB fromthe bus on ↑.
17	CK	HLT	Blocks oscillator pulses following a ↑.
18	PC	PPL	Loads the PC (program counter) from the bus when 1.
19		PCP	Increments the contents of the PC by 1 on ↑.
20	MAR	BUS / $\overline{\text{PC}}$	Connects memory address register input to bus when 1, to PC when 0.
21		MCP	Loads MAR on ↑.
22	M	WCP	Writes into memory location addressed on ↑ (and into OR if address is 11111111).
23		MOE	Puts contents of addressed location onto the bus when 1.
24		- not used -	
25	E	EOE	Connects output of register E to bus when 1.
26		ECP / RCP	Loads register E from an external source on ↑.
27	KP	$\overline{\text{KOE}}$	Copys the contents of the keypad register onto the bus when 0.
28		- not used -	Disconnects IR output when 1 (decoder outputs pulled down to zero).
29	IR	$\overline{\text{IOE}}$	

Table 19.1: Control line functions. ↑ means 0-1 transition (leading edge) or (narrow) clock pulse.

As well as the 8-bit connections just mentioned the registers have a number of 1-bit control inputs (shown in the more detailed figure 19.2). The transfers of bytes and other operations are orchestrated by the logic levels (or changes of level) on these inputs. E.g., by placing a logic 1 onto its relevant control input, register **A** can be told to put a copy of its contents onto the bus; then by placing a 1 on one of its control inputs register **X** can be told to accept the byte on the bus. The effect is to put a copy of the contents of **A** into **X** , overwriting (replacing)

whatever **X** held before. These are the kinds of operation that take place as instructions are executed. Of course, it must be ensured in the operation of the machine that only one register at a time puts a byte onto the bus.

19.3 The instruction set

The operations the hardware of the machine can perform when running a program are called *instructions*. A list of all such operations is called the *instructions set* of the machine. The instruction set of the EC1 is shown in Table **??**.

The instructions are of two types. Those that need one byte to specify them, and those that need two bytes. The only bytes, or the first byte in the case of a two-byte instruction, tells the machine which instruction it is, and is called the opcode. The second byte, if there is one, is an address. An example of a one-byte instruction is 01, which is the opcode for set the accumulator contents to zero. An example of a two-byte instruction is E0 03, which adds the contents of memory location 03 to the contents of the accumulator. The EC1 has eight two-byte instructions and five one-byte instructions. There are also three manually initiated instructions, which are used for loading and examining the contents of memory.

19.4 The clock subsystem

The pulses needed on the control lines to execute the instructions are constructed from these waveforms, the mode switch settings, and the output of the instruction decoder, using AND and OR logic gates. The pulses generated are either one clock period or one clock pulse long. The longer pulses, J1–J8 are used to set up pathways e.g. setting the Memory Address register to take its input from the bus, the shorter pulses (CP) are used to initiate data transfer or increment counters, e.g. to load a byte on the bus into the MA register. Generally things happen on the leading edges of the long and short pulses so it is important these are not coincident. (The drawbridge must be down before the cart is pushed over.) That is why in a given clock period the leading edges of the short pulses occur later than the leading edges of the long pulses.

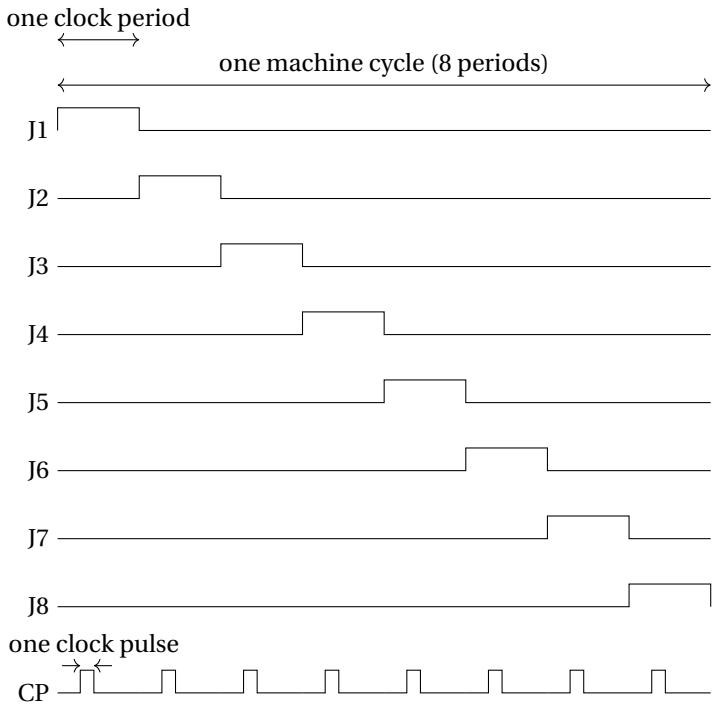


Figure 19.3: EC1 clock waveforms

Instruction bytes[1]		**Function**	**ECAL mnemonic**[8]
hex	**binary**		
EO XX	1110 0000 XXXX XXXX	Add the contents of location XX to the contents of A (the accumulator) and update AF.[2,3]	AD (add)
60 XX	0110 000 XXXX XXXX	Subtract the contents of location XX from the contents of A and update AF.[2,3,9]	SU (subtract)
C0 XX	1100 0000 XXXX XXXX	Bitwise AND the contents of A with the contents of location XX and update AF.[2,3]	AN (and)
A0 XX	1010 000 XXXX XXXX	Copy the contents of A into location XX.	SA (store accumulator)
40 XX	0100 000 XXXX XXXX	Copy the contents of the RAC (return address counter) to location XX.	SS (store subroutine return address)
10 XX	0001 0000 XXXX XXXX	Load register E from the external source and copy the contents to location XX.	SE (store external)
80 XX	1000 0000 XXXX XXXX	Jump to location XX (also sets RAC contents to 2 + address of location containing 80).	JU (jump unconditional)
20 XX	0010 0000 XXXX XXXX	Jump to location XX if AF=1 (also sets RAC contents to 2 + address of location containing 20).	JC (jump conditional)
08	0000 1000	Shift the contents of A one place to the left and update AF.[4]	SL (shift left)
04	0000 0100	Execute a machine cycle but do nothing else.[5]	NO (no operation)
02	0000 0010	Shift the contents of A one place to the right and update AF.[6]	SR (shift right)
01	0000 0001	Set the contents of A to 00 and of AF to zero.	CA (clear accumulator)
00	0000 0000	Halt.[7]	HT (halt)
Load Address		On initiate, copy the contents of the KPR (on the bus) into the MAR.	
Load Memory		On initiate, copy the contents of the KPR (on the bus) into the addressed location and then increment the MAR contents by 1.	
Examine		Put the contents of the addressed location onto the bus. On initiate, increment the MAR contents by 1.	

1 First (opcode) byte tells machine which instruction it is, second byte XX (if any) is an address.
2 The result is left in the accumulator.
3 AF is the Accumulator Flag, 1 bit formed from ORing all the accumulator bits (= 0 if contents of A are 00, = 1 otherwise). Used with an AND mask it allows any accumulator bit to be used as the flag.
4 LSB is loaded with 0, MSB is lost.
5 i.e. a one cycle wait.
6 MSB is loaded with 0, LSB is lost.
7 Stops flow of clock signals to Instruction Logic.
8 Not needed until you start writing programs.
9 Subtraction is performed by adding the 2's complement of the number in location XX.

Table 19.2: Instruction set of the EC1.

19.5 The loading process

The machine has two major modes of operation, LOAD/EXAMINE and RUN. Here we describe loading a byte in LOAD/EXAMINE mode.

The first step is to load the keypad register KP. Suppose the byte to be loaded is 10110111 (most significant bit on left). In hex notation this is B7. A stroke on the B key on the keypad loads the four bistables in the most significant half of KP with 1011. A stroke on the 7 key then loads the four bistables in the least significant half of KP with 0111.

The next thing to do is to load the byte into a memory register. Now, the 256 registers in the memory all use the same bi-directional connection to the bus so we need to be able to select the register we wish to load. This is called *addressing*.

The 256 registers on the memory chip are permanently assigned the addresses 0–255 decimal (or 00 to FF hex) and may be thought of as being in a stack. Addressing works as follows. A byte is fed to a separate address input on the memory chip and logic circuits on the memory chip use this byte to make the connection between the chosen register and the bus.

The address byte is held in the MA (memory address) register, the outputs of whose bistables are permanently connected to the address input of the memory. Addressing a given memory register therefore comes down to putting its address into the MA register. This is what the manual LA (Load Address) instruction does, it first copies the byte in the keypad register KP onto the bus then copies it from the bus into the MA register.

To load the addressed register with a byte the manual LM (load memory) instruction is used. This copies the byte in KP to the bus and then into the addressed register. As it finishes its execution LM also increments the memory address byte by 1 (see below).

Bytes can also be loaded from the outside world into a specified memory location while the machine is running a program by the execution of an SE instruction.

19.6 Loading programs and data

A program without any jump instructions (JC or JU) is always loaded into a block of consecutively numbered memory registers. The procedure is as follows:- the address of the register chosen to hold the first byte of the program is loaded, (this will usually be address 00, often called the *top of memory*). The first byte of the program is typed into KP and LM (load memory) instruction initiated. This puts the first byte of the program into memory register 00. To simplify the loading of further bytes into consecutive memory registers the LM instruction also increments the address by one after writing to memory has taken place. Any data required by the program will be loaded into separate block(s) of memory whose starting address(es) will have to be entered. A number of blocks will also be used if a program contains jump instructions.

As data and programs are simply lists of bytes, how, you might ask, does the machine know whether a byte is an opcode, an address, or data? This question is answered below.

19.7 What happens as a program runs

The address of the first byte of a program is known to the operator (because he/she loaded it). When loading is finished the memory address is set back to that of the first program byte (usually address 00), and RUN mode is initiated. The program can be run one instruction at a time or continuously. In the latter case pulses from the clock subsystem continue to flow into the Instruction Logic until a Halt instruction is encountered. (Halt stops the clock.)

Let us assume that the first instruction in a program is Clear Accumulator, a 1-byte instruction, and that this byte is in location 00 of memory. The machine assumes that the first byte of any program is an opcode and copies it (via the bus) into the instruction register IR where it is held for the duration of the instruction. As soon as the opcode is stored in the IR it passes to the instruction decoder ID which sends a signal to the instruction logic IL causing it to generate the pulses which go out to the registers involved to implement the instruction. In this case all it appears the control logic has to do is to set the contents of register A to all zeros, which simply requires a pulse to be put on the control line which goes to the reset pin on the A register chip.

Actually this is not all the control logic has to do, it also has to enable the machine to move on to and implement the next instruction whose opcode byte is in memory register 01. To get a copy of this opcode byte to pass to the instruction register it is necessary first to address register 01. This is done as follows. Referring to Figure 19.1 it can be seen that the MA register can be set up to receive a byte from either the bus or the Program Counter PC (depending on whether there is a 1 or a 0 on one of its control lines). At the time we are considering

(and indeed most of the time) it is set up to receive a copy of the byte in the PC. At the start of the execution of the CA instruction the byte in the PC and in the MA register is the address of the first byte of the program, 00 in this case. During the Clear Accumulator instruction not only is a pulse sent out to reset register A but a pulse is also sent to the PC to cause it to count up by 1 and this is followed by a pulse to the MA register telling it to accept a copy of the contents of the PC. The effect of these two extra pulses is to address location 01 as desired. Execution of the program continues with the opcode of the second instruction being passed to the IR.

Let us suppose that the second instruction is an ADd instruction. Table **??** shows this to be a 2-byte instruction which adds to the contents of the Accumulator A the data byte in the memory register whose address is the second byte of the instruction. The first byte is already in the IR. The second byte is the address of the register holding the data byte. At some time during the execution of the ADd instruction this address must occupy the MAR so that the data byte can be read. The sequence of events during execution (refer to Figure 1) is as follows:-

(a) First (opcode) byte of instruction copied to bus and then to Instruction Register

(b) PC incremented by 1

(c) PC contents copied to MAR

(d) Second (address) byte of instruction copied to bus and then to MAR

(e) Data byte in addressed memory location copied to bus and then to X

(f) Byte in A copied to Y

(g) Output of arithmetic and logic unit ALU (set to add mode by a control input) copied into A

(h) PC contents incremented again by 1

(i) PC contents copied to MAR (thereby addressing the opcode of the next instruction in the program)

Note that in order to address the location of the next instructions opcode, the PC has to increment by 1 during the execution if the current instruction consists of a single byte and by 2 if the current instruction consists of two bytes. The way the PC and the MAR work together should now be clear, the task of the MAR is to hold the relevant memory address at all times, the task of the PC is to hold the address of the next opcode. Their contents are often the same but they are different while a data byte is being read - as occurs during the ADd instruction.

19.8 Programming in machine code

It is a good idea to write down some sort of flow chart outlining what your program has to do before producing the list of instructions (code).

If the same piece of code is used more than once in a program you can economise on memory and time spent loading if you make it a subroutine, i.e. a piece of code you can jump to, execute, and return from as often as needed. The last instruction in a subroutine must be JU with an address which has been entered by an SS instruction placed at the start of the subroutine.

When writing programs you need to list at least the memory locations and the code. You may find it helpful to have a third column showing the opcode mnemonics to remind you what the instruction is.

19.9 Programming in assembly language

Working with machine code is very tedious, mainly because it is necessary to keep track of addresses when programs are modified. It would be nice to be able to write and edit programs in a more friendly “higher level” language and then get the conversion to machine code done automatically.

The first higher level language above machine language, usually called *assembly language*, allows the use of the instruction mnemonics (SU *etc.*) instead of opcodes and labels instead of numerical addresses. Programs which convert higher level languages to machine code are generally called *compilers*. In the case of an assembly language the compiler is usually called an *assembler*.

An assembler has to translate mnemonics into opcodes, calculate addresses from the positions of labels in a program, and produce machine code for the machine in question. To produce the code, assemblers usually make two passes through an assembly language program. On the first pass the labels are listed and their addresses worked out. On the second pass the label references are replaced with the addresses and the instruction mnemonics with opcodes. Any data used by the program is usually placed immediately after the halt instruction.

Assembly language programs are written in plain text and the editor and the assembler are usually installed on a support machine. The assembler is run after each edit and the machine code output is then downloaded from the support machine to the target machine.

19.10 ECAL language definition

The assembly language for the EC1 machines, called ECAL (Educational Computer Assembly Language), was originally defined by E J Williamson. The current version was written by Hiro Yamazaki.

19.10.1 Numbers

Numbers may be written in hex or decimal. Hex numbers are preceded by 0x (zero x) e.g. `0xE6`, decimal numbers have no prefix and must not have leading zeros.

19.10.2 Instructions

The two-character string mnemonics in the left hand column of Table **??** are used instead of the opcodes, and label references (see below) are used instead of addresses. Instructions are written one to a line, (i.e. terminated by a return).

An example of a one-byte instruction i.e. one without a label reference, is:

```
CA
```

Where there is a label reference (two byte instructions) it is separated from the mnemonic by one or more spaces. e.g.:

```
JU begin
```

19.10.3 Labels and label references

A label consists of a string of up to 100 characters the first of which must be a letter, followed by a colon (:). Allowed characters are the letters `A`–`Z` (either case), the digits `0`–`9`, and underscores (`_`) which can be used as spacers instead of actual space characters which are illegal. Note also that `OR` is a label reserved for the output register.

Labels are placed first on a line. Label references have the same characters as the label but no colon. Following a label reference with `+1` references the next memory location after the labelled location. (This feature is used to reference the second byte of a two byte instruction.)

19.10.4 Setting the start address

The statement `LOC=n` (return) at the beginning of a program forces the first address to be `n`. If no such statement is made a starting address of `00` is assumed.

19.10.5 Comments

Programs should be sprinkled with enough comments to make it clear what they do. Comments are ignored by the assembler. A comment is preceded by a semi-colon (`;`) and terminated with a return. (Inserting a semi-colon at the beginning of a line is a convenient way of disabling an instruction without deleting it.)

19.10.6 Style

To make programs easy to read the following layout rules are suggested. (ECAL is case insensitive and tabs are ignored by the assembler.)

- Upper case for instruction mnemonics
- Lower case for labels and label references
- No tabs for labels and comments
- One tab for instruction mnemonics and data
- Two tabs for label references
- Three or more tabs for subroutines

19.11 Examples of style and the use of labels

```
; Put a program's first instruction, clear accumulator, into memory location hex 3E.
loc=0x3E
        CA
        -
; Increment and then decrement the accumulator contents by 1 then mask with 1
        AD      one
        SU      one
        AN      one
        -
one:    1       (or 0x01)
        -
```

(The `one:` is a label, the ones are label references. The contents of the labelled location are 1, that's why we chose one for the label.)

```
; Jump unconditionally to a SR instruction
        JU      there
        -
        -
there:  SR
; Store accumulator contents at label
        SA      box
        -
        -
box:    0x00
```

(Something must be put into the location labelled `box:` before the return, we have entered `0x00`. It will be overwritten with the accumulator contents.)

```
; conditional jump to subroutine and return
JC      subrt
        subrt:  SS      addju +1
                -
                -
                -
        addju:  JU      0x00
```

(The address in the `JU` instruction, initially loaded with `0x00`, will be overwritten)

19.12 Error messages

Error messages indicating the location of the error are generated in the following situations:

- Illegal character anywhere in the program
- Label referred to which is undefined
- Duplicate label
- LOC not followed by =
- Numerical address detected

20 Noise

20.1 Introduction

Discounting interference, the most significant processes giving rise to noise in circuits are:

(a) random thermal motion of charge carriers in resisting conductors giving rise to fluctuating potentials (Johnson, Nyquist or thermal noise)

(b) random emission of charge carriers over a potential barrier e.g. at a PN-junction or emissive cathode (shot noise)

(c) generation and recombination of charge carriers in semiconductors (gr noise)

(d) slow variations in effects influencing currents e.g. charge carrier traps on the surfaces of semiconductors (flicker noise).

20.2 Discussion of noise sources

20.2.1 Johnson noise

Looking into its terminals, a short $\left(\ll \lambda = c/f\right)$ dipole aerial of length δl looks like a real impedance R_r (its radiation resistance) given by:

$$R_r = \frac{2\pi\delta l^2 f^2}{3\epsilon_0 c^3} \qquad \text{(see e.g. Bleaney 8.9)} \tag{20.1}$$

Consider such a dipole connected to a resistance R via a lossless narrow bandpass filter F of width Δf at f the whole assembly being in a cavity with isothermal walls at temperature T. A perfectly reflecting radiation shield S surrounds the filter and the resistance.

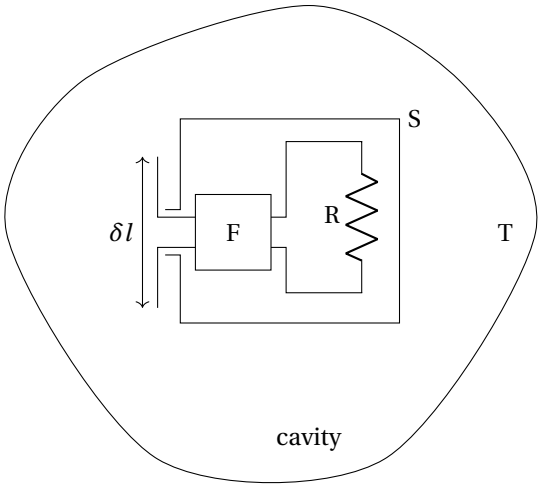


Figure 20.1: Johnson noise source

The mean power originating in the resistance which is radiated from the aerial is:

$$\overline{\left\{V(f,\Delta f)\right\}^2}\left(\frac{R_r}{R+R_r}\right)^2\frac{1}{R_r} \tag{20.2}$$

where $\overline{\{V(f,\Delta f)\}^2}$ is the mean square noise emf developed in R that we seek.

The x-component of the electric field in the cavity due to the Planck radiation in a bandpass Δf at f is $E_x(f,\Delta f)$. The emf generated in the dipole (oriented along the x-axis) is then $E_x(f,\Delta f)\delta l$ and the average power dissipated in the resistance is:

$$\overline{\{E_x(f,\Delta f)\delta l\}^2}\left(\frac{R_r}{R+R_r}\right)^2\frac{1}{R_r} \tag{20.3}$$

In thermal equilibrium these must be equal so:

$$\overline{\{V(f,\Delta f)\}^2}R_r = \overline{\{E_x(f,\Delta f)\delta l\}^2}R \tag{20.4}$$

Now for isotropic cavity radiation the energy density:

$$U(f,\Delta f) = 3\epsilon_0\overline{\{E_x(f,\Delta f)\delta l\}^2} \tag{20.5}$$

so:

$$\overline{\{V(f,\Delta f)\}^2} = U(f,\Delta f)\frac{\delta l^2}{3\epsilon_0}\frac{1}{R_r}R \tag{20.6}$$

Substituting the Planck function for U and inserting the expression for R_r yields

$$\overline{\{V(f,\Delta f)\}^2} = \frac{8\pi f^3\Delta f}{c^3\left(e^{hf/kT}-1\right)}\frac{\delta l^2}{3\epsilon_0}\left\{\frac{2\pi\delta l^2 f^2}{3\epsilon_0 c^3}\right\}^{-1}R \tag{20.7}$$

which reduces to:

$$\overline{\{V(f,\Delta f)\}^2} = \frac{4Rhf\Delta f}{e^{hf/kT}-1} \tag{20.8}$$

At room temperature $hf/kT = 1$ at $f = 6\times10^{12}$ Hz so at circuit frequencies hf/kT is small and:

$$\overline{\{V(f,\Delta f)\}^2} = 4kTR\Delta f \tag{20.9}$$

which is proportional to the bandwidth but independent of f. Such noise is described as *white*.

This derivation shows, as did the original transmission line treatment by Nyquist, the close connection between black body radiation and Johnson noise. The result does not depend on the mechanism of conduction so long as it is linear in the applied voltage at low excitation levels. Also the result would be the same if the conduction was due to a charged fluid continuum rather than discrete charge carriers.

20.2.2 Shot noise

Shot noise is the name given to the fluctuations in the electric current due to a stream of non-interacting charged particles that have surmounted a potential barrier. Such a stream may be represented by a random sequence of short current pulses (impulses) carrying change q, often $q = -|e|$. The most important devices exhibiting shot noise are the bipolar transistor, the vacuum photocell under typical illumination and the temperature limited thermionic diode.

(i) Random sequence of impulses applied to linear system defined by an impulse response function

We make use of Campbell's theorem (1909). If a device responds to an ideal impulse (δ function) of charge of magnitude q at $t = 0$ by a response $R(t) = qF(t)$ then its mean response to a random series of such impulses occurring at a mean rate ν is

$$\overline{R} = \nu q\int_{-\infty}^{\infty}F(t)\,\mathrm{d}t \tag{20.10}$$

with fluctuations:

$$\overline{\delta R^2} = \nu q^2\int_{-\infty}^{\infty}F^2(t)\,\mathrm{d}t \tag{20.11}$$

For proof of Campbell's theorem see e.g. Robinson, *Noise and Fluctuations*.

(ii) Random sequence of impulses applied to linear system defined by a frequency response

The impulse response $F(t)$ is related to the frequency response $G(f)$ by the Fourier transform

$$F(t) = \int_{-\infty}^{\infty} G(f)\, e^{j2\pi f t}\, \mathrm{d}f \tag{20.12}$$

This allows us to use the Fourier integral theorem (Parseval's form):

$$\int_{-\infty}^{\infty} F^2(t)\, \mathrm{d}t = \int_{-\infty}^{\infty} G(f)\, G^*(f)\, \mathrm{d}f \tag{20.13}$$

to express the second part of Campbell's theorem in terms of frequency response:

$$\overline{\delta R^2} = \nu q^2 \int_{-\infty}^{\infty} G(f)\, G^*(f)\, \mathrm{d}f \tag{20.14}$$

Since $F(t)$ is real $\left(G^*(f) = G(-f)\right)$, $G^*(f)\, G(f)$ is an even function of f so we may write, noting that νq is the mean current I_0,

$$\overline{\delta R^2} = 2qI_0 \int_{0}^{\infty} G(f)\, G^*(f)\, \mathrm{d}f \tag{20.15}$$

If the frequency responses are dimensionless the mean square fluctuation in the response is simply δi^2 so

$$\overline{\delta i^2} = 2qI_0 \int_{0}^{\infty} G(f)\, G^*(f)\, \mathrm{d}f \tag{20.16}$$

For $G(f)$ a square bandpass of width Δf this reduces to

$$\overline{\delta i^2} = 2qI_0 \Delta f \tag{20.17}$$

For ideal impulses the noise is therefore white. We see that, unlike Johnson noise, the fluctuations are directly proportional to the charge carried by the particles. For finite width impulses it can be shown that the noise spectrum cuts off in the neighbourhood of the reciprocal of the width.

Shot noise can be partially suppressed e.g. by a space charge around the cathode in a thermionic vacuum valve. (A space charge is not allowed to form in thermionic diodes used as standard noise sources.)

20.2.3 Generation recombination (gr) noise

When a current is flowing in a uniform semiconductor an increase of noise above the Johnson level is seen. The origin of some of this extra noise is the random thermal generation and recombination of charge carriers. When the distance travelled by a carrier with average lifetime is much less than the distance between the electrodes gr noise can be shown to have the form:

$$\frac{\overline{\delta i^2}}{I^2} = \frac{4\tau \Delta f}{N\left(1 + 4\pi^2 f^2 \tau^2\right)} \tag{20.18}$$

where τ is the carrier lifetime and N is the number of free charge carriers. The noise is white up to a frequency of the order of the reciprocal of the lifetime.

20.2.4 Flicker noise

Flicker noise is the name given to excess noise which increases at low frequencies and is found in almost all the devices used in electronics and optoelectronics when a current is flowing. Typically it is represented by:

$$\overline{\delta i^2} = \text{const.} \frac{i^x}{f^y} \Delta f \tag{20.19}$$

The theory of such noise is not well developed.

20.3 Noise in some devices

20.3.1 Resistors

Johnson noise is generated in all resistors and loss mechanisms representable by resistances. Semiconductor resistors in addition exhibit gr noise, carbon composition resistors (now thankfully obsolete) exhibited enormous flicker noise. For such resistors it was found that $y \approx 1$ in the expression above, explaining the name 'one-over-f noise' by which this excess noise is commonly known, x typically lies in the range 2–2.5. For small carbon resistors the noise was shown to be 'one-over-f' down to 1×10^{-3} Hz (Rollin and Templeton at the Clarendon in 1953) It is negligible in modern metal film resistors.

Flicker noise is often associated with the condition of surfaces. The frequency at which $1/f$ noise rises above the white noise in a particular system is called the '$1/f$ knee frequency'. It ranges from less than 1 Hz in some JFETs to a few hundred Hz in some infrared photoconductors to a few MHz in point contact semiconductor diodes.

Either voltage or current generators may be used to represent the noise as shown below. We use square symbols to depict the generators to remind us that the voltages and currents are mean squares.

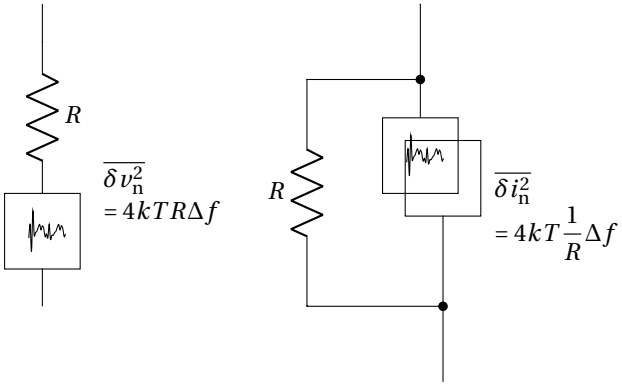


Figure 20.2: Representations of resistor noise

20.3.2 Lossless inductors and capacitors

Consider an inductor with a resistor at the same temperature connected between its terminals. If the equivalent circuit of the inductor contains no resistances the mean square noise current $\overline{\delta i^2}$ flowing through it (due to the Johnson noise in the resistor) will store energy $0.5L\overline{\delta i^2}$ in its magnetic field but cannot heat it. Therefore, such an inductor can have no noise generator associated with it (because if it had it would cool down as the generator dissipated power in the resistor contrary to the second law of thermodynamics). For the same reason, lossless capacitors also have no noise generator associated with them.

20.3.3 JFETs

The principle noise source is Johnson noise in the channel, also present are a noise current in the gate lead arising from and partially correlated with the Johnson noise in the channel and shot noise on the gate-channel diode leakage current.

If the source end of the channel is connected to 0 volts there will be a fluctuating voltage (Johnson noise) in the channel increasing in magnitude towards the drain end. Expressing the total noise voltage appearing across the gate junction (assuming the voltage of the gate terminal does not fluctuate) in terms of an effective resistance R_1 we obtain:

$$\overline{\delta v_1^2} = 4kTR_1\Delta f \tag{20.20}$$

It turns out that R_1 is approximately equal to $1/g_m$ so it is the unpinched off part of the channel which generates most of the noise. We represent it by a voltage generator in series with the gate lead.

The fluctuating voltages in the channel drive a fluctuating current through the gate channel capacitance C_g and the impedance Z_s of the signal source. The effective noise voltage driving the current results from a different average over the channel length and we may express it as:

$$\overline{\delta v_2^2} = 4kTR_2\Delta f \tag{20.21}$$

Detailed calculation (see e.g. Robinson) shows that R_2 is a little less than $1/g_\mathrm{m}$. The different average reduces the correlation of the noise current in the gate lead with the Johnson noise in the channel to the point where it is usually taken to be uncorrelated.

$$\overline{\delta i_{g_J}^2} = 4kTR_2\Delta f \frac{4\pi^2 f^2 C^2}{1+4\pi^2 f^2 C^2 Z_\mathrm{s}} \tag{20.22}$$

The reverse bias applied to the gate-channel junction of the FET in normal operation as an amplifier gives rise to a small dc current ($I_\mathrm{g} = I_0$ of the diode). This shows full shot noise so there is an additional noise current in the gate lead of:

$$\overline{\delta i_{g_s}^2} = 2|e|\, I_\mathrm{g}\Delta f \tag{20.23}$$

This is usually greater than $\overline{\delta i_{g_J}^2}$ at frequencies below 1 MHz. The disposition of the noise generators in the equivalent circuit (Figure 11.8) is shown in Figure 20.3.

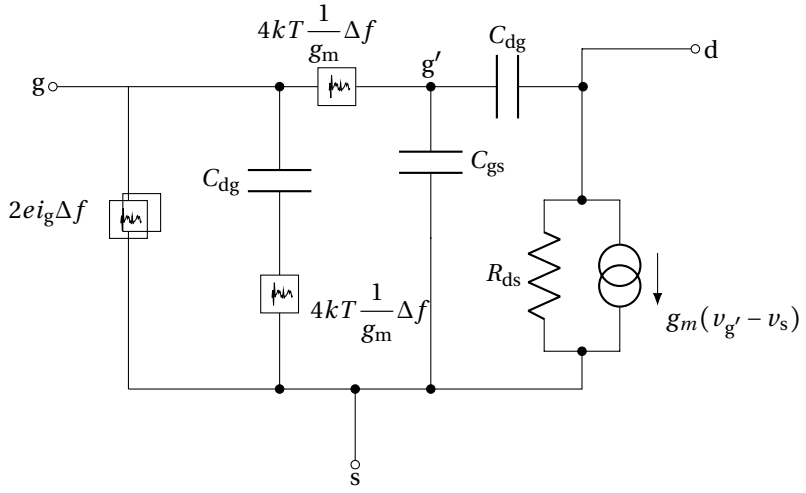


Figure 20.3: JFET noise

20.3.4 Diffusion transistors at low frequencies

The collector current arises from the forward current crossing the emitter-base junction and diffusing to the collector. It shows full shot noise and is given by

$$\overline{\delta i_c^2} = 2|e|\, I_\mathrm{c}\Delta f \tag{20.24}$$

$$= 2kTg_\mathrm{m}\Delta f \tag{20.25}$$

We represent it by a mean square voltage $\left(2kT/g_\mathrm{m}\right)\Delta f$ in series with the base lead. (Because of the insensitivity of the collector current to collector-base voltage this noise current is unaffected by the noise voltage appearing at the collector.)

The part of the base resistance $r_{\mathrm{bb}'}$ is a source of Johnson noise $4kTr_{\mathrm{bb}'}\Delta f$.

The mean base current arises from recombination in the base and the current flow across the emitter-base junction due to the non-unity emitter efficiency. It shows full shot noise which is uncorrelated with the noise in the collector current. Its mean square is given by

$$\overline{\delta i_b^2} = 2|e|\, I_\mathrm{b}\Delta f \tag{20.26}$$

$$= \frac{2kTg_\mathrm{m}}{\beta}\Delta f \tag{20.27}$$

This noise current affects the output by developing a voltage across the input impedance in parallel with the $r_{\mathrm{bb}'}$ in series with the impedance of the signal source. In Figure 20.4 the noise generators are shown added to the equivalent circuit (Figure 10.8).

20.4 Noise calculations

There is no particular difficulty in dealing with uncorrelated noise generators, some examples should make this clear.

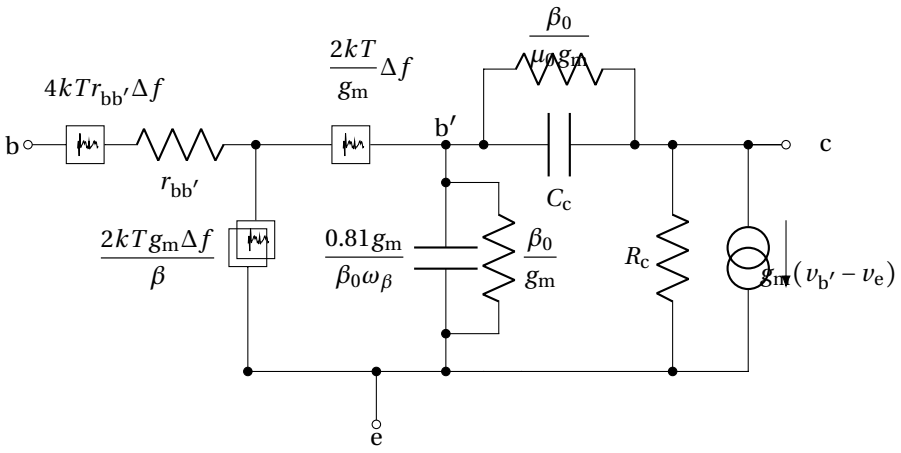


Figure 20.4: Diffusion transistor with noise sources.

(i) Two resistances in parallel

Consider two resistances in parallel as shown in Figure 20.5. Then

$$\overline{V_n^2} = \left(\frac{R_2}{R_1+R_2}\right)^2 4kT_1R_1\Delta f + \left(\frac{R_1}{R_1+R_2}\right)^2 4kT_2R_2\Delta f \tag{20.28}$$

The potential divisions appear squared because we are dealing with the squares of voltages. When the resistances are at the same temperature this reduces to:

$$\overline{V_n^2} = 4kT\frac{R_2R_1}{R_1+R_2}\Delta f \tag{20.29}$$

the Johnson noise due to the two resistances in parallel.

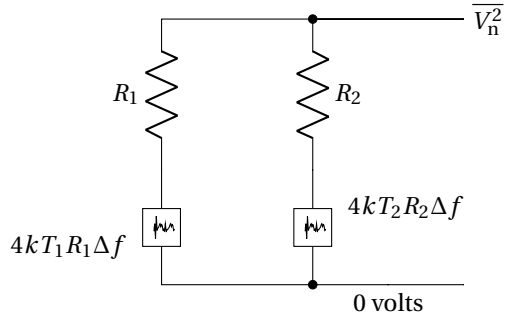


Figure 20.5: Noise for two resistances in parallel

(ii) Common emitter amplifier stage at low frequencies

To see how to maximise a signal to noise ratio (the usual reason for doing a noise calculation) it is necessary to derive expressions for the square of the signal voltage and the mean square noise voltage at a suitable point in the circuit. As as example we treat the low frequency common emitter amplifier a fragment of which is shown in Figure 20.6.

The square of the signal voltage at b′ is:

$$\left\{\frac{\frac{\beta_0}{g_m}}{R_S+r_{bb'}+\frac{\beta_0}{g_m}}\right\}^2 v_S^2 \tag{20.30}$$

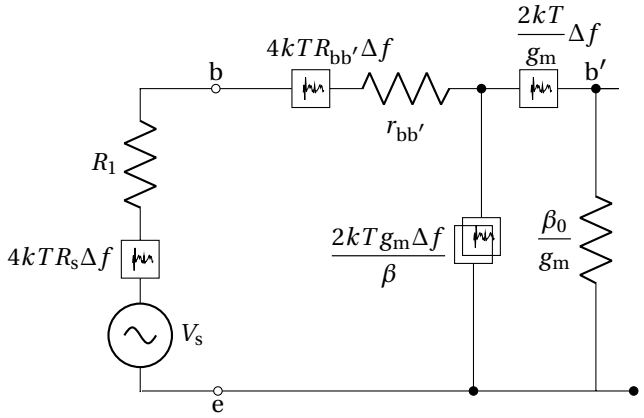


Figure 20.6: Common emitter amplifier stage at low frequencies

and the mean square noise voltage at b′ is:

$$\left[4kTr_{bb'}+\frac{2kT}{g_m}+\frac{2kTg_m}{\beta}\left\{\frac{(r_{bb'}+R_S)\,\frac{\beta_0}{g_m}}{R_S+r_{bb'}+\frac{\beta_0}{g_m}}\right\}^2+4kTR_S\left\{\frac{\frac{\beta_0}{g_m}}{R_S+r_{bb'}+\frac{\beta_0}{g_m}}\right\}^2\right]\Delta f \tag{20.31}$$

The quotient of these is the *signal to noise ratio* (SNR), it is usually expressed in decibels (2.5.3).

Another measure of the performance of an amplifier is the *noise figure, F*, defined as the ratio of the total noise power to the noise power originating from the signal source alone (perfect is $F = 1$ or 0 dB).

20.5 Aerial temperature

Referring back to Figure 20.1 we see that a perfectly conducting short aerial in a cavity at temperature T behaves as a Johnson noise source at temperature T with a resistance equal to its radiation resistance. (The actual temperature of the aerial is immaterial.) This is important for communication via satellites. The path to and from the satellite passes through the Earth's atmosphere which is significantly absorbing (and hence emitting) in some regions of the microwave spectrum and is also hot at high altitudes. The noise received is described in terms of an effective *aerial temperature*.

For these reasons communication channels are positioned in spectral regions of low atmospheric absorption to allow the antenna as unrestricted a view to cold space (4 K) as possible. The aerial's view is limited to a narrow beam (by providing a parabolic reflector) to prevent it seeing the warm Earth, this also greatly increases its gain in the beam direction.

20.6 Extraction of signals from noise

20.6.1 Signal much larger than the noise

Consider an ac signal at frequency f with unknown phase and with amplitude variable but large compared with the rms noise in the bandpass Δf defined by a filter centred on f, see Figure 20.7. To discover the amplitude of the signal and how it varies with time the signal and noise are fed into a half wave rectifier (Figure 15.16) in which the smoothing time constant RC is made as long as possible without it smoothing out the highest frequency variations of amplitude of the signal which are of interest. Usually $2/RC \ll \Delta f$.

Considered in the time domain the rectifier produces an output close to the peak value of the input signal with superimposed ripple due to the noise and the positive peaks of the signal. Viewed in the frequency domain the major components in the output due to the signal are at dc and at frequencies associated with the amplitude variation. The output due to the noise lies in the frequency band 0–$1/RC$ and arises from the mixing the signal with the noise components in the input which are within $\pm 1/RC$ of the signal frequency (but note that noise components in quadrature with the signal do not contribute to the output). The output due to noise components further away from f are filtered out by the time constant.

Input noise components present in the rest of Δf are also mixed with each other and give rise to outputs in the frequency range 0–$1/RC$ but they are of second order in magnitude compared with those that mix with the signal. The effective bandwidth of the whole system for noise calculations is $\approx 1/RC$, independent of Δf.

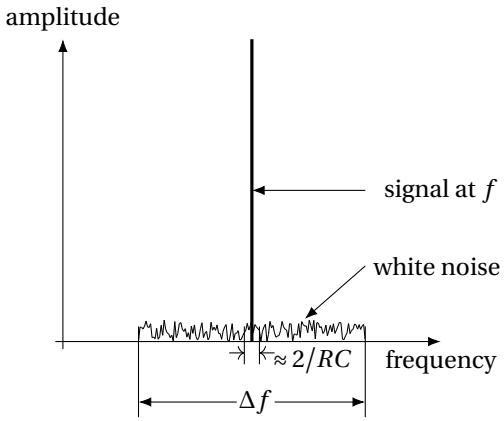


Figure 20.7: Spectrum of a large signal and white noise

20.6.2 Signal comparable with or smaller than the noise

Consider now a signal as above but with an amplitude comparable with that of the noise. The difference components of the input noise at frequencies greater than $1/RC$ from f, previously negligible, are now significant. The noise bandwidth of the system depends not only on $1/RC$ but also on Δf. Typically the noise bandwidth is $\sqrt{\Delta f/RC}$.

20.6.3 Lock-in system

The situation in 20.6.2 can be improved on that in 20.6.1 if an additional large signal of constant amplitude and in phase with the signal (called the reference) is available. This is the arrangement referred to in section 15.10.

21 Derivation of Kirchhoff's Laws

21.1 Theoretical Basis

21.1.1 The laws of classical electromagnetism

The force **F** on a charge q subjected to an electric field **E** and a magnetic field **B** is given by:

$$\mathbf{F} = q\left(\mathbf{E} + \mathbf{v} \times \mathbf{B}\right) \tag{21.1}$$

where **v** is the velocity of the charge. The fields **E** and **B** satisfy Maxwell's equations:

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0} \tag{21.2}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{21.3}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{21.4}$$

$$c^2 \nabla \times \mathbf{B} = \frac{\mathbf{j}}{\epsilon_0} + \frac{\partial \mathbf{E}}{\partial t} \tag{21.5}$$

$$\tag{21.6}$$

where ρ is the (total) charge per unit volume and **j** is the (total) current per unit area.

21.1.2 The macroscopic Maxwell equations

The charge density ρ is a field which is zero everywhere except within the sharp peaks associated with the charged particles. The **E** and **B** fields have corresponding complexity at the inter-particle scale. If we give up our interest in what happens at this scale and spatially smooth (instantaneously) all the fields over volumes a suitable number of inter-particle spacings across, the *macroscopic* Maxwell equations are obtained. If, in addition, we add the fact that all the charged particles can be classified as either bound or free, the smoothed charge and current densities can be expressed (correct to dipole terms) as:

$$\rho_{\text{free}} + \rho_{\text{fixed}} - \nabla \cdot \mathbf{P} \tag{21.7}$$

and

$$\mathbf{j}_{\text{free}} + \frac{\partial \mathbf{P}}{\partial t} + \nabla \times \mathbf{M} \tag{21.8}$$

respectively where ρ_{free} and ρ_{fixed} are the smoothed free and fixed charge densities (which are equal and opposite in the interior of a good conductor), **P** is the polarisation, $\mathbf{j}_{\text{free}}$ is the smoothed free current density, and **M** the magnetisation.

In what follows all the fields are to be considered as spatially smoothed.

21.1.3 E and B in terms of potentials

It follows from the third and second macroscopic Maxwell equations that it is possible to write

$$\mathbf{B} = \nabla \times \mathbf{A} \tag{21.9}$$

$$\text{and} \quad \mathbf{E} = \pm\nabla\phi - \frac{\partial \mathbf{A}}{\partial t} \tag{21.10}$$

where **A** is a macroscopic (spatially smoothed) vector field (the magnetic potential) and ϕ is a macroscopic scalar field (the electric potential). The convention is adopted that work must be done by an external force to move a positive charge to a region of higher potential. This requires that the − sign be taken in eqn 21.10.

21.1.4 The potentials in terms of the charge and current distributions

Choice of the Lorentz gauge:

$$\nabla \cdot \mathbf{A} = -\frac{1}{c^2}\frac{\partial \phi}{\partial t} \tag{21.11}$$

leads to similar second order differential equations for the macroscopic potentials ϕ and $\mathbf{A}$ which have solutions:

$$\phi(\mathbf{r};t) = \frac{1}{4\pi\epsilon_0}\int_{\text{all space}} \frac{\rho\left(\mathbf{r}', t-\left|\mathbf{r}-\mathbf{r}'\right|/c\right)}{\left|\mathbf{r}-\mathbf{r}'\right|}\,\mathrm{d}^3\mathbf{r}' \tag{21.12}$$

and:

$$\mathbf{A}(\mathbf{r};t) = \frac{1}{4\pi\epsilon_0 c^2}\int_{\text{all space}} \frac{\mathbf{j}\left(\mathbf{r}', t-\left|\mathbf{r}-\mathbf{r}'\right|/c\right)}{\left|\mathbf{r}-\mathbf{r}'\right|}\,\mathrm{d}^3\mathbf{r}' \tag{21.13}$$

The macroscopic potentials at $(\mathbf{r},t)$ are determined by the *retarded* values of ρ and $\mathbf{j}$, that is the values they had at the (earlier) time $t-\left|\mathbf{r}-\mathbf{r}'\right|/c$. This is because, travelling at the speed of light, it takes a time $\left|\mathbf{r}-\mathbf{r}'\right|/c$ for information about a charge at $\mathbf{r}'$ to reach the point $\mathbf{r}$.

The $\mathbf{E}$ and $\mathbf{B}$ fields derived from the potentials have parts that remain in the vicinity of the sources ρ and $\mathbf{j}$ (induction fields) and parts that are travelling waves which can carry energy away from the sources (radiation fields). Mathematically the radiation fields arise from the retardation $\left|\mathbf{r}-\mathbf{r}'\right|/c$ in the expressions for the potentials and disappear if it is ignored.

21.2 Circuits

Electromagnetic energy supplied to an arrangement of conductors, insulators, and magnetic material is either stored, degraded into heat, or radiated away. We define a circuit as *an arrangement which can be adequately described without allowing for radiation.*

As mentioned above, radiation fields arise from retardation and ignoring them is permissible only if it is permissible to ignore the retardation. This is the case if the extent of the arrangement is very much smaller than the wavelength of radiation at the frequency being considered, or there is something special about the arrangement which makes any significantly retarded fields negligible. In the first case, given the dimensions of real hardware, this means systems which are excited at frequencies less than a few GHz (wavelengths longer than a few cm).

The most striking example of the second case is the transmission line, which can be treated successfully as a circuit (via the telegraph equations). Transmission lines are almost by definition more than a wavelength long so some of the contributions to the fields at a particular cross section will arise from charges and currents which are more than a wavelength away. However the charges and currents at any point along a line are equal and opposite on the two conductors and their fields are dipole fields which fall off rapidly with distance becoming negligible in magnitude before their retardation becomes significant.

21.3 Linear conduction

In most conducting materials the current density $\mathbf{j}_{\text{free}}$ exhibits a very close to linear dependence on the force $\mathbf{F}$ applied to the mobile charges (usually electrons). Further, this near-linear dependence is usually found to be nearly *isotropic* (same in all directions in the material), and nearly *homogeneous* (same at all places in the material) as well.

Given this observed behaviour, the theoretical concept of a conductor with a precisely linear, isotropic, and homogeneous (LIH) dependence of $\mathbf{j}_{\text{free}}$ on $\mathbf{F}$ will be useful. We express such a model as:-

$$\mathbf{j}_{\text{free}} = \frac{\sigma}{q}\mathbf{F} \qquad \text{(LIH conductor, Ohms law)} \tag{21.14}$$

where σ is a positive scalar constant called the (electrical) *conductivity* and q is the charge carried by each mobile particle.

We picture the conduction as follows. The effect of F alone would be to accelerate the charges and cause a current which increased continuously. That observed currents are steady we attribute to the momentum being gained by the charges from the action of F being destroyed at the same rate by collisions with the fixed atoms of the material. The proportionality arises from the details of the mechanisms by which the momentum is gained and lost.

21.4 Kirchhoff's laws in stationary circuits with linear conductors

21.4.1 Kirchhoff's first law

From the macroscopic Maxwell equations the conservation equation for the free charge ρ_{free} is easily derived as

$$-\int_{\mathbf{S}} \mathbf{j}_{\text{free}} . \mathrm{d}\mathbf{S} = \int_{\tau} \frac{\partial \rho_{\text{free}}}{\partial t} \mathrm{d}\tau \tag{21.15}$$

where $\mathbf{S}$ is the surface of the volume τ. This may be written:

$$\sum_{r=1}^{n} I_r = \frac{\partial q}{\partial t} \tag{21.16}$$

where I_r is one of the n currents entering τ and q is the charge in τ.

This statement of the conservation of charge is the general form of Kirchhoff's first law. When the charge q is negligible (or constant) it states that the sum of the currents entering τ is zero.

21.4.2 Kirchhoff's second law

When attempting to work out the behaviour of an electromagnetic system using Maxwell's equations, it is not usually helpful to imagine the system divided into driving and driven parts - we have to seek self consistent solutions for the whole system. In the realm of circuits however it is often possible to identify a part of a system which has a strong influence on the rest but is not itself significantly influenced in return, e.g. a system consisting of a signal generator and an amplifier under test. We can take a cause and effect view.

In this spirit we make explicit the contribution by the driving part of a system to the forces exerted on the charges in the driven part. We express the force on each mobile charge in the driven part, which we shall refer to as the circuit, as the sum of the external applied force and the force due to all the charge and current in the circuit (except that due to itself), i.e.

$$\mathbf{F}(\mathbf{r}, t) = \mathbf{F}_{\text{ex}}(\mathbf{r}, t) + \mathbf{F}_{\text{ct}}(\mathbf{r}, t) \tag{21.17}$$

The average drift velocity of the charges in the circuit, $\mathbf{v}_{\text{drift}}(\mathbf{r}, t)$, is given by $\mathbf{j}_{\text{free}}(\mathbf{r}, t) = \text{n}(\mathbf{r})\, \mathbf{v}_{\text{drift}}(\mathbf{r}, t)$ where $n(\mathbf{r})$ is the number of mobile charges per unit volume. Then following from section 21.3 we can write:

$$\frac{n(\mathbf{r})\, \mathbf{v}_{\text{drift}}(\mathbf{r}, t)}{\sigma(\mathbf{r})} = \frac{\mathbf{F}_{\text{ex}}(\mathbf{r}, t) + \mathbf{F}_{\text{ct}}(\mathbf{r}, t)}{q} \tag{21.18}$$

Now:

$$\frac{\mathbf{F}_{\text{ct}}}{q} = \mathbf{E}_{\text{ct}} + (\mathbf{v}_{\text{drift}} + \mathbf{v}_{\text{cond}}) \times \mathbf{B}_{\text{ct}} \tag{21.19}$$

where $\mathbf{E}_{\text{ct}}$ and $\mathbf{B}_{\text{ct}}$ are the electric and magnetic fields at $\mathbf{r}$ due to the charge and current in the circuit and $\mathbf{v}_{\text{cond}}$ is the contribution to the velocity of the charges from any motion of the conductor. (We have assumed that the random thermal motion of the charges gives rise to effects only at the level of Johnson noise, see 20.2, and have ignored it.)

Using:

$$\mathbf{E}_{\text{ct}} = -\nabla \phi_{\text{ct}} - \frac{\partial \mathbf{A}_{\text{ct}}}{\partial t} \tag{21.20}$$

yields:

$$\frac{\mathbf{F}_{\text{ex}}}{q} = \frac{\partial \mathbf{A}_{\text{ct}}}{\partial t} - (\mathbf{v}_{\text{drift}} + \mathbf{v}_{\text{cond}}) \times \mathbf{B}_{\text{ct}} + \frac{\mathbf{j}_{\text{free,ct}}}{\sigma} + \nabla \phi_{\text{ct}} \tag{21.21}$$

Next we restrict the discussion to stationary circuits ($\mathbf{v}_{\text{cond}} = 0$) and form the line integral of this equation from point a to point b in the circuit along a path $\mathbf{l}$ which is parallel to $\mathbf{v}_{\text{drift}}$ (and therefore to $\mathbf{j}_{\text{free}}$). For such paths the term in $\mathbf{v}_{\text{drift}} \times \mathbf{B}_{\text{ct}} . \mathrm{d}\mathbf{l}$ disappears (because $\mathbf{v}_{\text{drift}}$ is along $\mathrm{d}\mathbf{l}$) and we have:

$$\int_a^b \frac{\mathbf{F}_{\text{ex}} \cdot \mathrm{d}\mathbf{l}}{q} = \int_a^b \frac{\partial \mathbf{A}_{\text{ct}}}{\partial t} \cdot \mathrm{d}\mathbf{l} + \int_a^b \frac{\mathbf{j}_{\text{free,ct}}}{\sigma} \cdot \mathrm{d}\mathbf{l} + (\phi_b - \phi_a)_{\text{ct}} \tag{21.22}$$

This is the general form of Kirchoff's second law for circuits with linear conductors. It describes a balance between (on the left hand side) a measure of the external forces applied to a circuit and (on the right hand side) a measure of the response of the circuit.

Each term in (21.22) has the units work/charge, a combination which occurs frequently in circuits and is given the name Volt (after Alessandro Volta, who in 1802 made the first battery and first produced continuous currents). The left hand side is the work done by the source of the applied force when it causes unit charge to traverse the path ab. It is called the *applied electromotive force* or *emf*. (An unfortunate name because its units are not those of force.)

Of the three terms on the right hand side of equation (21.5) the first is called the *inductive voltage drop*, and depends on the free, polarisation, and magnetisation currents (that are caused in the circuit by the emf), the second is called the *resistive voltage drop* and describes the hindering of the free current in the circuit, and the third term, which depends on the free, fixed, and polarisation charge distributions set up in the circuit by the emf, is called the *capacitive voltage drop*. Of the three terms only the third involves the electric potential. In some circumstances contributions to the first term will be called *induced emfs*.

An example of the balance between an applied emf and a capacitive voltage drop is the following. Consider a length of neutral copper wire to which a longitudinal emf is applied by some external source. The effect of the emf is to push the free charges towards the ends of the wire. The resulting charge distribution, which appears almost instantaneously following the application of the force, is such that the effect of its electric field cancels that of the applied force throughout the wire. The integral of this electric field along the wire is the voltage drop.

Note, it is commonly stated that the electric field inside a perfect conductor is zero, this is not always the case. What happens is that the free charges in a perfect conductor rapidly rearrange themselves to maintain the net force on them at zero. If the emf applied to the free electrons in our copper wire is due to its motion in a constant magnetic field the only **E** field in the system is that due to the charge distribution which is created and this is not zero. Only if the applied emf is due solely to an applied **E** field (which it often is) is the total **E** field zero.

21.5 Sources of applied emf

We have resolved the forces experienced by charges in circuits into externally applied and local components. We view the electromotive force applied to a circuit as what makes things happen. The applied force is transmitted throughout the circuit by mutual repulsion of the free charges in its conductors. Sources of emf may be relatively localised (e.g. in the electrolyte of a chemical cell) or distributed (e.g. along a winding in a dynamo). We discuss briefly some sources of emf.

21.5.1 Chemical cells

An example of how electrochemical emfs can arise is the following:- In an electrolyte between two plane parallel electrodes, gradients of the densities of singly charged positive and negative ions are maintained by some means. The current density normal to the plates, the z direction, has conduction and diffusion components and is given by:

$$j_z = ne\mu_n E_z - D_n(-e)\frac{\partial n}{\partial z} + pe\mu_p E_z - D_p e\frac{\partial p}{\partial z} \tag{21.23}$$

where n and p are the number densities of the negative and positive ions, μ_n and μ_p are their mobilities, D_n and D_p are their diffusion coefficients, and E_z is the electric field. As the electrolyte is a good conductor it will be neutral everywhere i.e. $n = p$. Assuming $\frac{\partial n}{\partial z}$ is constant $\left(= \Delta n/d\right)$ where d is the distance between the plates, using the Einstein relation:

$$D = \frac{\mu kT}{e} \tag{21.24}$$

and setting the current to zero in equilibrium we obtain the open circuit voltage between the electrodes:

$$E_z d = \frac{-kT\left(\mu_n - \mu_p\right)}{e\left(n\mu_n + p\mu_p\right)}\Delta n \tag{21.25}$$

which is equal to the emf. Δn is maintained by chemical reactions at the surfaces of the plates. The familiar zinc carbon (Léclanché) cell has an emf of about 1.5 V.

21.5.2 Emf induced in a loop

We consider a conducting thin-wire closed loop with free charges q moving and/or changing its shape in a time-dependent magnetic field. From section 21.4.2 the emf induced in the loop is $\oint \frac{\mathbf{F}_{\text{ext}} \cdot \mathbf{dl}}{q}$. We take $\mathbf{F}_{\text{ext}}$ to be due to external electric and magnetic fields i.e. $\mathbf{F}_{\text{ext}} = q(\mathbf{E}_{\text{ext}} + \mathbf{v} \wedge \mathbf{B}_{\text{ext}})$ where $\mathbf{v}$ is the velocity of the charges q. Then the emf $\mathcal{E}$ in the direction of d$\mathbf{l}$ is given by

$$\mathcal{E} = \oint (\mathbf{E}_{\text{ext}} + \mathbf{v} \wedge \mathbf{B}_{\text{ext}}) \cdot \mathbf{dl}$$

Now $\mathbf{v}$ is made up of the velocity of the conductor $\mathbf{v}_{\text{cond}}$ and the drift velocity of the charges along the (thin) conductors, but the latter component does not contribute to the triple product so

$$\mathcal{E} = \oint (\mathbf{E}_{\text{ext}} + \mathbf{v}_{\text{cond}} \wedge \mathbf{B}_{\text{ext}}) \cdot \mathbf{dl}$$

Next we split the integral form of equation 21.3 into external and internal parts, giving us

$$\oint \mathbf{E}_{\text{ext}} \cdot \mathbf{dl} = -\int \frac{\partial \mathbf{B}_{\text{ext}}}{\partial t} \cdot \mathbf{ds}$$

$$\text{So} \quad \mathcal{E} = -\int_s \frac{\partial \mathbf{B}_{\text{ext}}}{\partial t} \cdot \mathbf{ds} + \oint \mathbf{v}_{\text{cond}} \wedge \mathbf{B}_{\text{ext}} \cdot \mathbf{dl}$$

The triple product on the right hand side can be rearranged into

$$-\int \mathbf{B}_{\text{ext}} \cdot \mathbf{dl} \wedge \mathbf{v}_{\text{cond}} = -\int_s \mathbf{B}_{\text{ext}} \cdot \frac{\partial s}{\partial t}$$

$$\text{So} \quad \mathcal{E} = -\int \left(\frac{\partial \mathbf{B}_{\text{ext}}}{\partial t} \cdot \mathbf{ds} + \mathbf{B}_{\text{ext}} \cdot \frac{\partial \mathbf{s}}{\partial t} \right)$$

$$\mathcal{E} = -\frac{\partial}{\partial t} \int_s \mathbf{B}_{\text{ext}} \cdot \mathbf{ds} \qquad \text{(in the direction of d}\mathbf{l}\text{)}$$

(The directions of d$\mathbf{s}$ and d$\mathbf{l}$ are related by the usual sign convention.) This expression tells us that whatever the comvination of moving wire and changing magnetic field the emf is given simply by the rate of change of the external magnetic flux $\int_s \mathbf{B}_{\text{ext}} \cdot \mathbf{ds}$ through the loop. Some people call this the *flux rule*.

21.5.3 Signal generators

Signal generators are electronic oscillator circuits which convert emfs arising from the mains alternator or batteries into alternating emfs whose frequencies, amplitudes, and often also waveforms, can be selected by the user. Oscillators are discussed in chapter 17.

21.5.4 Solar cells

A solar cell is a semiconductor device in which there is a region of strong electric field where photogenerated electron-hole pairs are separated. The I–V characteristic of a PN junction type cell is:

$$I = I_0 \left(e^{\frac{eV}{kT}} - 1 \right) - I_p \tag{21.26}$$

where I_0 is a constant and I_p is the photo generated current. On open circuit $(I = 0)$ the emf V is given by:

$$V = \frac{kT}{e} \ln \left(\frac{I_p}{I_0} \right) \tag{21.27}$$

On short circuit $(V = 0)$ the current is $-I_p$.

The load is usually chosen to maximise the power delivered, typically up to 80% of $(I_p V)$ can be obtained.

21.5.5 Thermal emfs

As a consequence of electrons being involved in both electric currents and the transport of heat *thermoelectric* effects occur.

Thomson emf The tendency of mobile charges to diffuse away from the hot end of a uniform conducting bar is opposed by the build up of an electric field. The strength of the field depends on the temperature gradient.

Peltier effect The tendency of mobile charges to cross the boundary between two different, originally neutral, conductors in contact is opposed by the build up of a potential difference. The name given to the fact that this *contact potential* is temperature dependent is the Peltier effect.

Thermocouple (Seebeck emf) A length of wire of one material is joined to a length of wire of a different material to form a ring. A break is made in one of the wires. When the two joints are held at different temperatures the combination of the Thomson and Peltier emfs gives rise to a potential difference (the Seebeck emf) at the break. It is rather small; for one of the strongest combinations of metals; copper and constantan, it is only 40 μV per K of temperature difference between the two joints. Of the order of $1\,\mathrm{mVK}^{-1}$ may be observed with a combination of a metal and a semiconductor.

21.6 References

R. P. Feynman, *Lectures on Physics Vol II*, Addison Wesley.
F. N. H. Robinson, *Macroscopic Electromagnetism*, Pergamon.
Ramo and Whinnery, *Fields and Waves in Modern Radio*, Wiley.

22 Equivalent Circuits of Passive Components

22.1 Introduction

In the previous chapter Kirchhoff's laws and the concepts of applied emf and voltage drop were introduced. Here we develop the concepts of capacitance, inductance and resistance and use them to construct models (*equivalent circuits*) of capacitors, inductors, transformers, and resistors.

22.2 Assumptions

In accordance with our definition of a circuit (section 21.2) we are concerned with a theory without radiation. Radiation fields disappear if the retardation $\left|\mathbf{r}-\mathbf{r}'\right|/c$ in the potentials is ignored. Accordingly we take the potentials for circuits to be:

$$\phi_{\text{ct}}\left(\mathbf{r},t\right) = \frac{1}{4\pi\epsilon_0}\int\limits_{\text{ct}} \frac{\left(\rho_{\text{free}}+\rho_{\text{fixed}}-\nabla\cdot\mathbf{P}\right)\left(\mathbf{r}',t\right)}{\left|\mathbf{r}-\mathbf{r}'\right|}\,\mathrm{d}^3\,\mathbf{r}' \tag{22.1}$$

and:

$$\mathbf{A}_{\text{ct}}\left(\mathbf{r},t\right) = \frac{1}{4\pi\epsilon_0 c^2}\int\limits_{\text{ct}} \frac{\left(\mathbf{j}_{\text{free}}+\nabla\times\mathbf{M}+\frac{\partial\mathbf{P}}{\partial t}\right)\left(\mathbf{r}',t\right)}{\left|\mathbf{r}-\mathbf{r}'\right|}\,\mathrm{d}^3\,\mathbf{r}' \tag{22.2}$$

where we have also assumed that the charges and currents are on or within the solid materials of the circuit i.e. that any effects due to the medium (usually air) in which the latter are immersed can be ignored.

We make the further assumptions:

(a) when the circuit has been quiescent for a long time the charge and current fields are zero everywhere, in other words there is no permanent magnetisation or polarisation. (Real dielectric materials which have a significant polarisation at zero E are rare. Real magnetic materials may have significant magnetisation when not under the influence of an external current loop.)

(b) all the materials can be considered linear. In such a circuit the magnitude of any field (e.g. $\{\mathbf{j}_{\text{free}}, \nabla\times\mathbf{M}, \frac{\partial\mathbf{P}}{\partial t}, \rho_{\text{free}}+\rho_{\text{fixed}}$, or $\nabla\cdot\mathbf{P}\}_{\text{ct}}$) at any point is proportional to the strength of the excitation so the ratio of the magnitudes of any two fields at any two points is independent of the level of excitation. The directions along which the vector fields oscillate do not change as the level of excitation is changed. In particular, the total charge density field ρ_{ct} is linearly related to $\rho_{\text{free}}+\rho_{\text{fixed}}$ and the total current density field $\mathbf{j}_{\text{ct}}$ is linearly related to $\mathbf{j}_{\text{free}}$ and we can write the potentials $\phi_{\text{ct}}\left(\mathbf{r},t\right)$ and $\mathbf{A}_{\text{ct}}\left(\mathbf{r},t\right)$ in a linear circuit as:

$$\frac{1}{4\pi\epsilon_0}\int\limits_{\text{materials}} \frac{h\left(\mathbf{r}',\omega\right)\left(\rho_{\text{free}}+\rho_{\text{fixed}}\right)_{\text{ct}}\left(\mathbf{r}',t\right)}{\left|\mathbf{r}-\mathbf{r}'\right|}\,\mathrm{d}^3\,\mathbf{r}' \tag{22.3}$$

and:

$$\frac{1}{4\pi\epsilon_0 c^2}\int\limits_{\text{materials}} \frac{\Xi\left(\mathbf{r}',\omega\right)\mathbf{j}_{\text{free}}\left(\mathbf{r}',t\right)_{\text{ct}}}{\left|\mathbf{r}-\mathbf{r}'\right|}\,\mathrm{d}^3\,\mathbf{r}' \tag{22.4}$$

where h and Ξ, which depend on the distribution of the materials, are functions of position and the frequency but not of the amplitude of the excitation.

We have already defined LIH conductors in section 21.3. We now add the definitions:

$$\mathbf{P} = (\epsilon_r - 1)\epsilon_0 \mathbf{E} \qquad \text{(LIH dielectric material)} \tag{22.5}$$

where ϵ_r is a constant called the relative permittivity or relative dielectric constant, and:

$$\mathbf{M} = \frac{\mu_r - 1}{\mu_r} \epsilon_0 c^2 \mathbf{B} \qquad \text{(LIH magnetic material)} \tag{22.6}$$

where μ_r is a constant called the relative permeability. The constants of proportionality are written as they are for historical reasons.

Real dielectric materials which exhibit significant non linearity at accessible levels of **E** are rare. Real magnetic materials may be readily brought close to a state of complete magnetisation (saturation).

22.3 Equivalent circuits of short segments of path

We consider a small volume $\delta\tau$ in a circuit. Applying Kirchhoff's first law, equation (21.16), yields:

$$\sum_{r=1}^{n} I_r = \frac{\partial q}{\partial t} \tag{22.7}$$

where I_r is one of the n currents entering $\delta\tau$ and q is the charge in $\delta\tau$.

We next consider a path of integration **l** passing through $\delta\tau$ whose surface marks off a short segment $\delta\mathbf{l}$ of the path. Applying Kirchhoff's second law, equation (21.22), to $\delta\mathbf{l}$ we have:

$$\mathcal{E}_{\delta\mathbf{l}} = \frac{\partial}{\partial t}(\mathbf{A}_{\text{ct}} \cdot \delta\mathbf{l}) + \frac{\mathbf{j}_{\text{free}} \cdot \delta\mathbf{l}}{\sigma} + (\phi_{\mathbf{l}+\delta\mathbf{l}} - \phi_{\mathbf{l}})_{\text{ct}} \tag{22.8}$$

where $\mathcal{E}_{\delta\mathbf{l}}$ is the emf applied to $\delta\mathbf{l}$. It is to be noted that $\mathbf{j}_{\text{free}}$ is purely local to the segment of path d**l** but that ϕ_{ct} and $\mathbf{A}_{\text{ct}}$ depend not only on the charge and current local to $\delta\mathbf{l}$ but also on the charges and currents in the rest of the circuit.

22.3.1 Conducting segments

Consider a thin round wire of length $2b$ and radius a made of material of conductivity σ. Further consider a short length $\delta\mathbf{l}$ near the middle of the length $2b$. Applying equations (22.7) and (22.8) to $\delta\mathbf{l}$ we find:

$$I_{\mathbf{l}} - I_{\mathbf{l}+\delta\mathbf{l}} = \frac{\partial q}{\partial t} \tag{22.9}$$

where $I_{\mathbf{l}}$ is the current at **l**, and q is the charge on $\delta\mathbf{l}$, and:

$$\mathcal{E}_{\delta\mathbf{l}} = \frac{\partial}{\partial t}(\mathbf{A}_{\text{ct}} \cdot \delta\mathbf{l}) + I\frac{\delta\mathbf{l}}{a\sigma} + (\phi_{\mathbf{l}+\delta\mathbf{l}} - \phi_{\mathbf{l}})_{\text{ct}} \tag{22.10}$$

The quantity $\delta\mathbf{l}/a\sigma$ in the second term on the right hand side is called the *resistance* of $\delta\mathbf{l}$ and will be denoted by δr. The resistive voltage drop may therefore be written $I\delta r$.

Next we assume that any charge on $\delta\mathbf{l}$ or on parts of the circuit outside $\delta\tau$ does not give rise to a significant gradient of potential along $\delta\mathbf{l}$ and ignore the third term. (Allowing for stored charge is considered in the following section.)

It remains to discuss the first term. Taking the simplest case of the current I confined to the surface of the wire, A at the centre of $\delta\mathbf{l}$ is given by:

$$\frac{2I}{4\pi\epsilon_0 c^2}\left(\ln\frac{\delta l + \sqrt{\delta l^2 + a^2}}{a} + \ln\frac{b + \sqrt{b^2 + a^2}}{\delta l + \sqrt{\delta l^2 + a^2}}\right)\frac{\mathrm{d}I}{\mathrm{d}t} \tag{22.11}$$

Making the crude assumption that **A** is the same all along $\delta\mathbf{l}$ the first term becomes:

$$\frac{2\delta l}{4\pi\epsilon_0 c^2}\left(\ln\frac{\delta l + \sqrt{\delta l^2 + a^2}}{a} + \ln\frac{b + \sqrt{b^2 + a^2}}{\delta l + \sqrt{\delta l^2 + a^2}}\right)\frac{\mathrm{d}I}{\mathrm{d}t} \tag{22.12}$$

which may be written

$$\delta L\frac{\mathrm{d}I}{\mathrm{d}t}+\delta M\frac{\mathrm{d}I}{\mathrm{d}t} \tag{22.13}$$

We come now to the question of naming inductive effects. The difference between the two terms just derived is that the first describes the voltage induced in δl by the current in δl while the second describes the voltage induced in δl by a current outside δl. This inside/outside current criterion determines how we refer to such effects and what representative symbols we use. "Inside" inductive effects are called *(self-induced) voltage drops*[1] and are represented by the coil symbol (see below) and the character L. "Outside" inductive effects are called *mutually induced emfs* and are represented by the usual emf symbol and the character M. (We used this convention without comment in Chapter 1.)

The mutually induced emf we have calculated is just that due to the current in the length $2b-\delta l$ of the wire. In practice there will be contributions from the currents in all the other wires in the circuit as well but the effects of distant currents will be smaller and may average out to a certain extent. We cannot say more about the total mutually induced emf without knowing the arrangement of the whole circuit but it can be of comparable magnitude to the self inductive term which we can easily estimate from the expression above. (A 1 cm length of 0.1 mm radius wire turns out to have a self inductance of about 10^{-8} Henries.)

Using these results, equation (22.8) becomes:

$$\mathcal{E}_{\delta\mathbf{l}}=I\delta r+\delta L\frac{\partial I}{\partial t}+\sum_r \delta M_r\frac{\partial I_r}{\partial t} \tag{22.14}$$

where we have indicated the presence of a number of mutually induced emfs. A diagrammatic representation of this is:

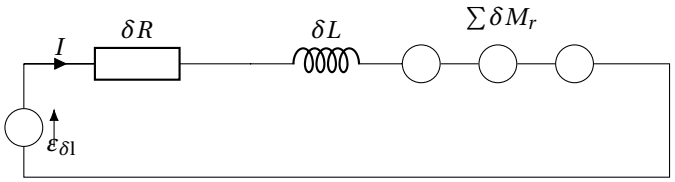


Figure 22.1: Representation of mutually induced emfs

Note that the three terms in equation (22.14) which represent *superimposed effects in* δl, are represented by symbols in series in the diagram. Each of the mutually induced emfs may either aid or oppose the applied emf. The plain lines represent connections having no voltage drops.

22.3.2 Dielectric segments

Now consider $\delta\tau$ to be a volume which just encloses a structure consisting of a thin layer of non magnetic, lossless, linear, dielectric sandwiched between a pair of very thin circular plane parallel perfectly conducting plates. Connections are made to the rest of the circuit from the centres of the plates. When the circuit is excited there are charges on the plates, free currents in the plates, polarisation charges on the faces of the dielectric and a polarisation current within the dielectric.

From the Coulomb-Gauss law we know that the charges on the inner surfaces of the two plates are equal and opposite at all times. The way charge is distributed on the outer surfaces of the plates will be influenced by charges on parts of the circuit outside $\delta\tau$, but the surface densities will generally be much smaller than those on the inner surfaces.

Taking $\delta\mathbf{l}$ to extend only infinitesimally beyond the outer surfaces of the plates along the connections to the rest of the circuit, the third term on the right hand side in equation (22.8), $(\phi_{\mathbf{l}+\delta\mathbf{l}}-\phi_{\mathbf{l}})_{\mathrm{ct}}$ is unaffected by the charges on the outsides of the plates and depends only on the charges on their inner surfaces.

The second term gives no contribution since the conductivity is infinite in the (perfect) conductors and there is no free current in the (perfect) dielectric.

The first term depends on the free currents in the plates, the polarisation current $\frac{\partial\mathbf{P}}{\partial t}$ in the dielectric, and the currents in the rest of the circuit. The free currents in the plates flow radially so give no contribution to $\mathbf{A}$ along the axis of the structure. Evaluating $\mathbf{A}$ on the axis due to the polarisation current yields a contribution to the first term of

$$2\pi^2(\epsilon_r-1)\frac{a\delta l}{\lambda^2}(\phi_{\mathbf{l}+\delta\mathbf{l}}-\phi_{\mathbf{l}})_{\mathrm{ct}} \tag{22.15}$$

[1] Some people call self-induced voltage drops *back emfs*.

where ϵ_r is the dielectric constant, a is the radius of the plates, and λ is the wavelength of electromagnetic radiation in the dielectric at the frequency of the excitation. This contribution is small compared with the potential difference $(\phi_{\mathbf{l}+\delta\mathbf{l}} - \phi_{\mathbf{l}})_{\text{ct}}$ if:

$$2\pi^2 (\epsilon_r - 1) \frac{a\delta l}{\lambda^2} \ll 1 \tag{22.16}$$

which it is for normal sized capacitors. The largest contribution to **A** from outside $\delta\tau$ will usually be from the connecting wires as they carry a free current $\epsilon_r/(\epsilon_r - 1)$ times the polarisation current in the dielectric and their ends are close to $\delta\tau$. An approximate evaluation for straight wires extending a distance b on both sides of $\delta\tau$ yields a contribution to the first term of:

$$2\pi^2 \epsilon_r \frac{a^2}{\lambda^2} \ln \frac{2b}{\delta l} (\phi_{\mathbf{l}+\delta\mathbf{l}} - \phi_{\mathbf{l}})_{\text{ct}} \tag{22.17}$$

It is not unreasonable to assume that:

$$2\pi^2 \epsilon_r \frac{a^2}{\lambda^2} \ln \frac{2b}{\delta l} \ll 1 \tag{22.18}$$

making this contribution also negligible. So if we restrict ourselves to arrangements in which these inequalities are satisfied (frequencies less than a few GHz) and in which the A in $\delta\tau$ due to currents outside $\delta\tau$ is not significantly greater than the A due to the local current, equation (22.8) reduces to:

$$\mathcal{E}_{\delta\mathbf{l}} = (\phi_{\mathbf{l}+\delta\mathbf{l}} - \phi_{\mathbf{l}})_{\text{ct}} \tag{22.19}$$

We see that the voltage drop between the plates is just their potential difference. Now the potential difference is related to the charges $\pm q$ on the inner surfaces of the plates by:

$$(\phi_{\mathbf{l}+\delta\mathbf{l}} - \phi_{\mathbf{l}})_{\text{ct}} = \frac{q}{\frac{\epsilon_0 \epsilon_r \pi a^2}{l_r}} = \frac{q}{C} \tag{22.20}$$

where C is the *internal capacitance* of the plates and dielectric. So:

$$\mathcal{E}_{\delta\mathbf{l}} = \frac{q}{C} \tag{22.21}$$

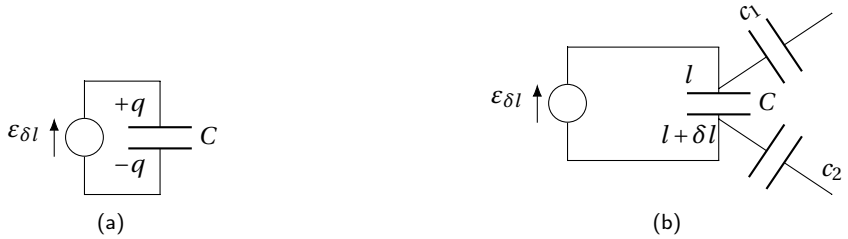


Figure 22.2: Internal capacitance

A diagrammatic representation of this is shown in Figure 22.2(a) where the adjacent edges of the parallel bars indicate places where charge is stored (in this case the inner surfaces of the plates). All charge storage is indicated with capacitor symbols, the plain lines simply indicate that the emf is applied to the capacitor, no charge is stored on them. As mentioned, the capacitance symbol in the figure above accounts only for the charges on the inner surfaces of the plates. To account for the charges on their outer surfaces additional capacitance symbols linking to parts of the circuit outside $\delta\tau$ must be added to the diagram as shown in Figure 22.2(b).

The total charge Q_l on all the plates connected to l is related to the current arriving at l from outside $\delta\tau$ by

$$\frac{\mathrm{d}Q_l}{\mathrm{d}t} = i_1 \tag{22.22}$$

A similar expression holds at $l + \delta l$.

22.4 The independent component assumption

Following the discussion in section 22.3 we know that there will be a dense web of capacitive and mutual inductive couplings between the parts of any circuit. At first sight it seems surprising therefore that circuit designers do not show them in their equivalent circuits. Clearly, an assumption that the components in circuits do not influence each other makes the task of design much easier but how is it possible to make the assumption?

Well, some of the interactions are inherently small, the charges and currents associated with inter-component capacitances (typically of the order of 1 pF) are usually negligible except at very high frequencies. Also many inductive components have ferromagnetic yokes which not only confine their own magnetic fields but also help to shield them from external fields.

Where inter-component capacitances would be troublesome circuit designers assume that the construction designer will insert metal screens between parts that must not "see" each other, e.g. the inputs and outputs of amplifiers. The unwanted capacitive links are thereby replaced by capacitances to screens which are usually less important. (The charges on these screens terminate the electric fields which would otherwise link the sensitive parts. The screens are connected by low resistance paths to the reference conductor (22.7.2) which is usually earthed, so that their voltages do not vary as the screening charges flow onto and off them.)

Unwanted inductive couplings are eliminated by orienting transformers appropriately, twisting pairs of wires, or providing magnetic shielding (high permeability flux bypass paths). Coils of open construction, used at radio frequencies and above, are spaced widely and oriented with their axes at right angles or isolated in closed metal compartments the currents on the surfaces of which terminate the magnetic fields.

Further discussion of the construction of hardware is given in section 22.7.

22.5 Equivalent circuits of passive components

Our aim in this section is to derive equivalent circuits of the common passive electronic components, by which we mean connecting wire, resistors, inductors, transformers, and capacitors. To do this we consider paths of integration extending from where the current enters a component to where it leaves.

Following the discussion in section 22.4 we will consider only couplings between segments of path within components and ignore interactions between components (except for capacitances to screens).

22.5.1 Connecting wire

A connecting wire (Figure 22.3(a) is taken to be a number of segments like those shown in Figure 22.1 in series. As our definition of "inside" now means the whole wire the mutually induced emfs due to the current in the wire are renamed voltage drops and described by coil symbols which may then be combined with the original self inductive symbols with which they are in series. We must also allow for each segment of the wire to carry some charge. Accordingly we represent the wire by the equivalent circuit shown in Figure 22.3(b)

The mutual inductive emfs are due to "outside" currents, which here means those in other wires (not, we reiterate, those in other segments of the same wire).

This is quite a complicated equivalent circuit and it is worth asking if all the complexity is necessary. In general the complexity required for a useful representation increases with frequency but this is partly mitigated by the fact that connections in "properly constructed" high frequency circuits are kept short. To set the scene we note that typical values for a 1 cm length of 0.1 mm radius copper wire are a resistance of the order of 1 mΩ, a capacitance to the surroundings of the order of 1 pF, and a self inductance of the order of 10^{-8} H.

At dc the series inductances and shunt capacitances have no effect so we can represent the wire solely by its total resistance. At audio frequencies it will usually be satisfactory to ignore the inductances but a single capacitance to the surroundings (screen) may need to be included in very high impedance circuits. At low radio frequencies and above it may be necessary to consider the equivalent circuit in Figure 22.3(c). The capacitance could equally well be shown connected to one end rather than as drawn.

In Chapter 1 we defined an *ideal* interconnection as one with no voltage drop or charge storage, i.e. one for which the strays l, c, and r are zero. The circuit diagram symbol for the wire is shown in Figure 22.3(d).

22.5.2 Film Resistors

Most resistors used in electronics today consist of insulating ceramic rods on which are deposited thin films of a high resistance conducting material which may be a metal alloy, a mixture of metal oxides or carbon. The value is set by cutting a spiral groove in the coated rod. After the fitting of end caps and wires the resistor is given a protective coat of varnish (Figure 22.4(a)) and coloured bands to indicate its value. (See Table 7.3.)

A resistor has all the strays possessed by a fragment of connecting wire but the inductive drop will be larger (due to the coiled nature of the conductor) and the capacitance between the ends may be significant. The equivalent circuit and the circuit diagram symbol are shown in Figure 22.4(b) and Figure 22.4(c) respectively.

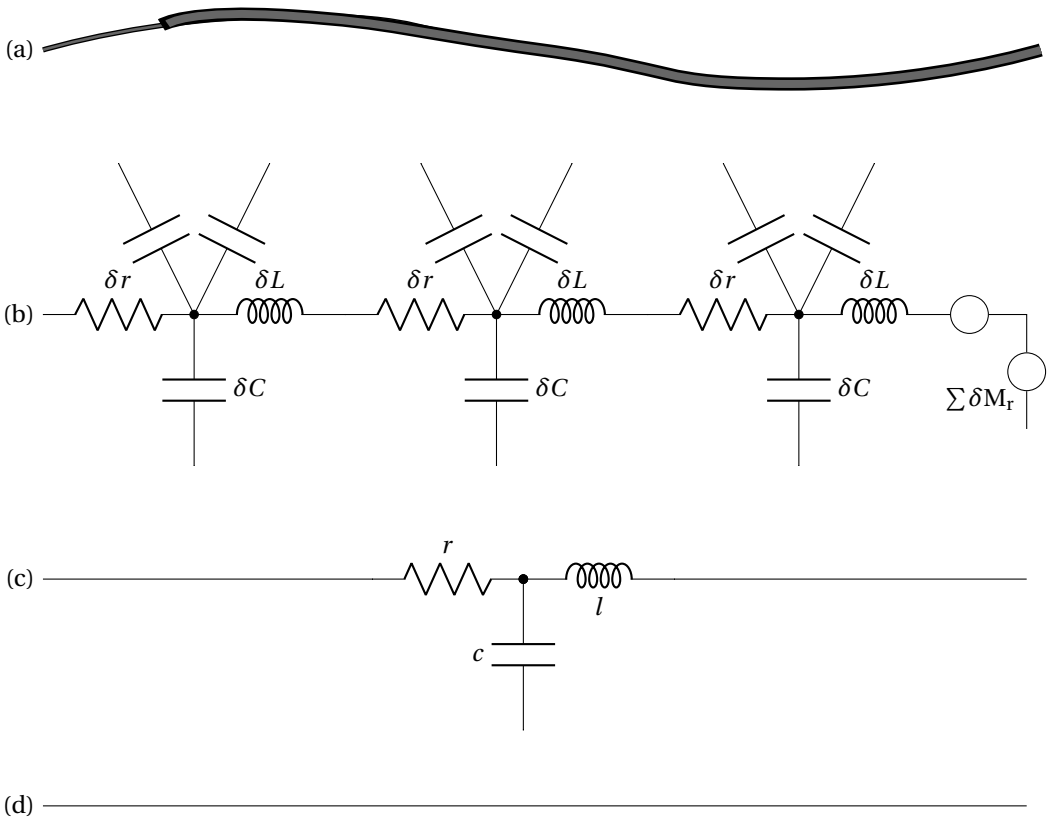


Figure 22.3: Representation of mutually induced emfs

22.5.3 Wirewound resistors

For higher powers resistance wire is wound onto a ceramic rod or tube and the assembly is coated with cement or china clay. The equivalent circuit is the same as the above but l is larger than in a film resistor.

22.5.4 Film capacitors

These capacitors are made from polymer films coated with metal evaporated from a hot source. Then either long strips of metallised film are rolled up or cropped lengths are stacked. To reduce inductance in the first method manufacturers make several or even continuous connections to each metallisation. This also reduces the series resistance due to the metallisation being thin.

In deriving an equivalent circuit the treatment for a short insulating path given in section 22.3.2 is nearly all we need. The capacitances to the surroundings, typically a few pF, need only be included at very high frequencies and it is usually all right to combine them and show only one. The plastics used are good insulators but exhibit dielectric losses (polarisation not quite in phase with electric field). Of the three commonly used film materials the least loss occurs in polystyrene, the most in polyester, with polycarbonate coming somewhere in between. This dielectric loss is represented by the resistance in the equivalent circuit shown in Figure 22.5 (See physics finals paper IIA 1992, Q2.) C_0 is the capacitance the plates would have without the dielectric, C is $(\epsilon_r - 1)/\epsilon_r$ of the total capacitance.

22.5.5 Aluminium electrolytic capacitors

Aluminium foils bearing a thin insulating oxide are separated by a spacing material impregnated with an electrolyte. When a voltage is applied the oxide film on one of the foils builds up, a process known as forming. The resulting structure has a high capacitance in a small volume. Electrolytic capacitors must only be used in circuits where the applied voltage does not reverse. Their chief application is in smoothing the outputs of rectifiers in power supplies and providing low impedance signal paths in low frequency circuits. They may have significant leakage current especially if no voltage has been applied for a long time but this will decrease as the main oxide layer is reformed. The most important stray is a parallel resistance representing this leakage.

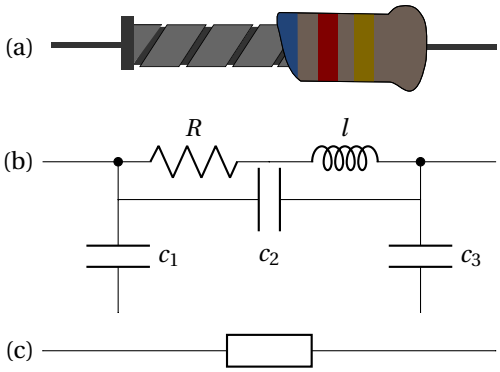


Figure 22.4: Film resistors

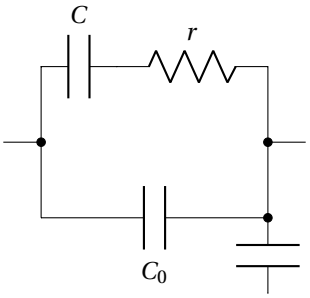


Figure 22.5: Film capacitor

22.5.6 Mica and ceramic capacitors

The usual construction is a single plate with coated electrodes although stacks of plates are made. A general characteristic is low stray inductance. Mica capacitors have good all round performance but are expensive and impractical for values above a few thousand pF. Ceramic capacitors fall into two groups, those made with high (1000s) permittivity, rather lossy, material for decoupling and bypassing applications in radio frequency circuits and those made with lower loss, lower permeability material for use in more demanding applications such as tuned circuits. In the latter case ceramic capacitors are often chosen for their negative temperature coefficient of permittivity which can be used to compensate for the (usually positive) coefficients of the other components in a tuned circuit.

22.5.7 Audio and power frequency inductors (chokes)

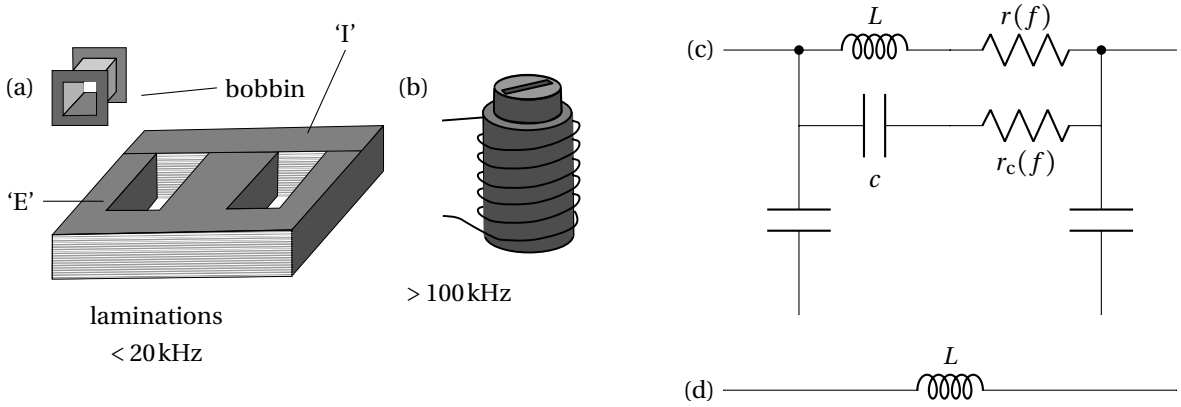


Figure 22.6: Inductors

Traditionally a coil of wire is wound on a former or bobbin and a closed magnetic circuit is then built up by inserting E and I (or T and U) shaped steel laminations, see Figure 22.6(a). Increasingly nowadays the coils are wound on prefabricated toroids (rings) of ferromagnetic material. At the higher audio frequencies and above magnetic circuits made from pieces of high resistivity ceramic material (ferrite) rather than steel laminations may be used. These closed magnetic circuits greatly increase the inductance obtainable with a given length of wire.

The coiling of the conductor enhances the mutual induced emf in each segment of the conductor due to the others and hence the total induced emf in the coil. Since all the segments are part of the same coil we consider all the inductive effects to be "inside" and we represent their total effect with the coil symbol.

The dominant element in the equivalent circuit Figure 22.6(c) is the inductance L. The major strays are the resistance of the wire ("copper loss"), a resistance representing eddy current heating of the steel, which laminating does not completely eliminate, and hysteresis ("iron loss"), the capacitances between the turns, and the shunt capacitances to the magnetic core and the surroundings.

The resistance representing the iron loss will be frequency dependent as will the resistance representing the copper loss because of the skin effect. The copper and iron losses have been represented by a single resistance.

The inter-turn capacitances can be reasonably well represented by a single capacitance in parallel with the inductance, exactly so if the inter-turn capacitance is uniformly distributed along the inductance. The distributed stray shunt capacitances may be lumped into one at each end if they are small. The wire insulation and the former on which the coil is wound may be poor dielectrics and these capacitances may have significant (and frequency dependent) losses associated with them.

The capacitances are responsible for the inductor having a parallel resonance at some frequency, the Q of the resonance depending on the losses.

This is a good point to relate the basic definition of inductance to a more elementary one. Under the assumptions leading to the equivalent circuit above the current is the same throughout the coil and the inductance of the coil may be written:

$$L_{\text{coil}} = \frac{\int_{\text{coil}} (\mathbf{A}_{\text{coil}} \cdot \mathrm{d}\mathbf{l})}{I_{\text{coil}}} \tag{22.23}$$

where I_{coil} is the current in the coil:

$$L_{\text{coil}} = \frac{\int_{\mathbf{S}} (\nabla \times \mathbf{A}_{\text{coil}} \cdot \mathrm{d}\mathbf{S})}{I_{\text{coil}}} \tag{22.24}$$

where **S** is the surface bounded by the path imagined closed by bridging the gap between the ends of the coil:

$$L_{\text{coil}} = \frac{\int_{\mathbf{S}} (\mathbf{B}_{\text{coil}} \cdot \mathrm{d}\mathbf{S})}{I_{\text{coil}}} \tag{22.25}$$

i.e. the self inductance of the coil = $\frac{\text{flux linked}}{\text{current}}$, the elementary definition of inductance.

22.5.8 Radio frequency inductors

At radio frequencies a core consisting of a short rod of magnetic material is often used not only to increase the inductance for a given number of turns but also because by changing its position in the coil the value of the inductance can be adjusted, see Figure 22.6(b). Ferrite materials are used up to about 1 MHz and iron dust in a binder is used up to about 30 MHz. Above this frequency power losses in the material tend to outweigh the advantages and 'air cored' coils are used.

22.5.9 Power and audio frequency transformers

The construction of power and audio frequency transformers is the same as that of the chokes described in section 22.5.8 except that more than one coil is wound on the bobbin. Similar considerations with regard to use of magnetic material apply as for (self) inductors. Nearly all the flux in the magnetic circuit links all the turns on all the coils. For clarity we will consider just two coils in what follows.

The direction of rotation of a transformer bobbin on a coil winding machine is invariably the same for all the windings. Knowing this and given that the start of each winding is identified we have enough information to choose the signs of M in section 1.2.2. On equivalent circuits starts are usually indicated by black dots, see Figure 3.12. On actual transformers, ends of windings are often labelled ip, op, is and os where i is inner, o outer, p primary and s secondary. (Starts are always on the *inner* layer of a multilayer winding!)

The relatively complicated construction of transformers compared with other components leads to them having a larger variety of strays which may need to be taken into account. A typical equivalent circuit for a power or audio frequency transformer with a primary winding and one secondary winding is shown in Figure 22.7.

Each of the coils has an equivalent circuit similar to that in Figure 22.6 with the addition of a mutual inductive emf. The value of M is given by $M = k\sqrt{L_p L_s}$ where k is the coefficient of coupling. $k = 1$ when all the magnetic flux links all the turns on both coils. In the equivalent circuit the inductance of each coil is shown divided into two parts, a part considered completely coupled to part of the other coil, and a part considered to have no coupling with the other coil. In audio and mains transformers the uncoupled inductances l_p, l_s are much smaller than the coupled parts and are called *leakage inductances*.

Designing transformers whose behaviour is close to ideal over a frequency range of 1000:1 (e.g. 20 Hz to 20 kHz) is quite difficult.

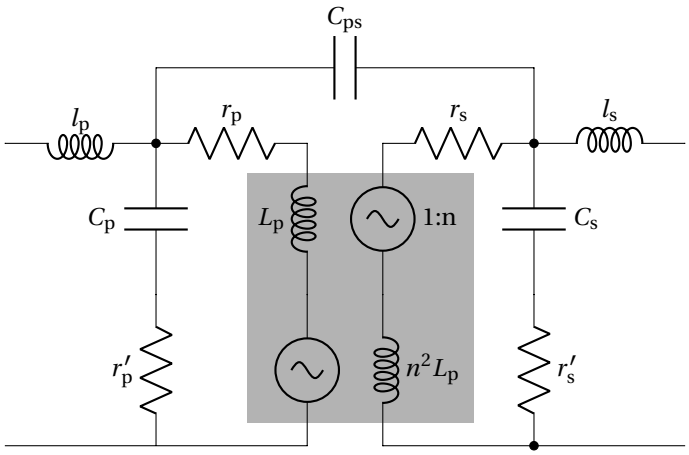


Figure 22.7: Inductors

22.5.10 Radio frequency transformers

At radio frequencies transformers often have coefficients of coupling as low as a few percent and windings that are tuned to be resonant at a desired frequency. Such transformers are used to couple the amplifying stages in the intermediate frequency sections of radio receivers, see section 15.9.4. The methods of tuning are (i) to fit adjustable capacitors in parallel with the windings and (ii) to fit fixed capacitors and vary the inductances by moving magnetic cores (if the frequency is not too high for cores to be used), see Figure 22.6(b). Moving the cores will also affect the coupling between the windings.

22.6 Further remarks on equivalent circuits of components

A component may be represented by an equivalent circuit in more than one way. For instance, given magnitude and phase measurements at a single frequency it is possible to represent a lossy (and known to be non-inductive) capacitor by either a series or parallel combination of one capacitance and one resistance. Now if the losses are due to a leaky dielectric but a series representation is used the magnitudes of the C and r in the equivalent circuit will be frequency dependent which is undesirable. In the more appropriate parallel representation C and r would be constant. The lesson is that choosing a good equivalent circuit of a component involves knowing about its construction and defects. It is a mistake to use a more complex equivalent circuit than the highest frequency and desired accuracy require because analysis will be more laborious.

While it is better to use equivalent circuits whose elements have constant values this is not always possible for the resistances representing losses. The resistance of a wire will increase with frequency due to the confining of the current to the surface (skin effect) and dielectric and magnetic losses in capacitors and inductors, which are represented by resistances, are invariably frequency dependent.

22.7 Notes on construction

22.7.1 Screening boxes and cable shields

A typical electronic system consists of a number of circuits enclosed in metal boxes and coupled together by cables covered with metal braids which are connected to the boxes at both ends usually via the shells on the connectors. The entire circuit is therefore surrounded by a continuous metal skin.

Some or all of the circuits will have their boxes connected to earth, see Chapter 5. Some may also have their own mains power units and connections to the mains. There will be small currents flowing in the walls of the boxes and in the earth wires, due to capacitive and leakage resistive couplings to the internal power wiring in each box and possibly also to external wiring, which will give rise to small voltage differences between the boxes due to the impedances of the earth connections. When different parts of such a metal skin are at different voltages it is an indication that it may not be providing complete isolation between internal and external electric fields (but the isolation is still likely to be very good). Another consequence of the voltage differences is that there will be currents flowing in the cable braids.

22.7.2 Reference conductors

Most circuits have a reference conductor, usually assigned a voltage of zero, throughout which it is intended that the voltage is constant. The facts that the conductor will have stray inductance and resistance and signal currents flowing in it mean that this can never be precisely achieved. This in turn means that currents in one part of a circuit may change voltages in another. Such effects must be kept to negligible levels if the circuit is to behave as intended.

The reference conductor in low frequency circuits is a wire or printed circuit board track which is connected to the metal box enclosing the circuit (and ultimately to earth) at the place where the signal voltage is lowest (in an amplifier this would be at the input socket). The required connections to the reference conductor are then made in order of ascending signal current so that large signal currents do not flow in it between points at which low level signal connections are made.

Ideally there should be only one connection between the reference conductor and the earthed screening skin to prevent voltage differences in the skin driving power frequency currents along the reference conductors. In a system containing several constituent circuits and several points at which low signal levels occur, or in which it is necessary to operate any of the constituent circuits in the system in isolation, more connections between the system reference conductor and earth will be required. These additional connections should contain resistances (typically 100–1000 Ω) to reduce power frequency currents in the reference conductor to negligible levels.

At high frequencies the same considerations apply but reducing the stray inductance of the reference conductor to low enough values usually requires it to be in the form of a sheet or *ground plane*. One side of a double sided printed circuit board can be used. The ultimate in this respect, permissible if the system does not respond to excitation at power frequencies, is to use the metal box as the reference conductor.

23 Time Domain Analysis Using Laplace Transforms

23.1 Introduction

Linear differential equations with constant coefficients arise frequently in electronics and other branches of engineering concerned with linear systems. As was shown in Chapter 2, the behaviour of simple circuits described by first order equations and excited by step emfs can be found by straightforward integration. For circuits described by second order equations and excited by more complicated applied emfs the labour involved can be considerable. Here a quicker method of analysis capable of dealing with such cases is described. It is based on Laplace transforms.

23.2 Using Laplace transforms

Taking the differential equation for the oscilloscope probe derived in Section 2.5.1 as an example, we begin by multiplying each term by e^{-st} and integrating with respect to t from zero to infinity. This is called forming the *Laplace transform* of each term. ($s = \sigma + j\omega$ is a complex quantity which we do not need to enquire into further here.) We shall assume that all the integrals converge. The equation becomes:

$$\int_0^\infty e^{-st}\left(\frac{V_{\text{in}}}{R_1} + C_1\frac{\mathrm{d}V_{\text{in}}}{\mathrm{d}t}\right)\mathrm{d}t = \int_0^\infty e^{-st}\left((C_1+C_2)\frac{\mathrm{d}V_{\text{out}}}{\mathrm{d}t} + V_{\text{out}}\left(\frac{1}{R_1}+\frac{1}{R_2}\right)\right)\mathrm{d}t \tag{23.1}$$

We represent the process of forming the Laplace transform of a function $f(t)$ by $\mathcal{L}\{f(t)\}$. It is straightforward to derive (by integration by parts) the useful property of the transform:

$$\mathcal{L}\left\{\frac{\mathrm{d}f(t)}{\mathrm{d}t}\right\} = s\mathcal{L}\{f(t)\} - [f(0)] \tag{23.2}$$

from which it follows also that

$$\mathcal{L}\left\{\frac{\mathrm{d}^2 f(t)}{\mathrm{d}t^2}\right\} = s^2\mathcal{L}\{f(t)\} - [f'(0) + sf(0)] \tag{23.3}$$

We assume the terms in the square brackets are zero for the applied emfs to be considered.

The relations just derived enable us to write the integral equation for the probe in the form:

$$\frac{\mathcal{L}\{V_{\text{out}}\}}{\mathcal{L}\{V_{\text{in}}\}} = \frac{\frac{1}{R_1} + sC_1}{s(C_1+C_2) + \frac{1}{R_1} + \frac{1}{R_2}} \tag{23.4}$$

We take the applied emf $\mathcal{E}_g$ to be $\mathcal{E}_1 u(t)$ and as the resistance of the signal source is being ignored $V_{\text{in}} = \mathcal{E}_g$. We can look up $\mathcal{L}\{V_{\text{in}}\}$ in Table 23.2 or work it out for ourselves. In this case it is easy,

$$\mathcal{L}\{V_{\text{in}}\} = \frac{\mathcal{E}_1}{s} \tag{23.5}$$

so our equation becomes:

$$\mathcal{L}\{V_{\text{out}}\} = \frac{\left(\frac{1}{R_1} + sC_1\right)\mathcal{E}_1}{s\left(s(C_1+C_2) + \frac{1}{R_1} + \frac{1}{R_2}\right)} \tag{23.6}$$

We now take the inverse transform, $\mathcal{L}^{-1}$, of both sides:

$$\mathcal{L}^{-1}\left\{\mathcal{L}\left\{V_{\text{out}}\right\}\right\} = \mathcal{L}^{-1}\left\{\frac{\left(\frac{1}{R_1}+sC_1\right)\varepsilon_1}{s\left(s\left(C_1+C_2\right)+\frac{1}{R_1}+\frac{1}{R_2}\right)}\right\} \tag{23.7}$$

The left hand side is simply $V_{\text{out}}(t)$, the right hand side we look up in the table. Actually this function is not listed; we must split it up into partial fractions first, i.e. the expression for V_{out} is:

$$\mathcal{L}^{-1}\left\{\frac{\varepsilon_1}{s\left(1+\frac{R_1}{R_2}\right)}+\frac{\left(C_1\frac{R_1}{R_2}-C_2\right)\varepsilon_1}{\left(1+\frac{R_1}{R_2}\right)\left(s\left(C_1+C_2\right)+\frac{1}{R_1}+\frac{1}{R_2}\right)}\right\} \tag{23.8}$$

Using the table and doing some manipulation we find V_{out} to be given by:

$$\frac{R_2\varepsilon_1 u(t)}{R_1+R_2}\left(1+\frac{C_1R_1-C_2R_2}{R_2\left(C_1+C_2\right)}\left(e^{-\left(\frac{1}{R_1}+\frac{1}{R_1}\right)\frac{t}{C_1+C_2}}\right)\right) \tag{23.9}$$

in agreement with the result found in section 2.5.1.

An advantage of using the transform method is that most of the work has been done for us by the people who generated the transform pairs; we simply use these standard results. The only algebra we may have to do is to split our expression into partial fractions.

23.3 Short cuts

There is an important short cut to setting up equations such as (23.4) above. Using the methods discussed in Section 3.4 the transmission at frequency ω of the scope probe circuit may be simply written down as:

$$\frac{V_{\text{out}}}{V_{\text{in}}} = \frac{\frac{1}{\frac{1}{R_2}+j\omega C_2}}{\frac{1}{\frac{1}{R_1}+j\omega C_1}+\frac{1}{\frac{1}{R_2}+j\omega C_2}} \tag{23.10}$$

which simplifies to:

$$\frac{V_{\text{out}}}{V_{\text{in}}} = \frac{\frac{1}{R_1}+j\omega C_1}{j\omega\left(C_1+C_2\right)+\frac{1}{R_1}+\frac{1}{R_2}} \tag{23.11}$$

We see that equation (23.4) may be obtained from equation (23.11) by the following recipe. *Write s for $j\omega$ and replace the input and output voltages by their Laplace transforms.* This saves all the labour of deriving the differential equation.

It is easy to see why the recipe works. The output of a linear circuit will be related to the input by a linear differential equation with constant coefficients such as:

$$A_0V_{\text{in}}+A_1\frac{\mathrm{d}V_{\text{in}}}{\mathrm{d}t}+A_2\frac{\mathrm{d}^2V_{\text{in}}}{\mathrm{d}t^2}+\ldots = B_0V_{\text{out}}+B_1\frac{\mathrm{d}V_{\text{out}}}{\mathrm{d}t}+B_2\frac{\mathrm{d}^2V_{\text{out}}}{\mathrm{d}t^2}+\ldots \tag{23.12}$$

(See the oscilloscope probe equation derived in Section 2.6.3.)

We take the Laplace transform of both sides:

$$\mathcal{L}\left\{A_0V_{\text{in}}+A_1\frac{\mathrm{d}V_{\text{in}}}{\mathrm{d}t}+A_2\frac{\mathrm{d}^2V_{\text{in}}}{\mathrm{d}t^2}+\ldots\right\} = \mathcal{L}\left\{B_0V_{\text{out}}+B_1\frac{\mathrm{d}V_{\text{out}}}{\mathrm{d}t}+B_2\frac{\mathrm{d}^2V_{\text{out}}}{\mathrm{d}t^2}+\ldots\right\} \tag{23.13}$$

Then using the properties of the transform and considering only inputs (and therefore for a causal system also outputs) which are zero for $t<0$ we find:

$$\mathcal{L}\left\{A_0+A_1s+A_2s^2+\ldots\right\} = \mathcal{L}\left\{B_0+B_1s+B_2s^2+\ldots\right\} \tag{23.14}$$

or:

$$\frac{\mathcal{L}\left\{V_{\text{out}}\right\}}{\mathcal{L}\left\{V_{\text{in}}\right\}} = \frac{A_0+A_1s+A_2s^2+\ldots}{B_0+B_1s+B_2s^2+\ldots} \tag{23.15}$$

Now in the special case that the input V_{in} is $V_{\text{in}0}e^{j\omega t}$ we know that (because the circuit is linear) the output V_{out} can only be of the form $V_{\text{in}0}e^{j(\omega t+\phi)}$. Substituting these into the differential equation yields

$$V_{\text{in}}\left(A_0 + A_1 j\omega + A_2 (j\omega)^2 + \ldots\right) = V_{\text{out}}\left(B_0 + B_1 j\omega + B_2 (j\omega)^2 + \ldots\right) \tag{23.16}$$

or:

$$\frac{V_{\text{out}}}{V_{\text{in}}} = \frac{A_0 + A_1 j\omega + A_2 (j\omega)^2 + \ldots}{B_0 + B_1 j\omega + B_2 (j\omega)^2 + \ldots} \tag{23.17}$$

Comparing equations (23.15) and (23.17) justifies the recipe for constructing the Laplace transform equation.

23.4 More examples

23.4.1 Bandpass filter using a parallel tuned circuit

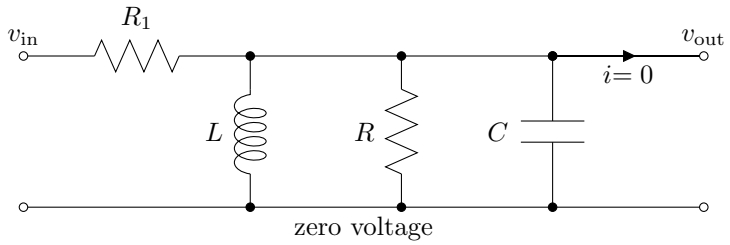


Figure 23.1: Equivalent circuit of a bandpass filter (from Figure 3.10)

The circuit treated in section 3.9.2 is shown again in Figure 23.1. Using the short cuts mentioned above we can immediately write down:

$$\frac{\mathcal{L}\{V_{\text{out}}\}}{\mathcal{L}\{V_{\text{in}}\}} = \frac{1}{1 + R_1\left(\frac{1}{sL} + sC + \frac{1}{R}\right)} \tag{23.18}$$

If V_{in} is again $\varepsilon_1 u(t)$, we have:

$$V_{\text{out}} = \mathcal{L}^{-1}\left\{\frac{\varepsilon_1 u(t)}{R_1\left(\frac{1}{L} + s^2 C + \frac{s}{R}\right) + s}\right\} \tag{23.19}$$

Finally we look up the inverse transform of the right hand side; the result is:

$$V_{\text{out}} = \frac{\varepsilon_1 u(t)\, e^{-\alpha t} \sin\omega_f t}{\omega C R_1} \tag{23.20}$$

where:

$$\alpha = \frac{1}{2C}\left(\frac{1}{R_1} + \frac{1}{R}\right) \tag{23.21}$$

and

$$\omega_f = \sqrt{\frac{1}{LC} - \frac{1}{4C^2}\left(\frac{1}{R_1} + \frac{1}{R}\right)^2} \tag{23.22}$$

V_{out} is a damped free oscillation at angular frequency ω_f with an amplitude proportional to the height V_1 of the voltage step input and a rate of decay depending on the resistances.

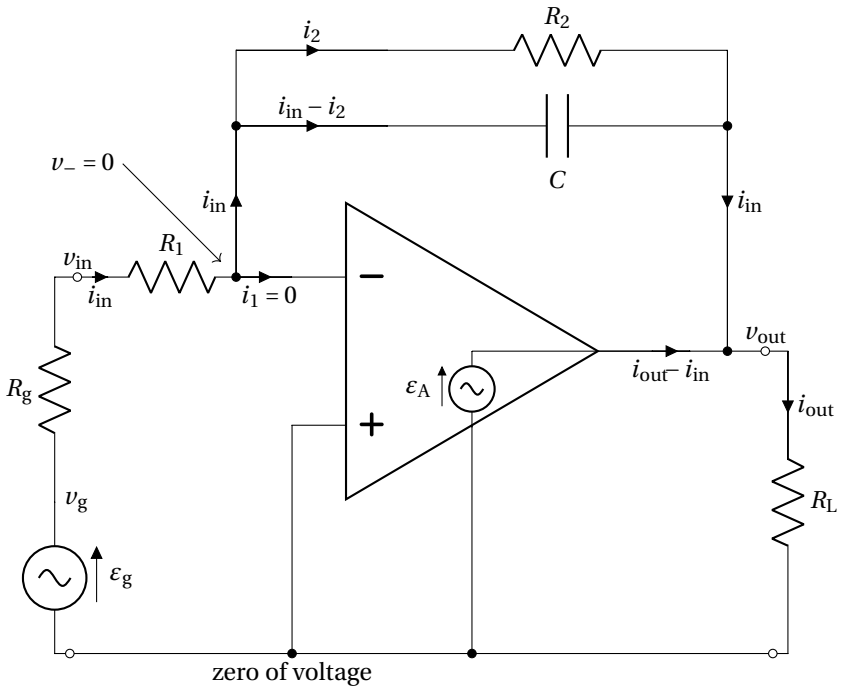


Figure 23.2: Amplifier with signal source

23.4.2 An amplifier

Consider the amplifier with signal source and load shown in Figure 23.2.

It is left as an exercise to show that in the frequency domain the voltage gain is given by:

$$\frac{v_{out}}{\varepsilon_g} = -\frac{R_2}{R_g + R_1}\frac{1}{1 + j\omega CR_2} \tag{23.23}$$

To find the response to a step input $\varepsilon_1 u(t)$ we write:

$$\frac{\mathcal{L}\{V_{out}\}}{\mathcal{L}\{V_{in}\}} = -\frac{R_2}{R_g + R_1}\frac{1}{1 + sCR_2} \tag{23.24}$$

where:

$$\mathcal{L}\{V_{in}\} = \frac{\varepsilon_1}{s} \tag{23.25}$$

so:

$$\mathcal{L}\{V_{out}\} = -\frac{R_2}{R_g + R_1}\frac{1}{(1 + sCR_2)\,s}\varepsilon_1 \tag{23.26}$$

$$V_{out} = -\frac{R_2}{R_g + R_1}\varepsilon_1 \mathcal{L}^{-1}\left\{\frac{1}{s(1 + sCR_2)}\right\} \tag{23.27}$$

$$\tag{23.28}$$

which after splitting the bracketed term into partial fractions and using the table of inverse transforms yields:

$$V_{out} = -\frac{R_2}{R_g + R_1}\varepsilon_1 u(t)\left[1 - e^{-t/CR_2}\right] \tag{23.29}$$

23.5 Table of Laplace transform pairs

Time domain	Laplace transform
$x(t)$	$X(s) = \int_{0^-}^{\infty} x(t)\, e^{-st}\, \mathrm{d}t$
$\alpha_1 x_1(t) + \alpha_2 x_2(t)$	$\alpha_1 X_1(s) + \alpha_2 X_2(s)$
$\frac{\mathrm{d}}{\mathrm{d}t} x(t)$	$sX(s) - x\left(0^-\right)$
$\frac{\mathrm{d}^n}{\mathrm{d}t^n} x(t)$	$s^n X(s) - \sum_{l=1}^{n} s^{n-l} x^{l-1}\left(0^-\right)$
$\int_{0^-}^{t} x(\tau)\, \mathrm{d}\tau$	$\frac{1}{s} X(s)$
$(-t)^n x(t)$	$\frac{\mathrm{d}^n}{\mathrm{d}s^n}\left\{X(s)\right\}$
$\frac{1}{t} x(t)$	$\int_{s}^{\infty} X(s)\, \mathrm{d}s$
$x(t - t_0)\, u(t - t_0)$	$e^{-st_0} X(s)$
$e^{\pm \alpha t} x(t)$	$X(s \mp \alpha)$
α is complex with non-negative real part	
$x(\alpha t)$ where $\alpha > 0$	$\frac{1}{\alpha} X\left(\frac{s}{\alpha}\right)$
$\lim_{t \to \infty} x(\alpha t)$	$\lim_{s \to 0} sX(s)$
poles of $X(s)$ in left hand plane	
$\lim_{t \to 0} x(\alpha t)$	$\lim_{s \to \infty} sX(s)$

Time domain	**Laplace transform**
$\left.\begin{array}{l} x_1(t) * x_2(t) \\ = \int_{0^-}^{t} x_1(t-\tau)\, x_2(\tau)\, \mathrm{d}\tau \end{array}\right\}$	$X_1(s)\, X_2(s)$
$\mathrm{Re}\left[x(t)\right]$	$\mathrm{Re}\left[X(s)\right]$
$\mathrm{Im}\left[x(t)\right]$	$\mathrm{Im}\left[X(s)\right]$
$\delta(t)$	1
$\frac{\mathrm{d}^n}{\mathrm{d}t^n}\delta(t)$	s^n
$u(t)$	$\frac{1}{s}$
$tu(t)$	$\frac{1}{s^2}$
$e^{-\alpha t}u(t)$	$\frac{1}{s+\alpha}$
$te^{-\alpha t}u(t)$	$\frac{1}{(s+\alpha)^2}$
$\frac{1}{b-a}\left(e^{-at}-e^{-bt}\right)u(t)$	$\frac{1}{(s+a)(s+b)}$
$\sin\omega t.u(t)$	$\frac{\omega}{s^2+\omega^2}$
$\cos\omega t.u(t)$	$\frac{s}{s^2+\omega^2}$
$e^{-\alpha t}\sin\omega t.u(t)$	$\frac{\omega}{(s+\alpha)^2+\omega^2}$
$e^{-\alpha t}\cos\omega t.u(t)$	$\frac{s+\alpha}{(s+\alpha)^2+\omega^2}$

Table 23.2: Table of Laplace transform pairs

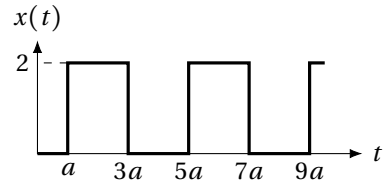
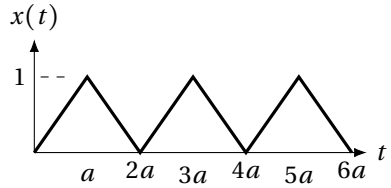
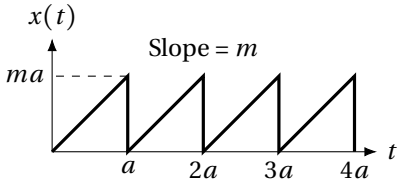
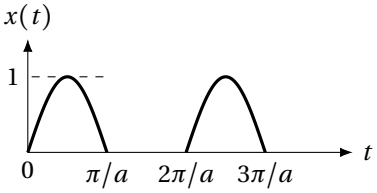
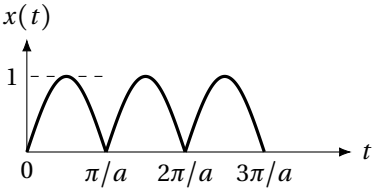
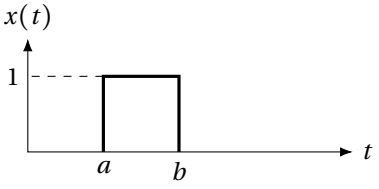
Time domain	$x(t)$	**Laplace transform**
Repeated pulse 	$2\sum_{k=0}^{\infty}(-1)^k u(t-(2k+1)a)$	$\frac{1}{s\cosh as}$
Triangular waveform 	$\frac{1}{a}\left[tu(t)+2\sum_{k=1}^{\infty}(-1)^k(t-ka)u(t-ka)\right]$	$\frac{1}{s^2}\tanh\left(\frac{as}{2}\right)$
Sawtooth waveform 	$mtu(t)-ma\sum_{k=1}^{\infty}u(t-ka)$	$\frac{m}{s^2}-\frac{ma}{2s}\left(\coth\left(\frac{as}{2}\right)-1\right)$
Half-wave rectification of sine wave 	$\sum_{k=1}^{\infty}\left[\sin a\left(t-\frac{k\pi}{a}\right)\right]u\left(t-\frac{k\pi}{a}\right)$	$\frac{a}{(s^2+a^2)(1-\mathrm{e}^{-\pi s/a})}$
Full-wave rectification of sine wave 	$\sin at.u(t)+2\sum_{k=1}^{\infty}\left[\sin a\left(t-\frac{k\pi}{a}\right)\right]u\left(t-\frac{k\pi}{a}\right)$	$\frac{a}{s^2+a^2}\coth\frac{\pi s}{2a}$
Single pulse 	$u(t-a)-u(t-b)$	$\frac{1}{s}\left(\mathrm{e}^{-as}-\mathrm{e}^{-bs}\right)$

Table 23.4: More Laplace transform pairs